Technische Universität München
TUM School of Computation, Information and Technology

3D Human Behavior Generation through Action and Interaction Synthesis

Christian Gerhard Hans Diller

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:
Prof. Dr. Stefan Leutenegger

Prüfende der Dissertation:
1. Prof. Dr. Angela Dai
2. Prof. Dr. Michael J. Black

Die Dissertation wurde am 22.05.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 16.10.2024 angenommen.

Acknowledgements

This dissertation marks the end of my Ph.D. journey which started about 4.5 years ago. During that time, I learned many things about academia and research – the most important of which is to keep going even in challenging and demotivating times. A crucial part of every Ph.D. journey is to learn how to manage time, motivation, and goals, and this is true especially in times of global pandemics. These insights made me grow as a person, from finishing my Master's thesis to finishing my Ph.D. dissertation.

Of course, a journey like this cannot be accomplished alone, and I am forever grateful to all the people who supported me along the way, be it during weekly meetings, during lunch or coffee breaks, or in spontaneous discussions about research and life.

First and foremost, I would like to thank my advisor **Prof. Angela Dai**. For giving me the opportunity to join the Ph.D. program in the first place, and for helping me get and stay on track during my projects. Her spirit for research is remarkable and incredibly inspiring. Without her guidance, I would not be where I am today. Along the same lines, I also thank **Prof. Thomas Funkhouser**, for his insights and discussions.

My colleagues certainly helped make my Ph.D. time more enjoyable and deserve great appreciation. I want to especially thank my office mates Norman Müller, David Rozenberszki, and Alexey Bokhovkin, for fruitful discussions and an overall enjoyable time. Many thanks also to Pablo Palafox, Yuchen Rao, Chandan Yeshwanth, Yueh-Cheng Liu, Quan Meng, Daoyi Gao, and Lei Li, as well as Armen Avetisyan, Aljaz Bozic, Ji Hou, Andreas Rössler, Dejan Azinović, Justus Thies, Dave Zhenyu Chen, Manuel Dahnert, Yawar Siddiqui, Andrei Burov, Guy Gafni, Shivangi Aneja, Artem Sevastopolsky, Yujin Chen, Lukas Höllein, Can Gümeli, Barbara Rössle, Jiapeng Tang, Haoxuan Li, Marc Benedí San Millán, Peter Kocsis, Shenhan Qian, Simon Giebenhain, Tobias Kirschstein, Ziya Erkoc, Yinyu Nie, and Alexey Artemov. To Susanne Weitz, Georgi Georgiev, and Assia Franzmann for their administrative support as well as Christoph Weiler and Sebastian Wohner for their technical support. To Minh Vo and Aayush Bansal for their guidance during my internship at Meta RL as well as my fellow interns David Bethge, Shubabrata Choudery, Edgar Sucar, and Baris Sen for making the time unforgettable.

During the time of my Ph.D., I was supported by two remarkable partners. I want to thank **Renée** (睿) for going through the early stages of this journey with me. A special thank you goes to **Jenny** (雪晨) for believing in me, being a happy spirit, and always knowing how to lift me up in times of hardship.

Lastly, I would like to highlight that none of this – from the first day I set foot onto the university campus in Augsburg to the final days of my Ph.D. in Munich – would have been possible without the support of my family. Thank you, **Angelika** and **Elmar**, **Niklas**, **Hildegard**, **Karl-Hans**, and **Uwe-Carsten**, for inspiring me, always and unconditionally believing in me, and supporting me throughout my endeavors.

Abstract

Computer vision algorithms have come a long way in recent years, from the simple classification of objects in 2D images to digitizing and understanding entire 3D environments. This has already enabled autonomous agents to confidently navigate in and interact with their surroundings. Building on this progress, the next challenge is to share the same space with humans while offering assistance and meaningful interactions. This will allow for truly autonomous assistive robotics, content creation, and mixed-reality applications. For example, a robotic surgical assistant could predict in advance where best to place a utensil to assist the surgeon's next action, and an animation tool could suggest plausible future human motion based on captured sequences, enabling rapid prototyping.

At the core of any such method is a comprehensive understanding of human behavior. Learning to forecast future human behavior in an analysis-by-synthesis manner is a promising direction for current research in this area: Realistic forecasting implies a high level of understanding of human actions and interactions with their environment.

Existing works on human motion forecasting are often insufficient for modeling long-term human behavior as well as realistic interactions between humans and their environment: They model human motion only in the immediate future, while human actions often last for multiple minutes. This observation is especially true for complex multi-step behavior (e.g., cooking), where the lack of training data is an additional constraint. Also, existing methods only consider interactions with environments on the human side – the generated human behavior is affected by its environment but not vice versa.

This dissertation aims to address such shortcomings: First, we introduce the notion of *Characteristic 3D Poses*. Instead of predicting a human pose sequence with poses at fixed time intervals, we observe that human motion is goal-oriented. Thus, we propose to forecast a single semantically meaningful *characteristic* action pose from an observed human motion sequence in a probabilistic manner.

Building on this notion, we present a method to forecast complex sequences of action labels and corresponding *Characteristic 3D Poses*. This approach allows for modeling complex composite human actions such as cooking or furniture assembly. As 3D data for such approaches is hard to acquire, we train our method on widely available 2D action datasets and an uncorrelated 3D pose database. This way, we can generate long-term human action sequences of 3D poses and actions from only 2D observations.

As humans usually interact with an environment, we present an approach to generate human-object interactions from input object geometry and a text prompt. In these sequences, both the human and the object are in motion. Using a contact-based human-object interaction representation allows for generating physically plausible sequences.

Finally, a discussion of potential research avenues aims to encourage future progress in the domain of 3D human behavior forecasting, generation, and understanding.

Zusammenfassung

Computer Vision Algorithmen sind in den letzten Jahren weit gekommen, angefangen bei der einfachen Klassifizierung von Objekten in 2D-Bildern bis hin zur Digitalisierung und zum Verstehen ganzer 3D-Umgebungen. Dies hat es autonomen Systemen bereits ermöglicht, sich sicher in ihrer Umgebung zu bewegen und mit ihr zu interagieren. Aufbauend auf diesen Fortschritten besteht die nächste Herausforderung darin, sich einen gemeinsamen Raum mit Menschen zu teilen und dabei Unterstützung und sinnvolle Interaktionen anzubieten. Dies wird autonome Assistenzroboter, die Erstellung von 3D-Inhalten und Mixed-Reality-Anwendungen ermöglichen. So könnte beispielsweise ein chirurgischer Assistenzroboter im Voraus vorhersagen, wo ein Werkzeug am besten platziert werden sollte, um den Chirurgen bei seiner nächsten Aktion zu unterstützen, und ein Animationstool könnte auf der Grundlage erfasster Sequenzen plausible zukünftige menschliche Bewegungen vorschlagen, was ein schnelles Prototyping ermöglicht.

Der Kern einer jeden solchen Methode ist ein umfassendes Verständnis menschlichen Verhaltens. Das Erlernen der Vorhersage zukünftigen menschlichen Verhaltens mittels Analyse-durch-Synthese ist eine vielversprechende Richtung in der aktuellen Forschung: Realistische Vorhersagen setzen ein umfassendes Verständnis menschlicher Handlungen und Interaktionen mit ihrer Umgebung voraus.

Bestehende Arbeiten zur Generierung menschlicher Bewegungen sind oft unzureichend, um langfristiges menschliches Verhalten sowie realistische Interaktionen zwischen Menschen und ihrer Umwelt zu modellieren: Sie generieren menschliche Bewegungen nur für die unmittelbare Zukunft, während menschliche Handlungen oft mehrere Minuten lang andauern. Diese Beobachtung gilt insbesondere für komplexes mehrschrittiges Verhalten (z. B. Kochen), bei dem der Mangel an Trainingsdaten eine zusätzliche Einschränkung darstellt. Außerdem berücksichtigen bestehende Methoden nur die Interaktionen mit der Umgebung auf der menschlichen Seite - das generierte menschliche Verhalten wird von seiner Umgebung beeinflusst, aber nicht umgekehrt.

Das Ziel dieser Dissertation ist, diese Schwächen anzugehen: Zunächst führen wir den Begriff der *charakteristischen 3D-Posen* ein. Anstatt eine menschliche Posenfolge mit Posen in festen Zeitintervallen vorherzusagen, stellen wir fest, dass die menschliche Bewegung zielorientiert ist. Daher schlagen wir vor, eine einzelne semantisch sinnvolle *charakteristische* Aktionspose aus einer beobachteten menschlichen Bewegungssequenz auf probabilistische Weise vorherzusagen.

Aufbauend auf diesem Konzept stellen wir eine Methode zur Vorhersage komplexer Sequenzen von Handlungsbezeichnungen und entsprechenden *charakteristischen 3D-Posen* vor. Dieser Ansatz ermöglicht die Modellierung komplexer, zusammengesetzter menschlicher Handlungen wie Kochen oder Möbelmontage. Da 3D-Daten für solche Ansätze schwer zu beschaffen sind, trainieren wir unsere Methode auf weithin verfügbaren 2D-Aktionsdatensätzen und einer unkorrelierten 3D-Posendatenbank. Auf diese Weise können wir langfristige menschliche Aktionssequenzen mit 3D-Posen und Aktionen aus nur 2D-Beobachtungen generieren.

Menschen interagieren normalerweise mit ihrer Umgebung. Wir stellen daher einen Ansatz vor, um Mensch-Objekt-Interaktionen aus einer Objektgeometrie und einem Textprompt zu generieren. In diesen Sequenzen sind sowohl der Mensch als auch das Objekt in Bewegung. Die Verwendung einer kontaktbasierten Darstellung der Mensch-Objekt-Interaktion ermöglicht die Generierung physikalisch plausibler Sequenzen.

Abschließend werden potenzielle Forschungsmöglichkeiten erörtert, um künftige Fortschritte auf dem Gebiet der Vorhersage und Generierung menschlichen Verhaltens in 3D zu fördern.

Contents

Part I

Introduction

1 Introduction

Over the past several decades, technological advances have led to unprecedented transformations, reshaping our capabilities to automatically capture, reconstruct, model, and understand the world around us. While this journey of innovation began early on with the first computers, the advent of fast and massively parallel compute capabilities enabled an efficient way to design, train, and deploy machine learning at scale in the form of neural networks. These advances soon led to a revolution in the area of computer vision, a discipline that tries to understand the world around us from vision, starting with 2D images and later expanding to more involved capturing methods such as 3D scanners. In 2D, neural networks allow for more accurate image classification, detection, and segmentation. Methods for modeling and synthesizing parts of the real world soon followed, generating photo-realistic images and digital assets. All these relatively recent advances allow for popular applications such as self-driving vehicles, photo-realistic VR teleconferencing, advanced assistive robotics, digital content creation for entertainment and simulation purposes, and more.

However, these methods and technologies were initially constrained to capturing 2D images, projections of the real world onto a flat image plane, and learning from this representation. The world around us is three-dimensional, and modeling a realistic version of it requires a solid understanding of 3D primitives and structures. With more widely available sensors and novel techniques for capturing, the field of 3D computer vision has recently emerged to work directly in 3D, from capture to generation. Nowadays, capturing the world in 3D is possible with hand-held sensors and advanced reconstruction algorithms. Operating directly in 3D without the limitations of 2D images has several advantages, such as no scale ambiguity, no occlusions, and independence from a pre-defined camera viewpoint. Consequently, many methods have been proposed to address understanding in 3D: Classification and segmentation of 3D objects and whole scenes, detection and localization of objects and object parts in larger structures, and how humans interact with 3D environments. On the other side, generative approaches have also emerged to automatically and directly generate 3D assets, being able to produce object and scene geometry from text conditioning and, more recently, realistic 3D human motion and interactions between humans and their environment (Fig. 1.1).

This last task in particular, understanding and modeling human behavior, is fundamental for applications such as autonomous robotics, assistive systems, and realistic animation creation. Being able to generatively model and forecast plays a significant role and is a foundational part of intelligence [4, 5, 6]. Consider the challenges for autonomous assistive systems: They need to be able to understand their environment in terms of the types of surrounding objects, their properties, and possible ways to interact with them. This allows assistive systems to perform actions in isolation, i.e., without sharing the

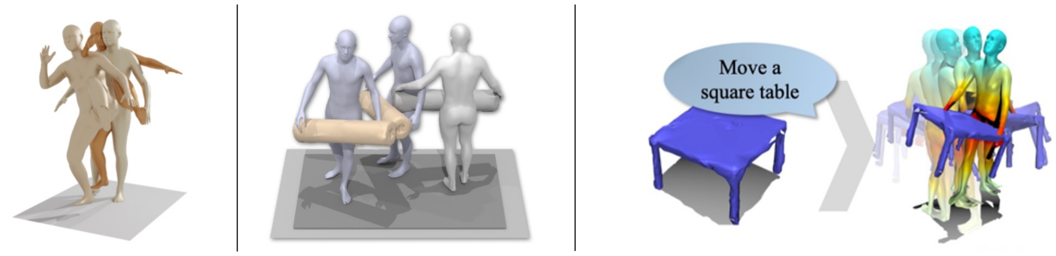


Figure 1.1: The task of modeling human actions and interactions in 3D. Left: Generating 3D human sequences in isolation from text ("kick with left leg") [1]. Middle: Forecasting future human motion (colored) from past observations (gray) [2]. Right: Generating dynamic human-object interactions from text and geometry conditioning [3].

space with human actors. However, to actively assist humans in a shared space, their understanding must extend to human behavior. This includes forecasting the next likely movements a human will perform, based on previous observations, to be able to move out of the way or to hand over objects in time. It also requires modeling how humans affect the shared environment, e.g., by moving a chair when sitting down or by changing the geometry of closets when opening them. This modeling of human behavior in terms of future actions and motion in 3D can be approached from different viewpoints: Considering only coarse high-level representations such as action class labels [7, 8, 9, 10, 11, 12], fine-grained representations like 3D human pose sequences [13, 14, 15, 16, 17], or interactions between humans and their environment [18, 19, 20, 21, 22, 23, 24, 25, 26].

Forecasting human poses in 3D lays the basis for a more fine-grained structural understanding of human behavior instead of purely relying on semantics, as is done when predicting action labels. With 3D poses, anticipatory action can be taken by an assistive robotic system; for example, a surgical robotic assistant should predict where best to put a tool to assist a surgeon's next action, which viewpoint to take to reduce obstructions, and where to position itself to be out of the way for future motion. Previously, there has been notable progress in the field of 3D human motion forecasting, addressing the specific task of predicting future human motion, given a short observation sequence [13, 27, 28, 29, 14, 30, 31, 15]. These existing approaches all take a temporal approach, predicting human poses at fixed time intervals, following the frame rate of camera capture. This makes it challenging to predict longer-term (multiple seconds) sequences, limiting the expressiveness of generated behavior. Instead, in this dissertation, we aim to focus on the high-level goal of the future action, decouple temporal and intentional behavior, and forecast one future *characteristic 3D pose* for a given short sequence observation. This characteristic pose depicts the human in an action-defining, semantically meaningful moment. Focusing on this moment enables many potential applications, including human-robot interactions, surveillance, visualization, simulation, and content creation. Specifically, the characteristic pose can be used to predict the exact hand-off point when a robot is passing an object to a person; to detect and display future poses worthy of alerts in a safety monitoring system; to coordinate grasps when assisting a person lifting a heavy object; to assist tracking through occlusions; or to predict future keyframes, as is done in video generation [32, 33]. Future characteristic 3D poses usually occur a longer time into the future (more than 1 second). Thus, an inherent ambiguity exists, and we aim to capture this multi-modality in our forecasting. Instead of deterministically predicting one future characteristic pose, we develop a method to generate a diverse set of plausible poses for a given short-sequence observation.

Forecasting a single characteristic 3D pose already allows for efficiently modeling human behavior but is limited to a single action. Realistic human behavior, such as when preparing a meal or assembling furniture, consists of multiple sub-actions that must be taken in a well-defined order. Thus, we propose to model such sequences in 3D with one characteristic pose per sub-action and its corresponding action label. We observe that these two tasks are coupled in nature. As opposed to previous work either predicting high-level action labels only [7, 8, 9, 10, 11, 12] or 3D pose sequences limited to short time frames [13, 14, 15, 16, 17], we show that the synergy between both allows for richer feature learning and ultimately, improved robustness and forecasting quality. This allows us to model multi-step composite human action sequences.

There are two significant challenges we address along the way. The first one arises from limited training data. As we aim to forecast future poses in 3D, having a dataset with ground-truth 3D pose and action annotations of complex action sequences would be ideal. However, no such dataset exists, and capturing one poses significant challenges due to occlusions and the required scale of the capturing setup. With only 3D datasets of single actions per sequence and 2D datasets of complex actions available, we instead aim to leverage these 2D datasets as input and supervision, with added weak supervision from an adversarial loss to ensure valid 3D pose predictions. The second challenge is the difficulty of predicting long (up to multiple minutes) action sequences. We propose the previously introduced characteristic pose representation to decouple actions and time, forecasting one semantically meaningful pose for each predicted action step. We formulate this approach autoregressively and are thus able to generate long composite action sequences of human behavior.

Finally, we observe that human behavior rarely happens in isolation. Meaningful motion arises from interaction with other humans in the vicinity or with the surrounding environment, using and manipulating objects in indoor and outdoor scenes. Existing works focus solely on generating dynamic humans, disregarding their surroundings [34, 35, 36, 37, 38, 39], or grounding interactions in a completely static environment that remains unchanged throughout the interaction [18, 19, 20, 21, 22, 23, 24, 25, 26]. However, real-world human interactions affect the environment, e.g. sitting down on a chair typically moves that chair, either to adjust it before sitting down or to move it away from other objects such as a table. Thus, we propose an approach to jointly generate human and object motion in 3D to generate realistic interactions between humans and objects. The key to our approach is to model both motions separately and bridge them by explicitly modeling contact. This helps encourage human and object motion to be semantically coherent and provides a constraint indicating physical plausibility (i.e., to discourage object floating and intersection). We additionally employ a contact weighting scheme based on the insight that object motion, when manipulated by a human, is most defined by the motion of the body part in closest contact.

In this dissertation, we present approaches focused on efficiently generating future human behavior in single-action and complex composite multi-action sequences. In an additional approach, we explore how to generate realistic human-object interactions, bridging the gap from isolated human behavior to interactions with an environment. In summary, we provide these contributions to the field of 3D human motion forecasting and generation:

- A probabilistic approach to forecasting future human behavior in an intent-driven manner, using the novel representation of *characteristic 3D poses*, efficiently modeling the multi-modality of long-term future human behavior.
- A method for forecasting complex sequences of human behavior, consisting of action labels and corresponding characteristic poses in 3D, while being trained on 2D action data and a database of uncorrelated 3D human poses.
- An approach to generate realistic, diverse, and physically plausible whole-body human-object interaction sequences, utilizing a contact formulation bridging human and object motion.

1.1 Dissertation Overview

This dissertation contains seven chapters across three parts, which are structured as follows:

- Part I: Introduction (Chapters 1-2)
  - Chapter 1 introduces the topic of 3D human behavior modeling and forecasting in terms of recent developments and the significance of our contributions to the research community
  - Chapter 2 explains fundaments concepts in 3D static geometry and 3D dynamic human representation, capture thereof, and modeling of human motion
- Part II: 3D Human Behavior Understanding (Chapters 3-5)
  - Chapter 3 introduces our work on forecasting characteristic 3D poses of human actions from a short input sequence
  - Chapter 4 introduces our work on complex long-term 3D human behavior forecasting from 2D video observations
  - Chapter 5 introduces our work on generating contact-guided 3D human-object interactions from object geometry and text description input
- Part III: Conclusion & Outlook
  - Chapter 6 summarizes our proposed methods and provides an overall conclusion
  - Chapter 7 discusses remaining limitations and gives an outlook for possible future research directions

1.2 Contributions

This dissertation proposes three novel approaches, each addressing one important aspect of human behavior modeling. We first introduce the notion of *Characteristic 3D Poses*, a more suitable representation for goal-directed human action sequences. By forecasting a single semantically meaningful 3D human action pose in a probabilistic manner, the inherent multi-modality of multiple-second human actions can be modeled efficiently and expressively. We then extend this 3D human pose representation to complex composite human action sequences, achieving long-term human action forecasting that was not possible with previous methods. Combined with its ability to jointly forecast actions and characteristic poses in 3D from only 2D observations, this method is highly applicable to easy-to-capture video observations. Finally, we propose a method to generate human-object interactions from text and object geometry to model more realistic human motion. Our contact-based representation of interactions between humans and objects in motion allows for generating realistic full-body human motion sequences without constraining an environment to remain static throughout the sequence. In summary, the contribution of this dissertation, structured by publication, are:

- We propose "Forecasting Characteristic 3D Poses of Human Actions" and introduce the task of forecasting characteristic 3D poses: from a short sequence observation of a person, predict a future 3D pose of that person in a likely action-defining, characteristic pose. Prior work estimates future poses at fixed time intervals. This frame-by-frame formulation confounds temporal and intentional aspects of human action. Instead, we define a semantically meaningful pose prediction task that decouples the predicted pose from time. To predict characteristic poses, we propose a probabilistic approach that models the possible multi-modality in the distribution of likely characteristic poses. We then sample future pose hypotheses from the predicted distribution in an autoregressive fashion to model dependencies between joints. The method development and implementation were done by the first author. Discussions with co-authors led to the final publication [40].

- We propose "FutureHuman3D: Forecasting Complex Long-Term 3D Human Behavior from Video Observations" and present a generative approach to forecast long-term future human behavior in 3D, requiring only weak supervision from readily available 2D human action data. We design our method to only require 2D RGB data while being able to generate 3D human motion sequences. We use a differentiable 2D projection scheme in an autoregressive manner for weak supervision and an adversarial loss for 3D regularization. Our method predicts long and complex behavior sequences (e.g., cooking, assembly) consisting of multiple subactions. We tackle this in a semantically hierarchical manner, jointly predicting high-level coarse action labels together with their low-level fine-grained realizations as characteristic 3D human poses. We observe that these two action representations are coupled in nature, and joint prediction benefits both action and pose forecasting. The method development and implementation were done by the first author. Discussions with co-authors led to the final publication [41].

- We propose "CG-HOI: Contact-Guided 3D Human-Object Interactions", the first method to address the task of generating dynamic 3D human-object interactions (HOIs) from text. We model the motion of humans and objects in an interdependent fashion, as semantically rich human motion rarely happens in isolation without any interactions. Our key insight is that explicitly modeling contact between the human body surface and object geometry can be used as strong proxy guidance, both during training and inference. Using this guidance to bridge human and object motion enables the generation of more realistic and physically plausible interaction sequences, where the human body and corresponding object move coherently. Our method first learns to model human motion, object motion, and contact in a joint diffusion process, inter-correlated through cross-attention. We then leverage this learned contact for guidance during inference synthesis of realistic, coherent HOIs. The method development and implementation were done by the first author. Discussions with co-authors led to the final publication [3].

1.3 List of Publications

Authored

- **Christian Diller**, Thomas Funkhouser, and Angela Dai. "Forecasting Characteristic 3D Poses of Human Actions" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 15914-15923
- **Christian Diller**, Thomas Funkhouser, and Angela Dai. "FutureHuman3D: Forecasting Complex Long-Term 3D Human Behavior from Video Observations" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024
- **Christian Diller** and Angela Dai. "CG-HOI: Contact-Guided 3D Human-Object Interactions" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024

Co-Authored

- Angela Dai, **Christian Diller**, and Matthias Niessner, "SG-NN: Sparse Generative Neural Networks for Self-Supervised Scene Completion of RGB-D Scans" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 849-858

2 Fundamentals and Methods

Over the past few decades, there has been tremendous work toward understanding and modeling three-dimensional environments, both in computer graphics and computer vision. While these works were initially focused on static 3D geometry, recently, there has been increasing interest in representing dynamic human bodies and their interactions with object and scene geometry. This chapter provides a broad overview of widely used methods for representing, capturing, and modeling geometry, human bodies, interactions between the two, and approaches to modeling human dynamics.

Section 2.1 introduces and discusses common ways to represent 3D geometry, specifically for objects and scene environments. How to represent humans in 3D is discussed in section 2.2. Section 2.3 then explains techniques for capturing such geometry and human body motion from real-world observations. Finally, section 2.4 discusses fundamental methods used to model and forecast human motion in 3D.

2.1 Representing 3D Geometry

In computer graphics and computer vision, choosing a suitable representation for 3D geometry is crucial. This is a major differentiating factor between 2D and 3D representations of reality: In the 2D domain, a sensor projects points in 3D space to a regular 2D grid of cells, called pixels, each of which representing the light intensity across different color channels. On the contrary, when working directly with 3D objects, no single representation fits all use cases.

This chapter explores the three most-used representations of 3D geometry: triangle meshes, point clouds, and voxel grids. Each representation has its own advantages and disadvantages, and each method has to weigh positive against negative properties, mainly in terms of memory consumption, flexibility, regularity, and level of detail.

2.1.1 Polygon Meshes

Polygon meshes are the most common and versatile representation of 3D geometry. Formally, a polygon mesh is a graph structure containing a set of vertices V and edges E. Polygons are formed between groups of N adjacent points connected by edges from E and act as a piecewise approximation of the actual geometry. Most commonly, the polygons are triangles, and as such, $N = 3$.

Meshes can be efficiently and compactly stored with two lists. First, a list of vertices, storing locations along x, y, and z axes in 3D space. Thus, in its most basic form, vertices are defined as $v_i \in \mathbb{R}^3$. Second, an index list for storing the edges of triangles between vertices. These triangles (or "faces") contain a triple of vertex indices (i, j, k), with each index referring to a vertex stored in the vertex list. Figure 2.1 shows an example of a triangle mesh.

Meshes are widely used in computer graphics for rendering and modeling objects due to their flexibility and ability to efficiently approximate complex surfaces. They are also utilized in computer-aided design (CAD), animation, and simulation applications.

As such, their most significant advantage is the applicability to downstream tasks and graphics pipelines for integration into existing workflows. In addition, they provide an efficient way to store and exchange geometric data due to their compact requirement of only storing two lists. Moreover, meshes support texture mapping and surface attributes, allowing for realistic material representations and visual effects. Their simplicity also allows for accelerated geometry operations (e.g., transforming a triangle's position in space by matrix-vector multiplication), which is exploited in modern graphics hardware by embedding specialized geometry instructions directly in hardware circuits.

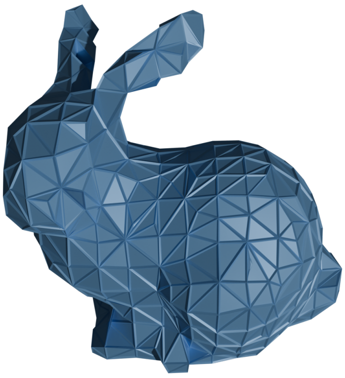


Figure 2.1: Stanford bunny in triangle mesh representation. Visualized are vertices, edges, and faces

However, triangle meshes have several limitations, preventing their widespread use in geometry generation and reconstruction. Representing smooth surfaces accurately with triangle meshes can be challenging, especially for highly curved or organic shapes. Meshes may require high tessellation levels to capture fine details accurately, leading to increased memory and computational costs. Furthermore, mesh data structures may lack explicit topological information, complicating boundary detection and mesh editing operations. Finally, processing meshes with neural networks is challenging due to the difficulty of computing gradients with respect to the geometry, which is often required for computer vision applications.

2.1.2 Point Clouds

Point clouds represent 3D surfaces as an unordered set of discrete points sampled on the geometry surface. Each point $P \in \mathbb{R}^3$ in the cloud typically contains spatial coordinates (x, y, z) and optionally additional attributes such as color or surface normals. As such, they can be efficiently stored as a single list of 3D points. Point clouds are generated using 3D scanning technologies such as LiDAR (Light Detection and Ranging) and structured light scanning, discussed in detail in section 2.3. From an existing mesh, a point cloud can be generated by uniformly sampling the surface, taking the area of faces into account. Figure 2.2 shows the result of uniformly sampling the mesh in figure 2.1.

Point clouds offer several advantages over other representations. Firstly, they directly represent object surfaces with high fidelity, capturing fine details and surface irregularities accurately.

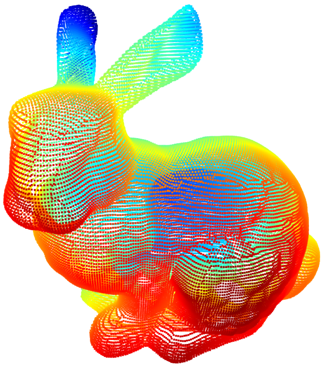


Figure 2.2: Stanford bunny in point cloud representation. Visualized are uniformly sampled surface points

They are flexible in terms of data acquisition, allowing for the integration of data from various sensing modalities and scanning techniques. Moreover, point clouds are suitable for capturing complex geometries and intricate structures that may be challenging to represent using other methods.

However, point clouds suffer from several limitations. One challenge is noise, which can arise from sensor inaccuracies, environmental factors, and imperfections in the scanning process. Cleaning and processing noisy point cloud data can be time-consuming and may require advanced filtering and reconstruction techniques. Additionally, neighborhood operations are more expensive since they are inherently just an unordered collection of points. Finally, they only offer a pointwise approximation of a surface without information in between, meaning that their usefulness highly depends on their quality and sparsity.

2.1.3 Voxel Grids

Voxel grids represent 3D shapes by discretizing 3D space into a regular grid of voxels, analogous to pixels in 2D space. Each voxel encodes information about the geometry (and, optionally, the color) at the voxel's location.

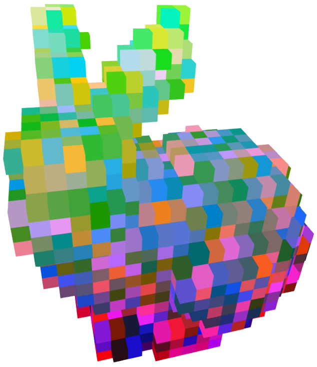


Figure 2.3: Stanford bunny in occupancy grid representation. Visualized are occupied voxels

The two most commonly used representations in voxel grids are occupancy and distance fields. Occupancy grids encode if a given voxel is occupied by the underlying geometry or not. Figure 2.3 shows the result of encoding a shape in this manner. Alternatively, implicit representations such as distance fields can be used. Here, each voxel encodes the distance to the closest surface, implicitly defining the geometry's surface as the zero-level set of a scalar function $f : \mathbb{R}^3 \to \mathbb{R}$ defined over the entire space. This way, surfaces can be represented continuously and smoothly, as opposed to binary occupancy grids. By using signed distance fields, the inside and outside of a shape can be defined, resolving ambiguities that arise from alternative representations. Finally, they allow for the use of constructive solid geometry operations, enabling the synthesis of complex shapes from simple primitives.

Voxel grids offer several advantages. Firstly, their regular structure makes spatial queries and neighborhood operations efficient and easy to perform. Secondly, they are easy to process and generate, especially in the domain of deep learning, where 3D convolutions can exploit cheap neighborhood operations.

Due to their grid structure, voxel grids suffer from several disadvantages. One challenge is the high memory consumption associated with fine resolutions and large grid sizes. Storing and processing voxel grid data can require significant computational resources due to their cubic growth, particularly for high-resolution volumetric datasets. Additionally, voxel grids may lack geometric detail compared to triangle meshes, especially for complex surfaces and fine structures. Finally, downstream applications often require an explicit geometry representation, such as meshes or point clouds, requiring a conversion step from regular voxel grids. The most common method is Marching Cubes [42], using a lookup table of triangles and vertex interpolation to extract a mesh from a voxel grid.

2.1.4 Neural Geometry Representations

Implicit representations, such as occupancy grids and signed distance fields mentioned above, are a way to encode 3D geometry in a voxel grid of any fixed size or resolution. However, they are simply functions mapping a point $P \in \mathbb{R}^3$ in 3D space to a scalar value representing occupancy or distance. Consequently, they are not limited to being stored in a grid. Several recent methods explored this aspect and proposed using a deep network to produce values corresponding to a given point P. This not only alleviates the need for a memory-intensive grid structure to store the values but also allows for shape representation at arbitrary scales, as networks can be queried at any continuous point in three-dimensional space.

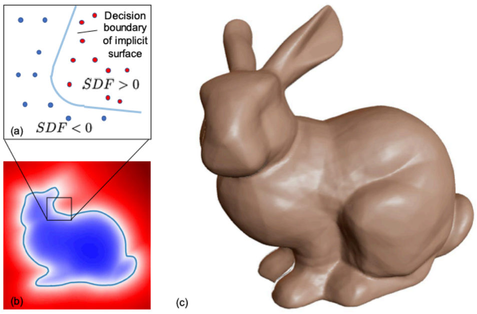


Figure 2.4: Stanford bunny as generated by DeepSDF [43]. Left: Visualization of points sampled inside and outside the shape, with the decision boundary. Right: The final shape

Figure 2.4 shows a shape generated by the popular approach DeepSDF [43]. It optimizes the parameters of the neural network to minimize the discrepancy between the predicted and actual SDF values for a set of training shapes. Its formulation enables it to learn complex geometries from sparse and irregular samples signed distance fields in a neural network and to reconstruct the original shape by querying the network at arbitrary locations during test time.

Such neural implicit representations combine the advantages of implicit fields mentioned above without being limited to storing values in a pre-defined voxel grid.

2.2 Representing Dynamic 3D Humans

While the above representations of 3D geometry can, in principle, also be applied for modeling human bodies, specialized representations are needed for efficiently representing dynamic 3D humans. When considering human motion and interaction, the focus lies less on a high-fidelity reconstruction of the human body surface but more on efficiently representing key elements of the human skeleton most affected by motion. This chapter explores skeleton-based and full-body representations of the human body in 3D, commonly used in 3D motion understanding and synthesis.

2.2.1 Skeleton-Based Representations

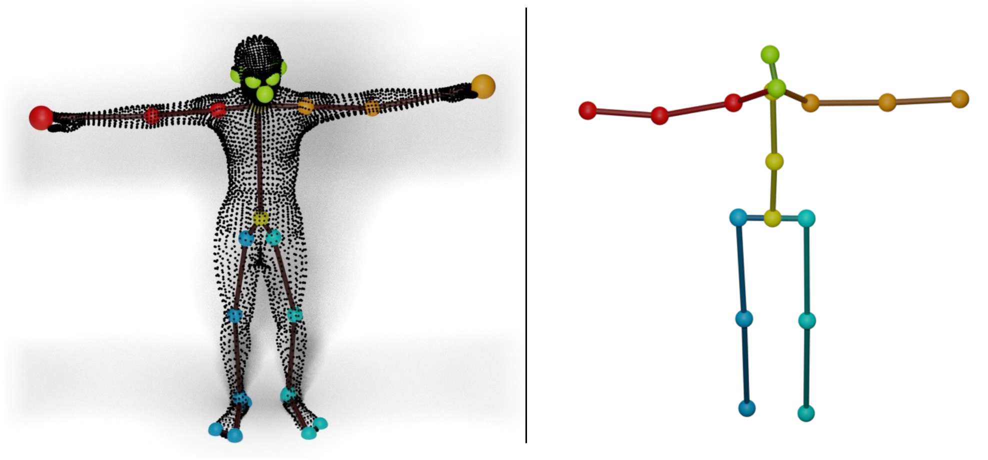


Figure 2.5: Surface point cloud of a human body from the GRAB [44] dataset and the corresponding extracted skeleton joints overlaid (left); native 17-joint skeleton from the Human3.6M [45] (right). Skeletons offer a simplified and easy-to-parse representation of the human body but suffer from strong simplification (e.g., not taking into account human body shape or disregarding body parts such as feet) and various assumptions (e.g., number of joints or anatomically incorrect placement of joints like the hip).

While a holistic, skinned model of the human body is crucial to applications like photorealistic rendering or human-object interaction, a simplified version is often desirable when modeling human motion. Here, the human body is often represented as an abstract version of the underlying skeleton by only considering the body joints and implicitly modeling the bones connecting pairs of joints. This abstraction makes it possible to efficiently represent, recognize, and forecast human motion in 3D. The exact amount of body joints and their topology depends on a given dataset's conventions – figure 2.5 shows two different examples from the GRAB [44] and Human3.6M [45] datasets – but representations typically contain the most salient body joints such as hands, shoulders, hips, and feet.

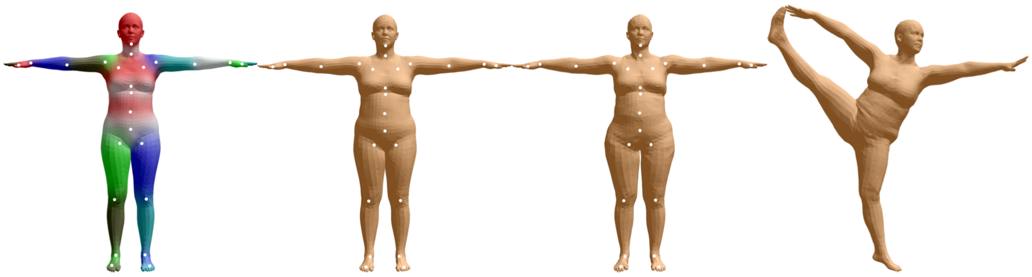
Another defining property of skeletal human body models is the representation of individual joints. The most obvious option is to assign 3D coordinates to each joint J. This can be done in global world space coordinates, allowing for more global reasoning of joint positions for understanding human motion within an environment or in local coordinates with respect to a root joint J_R. In the latter case, the 3D coordinates describe an offset of the current joint J to the root joint J_R, usually represented by a central joint in the human body such as the hip joint in Human3.6M [45].

Beyond simple 3D coordinate representation, joint angles offer a nuanced way to describe body posture through the orientation and rotation of skeleton bones. These angles, which can be expressed as Euler angles, quaternions, or exponential maps, represent the angle between the two adjacent bones. Euler angles, though straightforward, are prone to gimbal lock, leading to position ambiguities. Quaternions circumvent this issue and are computationally favorable for smooth animations, while exponential maps use a vector to denote rotation axis and magnitude, which is ideal for differential calculations in physics simulations. The transition from angular representations to precise joint locations in 3D employs forward kinematics. This technique calculates the position and orientation of each chain part, like a limb, given the joint angles, starting from the root of the skeleton, to sequentially apply rotations and compute the positions of all subsequent joints. This allows for the reconstruction of the body's posture from angular data. One major advantage of using joint angles instead of 3D coordinates is their invariance to body parameters such as height or width. When reconstructing spatial coordinates from angles, a template of bone lengths can be used. This simplifying assumption benefits many approaches for motion generation, allowing them to focus on the motion without the need to model differences in body shape.

Recent advancements have led to more sophisticated skeleton-based models like HumanML3D [46], which surpass earlier methods by incorporating features crucial for realistic human motion simulation. These models add elements such as foot contact flags and modeling joints in local space in terms of 3D coordinates and joint angles. Foot contact flags are vital for simulating actions like walking or running, where an accurate depiction of foot-ground interaction impacts motion realism.

2.2.2 Full-Body Representations

While a simplified abstraction of the human body is often desirable and sufficient for generating dynamic humans in 3D, as mentioned above, several applications can benefit from considering the whole human body, including skinned surface geometry, in their methodology. The most prominent examples are methods that aim to model interactions of humans with their environment, containing actions such as grasping a cup or sitting on a chair. For a realistic interaction, the internal skeleton of a human is not sufficient – it is the human body surface that is interacting with these objects. However, directly modeling the human body surface as a point cloud or polygon mesh has some serious drawbacks. For the necessary level of detail, such a mesh would require > 1000 individual vertices, which are orders of magnitude larger than the ≈ 20 joints typically modeled with skeleton representations. Additionally, this would not take topological properties



(a) $\bar{\mathbf{T}}, \mathcal{W}$ **(b)** $\bar{\mathbf{T}} + B_S(\vec{\beta}), J(\vec{\beta})$ **(c)** $T_P(\vec{\beta},\vec{\theta}) = \bar{\mathbf{T}}+B_S(\vec{\beta})+B_P(\vec{\theta})$ **(d)** $W(T_P(\vec{\beta},\vec{\theta}), J(\vec{\beta}), \vec{\theta}, \mathcal{W})$

Figure 2.6: Visualization of the skinning procedure from the original SMPL [47] paper: (a) Template mesh $\overline{\mathbf{T}}$ and colored blend weights $\mathcal{W}$; (2) Addition of identity-dependent blend shapes B_S with shape vector $\vec{\beta}$, also affecting regressed joint positions J; (c) Addition of pose-dependent blend shapes B_P; (d) Final reposed deformed vertices

into account: The topology of the human body cannot change by nature, and directly enforcing such a constraint would negatively impact any approach to generating human motion.

Several approaches have been developed over the years to tackle these issues. The most widely used one is SMPL (short for "Skinned Multi-Person Linear Model") [47]. It is a data-driven approach that utilizes linear blend shapes for an anatomically correct statistical model of the human body surface based on shape identity and pose parameters. Figure 2.6 shows an overview of their methodology. At its core, SMPL is a parametric model that defines the human body through a set of shape and pose parameters. The shape parameters β capture individual body variations, such as height, weight, and body proportions, while the pose parameters θ (represented by joint angles) control the articulations of the body joints. SMPL utilizes a learned model of vertex displacements, applied to a template mesh $\overline{\mathbf{T}}$, producing the final body surface geometry. This approach allows for the realistic modeling of muscle deformations and skin stretching, providing a highly accurate representation of human anatomy and motion. SMPL uses linear blend skinning (LBS) and corrective blend shapes $B_S(\beta)$ for the shape and $B_P(\theta)$ for the pose. LBS is a technique that uses transformations of the skeletal joints to deform the mesh, with each mesh vertex being influenced by one or more joints, as defined by a weight matrix $\mathcal{W}$. However, LBS alone can lead to unrealistic deformations, particularly around joint areas, e.g., visible soft tissue deformations when bending the knee. SMPL addresses this issue by incorporating corrective blend shapes based on the pose parameters. These blend shapes adjust the mesh deformations to mimic the natural movement of muscles and skin, resulting in significantly improved visual fidelity.

Figure 2.7 shows the influence of shape and pose parameters, with shape parameters encoding different identities. Note that varying the shape parameters does not alter the underlying skeleton of the body pose but influences the final body surface geometry.

In practice, a human body can be represented using SMPL with a tuple of vectors (θ, β, R, t) with pose parameters $\theta \in \mathbb{R}^{21\times3}$ in joint angle form, shape parameters $\beta \in \mathbb{R}^{16}$,

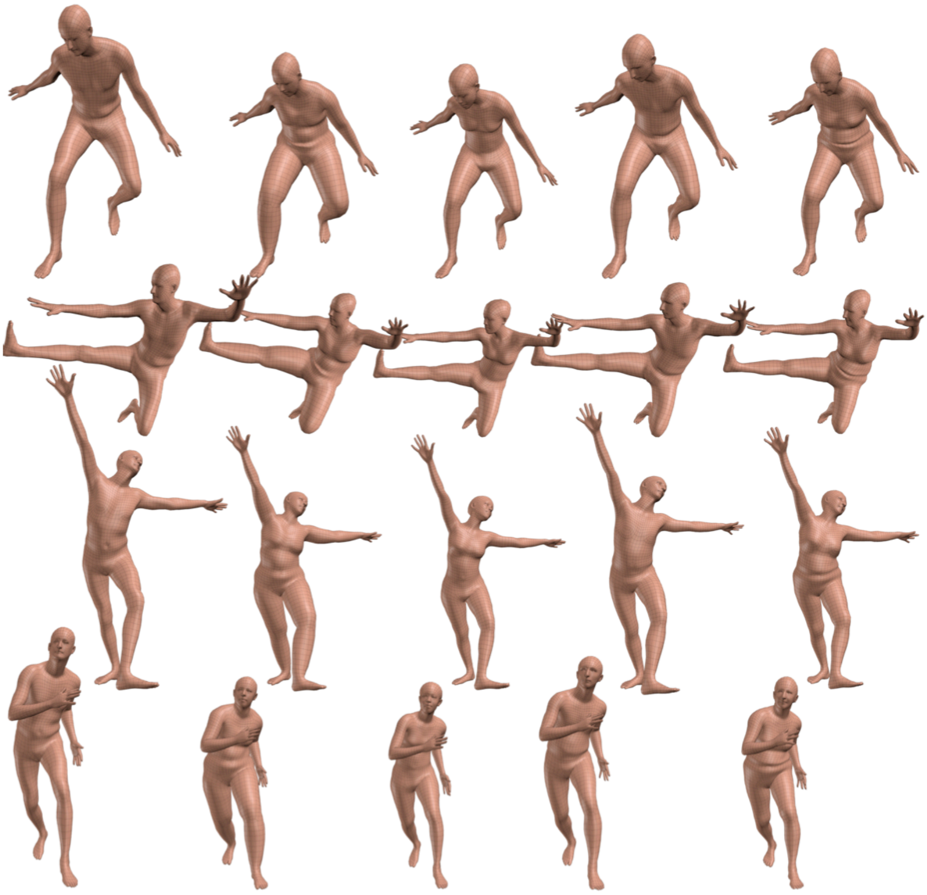


Figure 2.7: Exploring the SMPL [47] shape space vs. pose space. From top to bottom: Different poses, same identity. Left to right: Same pose, different identities. By compactly representing the human body surface in terms of identity and pose, SMPL is useful for many approaches aiming to model realistic human motion and interactions.

and global rotation $R \in \mathbb{R}^3$ as well as global translation $t \in \mathbb{R}^3$. This compact and fully differentiable representation is highly desirable for motion generation applications, making it possible to optimize for realistic human motion while generating holistic 3D human body geometry.

Following the success of SMPL, several extensions and improvements have been proposed, each designed to address specific limitations or to extend the model's capabilities. SMPL-H [48], for example, adds articulated hands to the SMPL model, providing a more comprehensive representation of human gestures and hand interactions. SMPL-X [49] further extends this by including facial expressions, resulting in a full-body model capable of conveying a wide range of human emotions. The STAR [50] ("Sparse Trained Articulated Human Body Regressor") model introduces improvements in the performance and generalizability of SMPL, offering better quality with fewer parameters, with learned sparse spatially local corrective blend shapes. Lastly, SUPR [51] ("Sparse Unified Part-Based Human Representation") offers a factorized representation of the human body, which can be separated into individual body part models for head, hands, and feet (visualized in figure 2.8).

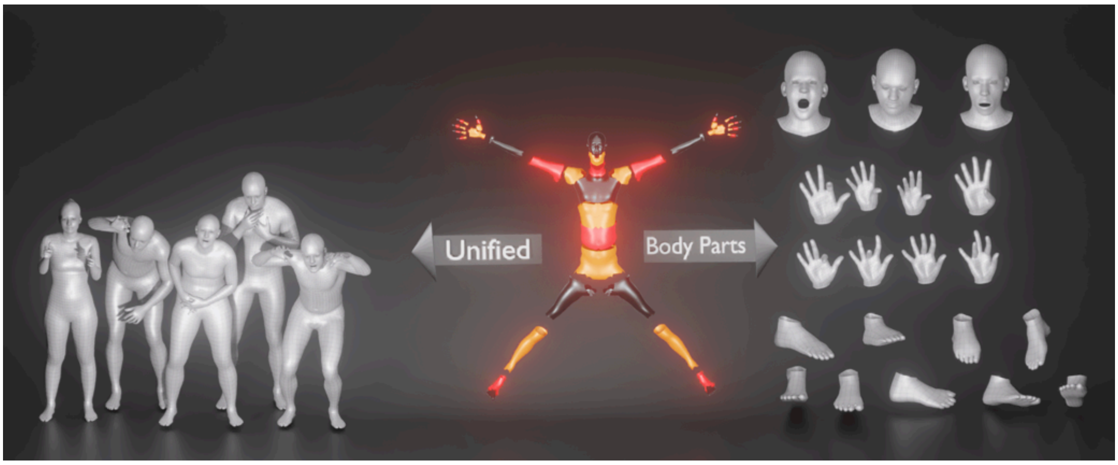


Figure 2.8: The SUPR [51] model for human body surface representation. Its part-based approach allows for the separate usage of heads, hands, and feet models as well as unifying them into one holistic human body representation.

2.2.3 Implicit Neural Representations

Parametric models such as SMPL [47] discussed above, while playing a crucial role in modeling human bodies and motion, have several limitations that more recent methods aim to address with neural representations. First, they usually require domain-specific annotations, such as the number of parts or the exact kinematic chain. Additionally, complex surface features such as wrinkles of clothing are often hard to represent with parametric vertex-based methods like SMPL.

Recent approaches like NPMs [52] ("Neural Parametric Models for 3D Deformable Shapes") and SPAMs [53] ("Structured Implicit Parametric Models") can be learned without manual annotation or expert knowledge in a specific domain by leveraging implicit functions, similar to methods for static 3D geometry (see section 2.1), while capturing more intricate surface detail. They first learn a latent space of shape identities in canonical pose by conditioning a shape MLP on a shape code. Afterward, a learned deformation field maps points from the shape's canonical space into a posed shape version. This deformation field is represented by another MLP conditioned on the shape and corresponding latent pose code to predict an offset vector for any given query point sampled in the canonical pose.

While still an active area of research, such neural representations promise more flexibility than traditional methods such as SMPL and are an exciting direction for future work on human motion generation.

2.3 Capturing Static Geometry and Dynamic Humans in 3D

Understanding and modeling realistic human motion and human interactions with their environment requires capturing real-world observations and datasets of such behavior. While it is possible to create synthetic data for static 3D geometry and dynamic humans, such generated data is often lacking in diversity and realism.

Thus, it is essential to devise methods to capture datasets of real-world behavior. This chapter will give an overview of fundamental computer vision approaches to capture both static 3D geometry and dynamic 3D human motion.

2.3.1 Image Formation Process

The image formation process is the most fundamental concept in computer vision, describing the behavior of light as it interacts with objects and optical systems. Among various models describing this interaction, the pinhole camera model is often used to understand the essential geometric aspects of image formation. This simplified model (visualized in figure 2.9) describes how light rays originating from a scene pass through a small aperture (the pinhole) to form an inverted image on the opposite side. The benefit of this model lies in its ability to provide a geometrically accurate representation of a scene without the complexity introduced by lens systems.

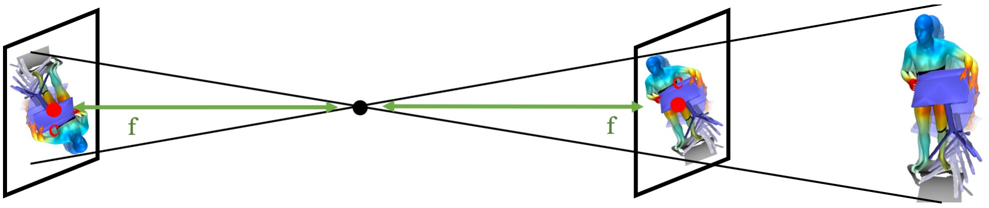


Figure 2.9: Pinhole camera model: The 3D model (right) is projected onto the image sensor (left), centered at principal point c, through the pinhole, which is a distance f away from the sensor plane. In many applications, a virtual image plane is considered to be in front of the pinhole at the same distance f, in which case the projection is not flipped.

Mathematically, the pinhole camera model is described by the equation $\mathbf{p} = \mathbf{K}[\mathbf{I}|\mathbf{0}]\mathbf{P}$, where $\mathbf{p}$ represents the homogeneous coordinates of a point in the image plane, $\mathbf{P}$ denotes the homogeneous coordinates of the corresponding point in the 3D scene, and $\mathbf{K}$ is the intrinsic camera matrix. The matrix $\mathbf{K}$ contains the parameters intrinsic to a given camera, namely focal length f and principal point (c_x, c_y), typically represented as

$$\mathbf{K} = \begin{bmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{2.1}$$

The focal length f quantifies the distance between the pinhole and the image plane, essentially controlling the scale of the projected image, while the 2D principal point (c_x, c_y) signifies the intersection of the optical axis with the image plane, serving as a reference for image coordinates.

In practical imaging systems, lenses are usually used to gather more light instead of a pure pinhole camera. Calibrated lens camera models then account for the refractive properties of the lens materials, which bend the light rays to focus them onto the image sensor. This mechanism enables the capturing of sharp images of objects at various distances which is inherently missing in the pinhole model due to its infinite depth of field. However, lenses introduce additional complexities that must be accounted for with additional parameters in the camera model. Thus, lens-based models are often simplified back to the pinhole model for many applications, reducing the computational complexity of image analysis and reconstruction tasks. This is possible in cases where the pinhole model provides a sufficiently accurate approximation, i.e., when lens distortions are minimal or can be corrected through post-processing.

While a single camera is usually used in 2D computer vision to project object and scene geometry into 2D images, there are several approaches, such as Structure from Motion (section 2.3.2) or multi-camera setups for human motion capture (section 2.3.4) which require the notion of a shared world coordinate system as well as individual camera coordinate systems. Thus, in addition to the intrinsic camera parameters above, camera models usually also contain extrinsic camera parameters, defining the position and orientation of the camera frame relative to a world coordinate system. Typically represented by a rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times 3}$ and a translation vector $\mathbf{t} \in \mathbb{R}^3$, these parameters are crucial for mapping a point $\mathbf{P_w}$ in the world coordinates to a point $\mathbf{P_c}$ in the camera coordinates, following the equation

$$\mathbf{P_c} = \mathbf{R}\mathbf{P_w} + \mathbf{t} \tag{2.2}$$

The rotation matrix $\mathbf{R}$ encodes the camera's orientation, while the translation vector $\mathbf{t}$ specifies its position in space. Together, they form the rigid transformation that aligns the world and camera coordinate systems, enabling the projection of three-dimensional real-world points onto the two-dimensional image plane of a given camera while maintaining a shared world coordinate system between cameras.

2.3.2 Reconstructing Static 3D from 2D RGB Data

Reconstructing 3D data from RGB images is a foundational challenge in computer vision, primarily addressed through Structure from Motion (SfM) and Multi-View Stereo (MVS) approaches. These methodologies leverage the spatial and temporal variance in 2D images to infer the depth and structure of the scene, allowing for reconstructing volumetric 3D models from a collection of 2D images.

The fundamental principle behind SfM lies in the observation that the relative motion between the observer and the scene across multiple images can be used to infer the 3D structure of the scene. Mathematically, this is achieved by estimating each camera's extrinsic parameters (rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$) and the scene's geometry in terms of 3D points $\{\mathbf{P}_i\}$ simultaneously. The process begins with feature detection and matching across images, using algorithms such as SIFT [54] or ORB [55], to identify distinctive and salient key points in all images. A descriptor is computed for each key point, which can then be used to match corresponding points across images. Figure 2.10

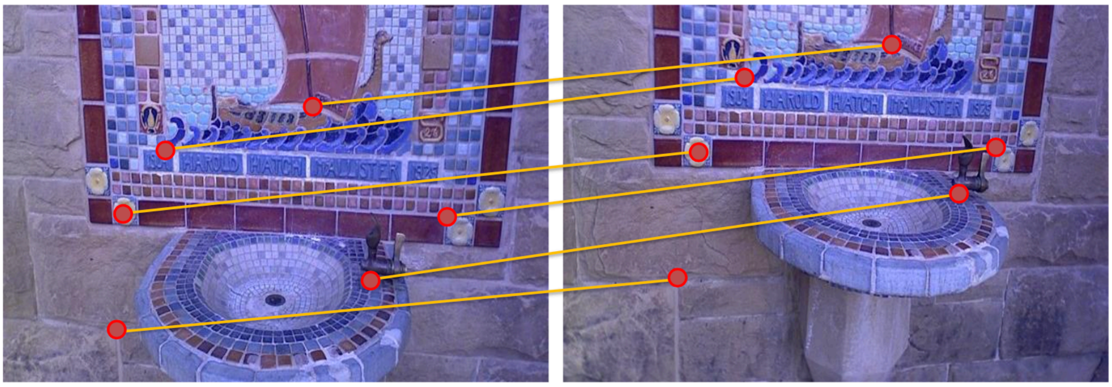


Figure 2.10: Salient feature point detection and matching across different views as part of a structure-from-motion pipeline. For optimal results, these feature points must be distinctive and detectable across views, and their descriptors sufficiently expressive to be matched to the corresponding point in different views.

visualizes this keypoint detection and matching process. Point correspondences are used to estimate the camera poses and a sparse 3D point cloud of the scene.

This estimation problem is inherently non-linear and is typically approached through iterative optimization techniques such as bundle adjustment, which refines the camera parameters and 3D point positions to minimize the reprojection error – the discrepancy between the observed feature positions in the images and the projected positions of the estimated 3D points. The reprojection error E for a point i in image j can be expressed as:

$$E_{ij} = \mathbf{p}_{ij} - \pi(\mathbf{K}, \mathbf{R_j}, \mathbf{t_j}, \mathbf{X_i}) \tag{2.3}$$

where $\mathbf{p}_{ij}$ is the observed position of point i in image j, π denotes the projection function, $\mathbf{K}$ is the camera's intrinsic matrix, $\mathbf{R_j}$ and $\mathbf{t_j}$ are the camera's extrinsic parameters consisting of rotation and translation for image j, and $\mathbf{X_i}$ is the 3D position of point i.

Following the establishment of a sparse 3D model via SfM, MVS techniques can then be used to densify this model by exploiting the texture and appearance information across the images. MVS assumes that the scene's geometric structure can be further detailed by examining the photometric consistency of multiple views. This is achieved by evaluating the similarity of image patches around the corresponding points across different images, considering variations in viewpoint and illumination. The core idea is to generate a depth map for each image by comparing its view with other images and then merge these depth maps into a dense 3D model.

The quality of both SfM and MVS decreases in the presence of occlusions, textureless surfaces, and varying lighting conditions, which significantly impact the accuracy of the final 3D reconstruction. Advanced implementations incorporate robust outlier detection mechanisms and global optimization frameworks to mitigate these issues.

Despite these challenges, reconstructing 3D data from 2D images has the advantage of not requiring special hardware since any digital camera can be used to capture data.

These reconstruction methods are thus most suitable in real-world environments where building a dedicated capture setup is impossible or would interfere with its naturalness, e.g., a confined kitchen setting where human actions are to be captured.

2.3.3 Reconstructing Static 3D from Monocular Depth Data

With the increased availability of specialized hardware sensors, directly capturing depth information alongside 2D RGB images (i.e., capturing RGB-D data) has recently become an alternative approach to Structure from Motion for 3D geometry reconstruction.

Capturing RGB-D data involves the acquisition of color (RGB) information along with depth (D) values for each pixel in an image. The depth, which indicates the distance of geometric surfaces from the sensor, can be acquired using various sensing technologies. Among the most popular are structured light sensors, time-of-flight (ToF) cameras, and stereo vision systems.

Structured light sensors project a known pattern of light onto the scene and observe the deformation of this pattern on surfaces. The sensor can infer the depth of objects in the scene by analyzing these deformations. The depth d at each pixel can be computed using triangulation between the known pattern, its deformation, and the angle of projection and observation. This method is sensitive to ambient light and may have difficulties with surfaces that are either too reflective or too absorbent.

On the other hand, time-of-flight cameras measure the time it takes for a light signal (often infrared) to travel from the camera to the objects in the scene and back. The depth d is calculated as

$$d = \frac{c \cdot \Delta t}{2} \tag{2.4}$$

where c is the speed of light and Δt is the measured time difference. ToF cameras can quickly acquire depth information over large areas but are limited by lower spatial resolution and potential inaccuracies due to multiple reflections or interference.

Stereo vision systems use two or more cameras spaced apart to simulate human binocular vision. The system computes the depth of points in the scene by matching features between the different camera views and using triangulation. The depth calculation relies on the disparity between corresponding points in the images, where

$$d = \frac{f \cdot B}{D} \tag{2.5}$$

with f being the focal length, B the baseline distance between cameras, and D the disparity. Stereo vision systems are versatile and can be used in scenarios where other sensors might fail (e.g., in outdoor settings) but require complex and computationally intensive image processing to resolve depth accurately.

Once RGB-D data is captured, reconstructing 3D data from these inputs typically involves volumetric integration techniques. A fundamental approach in this domain is using a Truncated Signed Distance Function (TSDF) to represent the scene. As outlined in section 2.1, a TSDF encodes the distance of a point in space to the nearest surface boundary; distances are negative if the point is inside the surface and positive if outside, with the function truncated at a certain threshold away from the surface to focus on near-surface regions.

The volumetric integration process accumulates depth measurements from multiple viewpoints into a unified 3D model. The corresponding TSDF value in a volumetric grid is updated for each depth measurement, effectively blending the information from all observed angles. The integration can be formalized as an iterative update of the TSDF values $S(\mathbf{x})$ at each voxel $\mathbf{x}$ in the grid, based on the new measurement d and the camera parameters. This iterative process is mathematically expressed as

$$S_{\text{new}}(\mathbf{x}) = \alpha S_{\text{old}}(\mathbf{x}) + (1 - \alpha) S(D(\mathbf{x})) \tag{2.6}$$

where α is a weighting factor determining the influence of new vs. existing data.

The resulting volumetric representation captures the continuous nature of the surfaces in the scene, allowing for the extraction of a mesh or surface model through techniques such as Marching Cubes [42].

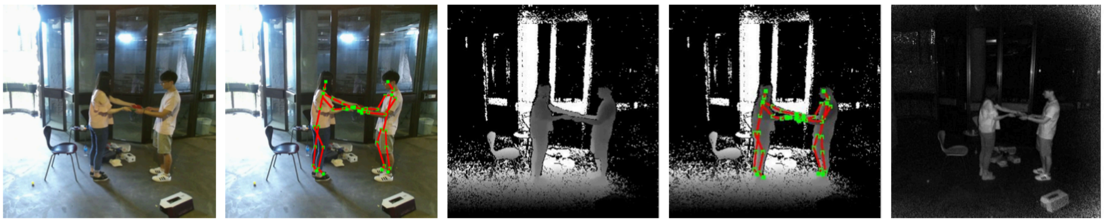


Figure 2.11: Human motion capture using monocular depth data, as used in the NTU RGB-D dataset [56, 57]. From left to right: RGB color image, overlaid detected poses in 2D, depth as captured by the sensor, joint detections from RGB transformed and overlaid on top of the depth image, and the infrared image captured by the sensor.

This technique is widely used for capturing and reconstructing static 3D geometry, and 3D datasets have been created using hand-held depth sensors and advanced reconstruction algorithms. Similar approaches have also been utilized for human motion capture. In SMPL [47], depth sensors were used for scanning human geometries to build their statistical human model. Figure 2.11 shows how the sensors can be used for motion capture specifically: From the 2D RGB image, a 2D skeleton can be detected using robust methods such as OpenPose [58] or AlphaPose [59]. Using the depth information captured for the same frame, these 2D joint detections can then be back-projected into 3D space. This way, human motion and even interactions with the environment can be efficiently captured in 3D.

2.3.4 3D Human Motion Capture

Realistic and accurate human motion capture is of interest in many domains, from creating realistic assets for movies and games to creating a baseline for human simulation and ground-truth data for human motion generation methods. The capturing itself poses multiple challenges, such as how to deal with occlusions, fast motion, and interactions with an environment.

As mentioned in section 2.3.3, it is possible to use commodity sensors for this purpose. However, these are limited in quality due to noise in the depth measurements and ambiguities in the 2D human pose estimation step. Thus, specialized setups exist with specific trade-offs between necessary tracking accuracy, cost, and applicability (i.e., the size of the captured area and the possibility of interacting with the environment). On a high level, the capturing methods employed can be categorized into marker-based systems and markerless systems, each with its distinct advantages, challenges, and computational frameworks.

Marker-based systems rely on physical markers placed at salient points on the human body. The positions of these markers are tracked using multiple calibrated cameras positioned around the subject. The 3D position of each marker is reconstructed through triangulation. The key challenge in marker-based systems is accurately tracking markers across frames, which is typically addressed through optimization algorithms that minimize the error between the observed marker positions and their predicted positions based on prior movements.

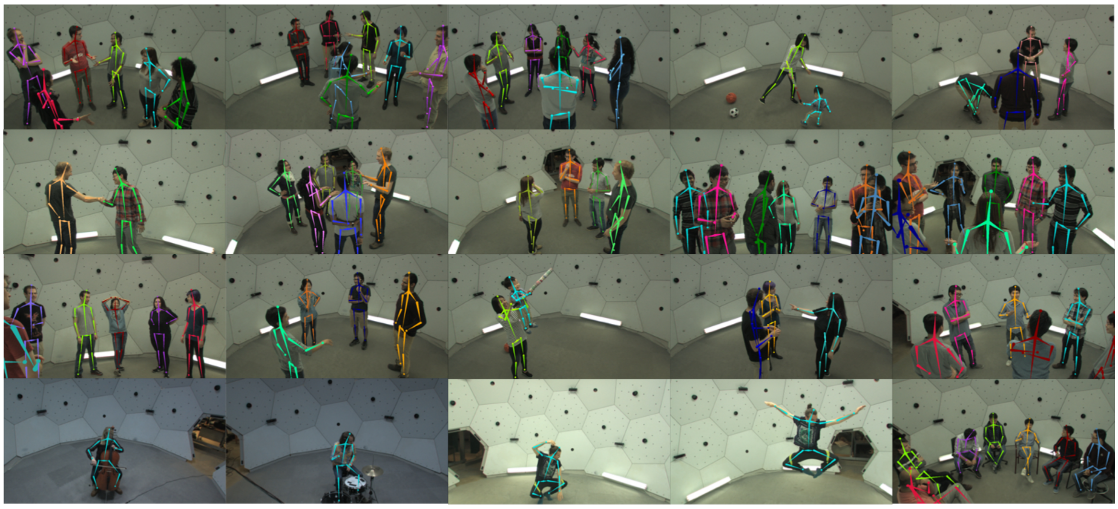


Figure 2.12: Specialized multi-camera setup for capturing human motion, e.g., the human motion capture dome for the CMU Panoptic Dataset [60].

On the other hand, Markerless systems do not require physical markers but instead utilize sophisticated image processing and machine learning algorithms to identify and track body parts directly from video data. These systems leverage techniques such as deep learning-based pose estimation, where a neural network model is trained to predict the 3D coordinates of body joints from 2D images. The complexity of markerless systems lies in their need for large amounts of training data and the computational power required to process video frames in real time.

Additionally, some approaches use Inertial Measurement Units (IMUs), small, wearable devices equipped with accelerometers, gyroscopes, and sometimes magnetometers. These devices measure linear acceleration, angular velocity, and magnetic fields, respectively, allowing for the calculation of orientation and displacement over time. IMUs are particularly valued for their portability and the ability to capture motion outside laboratory settings. However, cumulative errors, known as drift, can affect their accuracy over time.

Data from several different sensors can be combined using special sensor fusion methods for more robust and accurate measurements. Temporal filters like the Kalman Filter further improve accuracy by minimizing the difference between predicted and observed values.

Figure 2.12 and 2.13 show examples of different motion capture systems. The motion capture dome used for the CMU Panoptic Dataset [60] (figure 2.12) is a highly complex system with 480 VGA camera views, 30+ HD views, 10 RGB-D sensors, hardware-based time synchronization and consistent calibration. It is a marker-based system and can capture multiple different people interacting with each other as well as with objects such as musical instruments. Human3.6M [45], on the other hand, followed a more straightforward approach and used four calibrated and synchronized HD cameras in a defined capture space. This dataset is focused on single human actions without human-object interactions.

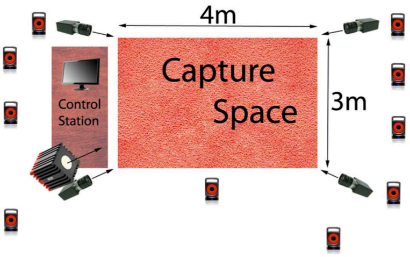


Figure 2.13: The capture setup of Human3.6M [45] uses multiple fixed cameras installed around an empty, pre-defined capture area.

2.4 Modeling 3D Human Motion

3D Human motion modeling is a problem at the intersection of computer graphics and vision that has gained increased attention in recent years. Following successes of deep learning in several computer vision tasks such as image classification and segmentation as well as machine learning problems such as time series forecasting, initial works focused on modeling human motion using recurrent neural network formulations [28]. These early experiments were followed by exploiting inherent dependencies of human motion, such as the human kinematic chain and interdependent joint movements, by modeling them with graph networks and attention mechanisms [14, 15, 61]. Finally, transformer-based methods and diffusion approaches have revolutionized the domain, allowing for longer and more natural human motion generation [62, 1].

This chapter focuses on these fundamental techniques used for human motion modeling and lays the groundwork for the methods presented in part II.

2.4.1 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a class of neural networks tailored for processing sequences, be it natural language texts, time-series data, or any other form of sequential information. Unlike feedforward neural networks, where the flow of information is unidirectional and does not retain any memory of past inputs, RNNs incorporate a mechanism to hold onto information from previous steps. This memory feature is achieved through loops within the network architecture, updating a hidden state with each step.

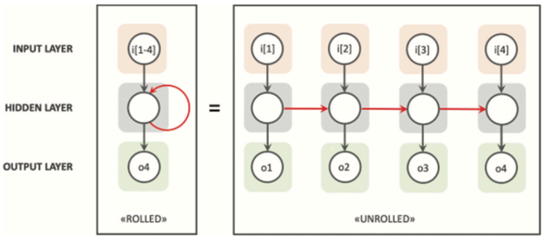


Figure 2.14: The basic idea of a recurrent neural network, comprised of an input, hidden, and output layer. Right: Unrolled along the temporal axis. Visualization from [63].

In the simplest form, an RNN can be mathematically described at each time step t as follows: Let $\mathbf{x}_t$ denote the input vector at time step t, $\mathbf{h}_t$ the hidden state, which acts as the network's memory, and $\mathbf{y}_t$ the output. The hidden state $\mathbf{h}_t$, is updated with all input states up to time step t as:

$$\mathbf{h}_t = f(\mathbf{W}_{hh}\mathbf{h}_{t-1} + \mathbf{W}_{xh}\mathbf{x}_t + \mathbf{b}_h) \tag{2.7}$$

where f is a non-linear activation function, typically tanh or ReLU. $\mathbf{W}_{hh}$ and $\mathbf{W}_{xh}$ represent the learned weight matrices from the previous hidden and current input state, and $\mathbf{b}_h$ is the bias. The output at each time step t can be calculated as:

$$\mathbf{y}_t = g(\mathbf{W}_{hy}\mathbf{h}_t + \mathbf{b}_y) \tag{2.8}$$

where g is the activation function for the output layer, $\mathbf{W}_{hy}$ the weight matrix from the hidden state, and $\mathbf{b}_y$ the bias. Figure 2.14 shows a visualization of this concept.

Despite their ability to model sequences, RNNs suffer from significant limitations, notably the difficulties in learning long-term dependencies. Long-Short-Term Memory (LSTM) [64] networks and Gated Recurrent Units (GRUs) [65] are two popular architectures that mitigate these issues.

LSTMs' [64] design goal is to capture long-term dependencies. This is achieved through a more sophisticated cell structure, including several gates: the input gate, the forget gate, and the output gate. These gates regulate the flow of information into and out of the cell and the retention of information across time steps, allowing the network to learn

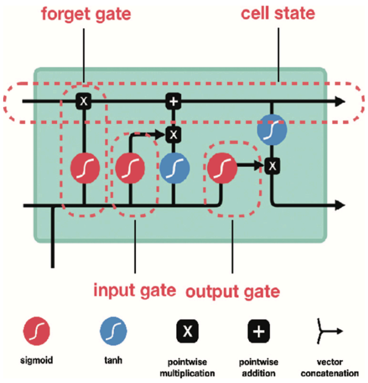


Figure 2.15: LSTM cells better capture long-term dependencies using explicit input, forget, and output gates. Visualization from [63].

when to forget previous information and when to update the hidden state based on new information. The LSTM updates for time step t are mathematically defined as:

$$\begin{aligned}
\text{Forget gate: } \mathbf{f}_t &= \sigma(\mathbf{W}_f[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f) \\
\text{Input gate: } \mathbf{i}_t &= \sigma(\mathbf{W}_i[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i) \\
\text{Output gate: } \mathbf{o}_t &= \sigma(\mathbf{W}_o[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o) \\
\text{Cell state update: } \tilde{\mathbf{C}}_t &= \tanh(\mathbf{W}_C[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_C) \\
\text{Final cell state: } \mathbf{C}_t &= \mathbf{f}_t * \mathbf{C}_{t-1} + \mathbf{i}_t * \tilde{\mathbf{C}}_t \\
\text{Hidden state update: } \mathbf{h}_t &= \mathbf{o}_t * \tanh(\mathbf{C}_t)
\end{aligned} \tag{2.9}$$

Here, σ denotes the sigmoid activation function, and $*$ denotes element-wise multiplication. **W** and **b** represent each gate's weight matrices and bias vectors, respectively. Figure 2.15 visualizes the gates in red as well as cell state updates.

The advantages of LSTMs over traditional RNNs lie in their ability to capture longer-range dependencies, making them more effective for various applications such as language modeling and machine translation. However, LSTMs come with increased computational complexity and parameter count, making them impractical for large models.

GRUs [65] offer a more streamlined alternative to LSTMs, with fewer parameters and gates, by combining the forget and input gates into a single "update gate" and merging the cell state and hidden state, thus reducing the complexity of the model without a significant drop in performance. Thus, their update rules look like this:

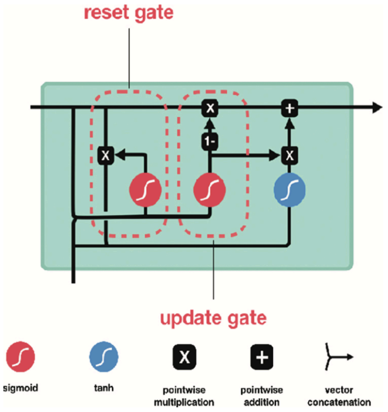


Figure 2.16: GRUs simplify the RNN architecture compared to LSTMs and only use reset and update gates. Visualization from [63].

$$\begin{aligned}
\text{Update gate: } \mathbf{z}_t &= \sigma(\mathbf{W}_z[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_z) \\
\text{Reset gate: } \mathbf{r}_t &= \sigma(\mathbf{W}_r[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_r) \\
\text{Candidate hidden state: } \tilde{\mathbf{h}}_t &= \tanh(\mathbf{W}_h[\mathbf{r}_t * \mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_h) \\
\text{Final hidden state: } \mathbf{h}_t &= (1 - \mathbf{z}_t) * \mathbf{h}_{t-1} + \mathbf{z}_t * \tilde{\mathbf{h}}_t
\end{aligned} \tag{2.10}$$

with σ the sigmoid activation function, and $*$ element-wise multiplication. Analogous to LSTMs, learned weight matrices **W** and biases **b** exist for each gate. Note the decreased complexity in figure 2.16 as compared to LSTMs' architecture.

The introduction of LSTMs and GRUs has significantly advanced the fields of time-series analysis and natural language processing. For 3D human motion forecasting and generation, these advances enabled realistic and plausible results [13, 27, 28].

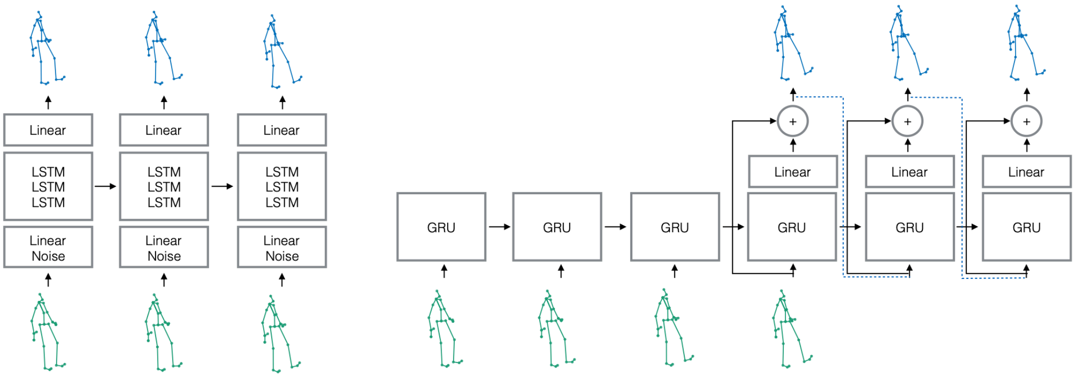


Figure 2.17: Using RNN architectures for forecasting 3D human pose sequences (blue) from ground-truth sequences (blue). Left: Three-layer LSTM architecture proposed in [13]. Right: Improved residual architecture based on GRUs proposed by [28]. Visualization from [28].

Human motion generation is inherently time-based, usually predicting a fixed number of human poses, given a condition such as text or past motion observation. Early works focused on representing the human body in its simplified skeleton representation, as detailed in section 2.2. One of the earliest works using RNNs [13] showed that LSTMs can be used to regress locations of joints in the human skeleton. Martinez et al. [28] later showed that simple combinations of RNN blocks are often outperformed by simple statistical baselines and proposed an alternative architecture based on GRU blocks for improved human motion forecasting, shown in figure 2.17. Despite the applicability of RNNs for human motion generation, limitations around long-term memory remain, and subsequent research has focused on alternative approaches, as discussed in section 2.4.2.

2.4.2 Graph Networks

Graph neural networks (GNNs) are a class of neural networks for learning representations on graphs, capturing the dependencies among nodes through the graph's structural information. This approach effectively models complex systems across various domains, including social networks, biological networks, and knowledge graphs. The foundational principle of GNNs lies in aggregating feature information from a node's neighborhood, facilitating the learning of node representations that incorporate local structure and feature information.

The node-edge update strategy forms the core of traditional GNN architectures. Consider a graph $G = (V, E)$, where V denotes the set of vertices or nodes and E represents the set of edges. Each node $v \in V$ is associated with a feature vector $\mathbf{x}_v$. The GNN iteratively updates the representation of each node by aggregating features from its neighbors, followed by a transformation through a neural network. The update rule for a node v in the k-th iteration is given by:

$$\mathbf{x}_v^k = \sigma\left(\mathbf{W}^k \cdot \mathrm{AGG}\left(\{\mathbf{x}_u^{k-1} : u \in \mathcal{N}(v)\}\right) + \mathbf{b}^k\right) \tag{2.11}$$

where $\mathbf{x}_v^k$ is the feature representation of node v at the k-th iteration, $\mathcal{N}(v)$ denotes the set of neighbors of v, $\mathbf{W}^k$ and $\mathbf{b}^k$ are the weight matrix and bias vector, respectively, σ is a non-linear activation function, and AGG is an aggregation function, such as sum, mean, or max, which combines the feature vectors of the neighboring nodes.

Graph convolutional networks (GCNs) [66, 67], a variant of GNNs, extend this concept by leveraging the convolution operation defined in the spectral domain. GCNs assume that graph structure can be captured by learning a function on the graph Laplacian, enabling filtering signals (node features) on the graph. The convolution operation in GCNs for a single layer can be expressed as:

$$\mathbf{H}^k = \sigma\left(\hat{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{k-1}\mathbf{W}^{k-1}\right) \tag{2.12}$$

where $\mathbf{H}^k$ is the matrix of node features in the k-th iteration, $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}_N$ is the adjacency matrix of the graph G with added self-connections represented by the identity matrix $\mathbf{I}_N$, $\hat{\mathbf{D}}$ is the diagonal node degree matrix of $\hat{\mathbf{A}}$, and $\mathbf{W}^k$ is the weight matrix at the k-th layer. This formulation enables the operation to scale with the size of the graph, making it computationally efficient.

Both graph update variants show how GNNs can use graph structure to learn node representations. The key difference lies in their approach to aggregating neighborhood information; while traditional GNNs rely on direct aggregation from spatial neighbors, GCNs implement a form of spectral convolution that leverages the graph's Laplacian. This allows GCNs to capture global graph properties through local operations, offering a powerful mechanism for learning representations in graph-structured data. Through iterative updates and the integration of neighborhood information, GNNs consider both the topological structure and feature information of graphs.

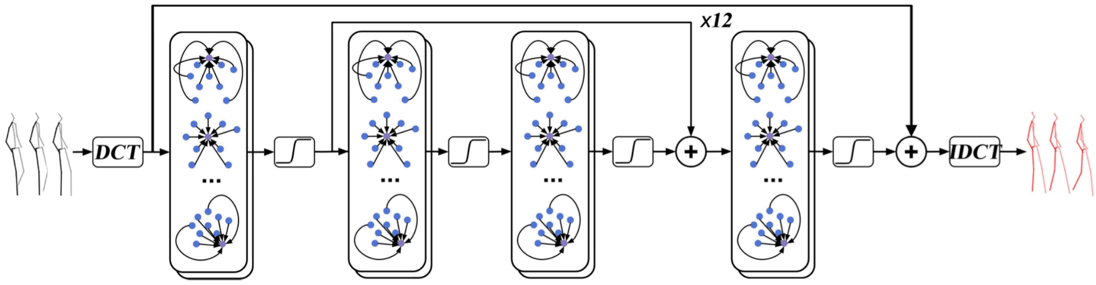


Figure 2.18: Forecasting human motion with graph convolution networks, as proposed by [14]. The location of joints in 3D space is first encoded with discrete cosine transforms before being processed with multiple blocks of residual graph convolutions over the human skeleton. This method can exploit not only explicit but also implicit connections between joints.

This property also makes GNNs useful for 3D human motion generation: The human skeleton exhibits an inherent graph structure, both through the explicit bone connections (e.g., the motion of the hand depends on the motion of joints along the arm and torso) as well as through implicit connections (e.g., when walking, the left hand moves forward when the right foot moves forward).

Mao et al. [14] proposed to use GCNs for human motion forecasting. Their architecture is depicted in figure 2.18. By learning the adjacency matrix **A** during training, not only explicit connections along the kinematic chain of the human body are utilized but also semantic ones, i.e., how different body joints influence each other for different action categories. This works well on a per-frame basis; however, one major disadvantage of graph structures is the quadratic memory requirement with large numbers of nodes (joints over time in this case). Thus, research has shifted towards alternative attention approaches in recent years, as detailed in section 2.4.3.

2.4.3 Attention Mechanisms

The concept of attention is used in many neural network architectures nowadays, significantly enhancing the capability of models to focus selectively on parts of the input data that are most important for performing a given task, allowing for more efficient allocation of computational resources. First formalized for natural language processing (NLP) tasks [68], this mechanism is now widely used in many machine learning and computer vision tasks.

Formally, attention can be described as a function that computes a weighted sum of values V, based on a set of queries Q, and keys K. The attention weights are obtained by computing the dot product of each pair of query and key, typically followed by a softmax function to ensure that the weights sum to one. This process can be mathematically represented as follows:

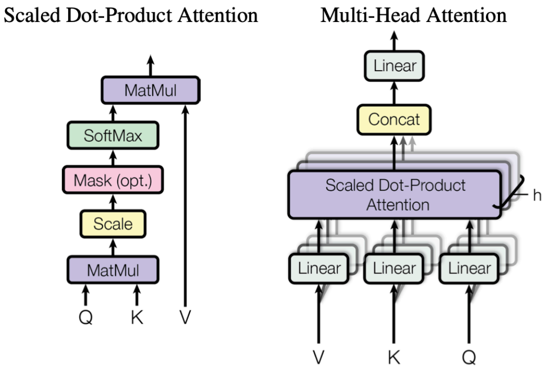


Figure 2.19: Left: Scaled Dot-Product Attention [68] based on query, key, and value representations. Right: Multi-head attention, consisting of multiple parallel attention layers.

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{2.13}$$

where d_k is the dimensionality of the keys, which is used to scale the dot product, improving gradient stability during training. This equation shows the most fundamental version of scaled dot-product attention. Multi-head attention was initially proposed for improved performance in [68]. It aims to capture information from different representation subspaces at different positions. In multi-head attention, the queries, keys, and values are linearly projected multiple times with different, learned linear projections to d_k, d_k, and d_v dimensions. Then, the attention function is applied in parallel to each set of projections, yielding multiple output vectors that are subsequently concatenated and linearly transformed. This process can be represented as:

$$\begin{aligned} \mathrm{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) &= \mathrm{Concat}(\mathrm{head}_1,\dots,\mathrm{head}_h)\mathbf{W}^O \\ \text{where } \mathrm{head}_i &= \mathrm{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^{\mathbf{K}}, \mathbf{V}\mathbf{W}_i^{\mathbf{V}}) \end{aligned} \tag{2.14}$$

where $\mathbf{W}_i^Q$, $\mathbf{W}_i^K$, $\mathbf{W}_i^V$, and $\mathbf{W}^O$ are weight matrices for the i-th head. Using multi-head attention allows a neural network to focus on multiple aspects of the input data instead of modeling everything with just one attention map, improving overall model performance. Figure 2.19 shows both single and multi-head attention visualizations.

Generating 3D human motion can benefit from these mechanisms similarly to the originally intended domain of natural language processing. Analogous to approaches using graph networks detailed above, attention can be used to process dependencies within the human body skeleton [40, 31] or even to attend to the most salient parts of a given motion sequence, applying attention along the temporal axis [15]. While these methods integrate the original attention formulation of [68], most recent approaches focus on using the more powerful Transformer models presented in section 2.4.4.

2.4.4 Transformer-Based Methods

The Transformer network architecture, introduced alongside and designed around Scaled Dot-Product Attention in [68] (section 2.4.3), can capture long-range dependencies and handle sequential data in parallel, a significant advantage over traditional RNN-based models.

In generative settings, Transformers have paved the way for models capable of producing high-fidelity images, music, and text. Models like DALL-E [69] and GPT (Generative Pre-trained Transformer) [70] leverage variations of the Transformer architecture to understand and generate content that is not only coherent but also creative. These models benefit from the scalability of Transformers, allowing them to be trained on vast datasets and capture a wide range of styles and patterns. The original Transformer architecture is visualized in figure 2.20; it consists of an encoder that uses self-attention among inputs to process incoming data and a decoder using mask self-attention as well as cross-attention to incorporate features generated by the encoder.

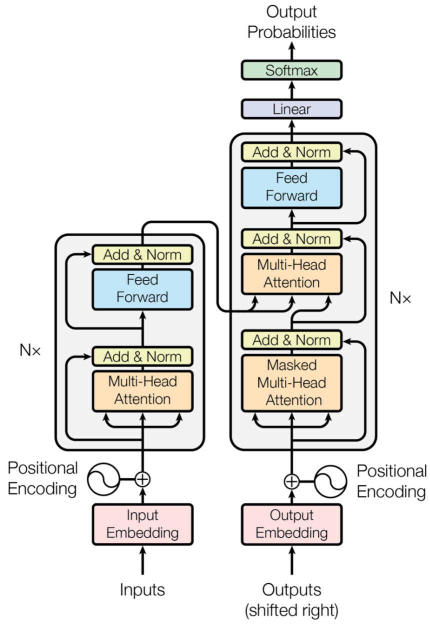


Figure 2.20: The full transformer architecture, as proposed by [68]. It consists of an encoder (left) using self-attention and a decoder (right) using both masked self-attention and cross-attention, additionally attending to features from the encoder.

This original architecture was shown to exhibit high performance on NLP-based tasks. Later adaptations to different domains also explored purely encoder-based and purely decoder-based architectures. Encoder-only architectures like BERT (Bidirectional Encoder Representations from Transformers) [71] excel at tasks that require understanding and interpreting input data due to their ability to encode the input into a rich, contextualized representation. This makes them suitable for tasks like classification, where the goal is to understand the input rather than generate new content. Decoder-only architectures, by contrast, are designed to generate sequences based on some given context. They shine in tasks that require creativity and extrapolation from given data, such as text generation and image synthesis. The absence of an encoder means these models typically operate by conditioning on an encoded representation of the input provided externally or by leveraging a pre-trained encoder model in a two-step process.

Transformer architectures started being widely used for computer vision tasks after the introduction of Vision Transformers (ViTs) [72]. ViTs treat an image as a sequence of patches and apply self-attention mechanisms across these patches to capture spatial hierarchies and relationships, analogous to how words in a sentence are treated in NLP applications. This approach has shown competitive performance with convolutional neural networks on various computer vision tasks, including image classification, object detection, and segmentation, by enabling the model to focus on the most informative parts of an image.

In 3D human behavior generation, transformers have been increasingly used to produce semantically and physically plausible human motion [61, 73, 74, 62, 75, 12, 3]. Especially for denoising diffusion approaches (section 2.4.5), Transformer architectures are nowadays often used as a drop-in replacement for U-Net models. The architecture of one such approach, Human Motion Diffusion, is shown in figure 2.22.

2.4.5 Denoising Diffusion Approaches

Denoising diffusion probabilistic models [76, 77] have emerged as a powerful class of generative models, drawing significant attention for their capability to produce high-quality samples, originally in the domain of image generation. The fundamental idea behind denoising diffusion models involves gradually adding noise to an image or data sample until it closely approximates Gaussian noise and then learning to reverse this process to generate new data samples directly from noise.

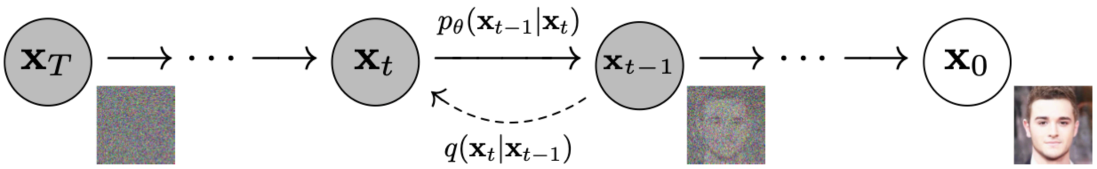

$\mathbf{x}_T$
$\mathbf{x}_t$
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$
$\mathbf{x}_{t-1}$
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$
$\mathbf{x}_0$

Figure 2.21: The denoising diffusion process as a directed graph: From Gaussian noise $\mathbf{x}_T$ at time step T, a reverse process p_θ us used to iteratively denoise towards valid representation $\mathbf{x}_0$ (here, an image of a face) at time step 0. Visualization from [78].

The process of diffusion in these models is divided into two main phases: the forward process (also known as the noising phase) and the reverse process (the denoising phase). The forward process is modeled as a Markov chain that incrementally adds Gaussian noise to the data over a series of time steps, T, effectively transforming the data into an isotropic Gaussian distribution. This can be formalized as:

$$\mathbf{x}_t = \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\epsilon_t \tag{2.15}$$

where $\mathbf{x}_t$ represents the data at time step t, β_t is a variance schedule that controls the amount of noise added at each step, and ϵ_t is sampled from a standard Gaussian distribution. The forward process is known to be Markovian, as each step depends only on the previous state.

The reverse process aims to learn the conditional distribution $p(\mathbf{x}_{t-1}|\mathbf{x}_t)$ to gradually denoise the data, eventually reconstructing the original data distribution from pure noise. This is achieved by training a neural network (usually a U-Net or Transformer architecture) to predict the noise ϵ_t added at each step of the forward process, effectively learning to reverse the diffusion process. This process can be described as:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\beta_t}}\epsilon_\theta(\mathbf{x}_t, t)) \tag{2.16}$$

where $\epsilon_\theta(\mathbf{x}_t, t)$ is the noise predicted by the neural network parameterized by θ. Figure 2.21 visualizes both processes as directed graphs for the task of image generation.

A critical aspect of denoising diffusion models is the choice of the variance schedule β_t, which significantly influences the quality of the generated samples. The schedule must be carefully designed to ensure the forward process creates a smooth transformation from data to noise and vice versa for the reverse process.

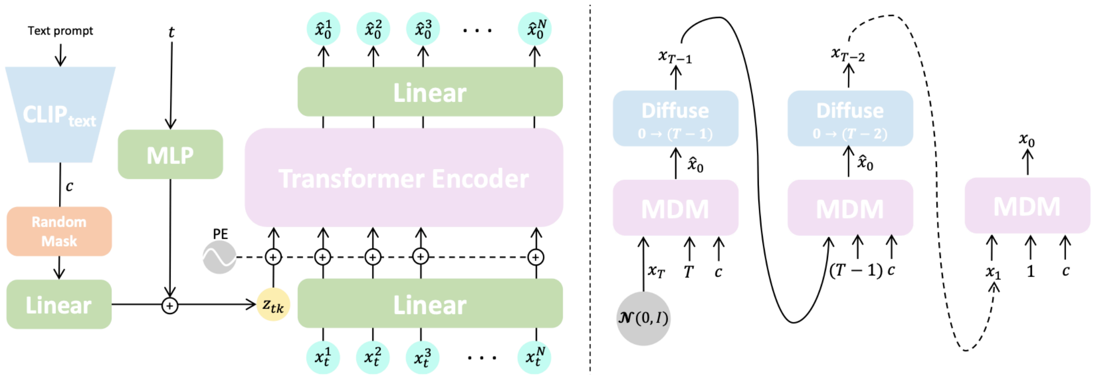


Figure 2.22: Human motion diffusion model [1] overview. Left: Network architecture, using a transformer encoder conditioned on text and the current time step. Right: Using this model during inference to iteratively sample human motion sequences from noise.

Recent advancements in denoising diffusion models have shown remarkable success in image generation tasks. These models have been able to generate images of unprecedented quality and diversity, outperforming previous generative models in many benchmarks. The success of diffusion models in image generation can be attributed to their ability to model complex, high-dimensional data distributions and their robustness to mode collapse. Recent research has also explored extensions to conditional generation tasks, where the model generates images based on specific attributes or text conditioning.

In the domain of 3D human motion generation, diffusion approaches have been used for a wide variety of applications, from the pure generation of human motion sequences to interactions with objects and whole scenes [3, 34, 35, 36, 37, 38, 39, 1, 75, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92]. Analogous to the setting of image generation, diffusion models make it possible to generate human motion sequences from noise, both unconditionally and conditioned on signals such as text or the surrounding objects and scenes. Figure 2.22 shows the architecture of the popular approach Human Motion Diffusion Model [1]. Here, a short text description is first encoded as the condition, and the motion sequence is subsequently generated from noise by a transformer encoder network.

2.4.6 Adversarial Regularization

Adversarial regularization leverages the concept inherent in Generative Adversarial Networks (GANs) [93], where two neural networks, namely the generator G and the discriminator D, are working against each other in a game-theoretic scenario. The generator aims to produce data indistinguishable from real data, while the discriminator evaluates the authenticity of the samples, thus engaging in a min-max optimization problem formalized as

$$\min_G \max_D V(D, G) \tag{2.17}$$

where $V(D, G)$ represents the value function indicating the discriminator's ability to distinguish real from generated data.

This framework is extended to adversarial regularization by applying the adversarial principle to enforce constraints or regularize models in supervised and unsupervised learning tasks beyond generating synthetic data. This is especially important in domains with limited data where the aim is to perform unsupervised or weakly supervised learning, e.g., to generate sequences of realistic 3D human motion with only 2D datasets for supervision available.

In this case, learning methods aim to minimize the objective

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda \mathcal{L}_{adv} \tag{2.18}$$

where $\mathcal{L}_{task}$ is the primary task loss, $\mathcal{L}_{adv}$ is the adversarial loss that penalizes the model for producing unrealistic samples of the target domain, and λ is a regularization coefficient. Using this training objective, it is possible to perform weakly supervised learning with an unseen target domain. For 3D human behavior forecasting, this has been shown to be effective for generating sequences of realistic 3D poses with only 2D action data and an uncorrelated database of 3D poses available [41].

Part II

3D Human Behavior Understanding

3 Characteristic 3D Poses of Human Actions

This chapter introduces the following paper:

Christian Diller, Thomas Funkhouser, and Angela Dai. “Forecasting Characteristic 3D Poses of Human Actions” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 15914-15923

Abstract of Paper We propose the task of forecasting characteristic 3D poses: from a short sequence observation of a person, predict a future 3D pose of that person in a likely action-defining, characteristic pose – for instance, from observing a person picking up an apple, predict the pose of the person eating the apple. Prior work on human motion prediction estimates future poses at fixed time intervals. Although easy to define, this frame-by-frame formulation confounds temporal and intentional aspects of human action. Instead, we define a semantically meaningful pose prediction task that decouples the predicted pose from time, taking inspiration from goal-directed behavior. To predict characteristic poses, we propose a probabilistic approach that models the possible multi-modality in the distribution of likely characteristic poses. We then sample future pose hypotheses from the predicted distribution in an autoregressive fashion to model dependencies between joints. To evaluate our method, we construct a dataset of manually annotated characteristic 3D poses. Our experiments with this dataset suggest that our proposed probabilistic approach outperforms state-of-the-art methods by 26% on average.

Contribution The method development and implementation was done by the first author. Discussions with the co-authors led to the final paper.

3.1 Introduction

Future human pose forecasting is fundamental towards a comprehensive understanding of human behavior, and consequently towards achieving higher-level perception in machine interactions with humans, such as autonomous robots or vehicles. In fact, prediction is considered to play a foundational part in intelligence [4, 5, 6]. In particular, predicting the 3D pose of a human in the future lays a basis for both structural and semantic understanding of human behavior, and for an agent to take fine-grained anticipatory action towards the forecasted future. For example, a robotic surgical assistant should predict in advance where best to place a tool to assist the surgeon's next action, what sensor viewpoints will be best to observe the surgeon's next manipulation, and how to position itself to be out of the way at critical future moments.

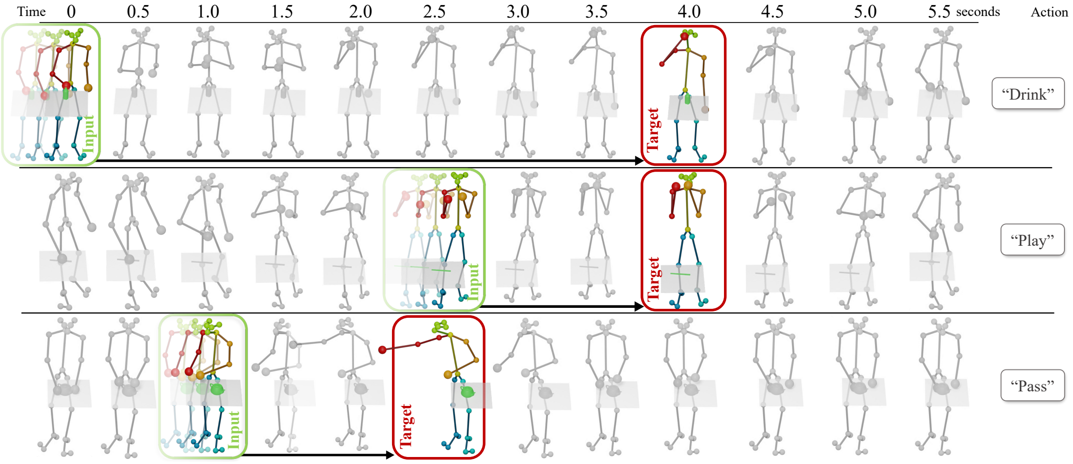


Figure 3.1: For a real-world 3D skeleton sequence of a human performing an action, we propose to forecast the semantically meaningful *characteristic 3d pose*, representing the action goal for this sequence. As input, we take a short observation of a sequence of consecutive poses leading up to the target characteristic pose. Thus, we propose to take a *goal-oriented* approach, predicting the key moments characterizing future behavior, instead of predicting continuous motion, which can occur at varying speeds with predictions more easily diverging for longer-term (>1s) predictions. We develop an attention-driven probabilistic approach to capture the most likely modes of possible future characteristic poses.

Recently, we have seen notable progress in the task of future 3D human motion prediction – from an initial observation of a person, forecasting the 3D behavior of that person up to ≈ 1 second in the future [13, 27, 28, 14, 15]. Various methods have been developed, leveraging RNNs [13, 27, 28, 29], graph convolutional neural networks [14, 30], and attention [31, 15]. However, these approaches all take a temporal approach towards forecasting future 3D human poses, and predict poses at fixed time intervals to imitate the fixed frame rate of camera capture. This makes it difficult to predict longer-term (several seconds) behavior, which requires predicting both the time-based speed of movement as well as the higher-level goal of the future action.

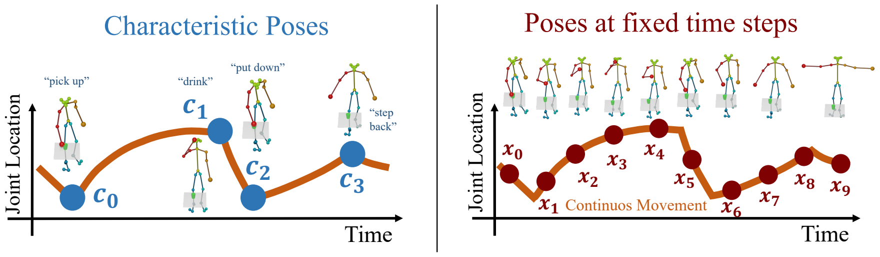


Figure 3.2: These plots show the salient difference between our new task (left) and the traditional one (right). The orange curve depicts the motion of one joint (e.g., hand position as a person drinks from a glass). It represents a typical piecewise continuous motion, which has discrete action-defining characteristic poses at cusps of the motion curves (e.g., grasping the glass on the table, putting it to ones mouth, etc.) separating smooth trajectories connecting them (e.g., raising or lowering the glass). Our task is to predict future characteristic poses (blue dots on left) rather than in-between poses at regular time intervals (red points on right).

Thus, we propose to decouple the temporal and intentional behavior, and introduce a new task of forecasting *characteristic 3D poses* of a person's future action: from a short pose sequence observation of a human, the goal is to predict a future pose of the person in a characteristic, action-defining moment. This has many potential applications, including HRI, surveillance, visualization, simulation, and content creation. It could be used to predict the hand-off point when a robot is passing an object to a person; to detect and display future poses worthy of alerts in a safety monitoring system; to coordinate grasps when assisting a person lifting a heavy object; to assist tracking through occlusions; or to predict future keyframes, as is done in video generation [32, 33].

Fig. 3.2 visualizes the difference between this new task and the traditional, time-based approach: our task is to predict a next characteristic pose at action-defining moments (blue dots) rather than at fixed time-intervals (red dots). As shown in Fig. 3.1, the characteristic 3D poses are more semantically meaningful and rarely occur at exactly the same times in the future. We believe that predicting possible future characteristic 3D poses takes an important step towards forecasting human action, by understanding the objectives underlying a future action or movement separately from the speed at which they occur.

Since future characteristic 3D poses often occur at longer-term intervals ($>$ 1s) in the future, there may be multiple likely modes of the characteristic poses, and we must capture this multi-modality in our forecasting. Rather than deterministic forecasting, as is an approach in many 3D human pose forecasting approaches [14, 15, 30], we develop an attention-driven prediction of probability heatmaps representing the likelihood of each human pose joint in its future location. This enables generation of multiple, diverse hypotheses for the future pose. To generate a coherent pose prediction across all pose joints' potentially multi-modal futures, we make autoregressive predictions for the end effectors of the actions (e.g., predicting the right hand, then the left hand conditioned on the predicted right hand location) – this enables a tractable modeling of the joint distribution of the human pose joints.

To demonstrate our proposed approach, we introduce a new benchmark on *characteristic 3d pose* prediction. We annotate characteristic keyframes in sequences from the GRAB [44] and Human3.6M [45] datasets. Experiments on this benchmark show that our probabilistic approach outperforms time-based state of the art by 26% on average.

In summary, we present the following contributions:

- We propose the task of forecasting *characteristic 3D poses*: predicting likely next action-defining future moments from a sequence observation of a person, towards goal-oriented understanding of pose forecasting.

- We introduce an attention-driven, probabilistic approach to tackle this problem and model the most likely modes for the next characteristic pose, and show that it outperforms state of the art.

- We autoregressively model the multi-modal distribution of future pose joint locations, casting pose prediction as a product of conditional distributions of end effector locations (e.g., hands), and the rest of the body.

- We introduce a dataset and benchmark on our *characteristic 3D pose* prediction, comprising 1535 annotated characteristic pose frames from the GRAB [44] and Human3.6M [45] datasets.

3.2 Related Work

Deterministic Human Motion Forecasting. Many works have focused on human motion forecasting, cast as a sequential task to predict a sequence of human poses according to the fixed frame rate capture of a camera. For this sequential task, recurrent neural networks have been widely used for human motion forecasting [13, 27, 28, 94, 95, 96, 97]. Such approaches have achieved impressive success in shorter-term prediction (up to $\approx$ 1s, occasionally several seconds for longer term predictions), but the RNN summarization of history into a fixed-size representation struggles to maintain the long-term dependencies needed for forecasting further into the future.

To address some of the drawbacks of RNNs, non-recurrent models have also been adopted, encoding temporal history with convolutional or fully connected networks [98, 99, 14], or attention [31, 15]. Li et al. [100] proposed an auto-conditioned approach enabling synthesizing pose sequences up to 300 seconds of periodic-like motions (walking, dancing). However, these works all focus on frame-by-frame synthesis, with benchmark evaluation of up to 1000 milliseconds. Instead of a frame-by-frame synthesis, we propose a goal-directed task to capture perception of longer-term human action, which not only lends itself towards forecasting more semantically meaningful key moments, but enables a more predictable evaluation: as seen in Fig. 3.1, there can be significant ambiguity in the number of pose frames to predict towards a key or goal pose, making frame-based evaluation difficult in longer-term forecasting.

Multi-Modal Human Motion Forecasting. While 3D human motion forecasting has typically been addressed in a deterministic fashion, several recent works have introduced multi-modal future pose sequence predictions. These approaches leverage well-studied approaches for multi-modal predictions, such as generative adversarial networks [101] and variational autoencoders [102, 16, 103]. For instance, Aliakbarian et al. [103] stochastically combines random noise with previous pose observations, leading to more diverse sequence predictions. Yuan et al. [16] learns a set of mapping functions which are then used for sampling from a trained VAE, leading to increased diversity in the sequence predictions than simple random sampling. In contrast to these time-based approaches, we consider goal-oriented prediction of characteristic poses, and model multi-modality explicitly as predicted heatmaps for body joints in an autoregressive fashion to capture inter-joint dependencies.

Goal-oriented Forecasting. While a time-based, frame-by-frame prediction is the predominant approach towards future forecasting tasks, several works have proposed to tackle goal-oriented forecasting. Recently, Jayaraman et al. [32] proposed to predict “predictable” future video frames in a time-agnostic fashion, and represent the predictions as subgoals for a robotic tasks. Pertsch et al. [33] predict future keyframes representing a future video sequence of events. Cao et al. [104] plan human trajectories from an image and 2d pose history, first predicting 2d goal locations for a person to walk to in order to synthesize the path. Inspired by such goal-based abstractions, we aim to represent 3D human actions as its key, characteristic poses.

3.3 Method Overview

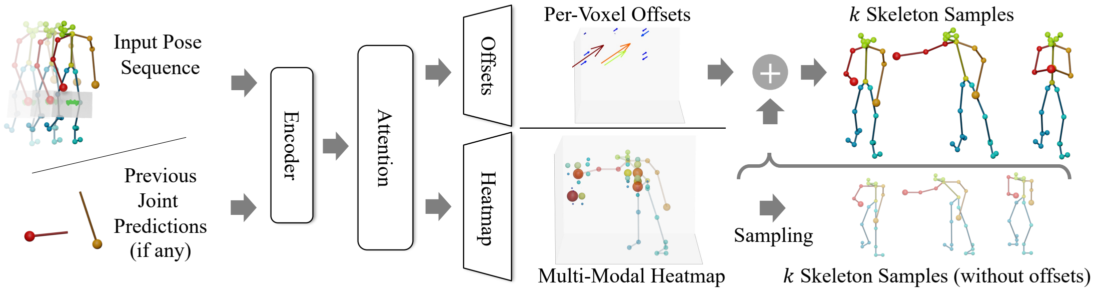


Figure 3.3: Overview of our approach for characteristic 3D pose prediction. From an input observed pose sequence, as well as any prior joint predictions, we leverage attention to learn inter-joint dependencies, and decode a 3D volumetric heatmap representing the probability distribution for the next joint to be predicted as well as a per-voxel offset field of same size for improved joint placement. This enables autoregressive sampling to obtain final pose hypotheses characterizing likely characteristic 3D poses.

Given a sequence of N 3D pose observations $\mathbf{X}_{1:N} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$ of a person, our aim is to estimate a characteristic 3D pose of that person, characterizing the intent of the person's future action. We take J joint locations (represented as their 3D coordinates) for each pose of the input sequence, i.e. $\mathbf{x}_i \in \mathbb{R}^{J\times 3}$. From this input sequence, we predict a joint distribution of J probability heatmaps $\mathbf{H}_j$ and finally, sample K output pose hypotheses $\mathbf{Y}_{1:K}$, characterized by their J 3D joints: $\mathbf{y}_i \in \mathbb{R}^{J\times 3}$. By representing probability heatmaps for the joint predictions, we can capture multiple different modes in likely characteristic poses, enabling more diverse future pose prediction. We note that we are the first to propose using volumetric heatmaps for future human pose forecasting, to the best of our knowledge, while previous work used them for the more deterministic task of pose estimation from multiple images [105, 106].

From the input sequence, we develop a neural network architecture to predict a probability heatmap over a volumetric 3D grid for each joint, corresponding to likely future positions of that joint. This enables effective modeling of multi-modality, but remains tied to a discrete grid, so we also regress a corresponding volume of per-voxel offsets, allowing for precise locations to be sampled. Fig. 3.3 shows an overview of our learned probabilistic predictions.

We model these predictions conditionally in an autoregressive fashion in order to tractably model the joint distribution over all pose joint locations. This enables a consistent pose prediction over the set of pose joints, as a set of joints may have likely modes that are unlikely to be seen all together (e.g., right hand moving forward while the right elbow moves to the side – both are valid independently but not together). To sequentialize the pose joint prediction autoregressively, we first predict probability heatmaps for the end effectors in our dataset – right hand first, then left hand conditioned on the right hand prediction, followed by the rest of the body joints.

3.4 Capturing Multi-Modality with Heatmap Predictions

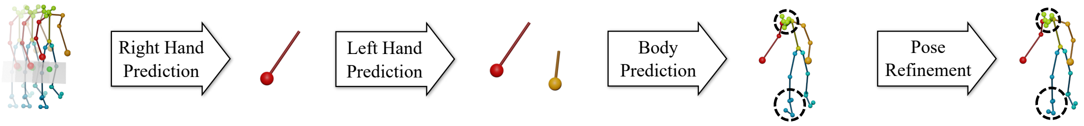


Figure 3.4: To model joint dependencies within the human skeleton, we sample joints in an autoregressive manner by first predicting the end-effectors (right and left hand), then the rest of the body; pose refinement then improves skeleton consistency.

We aim to learn to predict likely future locations for an output pose joint j, characterized by a probability heatmap $\mathbf{H}_j$ over a volumetric grid of possible pose joint locations. From the input sequence of N pose observations of J joints, and conditioned on any already predicted joints, we construct an attention-driven neural network to learn the different dependencies between human skeleton joints to inform the final heatmap prediction.

Attention-Driven Sequence Encoding. We represent the body joints of the input sequence $\mathbf{X}_{1:N} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$ as an $N \times J \times 3$ ($N = 10$ as well as $J = 25$ for the GRAB dataset and $J = 17$ for Human 3.6M, respectively) concatenation of the joint locations over time. Features are first extracted with a single-layer GRU [65]. We then compute an attention map from these features, representing dependencies to the input set of pose joints. This way, the network learns not only how different joints in the skeleton affect each other directly (e.g., kinematic relationships) but also learns to exploit more subtle correlations such as likely positions of one hand with respect to the other. Following the formalism of Scaled Dot-Product Attention [68], popularized in natural language processing, our attention maps are computed from a query $\boldsymbol{Q}$ and a set of key-value pairs $\boldsymbol{K}$ and $\boldsymbol{V}$. During training, representations for $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are learned which are shared between all joints. This allows us to project all joints into the same embedding space where we can then compare the joint of interest (represented by $\boldsymbol{Q}$) with all other joints ($\boldsymbol{K}$) to inform which parts of $\boldsymbol{V}$ (the learned latent representation for all joints which will be passed to the decoder) are relevant for this joint of interest.

$$\mathrm{Attn}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{D}}\right)\boldsymbol{V} = \boldsymbol{A}\boldsymbol{V}, \tag{3.1}$$

Intuitively, the similarity between key and query defines which parts of a learned pose skeleton representation are important for the desired prediction. Formally, this is defined in Eq. 3.1: The value representation $\boldsymbol{V}$ is weighed per-element by the result of the dot-product between $\boldsymbol{Q}$ and $\boldsymbol{K}$ (scaled by the dimension of the embedding vector D and a softmax operation). In our case, the attention map $\boldsymbol{A}$ has a dimensionality of $J' \times N$ with J' indicating the number of joints to be predicted. Any prior joint predictions for autoregressive prediction are considered as an additional node to our attention map, giving the attention map dimension $J' \times (N + n_p)$ for n_p prior joints.

Heatmap Prediction. Based on the attention scoring, we then use a series of nine 3D convolutions to decode an output probability heatmap $\mathbf{H}_j$ for each body joint j. The grids are centered at the skeleton's hip joint; we use a grid size of 16^3 voxels, spanning $2m^3$. A value in the grid of $\mathbf{H}_j$ at location $\mathbf{H}_j(x, y, z)$ corresponds to a probability of joint j being at location (x, y, z) in the future characteristic pose. Instead of directly regressing the probability values, we predict $\mathbf{H}_j(x, y, z)$ as a classification problem by discretizing the output values into $n_{discr} = 10$ bins in the $[0, 1]$ space. We then use a cross entropy loss with the discretized target heatmap to train our heatmap predictions. In our experiments, we found that this classification formulation for $\mathbf{H}_j$ produced better results than an ℓ_2 or ℓ_1 regression loss, as it mitigated tending towards the average or median.

Offset Prediction. Since predicting joint locations in a discrete grid inherently leads to grid artifacts in sampled output poses, we additionally learn an offset field $\mathbf{O}_j$ over the same volumetric grid. Here, each voxel $\mathbf{O}_j(x, y, z) \in \mathbb{R}^3$ represents the shift to be added after sampling a joint from the heatmap at $\mathbf{H}_j(x, y, z)$. We predict these offsets similarly to the heatmap volume, with a series of nine 3D convolutions, and clamp each offset vector $\mathbf{O}_j(x, y, z)$ to move the joint at most one voxel length. Output poses are then estimated by sampling the heatmap, followed by refinement using the corresponding predicted offset.

3.4.1 Training Details

Note that for real-world data captured of human movement, we do not have a full ground truth probability distribution for the future characteristic pose, but rather a set of paired observations of input pose to the target pose. Thus, we generate target heatmap data from a single future observation in the training data by applying a Gaussian kernel (size 5, $\sigma = 2$) over the target joint location. At test time, we apply softmax scaling to the predicted heatmaps with a temperature of 0.025 and from there, sample our final joint locations. We learn multi-modality by generalizing across train set observations which results in seeing multiple possibilities for similar inputs (e.g., right vs. forward pass), encouraging learned heatmaps to represent multiple modes. We show that our formulation can effectively model multi-modal heatmaps in Section 3.7.

We train our models on a single NVIDIA GeForce RTX 2080Ti. We use an ADAM optimizer with a weight decay of 0.001 and a linear warmup schedule for 1000 steps; learning rate is then kept at 0.001. We use a batch size of 100, as a larger batch size helps with training our attention mechanism. Our model trains for up to 8 hours until convergence. During training, we apply teacher forcing, i.e. pose joint predictions conditioned on prior joint predictions are trained using the ground truth locations of the prior joints. For a detailed specification of our network architecture, please refer to the appendix.

3.5 Autoregressive Joint Prediction

Given a set of heatmaps for each pose joint location, the next step is to predict specific joint locations. Since they are not independent of one another, we cannot simply sample joint locations from each heatmap independently. Instead, we must model the interdependencies between pose joints.

To do this, we model the joint distribution of pose joints autoregressively, as visualized in Fig. 3.4: we first predict end effector joints, followed by other body joints. For our experiments, we find that the right and left hands tend to have a large variability, so we first predict the right hand, then the left hand conditioned on the right hand location, followed by the rest of the body joints. Empirically, we found that the hands tended to define the body pose, while the order of the rest has little impact. To sample from a joint heatmap, we use temperature scaling to concentrate the heatmap near its local maxima, followed by random sampling.

Pose Refinement. While our autoregressive pose joint prediction encourages a coherent pose prediction with respect to coarse global structure, pose joints may still be slightly offset from natural skeleton structures. Thus, we employ a pose refinement optimization to encourage the predicted pose to follow inherent skeleton bone length and angle constraints while keeping all joints in areas of high probability and the end-effectors close to their original prediction, as formulated in the objective function:

$$\begin{aligned} \mathrm{E}_R(\mathbf{x}, \mathbf{e}, \mathbf{b}, \mathbf{x_0}, \theta, H) = w_e \|\mathbf{x}_e - \mathbf{e}\|_2 + w_b \|\text{bonelengths}(x) - \mathbf{b}\|_1 \\ + w_a \|\text{angles}(x) - \theta\|_1 + w_c \|x - x_0\|_1 + w_h \Sigma_j (1 - H_j) \end{aligned} \tag{3.2}$$

where $\mathbf{x}$ the raw predicted pose skeleton as a vector of N 3D joint locations; $\mathbf{b}$ and θ the bone lengths and joint angles, respectively, of the initially observed pose skeleton; x_0 the joint locations of the last skeleton in the input sequence; H_j the heatmap probability for each joint; $\mathbf{e}$ the sampled end effector locations; and w_e, w_b, w_a, w_h, w_c weighting parameters (in all our experiments, we use $w_e = 0.2, w_b = 1.0, w_a = 0.4, w_h = 0.1, w_c = 0.1$). We then optimize for $\mathbf{x}$ under this objective to obtain our final pose prediction.

3.6 Characteristic 3D Pose Dataset

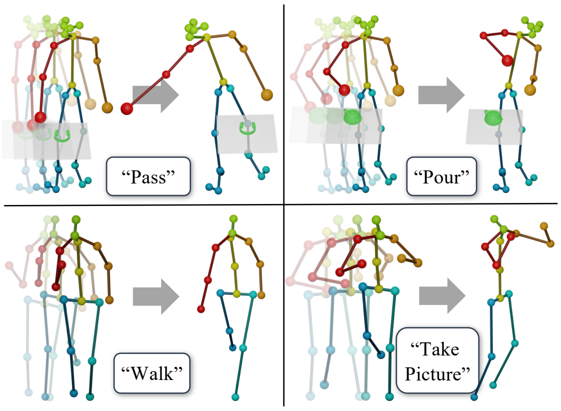


Figure 3.5: Example input observations and target characteristic 3D poses from our annotated datasets, based on GRAB (top) and Human3.6M (bottom).

To train and evaluate the task of characteristic 3D pose forecasting, we introduce a dataset of annotated characteristic poses, built on GRAB [44] and Human3.6M [45].

- **Human3.6M** is a commonly used dataset for human pose forecasting, comprising 210 actions performed by 11 professional actors in 17 scenarios for a total of 3.6 million frames. 3D locations are obtained for 32 joints via a high-speed motion capture system; we use a reduced 17-joint layout in our method, removing redundant and unused joints, following [16].
- **GRAB** is a recent dataset with over 1 million frames in 1334 sequences of 10 different actors performing a total of 29 actions with various objects. Each actor

starts in a T-Pose, moves towards a table with an object, performs an action with the object, and then steps back to the T-Pose. The human motions are captured using modern motion capture techniques, with an accuracy in the range of a few millimeters. GRAB provides SMPL-X [49] parameters from which we extract the 25 most defining body joints. For more details, we refer to the appendix.

We then annotate the timesteps of the captured sequences corresponding to characteristic poses. Input sequence start frames are randomly sampled, up until the characteristic pose frame. Several example input sequence-characteristic pose pairs are visualized in Fig. 3.5. Annotations were performed by the authors, within a time span of one day. This is the total time for annotating more that 1000 sequences across two datasets, with each annotation taking 10-30 seconds; this annotation efficiency enables quick and easy adoption of new datasets in the future. We define a characteristic pose as the point in time when the action is most articulated, i.e. right before the actor starts returning back to another pose (e.g., when the hand is furthest from the person when passing, most tilted when pouring, etc.). For sequences containing multiple occurrences of the same action, like lifting, we chose the repetition with most articulation, e.g. when the object is lifted highest. In the case of Human3.6M, where there are sometimes multiple possible options for characteristic poses, we pick the first one that is representative of the action, e.g., the first sitting pose.

Characteristic 3D Pose Prediction. For the task of characteristic 3D pose prediction, we consider an input sequence of $N = 10$ 3D pose observations of a person, represented as $J = 25$ 3D joint locations for the GRAB dataset and $J = 17$ for the Human3.6M dataset (in their native joint layouts; for more details we refer to the appendix). From this observation, the next characteristic pose is predicted as J 3D joint locations. All poses are considered in their hip-centered coordinate systems. Note that while we have action labels in the annotated dataset, we do not use them for this task.

The N input pose observations can occur at any time, so methods are trained with random input sequences up to the characteristic 3D pose. At test time, five input points are evaluated for each method, with the five input points selected to evenly distribute between the beginning of the sequence to N frames before the characteristic pose.

Evaluation. We use a train/val/test split by actor in each dataset. For GRAB we have 8/1/1 train/val/test actors, resulting in 992/197/136 train/val/test sequences. For Human3.6M, we follow the split of [15]: 5/1/1 and 150/30/30 train/val/test actors and sequences, respectively.

To evaluate our task of characteristic 3D pose prediction, we aim to consider the multimodal nature of the task. Since we do not have ground truth probability distributions available, and only a single observed characteristic pose for each input pose observation, we follow previous work on multi-modal human pose sequence predictions [102, 101, 16, 103]: At test time, we consider $k = 10$ hypotheses from each method. To characterize these hypotheses holistically, we consider several metrics to assess accuracy, diversity, and quality of predictions.

Accuracy. First, we evaluate the sampling error using the mean per-joint position error (MPJPE) [45] by comparing the most similar prediction p' to the ground-truth pose p:

$$\mathrm{E}_{\mathrm{MPJPE}} = \frac{1}{N} \sum_{j=1}^{N} ||p'_j - p_j||_2^2 \tag{3.3}$$

This evaluates whether the predicted hypotheses capture the target well and allows for comparison with deterministic baselines (where all hypotheses are identical).

Diversity. We evaluate the diversity as the MPJPE between all sampled poses for the same sequence. This evaluates the multi-modality of predicted distributions.

Quality. Finally, we evaluate quality of our multi-modal predictions with the Inception Score [107] (IS) over the set of predicted hypotheses for all test sequences. The Inception Score is widely used to measure the quality generative model outputs. More specifically, we use the conditional formulation first introduced in [108]. Similar to [103], we adapt this idea to our use case by training a simple skeleton-based action classifier on ground-truth samples from our datasets. Overall, this metric estimates how well the predictions capture an action while still producing diverse poses.

3.7 Experimental Evaluation

We evaluate the task of characteristic 3D pose prediction, using our annotated dataset built from the real-world GRAB [44] and Human3.6M [45] datasets.

		GRAB			Human3.6m		
	Method	MPJPE ↓	Diversity ↑	IS ↑	MPJPE ↓	Diversity ↑	IS ↑
Statistical	Random Sampling	1.018	-	-	1.159	-	-
	Average Train Pose	0.146	-	-	0.179	-	-
	Zero Velocity	0.063	-	-	0.166	-	-
Algorithmic	Learning Trajectory Dependencies [14]	0.077	-	-	0.165	-	-
	History Repeats Itself [15]	0.071	-	-	0.116	-	-
	DLow [16]	0.071	0.089	1.257 ±0.02	0.119	0.104	1.623 ±0.08
	Ours	**0.054**	**0.105**	**4.153** ±0.87	**0.092**	**0.189**	**3.139** ±0.32

Table 3.1: Characteristic 3D pose performance, in comparison with state of the art and statistical baselines. We evaluate MPJPE for all methods and additionally, the diversity of multimodal methods in terms of MPJPE between samples as well as their quality with the Inception Score, similar to [103].

Comparison to time-based state-of-the-art forecasting. In Tab. 3.1, we compare to state-of-the-art multi-modal sequence forecasting approach DLow [16], which is based on a conditional VAE, as well as to recent deterministic approaches for frame-based future human motion prediction, Learning Trajectory Dependencies [14] and History Repeats Itself [15], which use a graph neural network and an attention-based model, respectively, to predict human pose sequences. We train all of these sequential approaches on our datasets, given the input sequence of N frames, to predict an output N_o-frame pose sequence, with $N_o = 100$ frames to ensure that the characteristic pose falls within each target sequence. Since these sequence-based approaches each predict output sequences, we additionally allow them to predict the time step of the characteristic pose with an MLP to obtain the final characteristic pose prediction (see the appendix for additional detail).

Since we aim to predict a characteristic 3D pose given an arbitrary sequence observation, we sample different start points for the input sequence, and analyze performance across varying distance from the goal pose.

Method	GRAB		Human3.6m	
	MPJPE ↓	IS ↑	MPJPE ↓	IS ↑
L. T. D. [14]	0.075	-	0.156	-
H. R. I. [15]	0.066	-	0.116	-
DLow [16]	0.059	1.567 ±0.02	0.108	1.418 ±0.14
Ours	**0.054**	**4.153** ±0.87	**0.092**	**3.139** ±0.32

Table 3.2: Characteristic 3D pose performance comparison. In contrast to Tab 3.1, baselines are provided with ground-truth characteristic time step information.

We report the MPJPE, Diversity, and IS metrics in Tab. 3.1; we first measure the performance for each of the five input sequence start times mentioned above and average over those for the final result. Our approach more accurately characterizes the future characteristic poses while also producing improved diversity and quality. For comparison, we also report baseline performance when given an oracle providing the ground-truth characteristic time step in Tab. 3.2. Even with this additional information, our characteristic pose formulation achieves improved results. Qualitative results are shown in Fig. 3.6; our probabilistic approach more effectively captures a realistic set of characteristic modes.

In Fig. 3.7, we visualize the diversity of our predictions in comparison with multi-modal baselines. Our predicted pose hypotheses show more diversity in both joint placement and action representation, while still capturing the target pose.

Comparison to statistical baselines. We also compare with three statistical baselines: full random sampling from an evenly distributed heatmap, the average target train pose over the entire dataset, and a zero-velocity baseline (i.e., the error of simply using the last input pose as prediction), which was shown by Martinez et al. [28] to be competitive with and sometimes outperform state of the art. Our approach outperforms these statistical baselines, indicating learning of strong characteristic pose patterns.

3.8 Ablation Studies

Does a probabilistic prediction help? In addition to comparing to state-of-the-art alternative approaches which make deterministic predictions, we compare in Tab. 3.3 with our model backbone with a deterministic output head (an MLP) replacing the

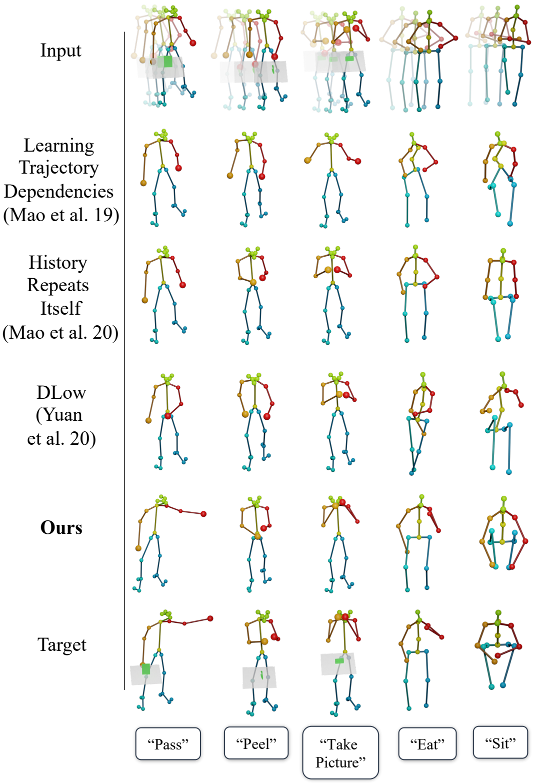

Input
Learning Trajectory Dependencies (Mao et al. 19)
History Repeats Itself (Mao et al. 20)
DLow (Yuan et al. 20)
Ours
Target
"Pass"
"Peel"
"Take Picture"
"Eat"
"Sit"

Figure 3.6: Qualitative results on characteristic 3D pose prediction. In comparison to deterministic [14, 15] (rows 2 and 3) and probabilistic [16] (row 4) approaches, our method more effectively predicts likely intended action poses. Note that action labels are only shown for visualization purposes.

volumetric heatmap decoder which regresses offset positions for each pose joint relative to the input positions. Removing our heatmap predictions similarly fails to effectively capture the characteristic modes; our probabilistic, heatmap-based predictions notably improve performance.

Does per-voxel offset prediction help? We analyze the effect of per-voxel offset prediction in Tab. 3.3, showing that they notably improve pose predictions. Applying pose refinement without offset prediction fails to achieve the same level of improvement.

Does autoregressive pose joint sampling help? We analyze the effect of our autoregressive pose joint sampling in Tab. 3.3. We compare against a version of our model trained to predict each pose joint heatmap independently, with pose joints sampled independently, which often results in valid individual pose joint predictions that are globally inconsistent with the other pose joints. In contrast, our autoregressive sampling helps to generate a likely, consistent pose.

		GRAB		Human3.6m	
	Ablation	MPJPE ↓	IS ↑	MPJPE ↓	IS ↑
Loss	ℓ_1 loss	0.132	1.132 ±0.01	0.198	2.246 ±0.24
	ℓ_2 loss	0.130	1.146 ±0.01	0.206	1.976 ±0.08
Model	Deterministic	0.064	-	0.108	-
	Not autoreg.	0.077	1.583 ±0.15	0.109	1.929 ±0.09
Sampling	No offsets	0.132	1.328 ±0.02	0.172	2.537 ±0.07
	↪ refined	0.127	1.509 ±0.03	0.163	2.978 ±0.14
	$k = 50$	0.049	1.222 ±0.02	0.082	1.845 ±0.19
	Not refined	0.057	3.989 ±0.95	0.098	2.418 ±0.11
	Ours	**0.054**	**4.153** ±0.87	**0.092**	**3.139** ±0.32

Table 3.3: Ablation study over varying heatmap losses, deterministic and non-autoregressive pose sampling, no offset prediction (with and without pose refinement), number of samples taken for the evaluation, and without pose refinement.

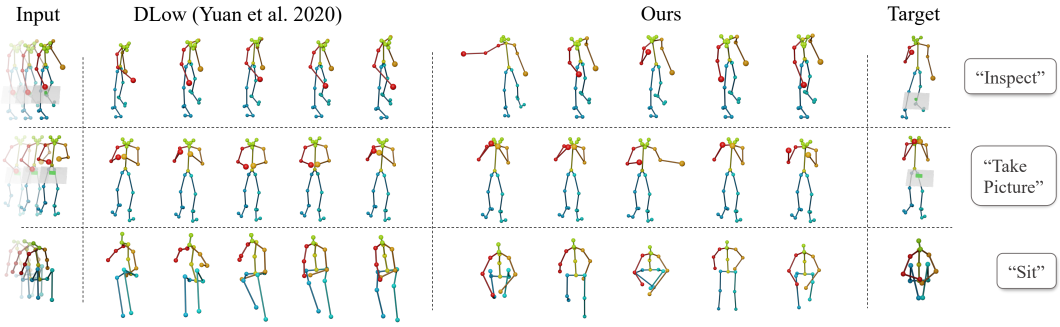


Figure 3.7: Qualitative results on characteristic 3D pose prediction, showing the diversity of our predictions in comparison with DLow [16].

How diverse are the sampled poses? We show qualitative examples of our multi-modal predictions in Fig. 3.7, outlining the diversity of both heatmap predictions and sampled skeletons. We also evaluate our prediction diversity as MPJPE between our sampled outputs as part of Tab. 3.1.

What is the effect of the number of pose samples? If we take more pose samples from our predicted joint distribution (from 10 to 50), we can, as expected, better predict the potential target characteristic pose, as seen in Tab. 3.1.

Do different heatmap losses matter? We evaluate our formulation for heatmap prediction as a discretized heatmap with a cross entropy loss against regressing heatmaps with an ℓ_1 or ℓ_2 loss, and find that our discretized formulation much more effectively models the relevant modes.

Limitations. Several limitations remain for our approach of characteristic 3D action pose forecasting. For instance, while our offset predictions help alleviate the ties to a volumetric heatmap grid, more precise modeling of smaller-scale behavior (e.g., detailed hand movement) would require more efficient representations such as sparse grids. In addition, our method relies on manually annotated characteristic 3D poses for supervision; while characteristic pose annotation is very efficient for new datasets, self-supervised formulations would also be an interesting future direction.

3.9 Conclusion

In this paper, we introduced a new task: predicting future *characteristic 3D poses* of human activities from short sequences of pose observations. We introduce a probabilistic approach to capturing the most likely modes in these characteristic poses, coupled with an autoregressive formulation for pose joint prediction to sample consistent 3D poses from a predicted joint distribution. We trained and evaluated our approach on a new annotated dataset for characteristic 3D pose prediction, outperforming deterministic and multi-modal state-of-the-art approaches. We believe that this opens up many possibilities towards goal-oriented 3D human pose forecasting and understanding anticipation of human movements.

3.10 Appendix

In this appendix, we show additional qualitative results, additional quantitative analysis, detail our network architecture specification, provide additional details regarding the dataset as well as our training setup, and discuss potential negative societal impacts of our method.

3.10.1 Additional Qualitative Results.

We show additional qualitative results of our method in Fig. 3.8, which demonstrate the diversity of our characteristic pose predictions for a given input sequence. Our approach not only effectively models the multi-modal nature of characteristic poses, but also captures the final target action pose (highlighted pose prediction).

In cases where the time between input sequence and target pose is longer, such as in 'sit' or 'greet', our approach produces a more diverse set of action poses, capturing the ambiguity in the future characteristic pose. When the input sequence is close to the target pose, our approach converges to a small set of probable poses (for example, in 'drink'), reflecting the reduced ambiguity.

3.10.2 Additional Quantitative Results.

MPJPE baseline comparison, by goal-normalized input time Fig. 3.9 shows MPJPE for varying input sequence start times in comparison with state of the art, goal-normalized from the start of each sequence (0) to N frames before the characteristic pose (1), with three steps inbetween.

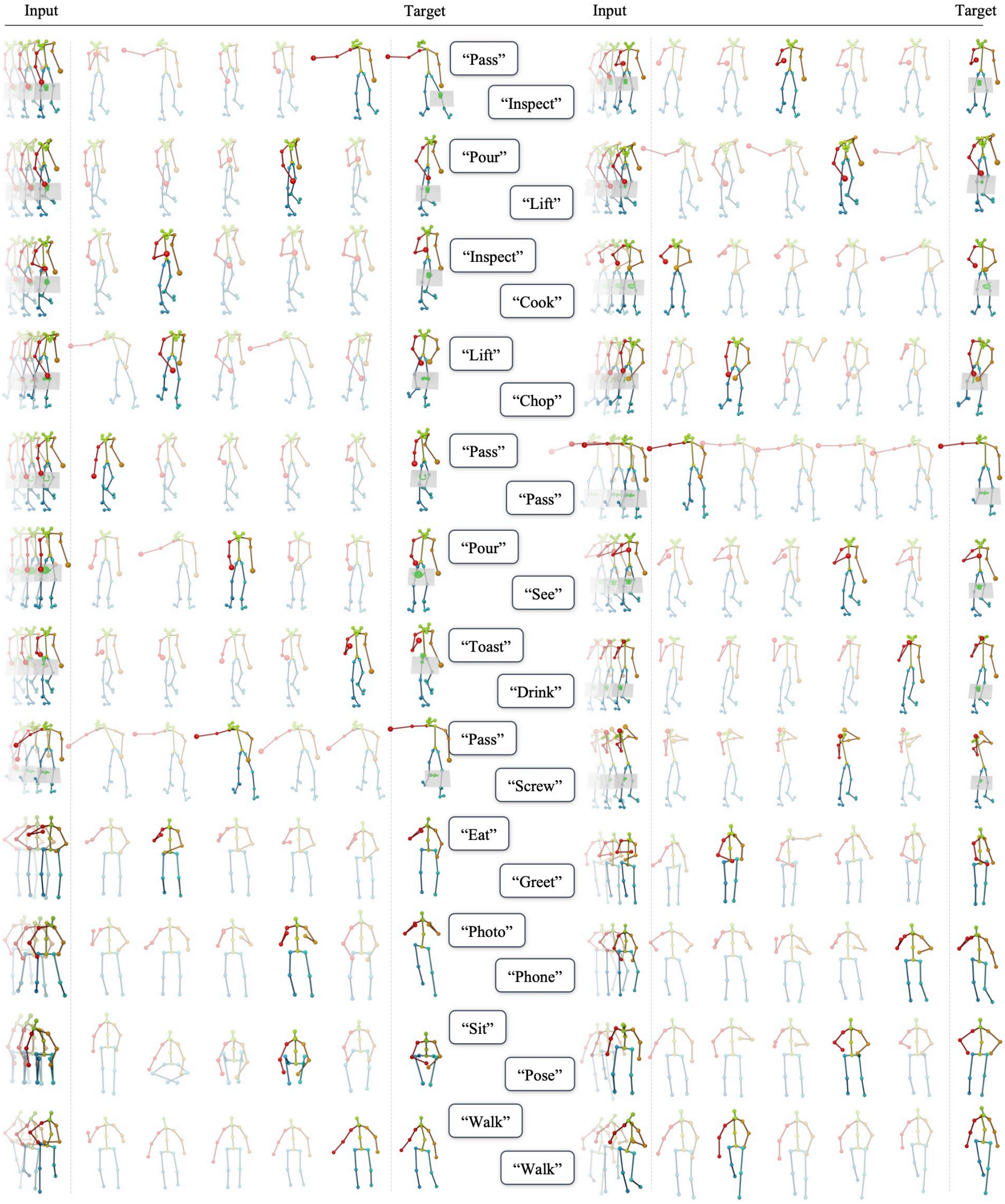


Figure 3.8: Additional qualitative results, showing the for each action sequence the inputs (left), our diverse set of predictions (middle) and the target action pose (right). Our final pose prediction is highlighted for each action sequence.

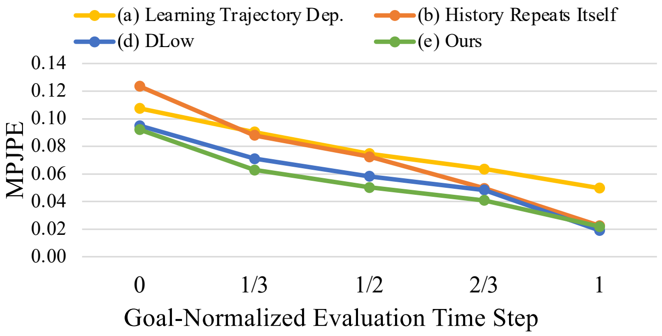


Figure 3.9: MPJPE comparison to baselines, evaluating with the input sequence at different points in time: from the start of the sequence (0) to N frames before the target characteristic pose (1).

Autoregressive Joint Order. We determined the order of the joints for the autoregressive prediction empirically; most ambiguity occurred in active end-effectors (i.e. right and left hands), whereas the rest of the body tended to have lower variability. In Tab. 3.4, we compare our original approach of (right hand, left hand, rest) with two alternatives: (left hand, right hand, rest), and (full autoregressive from human kinematic chain following left/right hands). Our method is robust to these orderings (though diversity of the rest of the body except hands decreases with autoregression through the kinematic chain).

Order	MPJPE ↓	Div. ↑	IS ↑
right hand → left hand → rest	**0.054**	**0.105**	**4.15** ±0.9
left hand → right hand → rest	0.057	0.049	4.09 ±1.6
following the kinematic chain	0.058	0.018	4.02 ±0.9

Table 3.4: Ablation analysis on autoregressive order on GRAB data.

Grid Resolution and Offset Prediction. We show additional ablations on the effect of grid resolution and offset prediction in Tab 3.5 on GRAB data; A resolution of 16^3 performs better than 8^3 or 32^3. Our offset prediction helps mitigate grid artifacts even at 32^3.

Per-Bodypart MPJPE. In Tab. 3.9, we show our final pose prediction performance in MPJPE, broken down per bodypart, as compared to sequential baselines.

Characteristic Pose Forecasting with Ground Truth Action Labels. In Tab. 3.6, we additionally evaluate our approach using ground truth action labels as input to provide additional contextual information.

Resolution	Offsets	MPJPE ↓	Diversity ↑	IS ↑
8^3	×	0.242	**0.189**	1.40 ±0.3
8^3	✓	0.092	0.068	1.71 ±0.1
16^3	×	0.127	0.081	1.51 ±0.1
$\mathbf{16^3}$	✓	**0.054**	0.105	**4.15** ±0.9
32^3	×	0.118	0.122	2.39 ±0.2
32^3	✓	0.066	0.058	1.91 ±0.2

Table 3.5: Ablation analysis on heatmap grid size and offset prediction on GRAB data.

The ground truth action label is processed as an additional attention node alongside input and previously predicted joint locations. This action label information reduces ambiguity in the possible set of output poses, resulting in reduced diversity, as is reflected in the diversity metric and inception score (as this directly considers diversity).

In our original action-agnostic scenario, our approach predicts plausible and diverse characteristic poses across all actions.

	GRAB			Human3.6M		
	MPJPE ↓	Div. ↑	IS ↑	MPJPE ↓	Div. ↑	IS ↑
×	0.054	0.105	4.153 ±0.87	0.092	0.189	3.139 ±0.32
✓	0.051	0.026	1.085 ±0.02	0.094	0.044	1.700 ±0.06

Table 3.6: Comparison of ours to an ablation with ground truth action labels as additional input.

3.10.3 Architecture Details

Fig. 3.10 details our network specification from input (left) to heatmap and offsets output (right). For each GRU layer, we provide the hidden dimension and number of layers in parentheses, for normalization layers the dimension to be normalized over, for dropout layers the dropout probability p, and for convolutions the number of input and output channels as well as kernel size (ks), stride (str), and padding (pad). We apply cross-entropy (CE) losses at a heatmap resolution of 8^3 and at the final resolution of 16^3; for the offsets prediction, we concatenate the offsets volume generated from the last input skeleton after 5 convolution blocks and supervise the final predictions with an ℓ_1 loss.

We take as input 25 joints in the case of GRAB and 17 joints for Human3.6M (#in_joints). The number of output joints (#out_joints) depends on whether the right or left hand is being predicted (#out_joints=1) or the rest of the body (#out_joints=23 for GRAB, #out_joints=15 for Human3.6M). In all our experiments, we use 10 as the number of probability bins.

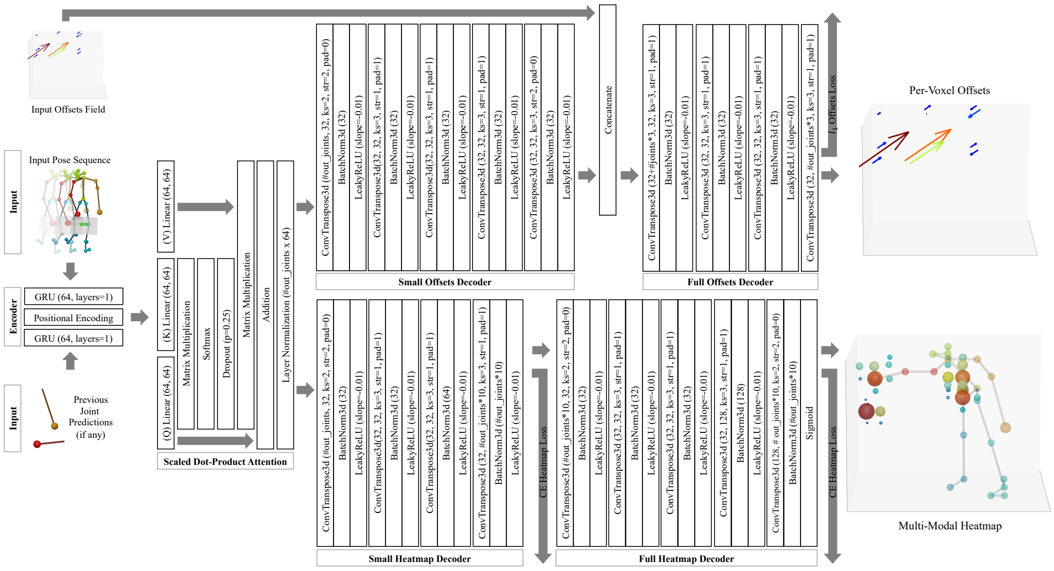


Figure 3.10: Our network architecture with details for encoder, scaled dot-product attention, as well as heatmap and offsets decoders.

3.10.4 Dataset

GRAB Pose Layout. Since GRAB [44] not only provides a human skeleton representation but full body shape parameters, we preprocess all pose sequences by first extracting relevant joints for our approach. For this, we chose the 3D OpenPose [58] layout as it describes the prevalent body joints and is widely used for representing 3D poses. Note that we do not apply the OpenPose method on 2d data; we only use their joint definitions in 3D. We extract 25 body joints from the SMPL-X body given by the GRAB dataset [44] using the correspondences shown in Tab. 3.8. Additionally, we denote in Tab. 3.8 the correspondences of joints to body parts, for the body part analysis in Tab. 3.9. Fig. 3.11 (left) visualizes our joint selection, overlaying the body shape given in GRAB as a point cloud over the 25-joint skeleton.

Human3.6M Pose Layout. For all our experiments on Human3.6M [45], we use 17 pose joints, visualized in Fig. 3.11 (right). Tab. 3.7 describes the exact joints used as well as the correspondences of joints to body parts, as used in Tab. 3.9.

Visualization Details. While our approach is agnostic to context or action, we visualize the context provided by GRAB [44, 109] (of the table and object) and action label provided by both GRAB and Human3.6M to help contextualize the pose visualizations. The context and action labels are not taken into account by the network or the evaluation, meaning that our approach infers plausible human action poses while being agnostic towards action and context.

	Ours (17-Joint)		Base (Human3.6M)	
	Idx	Label	Label	Idx
R. Leg	1	R. Hip	R. Hip	1
	2	R. Knee	R. Knee	2
	3	R. Foot	R. Heel	3
L. Leg	4	L. Hip	L. Hip	6
	5	L. Knee	L. Knee	7
	6	L. Foot	L. Heel	8
R. Arm	14	R. Shoulder	R. Shoulder	25
	15	R. Elbow	R. Elbow	26
	16	R. Hand	R. Hand	27
L. Arm	11	L. Shoulder	L. Shoulder	17
	12	L. Elbow	L. Elbow	18
	13	L. Hand	L. Hand	19
Spine	7	Spine	Spine	12
	0	Hip	Hip	0
Head	9	Nose	Nose	14
	10	Head	Head	15
	8	Thorax	Thorax	13

Table 3.7: Joint Correspondences for Human3.6M

Additional Characteristic 3D Pose Details. We show additional characteristic 3D poses in their original sequences in Fig. 3.12, and note the strong time differences at which the characteristic poses occur.

Furthermore, Fig. 3.13 and Fig. 3.14 show the times during the sequences at which the characteristic 3D poses are annotated for GRAB and Human3.6M; these characteristic poses are distributed across a wide range (0-12 seconds and 0-40 seconds, respectively) of time.

3.10.5 Additional Training Details

Cross Entropy Loss. Since our approach learns to predict the probabilities of a Gaussian-smoothed target point during training, we observe a very large class imbalance between the no-probability bin (bin 0) and the rest of the bins. We thus weigh the classes in the cross entropy loss to account for the class imbalances, by the inverse of their log-scaled occurrence, and a weight of 0.1 for the no-probability bin.

State-of-the-art comparisons. We use the official code with default settings of the methods we compare to ([14], [15], and [16]). We train all methods from scratch on our characteristic 3D pose dataset, setting the number of input frames to 10 and the number of output frames to 100. From the predicted sequence, we evaluate the pose at a timestep predicted by the baselines themselves as characteristic pose and compare it to the target. This scenario is the closest to our approach, as predicting characteristic 3D poses involves which pose is the characteristic pose.

	Ours (OpenPose [58])		Base (SMPL-X [49])	
	Idx	Label	Label	Idx
R. Arm	2	Right Shoulder	Right Shoulder	17
	3	Right Elbow	Right Elbow	19
	4	Right Finger	Right Index 3	42
L. Arm	5	Left Shoulder	Left Shoulder	16
	6	Left Elbow	Left Elbow	18
	7	Left Finger	Left Index 3	27
Right Leg	9	Right Hip	Right Hip	2
	10	Right Knee	Right Knee	5
	11	Right Ankle	Right Ankle	8
	22	Right Big Toe	Right Big Toe	63
	23	Right Small Toe	Right Small Toe	64
	24	Right Heel	Right Heel	65
Left Leg	12	Left Hip	Left Hip	1
	13	Left Knee	Left Knee	4
	14	Left Ankle	Left Ankle	7
	19	Left Big Toe	Left Big Toe	60
	20	Left Small Toe	Left Small Toe	61
	21	Left Heel	Left Heel	62
Head	0	Nose	Nose	55
	1	Neck	Neck	12
	15	Right Eye	Right Eye	24
	16	Left Eye	Left Eye	23
	17	Right Ear	Right Ear	58
	18	Left Ear	Left Ear	59
	8	Mid-Hip	Pelvis	0

Table 3.8: Joint Correspondences for GRAB

Therefore, we modified each baseline with a small prediction head to predict the characteristic pose frame within all 100 frames of the predicted sequence. In all cases, we supervise this prediction as a classification problem with a cross entropy loss and train the additional head together with the rest of the model.

For DLow [16], we add one linear layer to the final feature output of each of the 100 steps, followed by a ReLU, reducing each step's output dimension to 10. Then, one additional linear layer summarizes the combined output of all steps $(100 * 10)$ down to a vector of size 100.

In the case of History Repeats Itself [15], we add a classification head consisting of one linear layer, a 1d batch norm, a ReLU, and one additional linear layer to the output of their last Graph Convolution Block (GCN). While the first linear layer keeps the original dimensionality of 100, the second linear layer reduces the dimension from $\#\text{graph\_nodes} * 100$ down to 100.

Finally, for Learning Trajectory Dependencies [14], we apply the same architecture and add a linear layer, a 1d batch norm, a ReLU, and a second linear layer after the final GCN. Here, we first reduce the per-node feature dimension from 256 to 100 and combine the features of all nodes with the second linear layer, going from $\#\text{graph\_nodes} * 100$ down to 100.

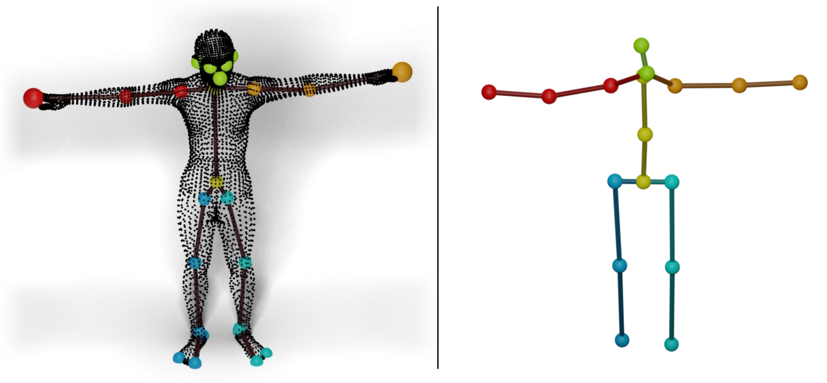


Figure 3.11: GRAB [44] body and our extracted skeleton joints overlaid (left); 17-joint skeleton based on Human3.6M [45] (right).

	GRAB						H3.6M					
Method	R. Arm ↓	L. Arm ↓	R. Leg ↓	L. Leg ↓	Spine ↓	Head ↓	R. Arm ↓	L. Arm ↓	R. Leg ↓	L. Leg ↓	Spine ↓	Head ↓
L. T. D. [14]	0.165	0.115	0.058	0.057	0.028	0.085	0.225	0.225	0.135	0.146	0.108	0.123
H. R. I. [15]	0.160	0.113	0.056	0.055	0.026	0.079	0.199	0.191	**0.079**	0.088	0.040	0.089
DLow [16]	0.146	0.109	0.052	0.050	0.024	0.068	0.174	0.169	0.108	0.112	0.044	0.096
Ours	**0.105**	**0.084**	**0.045**	**0.045**	**0.020**	**0.057**	**0.147**	**0.122**	0.091	**0.085**	**0.033**	**0.066**

Table 3.9: Characteristic 3D pose prediction performance comparison to baselines, broken down by body part MPJPE.

In the main paper, we additionally evaluated against these baseline approaches when given ground-truth time steps instead; in this scenario, our predictions also outperform the baselines given ground truth times for characteristic poses.

To evaluate the diversity and quality of multi-modal outputs, 10 samples are taken from a probabilistic method for each input sequence, and we report diversity in terms of MPJPE between samples as well as the Inception Score, following [103].

3.10.6 Potential Negative Societal Impacts

As we aim to study human pose behavior, we must take care to ensure that datasets used represent notable diversity in those represented. Our approach currently operates on skeleton abstractions that do not characterize finer-scale appearance differences; in possible future studies that may aim to characterize fine-scale interactions, diversity in body shape representations which must be taken into account for data collection and analysis.

In particular, in our scenario of forecasting probable future human behavior, we must also ensure that this possibility cannot be easily used for generating fraudulent motion video of a person. Such usage is currently severely limited in our proposed approach, as it does not target individual people, and does not model photo-realistic characteristics of people.

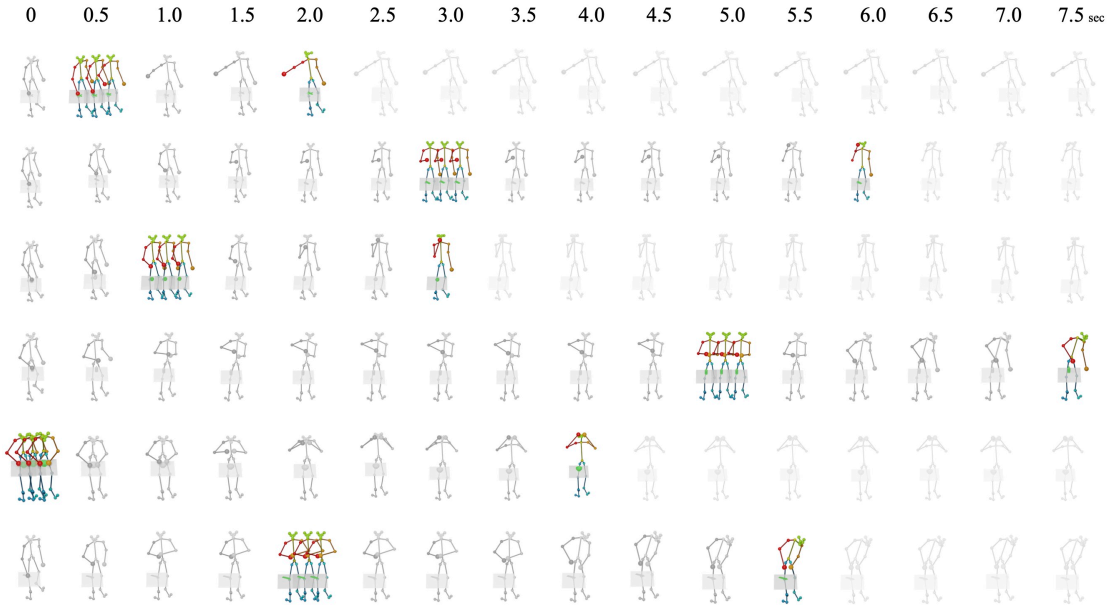


Figure 3.12: Sample input-target pairs (colored) for our characteristic 3D pose forecasting task, with temporal snapshots along the sequence (grayscale). Each snapshot is half a second apart. Depicted as input is the last frame of the respective input sequence.

Another concern might arise with the possibility of surveillance, in the context of predicting specific actions from only a short and possibly ambiguous observation of a person. The types of actions are currently limited by the training data to everyday activities such as eating or walking. With modified datasets, the prediction of various specific action sub-categories might be possible (e.g., forecasting possible malicious actions). While simpler methods may be more suitable for this kind of task, here we look to efforts in data transparency; we will provide our annotations and various statistics to characterize the everyday activities in our considered data.

Another axis to consider is that of environmental impact, in the cost of training deep neural networks. Our training time is relatively short with only a few hours until convergence and a moderately sized neural network. Additionally, adversarial attacks are a possibility to disrupt future predictions, but do not induce security concerns for our approach directly.

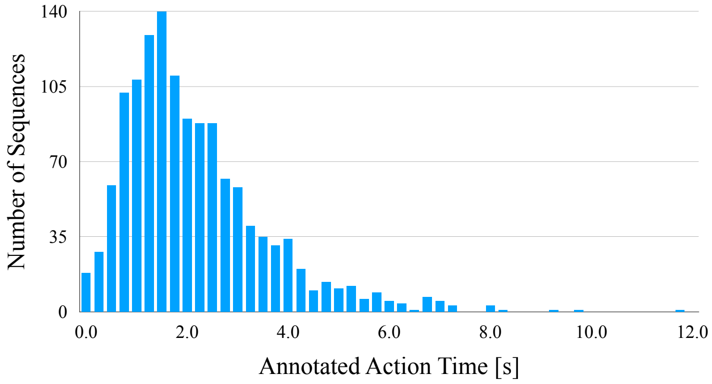


Figure 3.13: Times at which characteristic poses occur for GRAB.

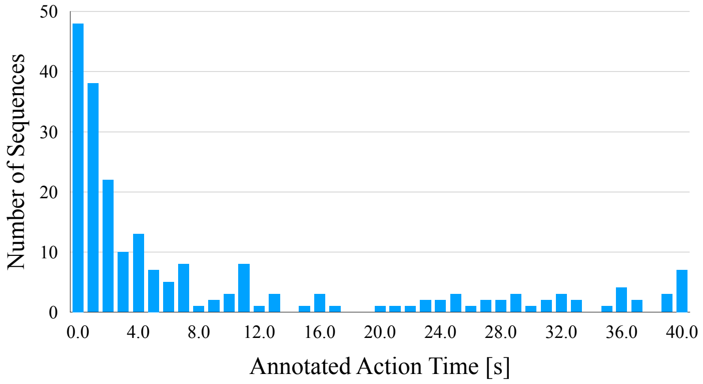


Figure 3.14: Times at which char. poses occur for Human3.6M.

4 Complex Long-Term 3D Human Behavior from Video Observations

This chapter introduces the following paper:

Christian Diller, Thomas Funkhouser, and Angela Dai. ”FutureHuman3D: Forecasting Complex Long-Term 3D Human Behavior from Video Observations.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Abstract of Paper We present a generative approach to forecast long-term future human behavior in 3D, requiring only weak supervision from readily available 2D human action data. This is a fundamental task enabling many downstream applications. The required ground-truth data is hard to capture in 3D (mocap suits, expensive setups) but easy to acquire in 2D (simple RGB cameras). Thus, we design our method to only require 2D RGB data at inference time while being able to generate 3D human motion sequences. We use a differentiable 2D projection scheme in an autoregressive manner for weak supervision, and an adversarial loss for 3D regularization. Our method predicts long and complex human behavior sequences (e.g., cooking, assembly) consisting of multiple sub-actions. We tackle this in a semantically hierarchical manner, jointly predicting high-level coarse action labels together with their low-level fine-grained realizations as characteristic 3D human poses. We observe that these two action representations are coupled in nature, and joint prediction benefits both action and pose forecasting. Our experiments demonstrate the complementary nature of joint action and 3D pose prediction: our joint approach outperforms each task treated individually, enables robust longer-term sequence prediction, and improves over alternative approaches to forecast actions and characteristic 3D poses.

Contribution The method development and implementation was done by the first author. Discussions with the co-authors led to the final paper.

4.1 Introduction

Predicting future human behavior is fundamental to machine intelligence, with many applications in content creation, robotics, mixed reality, and more. For instance, a monitoring system might issue early warnings of potentially dangerous behaviour, and a robotic assistant can use forecasting to place tools at the right place and time they will be needed in the future. Consider the specific scenario of an assembly line monitoring system deployed to issue early warnings of behavior that could be harmful in the near future: The system needs to have a long-term understanding of the future, enabling it to forecast multiple action steps ahead so that it can act in time before a harmful action occurs. However, simply understanding the next action steps on a high level is not sufficient: it must also reason about *where* the action will occur. Actions such as "grab a tool" are likely harmless when performed in a toolbox; they become dangerous when done next to an active table saw or moving robot arm. The monitoring system thus also needs to be able to reason about spatial relations in 3D – for both the location and body pose of involved humans.

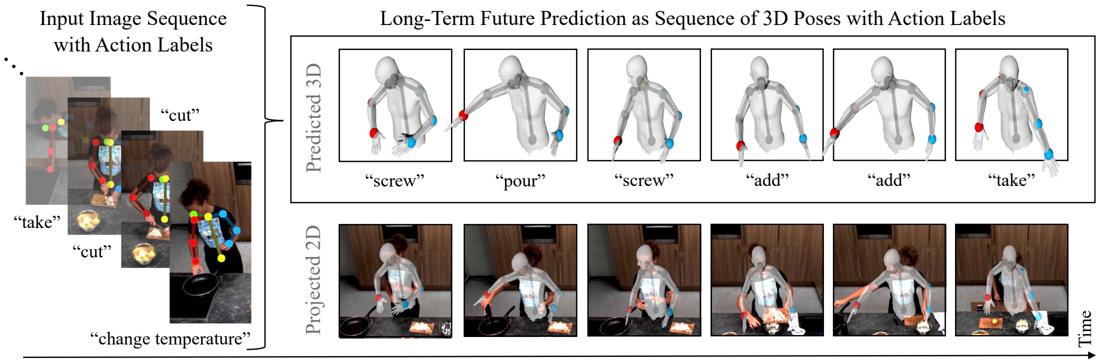

Input Image Sequence with Action Labels
"cut"
"take"
"cut"
"change temperature"
Long-Term Future Prediction as Sequence of 3D Poses with Action Labels
Predicted 3D
"screw"
"pour"
"screw"
"add"
"add"
"take"
Projected 2D
Time

Figure 4.1: We propose a novel generative approach to model long-term future human behavior by jointly forecasting a sequence of coarse action labels and their concrete realizations as 3D body poses. For broad applicability, our autoregressive method only requires weak supervision and past observations in the form of 2D RGB video data, together with a database of uncorrelated 3D human poses.

To support these types of applications, we must address two tasks: 1) forecasting long-term action sequences, and 2) predicting future 3D human poses. Prior work has focused on each of these tasks separately: activity forecasting predicts future action labels without considering the 3D poses [7, 8, 9, 10, 11, 12], while 3D pose forecasting focuses on fixed frame rate sequence prediction limited to single actions in short-term time frames without considering longer-term action sequences [13, 14, 15, 16, 17].

We propose that these two tasks are coupled in nature: predicting action labels with realized 3D poses helps to encourage richer feature learning and can materialize subcategory level differences in actions for predicting future activities, and grounding 3D poses with actions provides global structure for longer-term forecasting.

Leveraging this insight, we design a method that takes in a sequence of recent RGB image observations and their action labels, and jointly predicts a sequence of future 3D characteristic poses and action labels (Fig. 4.1). In our design, we had to address two significant research challenges: 1) forecasting 3D poses from 2D images without any paired 3D training data, and 2) forecasting long sequences of actions comprising several discrete action steps.

The first challenge arises from limited training data. It would be ideal to have a dataset with ground truth 3D pose and action annotations for complex sequences of actions. Unfortunately, no such dataset exists. There are RGB video datasets with tracked 3D poses for limited types of actions (e.g., walking or waving); and there are video datasets with action labels for complex sequences of actions (e.g., cooking or assembly). However, there is no single dataset that has both types of annotations, and capturing one would be difficult due to the challenges of setting up 3D pose trackers in settings where people typically perform complex sequences of actions (e.g., cooking in a kitchen). Instead, we have to learn to use 2D video observations for 3D pose and action label forecasting without paired data. We achieve this by weakly supervising our pose forecasting in 2D using readily available 2D action datasets [110, 111] and formulate an adversarial loss encouraging likely 3D characteristic poses with respect to a distribution learned from 3D pose datasets [45, 112, 44]. Crucially, this does not require any correspondence between the 2D video and 3D pose data.

The second challenge arises from the difficulties of predicting long sequences of discrete events. One option would be to train a model to output a multi-step sequence of actions and poses all at once – however, this is impossible given the exponential growth of multi-step sequences and the limited amount of available training data. Another option would be to train a model that predicts the next future poses and actions at fixed time points in the future (e.g., 1s in advance) and then recurrently make long-term predictions – however, this time-based forecasting approach produces sequences that tend to “drift” over the long-term, since the intermediate poses at fixed time steps are usually “in between” semantically meaningful actions and thus do not provide a distinctive input representation for the next prediction. To address this issue, we train our autoregressive approach to iteratively generate the next discrete action label along with the *3D characteristic pose* for that action. A 3D characteristic pose [40] is the set of 3D joint positions corresponding to the most distinctive moment a semantic action is being performed (e.g., when a hand grasps an object, when two objects are first brought together, etc.). By training our method to produce these poses as intermediate outputs (and inputs to the next step), we are able to generate more semantically plausible forecasts over longer action sequences.

Our experiments with two RGB video datasets demonstrate that our approach for joint prediction of action behaviors and 3D poses outperforms state-of-the-art methods applied separately to each task. Additionally, we find that predicting actions and their 3D characteristic poses enables more robust autoregressive prediction for longer-term forecasting. Overall, our contributions are:

- The first method to learn forecasting of future 3D poses from datasets with only 2D RGB video and action label data (i.e., without any paired 3D data).
- The first method to forecast future 3D poses jointly with action labels from commonly available video input.
- The first method to forecast future characteristic 3D poses and action labels for long-term and complex behaviors.

4.2 Related Work

3D Human Pose Forecasting. Forecasting 3D human poses has been studied in many previous works and is commonly formulated as a 3D sequential motion prediction task, taking an input 3D sequence of poses and generating an output 3D sequence of poses. For short-term future prediction (up to $\approx$ 1 second), RNN-based approaches have achieved impressive performance [13, 27, 28, 94, 95, 96, 97, 29, 113]. As RNNs tend to struggle to capture longer-term dependencies with a fixed-size history, graph-based [14, 114, 115, 30, 116, 117, 118, 119, 120, 17] and attention-based [31, 15, 61, 121, 73] approaches have been proposed to encode temporal history. Some methods also explored the applicability of temporal convolutions [99, 122] and MLP-only architectures [123, 124] to the task of human motion forecasting. Additionally, various approaches have been proposed to model future human motion stochastically to produce diverse future sequence predictions, either with adversarial GAN formulations [101, 125], conditional variational autoencoders (VAEs) [102, 126, 103, 17, 127, 128, 129, 130, 131], or diverse sampling [16, 132]. More recently, diffusion methods [76, 77] have been used for human motion generation and forecasting [133, 2, 134, 84, 1, 135, 37, 91]. These methods require 3D ground truth sequences for training, limiting applicability to scenarios where 3D inputs and ground-truth are not available. Ours requires only 2D training data for the action sequences, which is far more plentiful and easier to obtain. We generate valid 3D poses by leveraging an adversarial loss formulation, operating on a database of uncorrelated 3D poses.

Human Action Forecasting. Action forecasting has been studied by many approaches to predict future actions from a sequence of observed actions [136, 137, 138, 139] or directly from an input video sequence [140, 141, 142, 143, 11, 144, 141, 145]. Various methods have been developed to learn effective representations, including Hidden Markov Models [8], RNNs [137, 146, 147, 140, 141, 136, 137, 142], transformer-based networks [11, 12, 74], and self-supervised feature learning [148, 149]. There are approaches that focus on the short-term future [139, 142, 143, 11, 144, 74, 141, 145] or on longer-term actions [12, 136, 137, 138, 139, 140, 141, 142, 143, 11, 144, 145]. Such method focus on characterizing anticipation with action labels only, while we aim to predict a richer characterization of the anticipated future by leveraging characteristic 3D poses, representative of future action goals in a sequence of action-pose predictions. Forecasting actions alongside human poses in 2D only has been studied in a few works, for 2D hand placement [150] or full-body 2D human poses at most 1 second into the future [151]. Our approach addresses the benefits of 3D reasoning in human motion forecasting, without requiring full 3D sequences for supervision.

Goal-Driven Future Prediction. Goal-driven forecasting has previously been explored beyond action label forecasting, and has been leveraged to predict goal locations for future human walking trajectories [104] and for future video sequences by predicting keyframes [32, 152, 33, 153]. Diller et al. [40] introduced the task of forecasting *characteristic 3D poses*, salient keyframe poses representing the next action. These goal-based poses are more semantically meaningful and consistent across different action sequences than time-based ones, and thus are better suited for long-term forecasting. We build upon these ideas by introducing a new goal-driven method for joint action anticipation and characteristic 3D pose forecasting in an auto-regressive system that can predict complex, long-term behavior sequences.

4.3 Method Overview

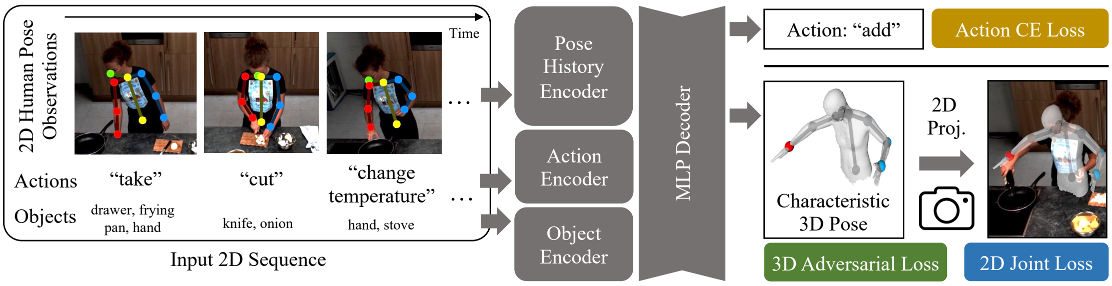


Figure 4.2: Our approach takes as input a sequence of RGB images, from which 2D poses are extracted, as well as their corresponding action label and initial set of objects. Each input is encoded into a joint latent space to jointly predict the next action label and characteristic 3D pose. While action labels are directly supervised, the 3D pose decoder is trained to match the next 2D action pose using differentiable projection, and an adversarial 3D loss encourages valid 3D pose prediction.

Our method aims to learn to jointly model future human actions along with the characteristic 3D poses representative of those actions. From a sequence of RGB image observations of a person performing a series of actions and the corresponding action labels, we predict a sequence of future action labels and 3D poses characteristic of these actions. This enables joint reasoning of not only global semantic behavior but also the physical manifestation thereof.

In the absence of 3D pose data of complex human actions, we weakly supervise forecasted 3D poses to align to future poses in 2D, and constrain the poses to be valid in 3D using an adversarial loss with a database of 3D poses. This does not require any correspondence between 3D pose data and 2D video, enabling action sequence supervision on commonly available 2D human action data together with carefully captured but unrelated human poses in 3D.

An overview of this approach is shown in Fig. 4.2. For an input sequence $S = \{(I_i, a_i, o_i)\}$ of N RGB images $\{I_i\}$ with corresponding actions $\{a_i\}$ and initially involved objects $\{o_i\}$, we aim to predict the future M actions $\{\hat{a}_k\}$ that will be taken along with their characteristic poses in 3D $\{\hat{Y}_k\}$. We define the human pose as a collection of J body joints at salient locations, so each output pose $\hat{Y}_k$ is predicted as a set of J 3D coordinates. We first extract information about the observed 2D pose movement by detecting 2D poses $\{X_i\}$, each with J 2D joints, with a state-of-the-art 2D pose estimator that seamlessly integrates into our pipeline in a pre-trained and frozen form.

Next, we encode this information along with previously observed action and object labels to predict the next future action label $\hat{a}_k$ and characteristic 3D pose $\hat{Y}_k$. We can then forecast a future sequence by autoregressively predicting a series, considering the 2D projections of the previously predicted 3D poses along with previously predicted actions as input to a new prediction.

4.4 Joint Forecasting of Actions and Characteristic 3D Poses

Our network takes as input the previous 2D observations $\{X_i\}$ extracted from the $\{I_i\}$ images, as well as action and object labels $\{a_i\}$ and $\{o_i\}$ as one-hot vectors. Since we only predict action labels, object labels are given from the objects seen at the beginning of the sequence, and subsequently re-used for the entire sequence. Each of these are encoded in parallel with three separate encoders; the actions and objects with MLPs while the poses are projected into latent space with a single linear layer and then processed with a stack of three residual blocks. These encoded features are then all concatenated together in latent space, and processed jointly with an MLP to produce a common latent code z. Finally, we decode both poses and actions in parallel based on z using an MLP decoder each, yielding the next action label class as a vector $\hat{a}_k \in \mathbb{R}^{N_a}$ and 3D characteristic pose $\hat{Y}_k \in \mathbb{R}^{J\times 3}$, with N_a the number of action classes. For a more detailed architecture specification, we refer to the appendix.

We jointly learn future action labels and characteristic 3D poses by supervising $\hat{a}_k$ and $\hat{Y}_k$ to match the observed future 2D video, and constrain $\hat{Y}_k$ to form a valid 3D pose by an adversarial loss, optimizing for the overall loss:

$$\mathcal{L} = \lambda_{action}\mathcal{L}_{action} + \lambda_{pose2d}\mathcal{L}_{pose2d} + \lambda_{adv3d}\mathcal{L}_{adv3d} \tag{4.1}$$

where $\mathcal{L}_{action}$ denotes the action loss, as described in Sec. 4.4.1, $\mathcal{L}_{pose2d}$ and $\mathcal{L}_{adv3d}$ constraining the predicted pose, as described in Sec. 4.4.2, and the λ weighting each loss.

4.4.1 Action Forecasting

Predicted future actions are decoded from the latent code z by an MLP decoder to predict the action class $\hat{a}_k$, supervised by cross entropy with the ground truth future action: $\mathcal{L}_{action} = \mathrm{CE}(\hat{a}_k, a_k^{\mathrm{gt}})$.

4.4.2 Characteristic Pose Forecasting

Our goal is to forecast complex action behavior not only in terms of action labels, but also manifested as a sequence of characteristic poses in 3D. Since we only have 2D pose annotations available, we first constrain these poses to represent future actions in 2D and make use of an adversarial regularization in 3D. This does not require any correspondence between 2D and 3D data, only a collection of valid 3D poses, which are readily available.

Differentiable 2D Projection Our generator network predicts the next characteristic action pose $\hat{Y}_k$ as a set of 3D joints. To constrain $\hat{Y}_k$ based on the target future 2D pose X^{gt} extracted from the ground truth future image, we differentiably project $\hat{Y}_k$ into the 2D image with intrinsic parameters K and extrinsic rotation and translation R, t:

$$\hat{X} = K(R\hat{Y}_k + t) \tag{4.2}$$

Since we learn from third-person video with a fixed camera, we can use the same camera parameters for all sequences used for training. We can then define the 2D pose loss as the mean squared error between the projected pose prediction and the ground truth:

$$\mathcal{L}_{pose2d} = ||X^{\text{gt}} - X_k||_2^2 \tag{4.3}$$

Note that we only predict the J joints that have been observed in the video data (excluding any joints that remain occluded in the observed video data), so this loss can be applied to all predicted joints.

Adversarial 3D Pose Regularization While the action and pose prediction losses provide effective predictions when considered in the 2D projections, the $\{\hat{Y}_k\}$ remain underconstrained in 3D and thus tend to exhibit large distortions and implausible bone lengths and angles, when trained with only 2D supervision. We thus constrain the predicted poses to form valid 3D poses by formulating an adversarial 3D loss from a critic network which is simultaneously trained to distinguish predicted poses from a database of real 3D skeleton samples. Note that there is no correspondence between these skeletons and the 2D poses extracted from the action video sequences – any database of 3D skeletons can be used. We can thus train our approach with an entirely uncorrelated 3D pose dataset without requiring 3D action pose correlations.

We then formulate $\mathcal{L}_{adv3d}$ as a Wasserstein loss [154], training the critic network in an alternating fashion with the generator. This enables effective forecasting of future 3D characteristic poses for predicted future action labels, without requiring any 3D observations as input.

In order to enable the critic network to learn effectively about likely intrinsic pose constraints (e.g., lengths, kinematic chains, or valid joint angles), the critic takes as input not only the 3D joint locations of $\hat{Y}_k$ but also their kinematic statistics as a matrix Ψ, following [155, 156].

Ψ encodes joint angles and bone lengths as $\Psi = B^T B$, where $B = (b_1, b_2, \ldots, b_b)$ is a matrix with columns $b_i = j_k - j_l$ representing the vectors between each joint j_k and

j_l. Ψ then contains bone lengths l_i^2 on its diagonal, and angular representations on the off-diagonal entries.

4.4.3 Sequence Prediction

In order to forecast longer-term future behavior, our 3D pose predictions enable a natural autoregressive sequence prediction by taking the predictions $\hat{X}_t, \hat{a}_t$ at time step t as part of the input for time step $t + 1$. We can thus predict a sequence of M future action labels $\{\hat{a}_k\}$ and characteristic 3D poses $\{\hat{Y}_k\}$; we use $M = 10$ for MPII Cooking II [110] and $M = 5$ for IKEA-ASM [111], respectively.

4.4.4 Training Details

We train our approach for the $J = 9$ joints commonly seen across the input observed video data, characterizing the upper body in MPII Cooking II [110] and IKEA-ASM [111].

Additionally, we use loss weights $\lambda_{action} = 1e^6$, $\lambda_{pose} = 1$, and $\lambda_{adv3d} = 1$, empirically chosen to numerically balance each individual loss with the others.

We train our approach on a single NVIDIA GeForce RTX 2080TI for $\approx$ 12 hours until convergence. We use ADAM with batch size 4096, weight decay 0.001, and a constant learning rate of 0.0001 for both generator and discriminator.

4.4.5 Datasets

We train and evaluate our approach on two datasets: MPII Cooking II [110] and IKEA-ASM [111]. Both datasets contain sequences of human actors performing complex, unscripted actions, and provide annotations of fine-grained sub-action steps. MPII Cooking II [110] is an action recognition dataset with 272 complex cooking sequences and an average sequence time of 182s (35 annotated sub-actions, each 5.2s on average). IKEA-ASM contains 370 sequences of actors assembling IKEA furniture, with an average of 74s per sequence (15 annotated sub-actions, each 4.9s on average).

In both datasets, each action sequence has been filmed from a fixed camera setup; the third-person point of view enables extraction of 2D poses with an off-the-shelf 2D pose estimator. We use OpenPose [58] in our experiments and note that our approach is agnostic to the concrete method of 2D pose detection. We provide more in-depth discussion and additional experiments in the supplemental material.

We consider the 9 upper-body joints of the OpenPose skeletons, as the other joints are almost always occluded in the video observations, and remove global translation by centering each 2D pose at the neck joint.

Characteristic poses, in contrast to an arbitrary pose within a labeled action range, are the most representative pose of that action, and are annotated for all sub-actions in each sequence as the most articulated pose of that sub-action, following the annotation protocol of [40]. Annotation can be done efficiently and was performed by the authors within just 32 hours, yielding a total of $\approx$18,000 characteristic poses ($\approx$12,000 for MPII Cooking II and ≈6,000 for IKEA-ASM). These poses are indicative of the action they represent as demonstrated in Tab. 4.3: Using such poses significantly improves performance, validating our annotation protocol.

For the 3D adversarial loss, we use ≈800,000 human poses from popular 3D pose datasets: Human3.6m [45], AMASS [112], and GRAB [44]. Note that none of these 3D poses have any correspondence with the 2D posed actions from the MPII Cooking II dataset, instead depicting various human skeletons in natural and diverse poses.

4.5 Results

We evaluate sequence forecasting of action labels and characteristic 3D poses on the MPI Cooking II [110] and IKEA-ASM [111] datasets, and 3D pose quality by comparing to our database of high-fidelity 3D poses.

4.5.1 Evaluation Metrics

2D Pose Error. Since we only have 2D ground-truth data available for complex action sequences, we first project predicted 3D poses back into 2D, and evaluate the 2D mean per-joint position error (MPJPE) [45], in comparison with 2D poses extracted from ground-truth future frames using [58]: $\mathrm{E}_{\mathrm{MPJPE}} = \frac{1}{M}\sum_{j=1}^{M}||\hat{X} - X^{\mathrm{gt}}||_2$.

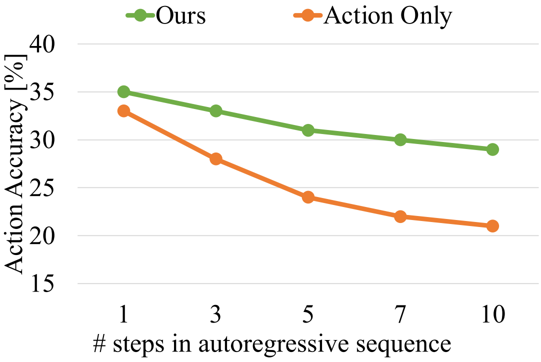


Figure 4.3: Action accuracy over time. Our joint action-characteristic pose forecasting enables more robust autoregressive action forecasting than action prediction without considering pose.

3D Pose Quality. In the absence of annotated ground truth 3D poses for the action video sequences, we measure the quality of predicted 3D poses as how distinguishable they are in comparison to a set of real 3D poses. We follow [103] and evaluate quality by training a binary classifier on 50,000 human poses generated at different training steps (representing examples of unrealistic 3D poses) and 50,000 real 3D pose samples. For classification accuracy a of this classifier, quality is measured as $1 - a$, with a quality of 1 indicating full indistinguishability from real poses. We refer to the supplemental for more details on this quality metric.

Action Accuracy. We report the action accuracy of the predicted sequences, as the mean over all sequences in the test set. We evaluate the top-n accuracy based on whether the ground truth action is among the n highest scoring predictions, for $n = 1$ and $n = 3$.

4.5.2 Comparison to Human Pose Forecasting

Tab. 4.1 compares our method to state-of-the-art 3D pose forecasting methods DLow [16], GSPS [127], STARS [17], and EqMotion [157]. These methods expect sequences of observed 3D human poses as input; we thus first apply a state-of-the-art weakly supervised 3D pose estimator [156] on our 2D input poses, producing inputs and supervision in 3D. This method estimates 3D poses using an adversarial formulation, requiring a database of 3D poses not correlated with the 2D pose inputs. To ensure a fair comparison, this database is exactly the same as the one our method uses.

	MPII Cooking II				IKEA ASM			
	2d	3d	Action Accuracy		2d	3d	Action Accuracy	
Approach	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
Zero Velocity	118	–	–	–	74	–	–	–
Train Average	166	–	–	–	91	–	–	–
AVT [11] RGB	–	–	19%	42%	–	–	22%	49%
AVT [11] RGB+Skeleton	–	–	20%	40%	–	–	23%	47%
FUTR [12] RGB	–	–	27%	48%	–	–	19%	45%
FUTR [12] RGB+Skeleton	–	–	27%	49%	–	–	20%	46%
RepNet [156] + DLow (min-10) [16]	72	**0.72**	–	–	45	0.31	–	–
RepNet [156] + GSPS (min-10) [127]	59	0.66	–	–	51	0.15	–	–
RepNet [156] + STARS (det.) [17]	70	0.62	–	–	54	0.27	–	–
RepNet [156] + EqMotion [157]	68	0.66	–	–	55	0.23	–	–
Joint 2D Pose & Action [151]	55	-	27%	43%	44	-	22%	46%
Ours	**50**	0.55	**29%**	**51%**	**40**	**0.31**	**29%**	**50%**

Table 4.1: Quantitative comparison with state-of-the-art action label and 3D pose forecasting. Our joint approach enables more accurate future action and pose predictions, compared to approaching both tasks separately, and outperforms joint action and 2D pose forecasting.

We chose the 3D pose estimator of [156] since its weakly supervised formulation is most comparable to our approach. An additional comparison to a fully supervised approach for 3D pose lifting (SPIN [158]) is provided in the supplemental.

We then train the 3D pose prediction methods from scratch on this generated data, using their original parameter settings. Stochastic methods DLow and GSPS are set to predict 10 possible future sequences; we report the minimum error across these. We use STARS in the method's deterministic mode. Each method takes as input a pose history of M poses and outputs a sequence of M poses, analogous to our setup where each pose is a characteristic pose corresponding to an action step ($M = 10$ for MPII Cooking II and $M = 5$ for IKEA-ASM). Our approach to lift 2D to future 3D poses and actions in an end-to-end fashion enables more effective pose forecasting than these state-of-the-art 3D pose forecasting approaches on both datasets.

In addition, we compare to the joint 2D action and pose forecasting approach of Zhu et al. [151]. Our approach of forecasting long-term sequences of 3D poses alongside actions is able to outperform their 2D MPJPE pose prediction and action accuracy performance, due to improved spatial reasoning when forecasting 3D poses.

Statistical 2D Baselines. We additionally compare with two statistical baselines in 2D, following [40]: the average target train pose, and a zero-velocity baseline which was introduced by Martinez et al. [28] as competitive with state of the art. We outperform both baselines, indicating that our method learns a strong action pose representation.

4.5.3 Comparison to Action Label Forecasting

We compare the action accuracy of our joint action-pose forecasting to AVT [11] and FUTR [12], two state-of-the-art action anticipation methods, in Tab. 4.1. We train and evaluate both AVT and FUTR on input RGB frames and their action and object labels, equal to our training setup, and use their original training settings initialized with a pre-trained vision transformer [72] for AVT and extracted I3D features [159] from our datasets for FUTR. Additionally, as we consider extracted 2D poses from the input RGB images, we also evaluate a variant that is trained and evaluated on RGB images overlaid with 2D poses ("+Skeleton"). Our approach outperforms these baselines in both scenarios, by jointly predicting future actions and characteristic 3D poses.

4.5.4 Ablation Studies

What is the effect of pose forecasting on long-term action understanding? Tab. 4.2 shows that there is a notable improvement in action accuracy between training only with an action loss vs. training action and 2D pose loss jointly. This becomes more apparent when training action only vs action and full pose prediction (2D and 3D losses). In addition, Fig. 4.3 shows the correspondence between autoregressive prediction length and action accuracy: jointly forecasting poses and actions enables more robust autoregressive forecasting over time. We conclude that pose forecasting is beneficial for long-term action understanding.

			MPII Cooking II				IKEA ASM			
Losses During Training			2D	3D	Action Accuracy		2D	3D	Action Accuracy	
Action	2D Proj.	3D Adv.	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
✓	×	×	–	–	21%	41%	–	–	24%	45%
✓	✓	×	62	0.10	26%	49%	46	0.05	27%	49%
×	✓	×	54	0.21	–	–	44	0.09	–	–
×	✓	✓	58	0.53	–	–	43	0.29	–	–
✓	✓	✓	**50**	**0.55**	**29%**	**51%**	**40**	**0.31**	**29%**	**50%**

Table 4.2: Ablation on the effect of the action, 2D projection, and 3D adversarial losses. Combining all together for joint forecasting enables complementary learning to produce the best performance.

How does action forecasting affect pose prediction performance? Tab. 4.2 demonstrates that pose forecasting trained jointly with action prediction is complementary and enables more accurate pose prediction.

What is the effect of characteristic pose forecasting? Since state-of-the-art pose forecasting focuses on fixed frame rate predictions independent of actions, we compare with such joint forecasting of action and pose where predicted poses are sampled at equally spaced points in time in Tab. 4.3 (uncoupled). Additionally, we consider alternative poses to forecast for each action rather than a characteristic 3D pose (middle of the annotated action range, and randomly selected within the action range). We keep the same pose representation for training and testing (i.e., evaluate on middle poses when trained on them, etc.), for a fair comparison. We observe the best performance when forecasting characteristic 3D poses along with action labels, showing their usefulness for forecasting long sequences of 3D poses and actions.

Poses		2D	3D	Action Accuracy	
Train	Test	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
Uncoupled	Uncoupled	75	0.29	28%	48%
Middle	Middle	58	0.45	26%	43%
Random	Random	67	0.37	22%	42%
Characteristic	**Characteristic**	**50**	**0.55**	**29%**	**51%**

Table 4.3: Ablation on pose forecasting on MPII Cooking II [110]. Our characteristic pose representation maximizes MPJPE and action performance: We consider pose prediction following state-of-the-art pose forecasting as decoupled from actions (uncoupled), as well as poses coupled to actions but in the middle of an action range, or at a random time therein, and our characteristic pose prediction. The same pose type is used for both train and evaluation.

4.5.5 Qualitative Results

Qualitative evaluations for the predicted poses are shown in Fig. 4.5 on data from MPII Cooking II [110] and in Fig. 4.4 on data from IKEA-ASM [111]. We compare our approach with state-of-the-art 3D pose forecasting of DLow [16], GSPS [127], and STARS [17]. For each method, we show a 3D body mesh in addition to the predicted 3D pose joints, to more comprehensively show the 3D structure of the forecasting results; we obtain body meshes by fitting SMPL [47] to each methods' predicted 3D body joints.

As there is no 3D ground truth available, we show the camera perspective with background for context as well as without background for a 3D pose only version. The two views demonstrate the fit to the ground truth 2D along with the quality of the 3D pose, respectively. Our approach leads to poses that better follow the ground-truth action poses in 2D compared to both previous methods while still maintaining a valid pose structure in 3D. Notably, this is true for both datasets, as our approach effectively forecasts the different data characteristics of both cooking as well as furniture assembly. In particular, our joint action-3D pose forecasting enables more accurate forecasting with diverse and accurate 3D pose structures.

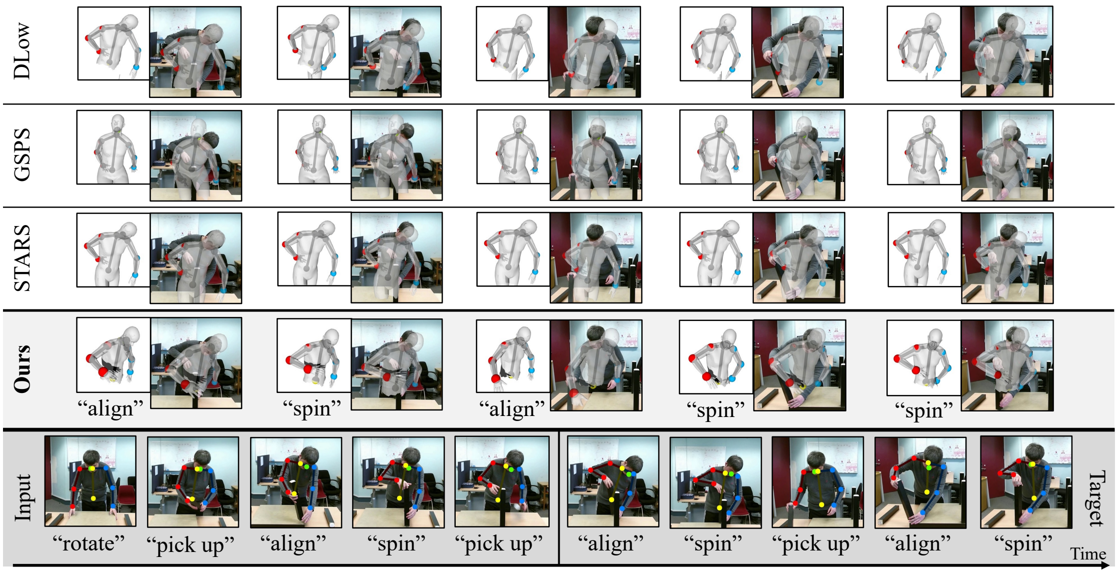


Figure 4.4: Qualitative comparison between DLow [16], GSPS [127], STARS [17], and our method on two sequences (left and right) from MPII Cooking II [110]. For each method, we show the 3D predicted pose projected into 2D, without background (small) and with background for context (full size). By considering both 3D pose and action forecasting together, we more effectively forecast the longer-term behavior.

4.5.6 Limitations

While we have demonstrated the potential of joint action and 3D pose forecasting, several limitations remain. For instance, our method leverages a separate 2D pose extraction as input to training, while an end-to-end formulation could potentially better leverage other useful signal in the input frames. Additionally, a more holistic body representation than pose joints would be important for finer-grained interactions that involve reasoning over small limbs (e.g., hands) and body surface contact.

4.6 Conclusion

In this paper, we proposed to forecast future human behavior by jointly predicting future actions alongside characteristic 3D poses. We do not require any 3D annotated action sequences, or 3D input data; instead, we learn complex action sequences from 2D action video data, and regularize predicted poses with an adversarial formulation against uncorrelated 3D pose data. Experiments demonstrate that our joint forecasting enables complementary feature learning, outperforming each individual task considered separately.

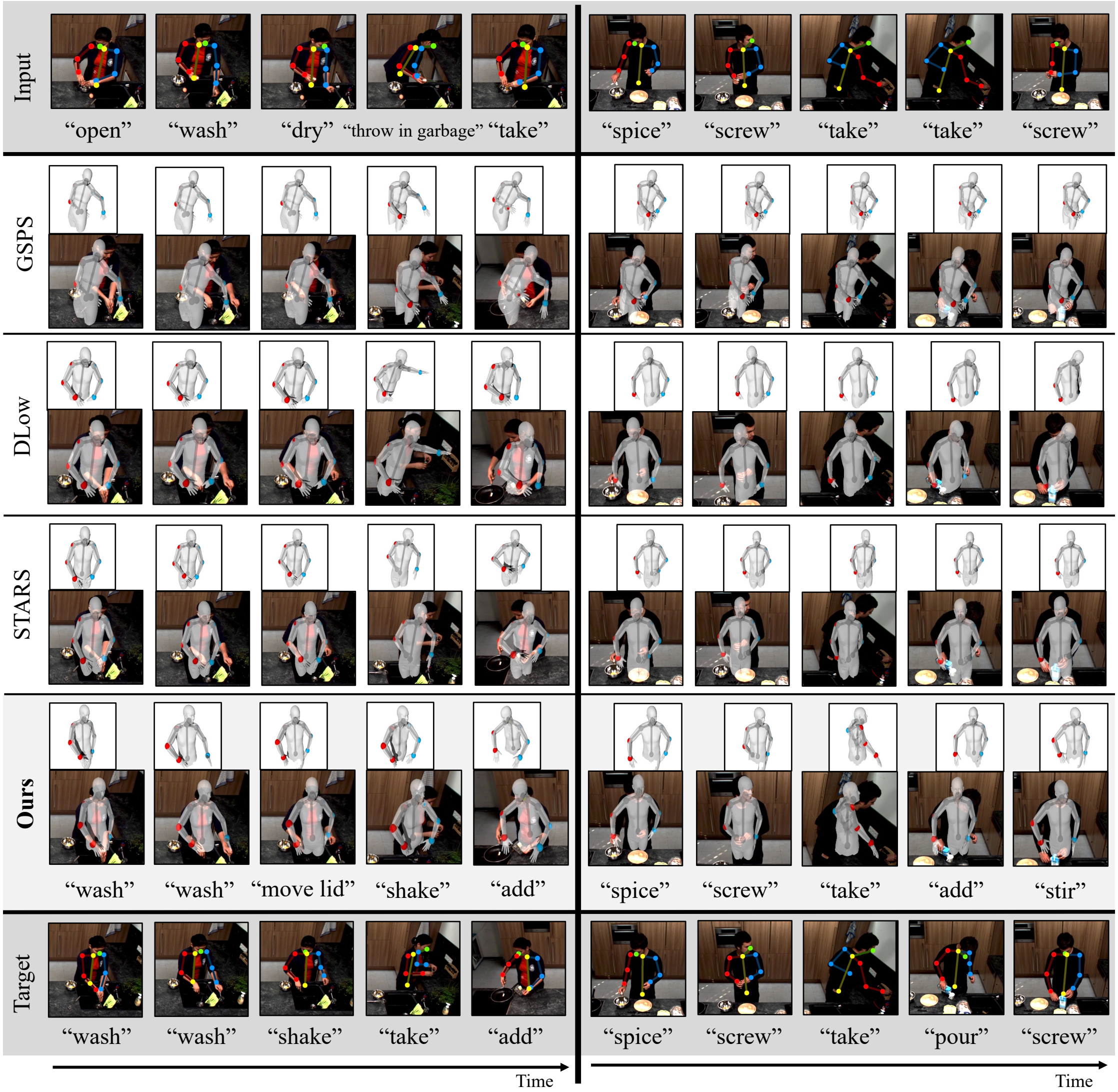


Figure 4.5: Qualitative comparison between DLow [16], GSPS [127], STARS [17], and our method on two sequences (left and right) from MPII Cooking II [110]. For each method, we show the 3D predicted pose projected into 2D, without background (small) and with background for context (full size). By considering both 3D pose and action forecasting together, we more effectively forecast the longer-term behavior.

4.7 Appendix

We show in this appendix additional qualitative and quantitative results, detail our baseline evaluation protocol, elaborate on the 3D quality metric, demonstrate the ability of our method to generalize to multi-actor scenarios, verify our method's robustness to 2D detection results, show the architecture used in our approach, and provide additional details regarding the data.

4.7.1 Additional Quantitative Results

Characteristic Poses Analogous to Tab. 2 in the main paper, Tab. 4.4 shows an ablation on pose timings and compares our approach of using characteristic poses to poses taken at regular time intervals ("uncoupled") as well as in the middle or at a random time of an action, on IKEA-ASM [111] data. To further illustrate this point, Tab. 4.5 shows additional ablations: Poses predicted at random points in the sequence, but at most 1s from the closest characteristic pose ("centered on the characteristic pose") and predicting characteristic poses but evaluating interpolated regularly spaced poses. Both demonstrate that the usage of characteristic poses improves performance compared to other approaches while still being outperformed by directly predicting characteristic poses.

	2D	3D	Action Accuracy	
Poses	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
Uncoupled	64	0.30	28%	48%
Middle	47	0.35	28%	47%
Random	49	0.24	28%	49%
Characteristic	**41**	**0.35**	**29%**	**50%**

Table 4.4: Ablation on pose forecasting, on the IKEA-ASM [111] dataset. We consider predicting poses following state-of-the-art pose forecasting in a decoupled fashion from actions (uncoupled), as well as poses coupled to actions in various fashions: middle (the middle pose of an action range), random (a random pose of the action), and our characteristic pose prediction, which benefits action prediction the most.

Input Noise Ablation Tab. 4.6 shows the effect using a noise vector as additional input to our method. It encourages more diversity in predictions, which benefits pose and action forecasting.

Input Objects Ablation Inputting initially observed objects slightly improves results (Tab. 4.6), due to added context for broad actions like "add," e.g."add ingredient" vs. "add water to pot.".

	2D	3D	Action Accuracy	
Poses	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
Uncoupled	75	0.29	28%	48%
Middle	58	0.45	26%	43%
Random	67	0.37	22%	42%
Centered on Char. Poses	69	0.33	28%	50%
Interp. from Char. Poses	62	0.13	29%	51%
Characteristic	**50**	**0.55**	**29%**	**51%**

Table 4.5: Ablation on pose forecasting on MPII Cooking II [110]. We consider pose prediction following state-of-the-art pose forecasting as decoupled from actions (uncoupled), as well as poses coupled to actions in various fashions: middle (the middle pose of an action range), random (a random pose of the action), random but at most 1s from the closest characteristic pose (centered), regularly spaced poses interpolated from characteristic pose prediction, and our characteristic pose prediction.

	MPII Cooking II				IKEA ASM			
	2d	3d	Action Accuracy		2d	3d	Action Accuracy	
Approach	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
No Objects	61	0.52	28%	51%	42	0.30	29%	50%
No Noise	55	0.49	29%	50%	48	0.29	**30%**	**51%**
Ours	**50**	**0.55**	**29%**	**51%**	**40**	**0.31**	29%	50%

Table 4.6: Ablations studies with no object input and no noise input.

4.7.2 Additional Qualitative Results

Fig. 4.6 shows additional qualitative results of our method, on both MPII Cooking 2 [110] (left column) and IKEA-ASM [111] (right column), as compared to pose baselines DLow [16], GSPS [127], and STARS [17].

4.7.3 Lifting 2D Predictions to 3D

In Tab. 4.1, we compare to first lifting input poses into 3D, then performing 3D motion prediction. Tab. 4.7 evaluates the other way around: Predicting 2D poses and action labels jointly with [151], then lifting the predicted 2D poses into 3D with RepNet [156] for evaluation. Our method outperforms both approaches.

	MPII Cooking II				IKEA ASM			
	2d	3d	Action Accuracy		2d	3d	Action Accuracy	
Approach	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
[151] + [156]	63	0.50	27%	43%	53	0.21	22%	46%
Ours	**50**	**0.55**	**29%**	**51%**	**40**	**0.31**	**29%**	**50%**

Table 4.7: Our approach of jointly forecasting 3D poses and actions achieves better performance compared to 2D pose + action forecasting [151] and then lifting forecasted 2D poses into 3D using [156].

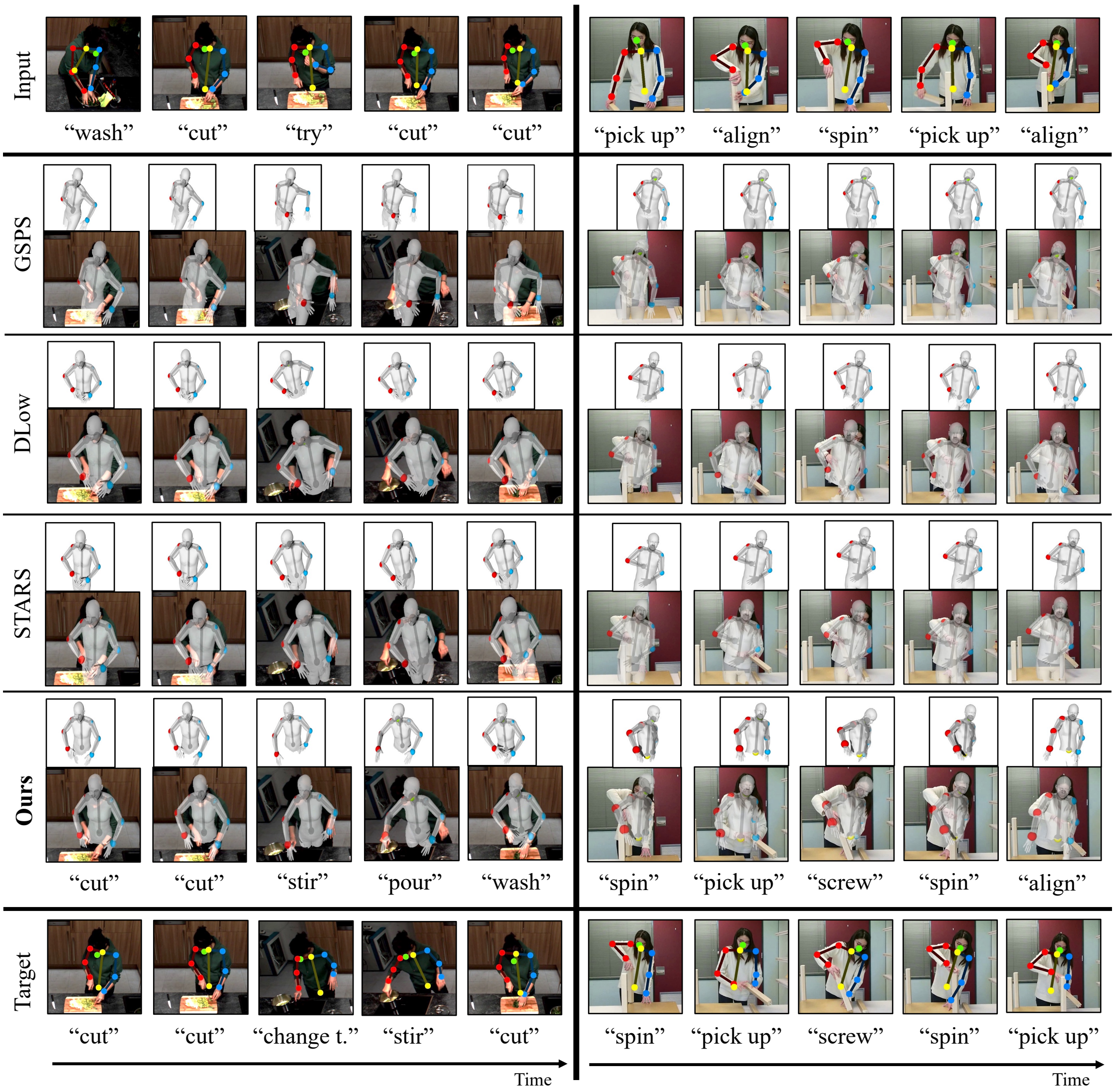


Figure 4.6: Additional qualitative comparison between DLow [16], GSPS [127], STARS [17], and our method on two sequences (left on MPII Cooking 2 [110], right on IKEA-ASM [111]). For each method, we show a the 3D predicted pose projected into the 2D target view, without background for a pose only version (small) as well as with background for context (full size).

4.7.4 Statistical Action Baselines

We additionally evaluate “Zero Velocity” and “Train Average” for action labels, analogous to forecasted poses, i.e. repeating the last action label and repeating the most frequent train action label, in Tab. 4.8. These baselines perform particularly poorly since actions are rarely repeated or fixed for entire sequences.

	MPII Cooking II		IKEA ASM	
Approach	top-1 ↑	top-3 ↑	top-1 ↑	top-3 ↑
Repeat Last Input	9%	43%	8%	35%
Most Common in Train	6%	10%	7%	26%
Ours	**29%**	**51%**	**29%**	**50%**

Table 4.8: Statistical action baselines: (1) Repeat the last input action label (2) Using the most common action label of the train set.

4.7.5 Baseline Evaluation Details

State-of-the-Art Pose Forecasting We evaluate the performance of our baselines using the same input data that is available to our method. Pose forecasting baselines DLow [16], GSPS [127], and STARS [17] are trained and evaluated on sequences of our manually annotated characteristic poses. Since there is no ground-truth 3D pose data available, we first use RepNet [156], a state-of-the-art 3D pose estimation method, to retrieve 3D skeletons from our 2D characteristic poses. We train this method from scratch using the same database of valid 3D poses that is available to our method, allowing for a fair comparison.

State-of-the-Art Action Label Forecasting We train action baselines AVT [11] and FUTR [12] using sequences of our characteristic pose frames together with the corresponding action labels as input. For AVT, we use their default parameters used by the original authors for their ablation on third-person dataset 50Salads/Breakfast, inputting our RGB frames instead. For a fair comparison, we also supply the action and object history for each step by encoding both label sequences with a small encoder (a single linear layer) each and fuse these features with the image features generated by the AVT encoder. For FUTR, we first generate I3D features [159] from our RGB frames and concatenate them with action and object history after encoding these in the same way as for AVT.

We then train two variants of both methods: One with the raw RGB frames, action history, and object history as input (“AVT RGB” and “FUTR RGB” in the main results figure), and one with additional 2D skeleton input (skeletons rendered on top of the RGB frames) from the skeletons that we extract with OpenPose [58] (“AVT RGB+Skeleton” and “FUTR RGB+Skeleton”).

4.7.6 Supervised 3D Pose Lifting

For better comparability, we used weakly supervised approach [156] for pose lifting. This is important, since there is no ground-truth coupling between 2D and corresponding 3D action poses in our setting.

	MPII Cooking II		IKEA ASM	
	2d	3d	2d	3d
Approach	MPJPE [px] ↓	Quality ↑	MPJPE [px] ↓	Quality ↑
SPIN [158] + DLow [16]	81	**0.89**	43	**0.43**
SPIN [158] + GSPS [127]	74	0.66	45	0.29
SPIN [158] + STARS [17]	66	0.80	41	0.40
Ours	**50**	0.55	**40**	0.31

Table 4.9: Comparison to pose baselines using fully-supervised pre-trained 3D pose estimation method SPIN [158]. In our main experiments, we instead compare to weakly supervised baseline RepNet [156] for a fair comparison.

Nevertheless, we compare to baselines [16, 127, 17] in Tab. 4.9 with poses lifted using fully supervised pre-trained SPIN [158]; our approach outperforms even these improved baselines in terms of 2D MPJPE.

4.7.7 3D Quality Metric Details

For our pose quality metric, we use a 3-layer MLP binary classifier of 3D poses. Training poses are randomly sampled from ground-truth (real) and predicted (fake) collected during the training process of our method and all baselines, producing a total of 100k real and fake poses each. Fake poses exhibit a range of small to large unrealistic deformations, depending on when they were sampled, ranging from random joint placements to widely inconsistent bone lengths to unnatural joint angles to only minor inconsistencies in the bone lengths. The classifier is trained once and then used to evaluate all methods, to ensure a fair comparison.

As an additional intuitive metric we show the mean absolute bone length difference of right and left body in 3D in Tab. 4.10. We observe that this metric correlates with our classifier-based quality.

	MPII Cooking II		IKEA ASM	
Approach	Symm. [mm] ↓	Quality ↑	Symm. [mm] ↓	Quality ↑
RepNet [156] + DLow [16]	**13**	**0.72**	45	0.31
RepNet [156] + GSPS [127]	18	0.66	56	0.15
RepNet [156] + STARS [17]	16	0.62	46	0.27
No 3D Adversarial Loss	75	0.10	66	0.05
2D Projection Loss Only	57	0.21	61	0.09
No Action Loss	22	0.53	39	0.29
Ours	22	0.55	**39**	**0.31**

Table 4.10: Additional quality metric and its correlation to our classifier-based metric: Absolute bone length difference btw. right and left body, compared to pose baselines and ablations.

4.7.8 Multi-Actor Interaction Scenario

In addition to our experiments with single human actors in the main paper, we show here that our approach is able to generalize to multi-actor scenarios, with minor modifications. We show this in Tab. 4.11 with additional dataset TICaM [160] where driver and passenger are interacting in an in-car driving scenario (actions include "talking", various handoffs). Our modifications are: **(1)** Additional encoder and decoder for the second person **(2)** Interaction pooling introduced in Social GAN [161]. Our modified method outperforms simple combinations of previous works, with and without interaction modelling, demonstrating the wide applicability of our method.

	2d	3d	Action Accuracy	
Approach	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
FUTR RGB + Skeleton	-	-	38%	64%
RepNet + STARS	89	0.34	-	-
Ours (No Interactions)	68	0.40	40%	67%
Ours (Interaction Modeling)	**58**	**0.41**	**48%**	**73%**

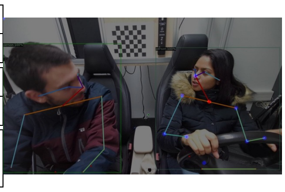


Table 4.11: Our approach can also be applied to multi-actor scenarios: We demonstrate improved performance on suitable dataset TICaM [160], with and without explicit interaction modeling.

4.7.9 2D Input Pose Quality

In Fig. 4.12, we replace OpenPose with AlphaPose [59] and Detectron2 [162], both only slightly changing the final results, indicating that our method does not depend on a specific 2D pose detector. We also experiment with added random noise to OpenPose: our method remains relatively robust. The coupled changes in pose and action accuracy further demonstrate the effectiveness of our joint feature learning.

MPII Cooking II	2d	3d	Action Accuracy	
Approach	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
OpenPose + max. 20px noise	59	0.45	26%	47%
OpenPose + max. 10px noise	57	0.47	26%	46%
Ours (using Detectron2)	47	**0.54**	28%	55%
Ours (using AlphaPose)	**46**	0.57	28%	**56%**
Ours (using OpenPose)	50	0.55	**29%**	51%

Table 4.12: Robustness of our method to different 2D pose detectors Detectron2 [162] and AlphaPose [59] as well as randomly added 2D noise. This only slightly affects our pose and action accuracy, further demonstrating the effectiveness of our joint feature learning.

4.7.10 Architecture Details

Generator Network Fig. 4.7 shows our generator architecture in detail with input and output dimensions for linear layers, and the slope for leaky ReLU layers.

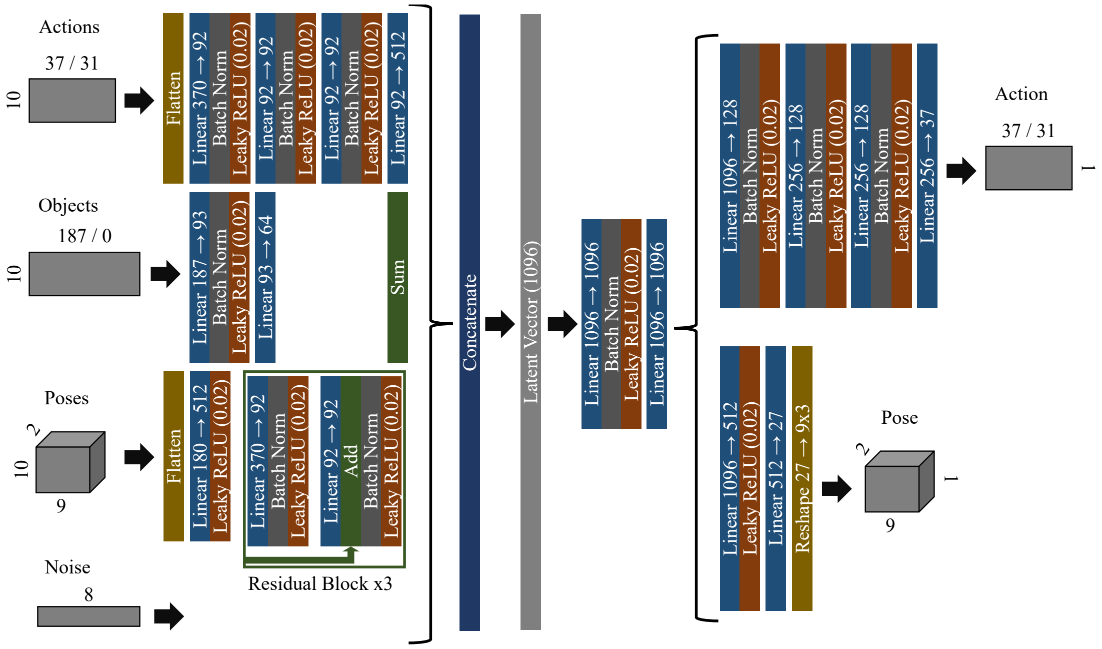


Figure 4.7: Network architecture specification.

Critic Network Our adversarial critic network processes generator outputs with 4 linear layers and 3 kinematic chain layers which are designed to encourage correct bone lengths (as shown in [156]), in parallel. 2 linear layers then combine both outputs and produce the final critic score.

4.7.11 Data Details

Camera Parameters While intrinsic camera parameters are often available in captured image data, the camera parameters for captured video were not available from the MPII Cooking 2 [110] dataset to use for pose projection. We thus optimized for intrinsic camera parameters from the video sequence data in correspondence with the 3D scene reconstruction of the empty kitchen environment, as given by [163]. For IKEA-ASM [111], we use the provided intrinsic camera parameters directly. Note that camera parameters are only required to be fixed within a sequence (i.e. no moving camera) but can change between sequences.

3D Pose Database Alignment We use popular 3D pose datasets Human3.6m [45], AMASS [112], and GRAB [44] for our database of uncorrelated valid 3D poses. All poses are pre-processed to follow the OpenGL coordinate system and aligned with respect to the neck joint.

Pose Joint Layout We use the 9 upper-body joints of the native OpenPose [58] joint layout for skeletons in 2D, and adapt skeletons in our 3D database to use the same

Ours		OpenPose		Human3.6m		SMPL-X	
Idx	Name	Idx	Name	Idx	Name	Idx	Name
0	head	0	nose	15	head	15	head
1	neck	1	neck	13	thorax	12	neck
2	right shoulder	2	right shoulder	25	right shoulder	17	right shoulder
3	right elbow	3	right elbow	26	right elbow	19	right elbow
4	right hand	4	right hand	27	right wrist	42	right index 3
5	left shoulder	5	left shoulder	17	left shoulder	16	left shoulder
6	left elbow	6	left elbow	18	left elbow	18	left elbow
7	left hand	7	left wrist	19	left wrist	27	left index 3
8	hip	8	mid-hip	0	hip	0	pelvis

Table 4.13: Human skeleton joint layout used in our experiments, for both 2D and 3D skeletons.

format. Tab. 4.13 shows the correspondence between our joint layout, OpenPose [58], Human3.6m [45], and SMPL-X [49]. 3D datasets AMASS [112] and GRAB [44] provide human bodies in SMPL-X format; we first extract their skeleton joints using their publicly available code and then convert it into our layout using the correspondences in Tab. 4.13.

MPII Cooking 2 Details We use action labels as annotated in the 2D cooking action dataset MPII Cooking 2 [110]. These annotations provide action labels (87 classes) for frame ranges in each sequence as well as the involved objects (187 classes). We first cluster similar actions together, yielding a total of 37 action clusters, which we then use as action classes in our experiments.

In addition, since our goal is to forecast upper-body actions with objects in the foreground, we remove instances of poses and corresponding actions that occur in the background - e.g., when taking out objects from the cupboard, or from the fridge.

In total, there are 272 cooking action sequences; we create a random train/val/test split along sequences with a ratio of 70% / 15% / 15%, yielding 190, 40, 40 sequences for each set.

IKEA-ASM Details We use action labels as annotated in the IKEA furniture assembly dataset IKEA-ASM [111]. These annotations provide action labels (31 classes) for frame ranges in each sequence; we use them without explicit object information since each action already encodes its associated object.

In total, there are 370 furniture assembly action sequences; we create a random train/val/test split along sequences with a ratio of 70% / 15% / 15%, yielding 227, 48, 48 sequences for each set.

5 Contact-Guided 3D Human-Object Interactions

This chapter introduces the following paper:

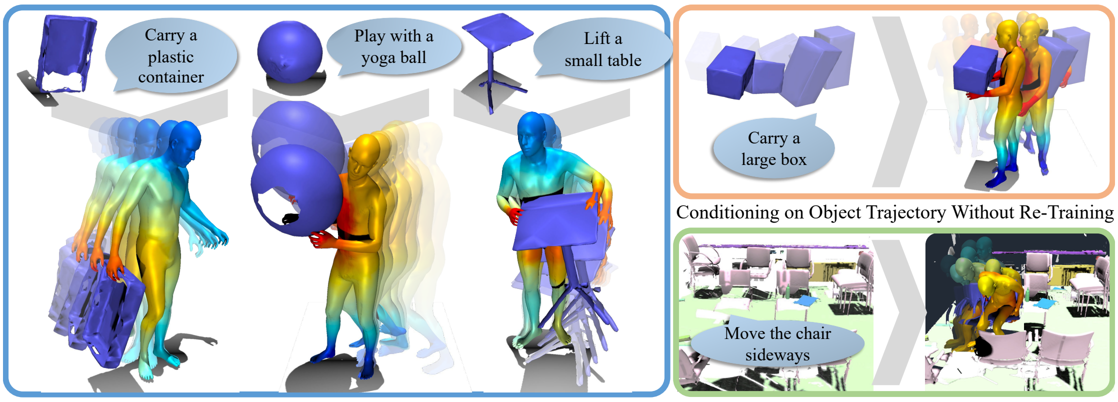
Christian Diller, and Angela Dai. ”CG-HOI: Contact-Guided 3D Human-Object Interactions.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Abstract of Paper We propose CG-HOI, the first method to address the task of generating dynamic 3D human-object interactions (HOIs) from text. We model the motion of both human and object in an interdependent fashion, as semantically rich human motion rarely happens in isolation without any interactions. Our key insight is that explicitly modeling contact between the human body surface and object geometry can be used as strong proxy guidance, both during training and inference. Using this guidance to bridge human and object motion enables generating more realistic and physically plausible interaction sequences, where the human body and corresponding object move in a coherent manner. Our method first learns to model human motion, object motion, and contact in a joint diffusion process, inter-correlated through cross-attention. We then leverage this learned contact for guidance during inference to synthesize realistic and coherent HOIs. Extensive evaluation shows that our joint contact-based human-object interaction approach generates realistic and physically plausible sequences, and we show two applications highlighting the capabilities of our method. Conditioned on a given object trajectory, we can generate the corresponding human motion without re-training, demonstrating strong human-object interdependency learning. Our approach is also flexible, and can be applied to static real-world 3D scene scans.

Contribution The method development and implementation was done by the first author. Discussions with the co-authors led to the final paper.

5.1 Introduction

Generating human motion sequences in 3D is important for many real-world applications, e.g. efficient realistic character animation, assistive robotic systems, room layout planning, or human behavior simulation. Crucially, human interaction is interdependent with the object(s) being interacted with; the object structure of a chair or ball, for instance, constrains the possible human motions with the object (e.g., sitting, lifting), and the human action often impacts the object motion (e.g., sitting on a swivel chair, carrying a backpack).



Conditioning on Object Trajectory Without Re-Training

Contact-Guided 3D Human-Object Interaction Synthesis from Text

Application to Objects in Static 3D Scene Scans

Figure 5.1: We present an approach to generate realistic 3D human-object interactions (HOIs), from a text description and given static object geometry to be interacted with (left). Our main insight is to explicitly model contact (visualized as colors on the body mesh, closer contact in red), in tandem with human and object sequences, in a joint diffusion process. In addition to synthesizing HOIs from text, we can also synthesize human motions conditioned on given object trajectories (top right), and generate interactions in static scene scans (bottom right).

Existing works typically focus solely on generating dynamic humans, and thereby disregarding their surroundings [34, 35, 36, 37, 38, 39], or grounding such motion generations in a static environment that remains unchanged throughout the entire sequence [18, 19, 20, 21, 22, 79, 24, 25, 26]. However, real-world human interactions affect the environment. For instance, even when simply sitting down on a chair, the chair is typically moved: to adjust it to the needs of the interacting human, or to move it away from other objects such as a table. Thus, for realistic modeling of human-object interactions, we must consider the interdependency of object and human motions.

We present CG-HOI, the first approach to address the task of generating realistic 3D human-object interactions from text descriptions, by jointly predicting a sequence of 3D human body motion along with the object motion. Key to our approach is to not only model human and object motion, but to also explicitly model contact as a bridge between human and object. In particular, we model contact by predicting contact distances from the human body surface to the closest point on the surface of the object being interacted with. This explicit modeling of contact helps to encourage human and object motion to be semantically coherent, as well as to provide a constraint indicating physical plausibility (e.g., discouraging objects to float without support).

CG-HOI jointly models human, object, and contact together in a denoising diffusion process. Our joint diffusion model is designed to encourage information exchange between all three modalities through cross-attention blocks. Additionally, we employ a contact weighting scheme, based on the insight that object motion, when being manipulated by a human, is most defined by the motion of the body part in closest contact (Fig. 5.3). We make use of this by generating separate object motion hypotheses for multiple parts of the human body and aggregating them based on that part's predicted contact. During inference, we leverage the predicted contact distances to refine synthesized sequences through our contact-based diffusion guidance, which penalizes synthesizing sequences with human-object contact far from the predicted contact distances.

Our method is able to generate realistic and physically plausible human-object interactions, and we evaluate our approach on two widely-used interaction datasets, BEHAVE [164] and CHAIRS [165]. In addition, we also demonstrate the usefulness of our model with two related applications: First, generating human motion given a specific object trajectory without any retraining, which demonstrates our learned human-object motion interdependencies. Second, populating a static 3D scene scan with human-object interactions of segmented object instances, showing the applicability of our method to general real-world 3D scans.

In summary, our contributions are three-fold:

- We propose an approach to generate realistic, diverse, and physically plausible human-object interaction sequences by jointly modeling human motion, object motion, and contact through cross-attention in a diffusion process.
- We formulate a holistic contact representation: Object motion hypotheses are generated for multiple pre-defined points on the surface of the human body and aggregated based on predicted contact distances, enabling comprehensive body influence on contact while focusing on the body parts in closer contact to the object.
- We propose a contact-based guidance during synthesis of human-object interactions, leveraging predicted contacts to refine generated interactions, leading to more physically plausible results.

5.2 Related Work

3D Human Motion Generation. Generating sequences of 3D humans in motion is a task which evolved noticeably over the last few years. Traditionally, many methods used recurrent approaches [13, 94, 95, 97, 27, 28] and, improving both fidelity and predicted sequence length, graph- and attention-based frameworks [14, 15, 31]. Notably, generation can either happen deterministically, predicting one likely future human pose sequence [13, 28, 14, 15, 41], or stochastically, thereby also modeling the uncertainty inherent to future human motion [103, 101, 127, 102, 16, 40, 166, 167].

Recently, denoising diffusion models [76, 77] showed impressive results in 2D image generation, producing high fidelity and diverse images [78, 77]. Diffusion models allow for guidance during inference, with classifier-free guidance [168, 169] widely used to trade off between generation quality and diversity. Inspired by these advances, various methods have been proposed to model 3D human motion through diffusion, using U-Nets [34, 35, 36, 37, 38, 39], transformers [1, 75, 79, 35, 80, 81, 82, 83, 84, 85, 86, 87], or custom architectures [88, 89, 90, 91, 92]. Custom diffusion guidance has also been shown to aid controllability [170, 171, 172] and physical plausibility [173].

In addition to unconditional motion generation, conditioning on text descriptions allows for more control over the generation result [1, 82, 84, 87, 36, 38]. In fact, generating plausible and corresponding motion from textual descriptions has been an area of interest well before the popularity of diffusion models [46, 39, 174, 175, 176, 177, 178].

These methods show strong potential for 3D human motion generation, but focus on a skeleton representation of the human body, and only consider human motion in isolation, without naturally occurring interactions. To generate realistic human-object interactions, we must consider the surface of the human body and its motion with respect to object motion, which we characterize as contact.

3D Human Motion in Scenes. As human motion typically occurs not in isolation but in the context of an object or surrounding environment, various methods have explored learning plausible placement of humans into scenes, both physically and semantically, [179, 180, 181, 182, 183, 184], forecasting future motion given context [104, 185], or generating plausible walking and sitting animations [18, 19, 20, 186, 187, 21, 22, 23, 24, 25, 26, 188]. This enables more natural modeling of human reactions to their environment; however, the generated interactions remain limited due to the assumption of a static scene environment, resulting in a focus on walking or sitting movements.

Recent methods have also focused on more fine-grained interactions by generating human motion given a single static object [189, 190, 191, 192, 193, 194, 195]. While these methods only focus on human motion generation for a static object, [196] generates human motion conditioned on object motion and [2, 197] generate full human-object interaction sequences directly from an initial sequence observation. Our approach also models both human and object motion, but we formulate a flexible text-conditioned generative model for dynamic human and object motion, modeling the interdependency between human, object, and contact to synthesize more realistic interactions under various application settings.

Contact Prediction for Human-Object Interactions. While there is a large corpus of related work for human motion prediction, only few works focus on object motion generation [198, 199, 200, 201]. Notably, these methods predict object movement in isolation, making interactions limited, as they typically involve interdependency with human motion.

Contact prediction has been most studied in recent years for the task of fine-grained hand-object interaction [202, 203, 204, 205, 206, 207, 208]. It is defined either as binary labels on the surface [202, 203, 204, 205, 207, 208] or as the signed distance to a corresponding geometry point [206]. In these works, predicting object and hand states without correct contact leads to noticeable artifacts. Contact prediction itself has also been the focus of several works [192, 209, 210, 211], either predicting contact areas or optimizing for them.

Applied to the task of generating whole-body human-object interactions, this requires access to the full surface geometry of both object and human. Only few recent motion generation works focus on generating full-body geometric representations of humans [24, 212, 62, 176, 213, 167, 214] instead of simplified skeletons which is a first step towards physically correct interaction generation. However, while several of these works acknowledge that contact modeling would be essential for more plausible interactions [62, 176, 24], they do not model full-body contact.

We approach the task of generating plausible human-object motion from only the object geometry and a textual description as a joint task and show that considering the joint behavior of full-body human, object, and contact between the two benefits output synthesis to generate realistic human-object interaction sequences.

5.3 Method Overview

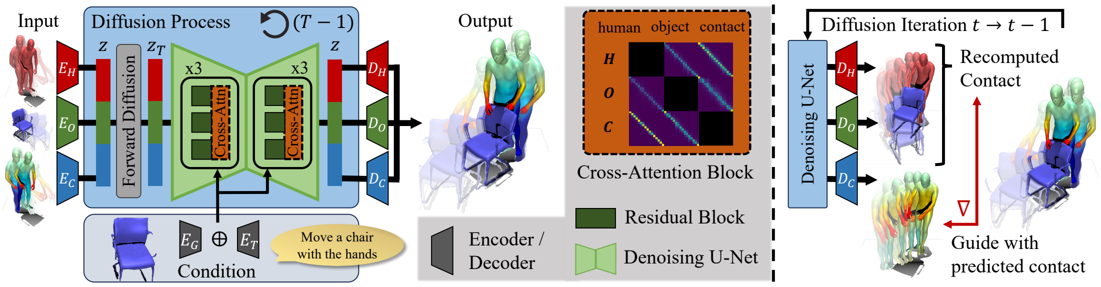


Figure 5.2: Method Overview. Given a text description and object geometry, CG-HOI produces a human-object interaction (HOI) sequence, modeling both human and object motion. To produce realistic HOIs, we additionally model contact to bridge the interdependent motions. Our method jointly generates all three during training (left), using a U-Net-based diffusion with cross-attention across human, object, and contact. During inference (right), we drive synthesis under guidance of estimated contact to sample more physically plausible interactions.

CG-HOI jointly generates sequences of human body and object representations, alongside contact on the human body surface. Reasoning jointly about all three modalities in both training and inference enables generation of semantically meaningful human-object interaction sequences.

Fig. 5.2 shows a high-level overview of our approach: We consider as condition a brief text description T of the action to be performed, along with the static geometry G of the object to be interacted with, and generate a sequence of F frames $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_F]$ where each frame $\mathbf{x}_i$ consists of representations for the object transformation o_i, for the human body surface h_i, and for the contact c_i between human and object geometry. We denote as $H = \{h_i\}$ the human body representations, $O = \{o_i\}$ the object transformations, and $C = \{c_i\}$ the contact representations.

We first train a denoising diffusion process to generate H, O, and C, using a U-Net architecture with per-modality residual blocks and cross-attention modules. Using cross-attention between human, object motion, and contact allows for effectively learning interdependencies and and feature sharing (Sec. 5.4). We use the generated contact to guide both training and inference: Instead of predicting one object motion hypothesis per sequence, we generate multiple, and aggregate them based on predicted contacts, such that body parts in closer contact with the object have a stronger correlation with the final object motion (Sec. 5.4.3). During inference, the trained model generates H, O, and C. For each step of the diffusion inference, we use predicted contact C to guide the generation of H and O, by encouraging closeness of recomputed contact and predicted contact, producing more refined and realistic interactions overall (Sec. 5.5).

5.4 Human-Object Interaction Diffusion

5.4.1 Probabilistic Denoising Diffusion

Our approach uses a diffusion process to jointly generate a sequence of human poses, object transformations, and contact distances in a motion sequence from isotropic Gaussian noise in an iterative process, removing more noise at each step. More specifically, during training we add noise depending on the time step ("forward process") and train a neural network to reverse this process, by directly predicting the clean sample from noisy input. Mathematically, the forward process follows a Markov chain with T steps, yielding a series of time-dependent distributions $q(\mathbf{z}_t|\mathbf{z}_{t-1})$ with noise being injected at each time step until the final distribution $\mathbf{z}_T$ is close to $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Formally,

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\sqrt{\beta_t}\mathbf{z}_{t-1} + (1-\beta_t)\mathbf{I}) \tag{5.1}$$

with the variance of the Gaussian noise at time t denoted as β_t, and $\beta_0 = 0$.

Since we adopt the Denoising Diffusion Probabilistic Model [215], we can sample $\mathbf{z}_t$ directly from $\mathbf{z}_0$ as

$$\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{z}_0 + \sqrt{1-\alpha_t}\epsilon \tag{5.2}$$

with $\alpha_t = \prod_{t'=0}^{t}(1-\beta_t)$, and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

For the reverse process, we follow [1, 2, 35], directly recovering the original signal $\tilde{\mathbf{z}}$ instead of the added noise.

Human-Object Interactions To model human-object interactions with diffusion, we employ our neural network formulation $\mathcal{F}$. $\mathcal{F}$ operates on the noised vector of concatenated human, object, and contact representations, together with the current time step t, and a condition consisting of object point cloud G, encoded by an encoder E_G, and text information T, encoded by encoder E_T. Formally,

$$\tilde{\mathbf{z}} = \mathcal{F}(\mathbf{z}_t, t, E_G(G) \oplus E_T(T)) \tag{5.3}$$

More specifically, in our scenario E_T extracts text features with a pre-trained CLIP [216] encoder. Encoder E_G processes object geometry G as a uniformly sampled point cloud in world coordinate space with a PointNet [217] pre-trained on object parts segmentation.

Object transformations o_i are represented as global translation and rotation using continuous 6D rotation representation [218]. In contrast to prior work [16, 40, 1, 82, 84, 194, 195] which focused on representing human motion in a simplified manner as a collection of J human joints, disregarding both identity-specific and pose-specific body shape, we model physically plausible human-object contacts between body surface and geometry. Thus, we represent the human body h_i in SMPL [47] parameters: $h_i = \{h_i^p, h_i^b, h_i^r, h_i^t\}$ with pose parameters $h_i^p \in \mathbb{R}^{63}$, shape parameters $h_i^b \in \mathbb{R}^{10}$, as well as global rotation $h_i^r \in \mathbb{R}^3$ and translation $h_i^t \in \mathbb{R}^3$. These body parameters can then be converted back into a valid human body surface mesh in a differentiable manner using the SMPL [47] model. This allows us to reason about the contact between human body surface and object geometry. We represent contact c_i on the human body as the distance between a set of $M = 128$ uniformly distributed motion markers on the body surface to the closest point of the object geometry, for each marker. Specifically, we represent contact for frame $\mathbf{x}_i$ and j-th contact marker ($j \in \{0..M-1\}$) c_i^j as its distance from the human body surface to the closest point on the same frame's object geometry surface.

5.4.2 Human-Object-Contact Cross-Attention

We jointly predict human body sequences $H = \{h_i\}$, object transformations $O = \{o_i\}$, and corresponding contact distances $C = \{c_i\}$ in our diffusion approach. We employ a U-Net backbone for diffusion across these outputs, with separate residual blocks for human, object, and contact representations, building modality-specific latent feature representations. As we aim to model the inter-dependency across human, object, and contact, we introduce custom human-object-contact cross-attention modules after every residual block where each modality attends to the other two.

We follow the formulation of Scaled Dot-Product Attention [68], computing the updated latent human body feature:

$$h_i = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}}\right)\boldsymbol{V}, \tag{5.4}$$

with query $\boldsymbol{Q} = H$, and key and value $\boldsymbol{K} = \boldsymbol{V} = O \odot C$ ($\odot$ denotes concatenation), i.e. $\boldsymbol{Q} \in \mathbb{R}^{F \times d}$ and $\boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{2F \times d}$. As in [68], d denotes the dimensionality of query and key. Applying this similarly to O and C yields the final features after each cross-attention module.

5.4.3 Contact-Based Object Transform Weighting

As visualized in Fig. 5.3, object motion is naturally most influenced by parts of the human body in very close contact to the object (as they are often the cause of that motion), and less impacted (if at all) by body parts further away. For instance, if a person moves an object with their hands, the object follows the hands but not necessarily other body parts (e.g., body and feet may remain static or walk in a different direction). Thus, instead of directly generating one object motion hypothesis o_i alongside the corresponding human motion h_i, we couple o_i to the M body contact points $j \in \{0..M-1\}$ and their predicted distances $\{c_i^j\}$ between human body surface and object geometry.

Formally, we predict object transformation hypotheses o_i^j for each contact point on the human body, and weigh them with the inverse of their predicted contact distance c_i^j:

$$o_i = \frac{1}{\sum_j \max(|c_i|) - |c_i^j|} \sum_{j=0}^{M-1} (\max(|c_i|) - |c_i^j|) o_i^j \tag{5.5}$$

5.4.4 Loss Formulation

During training, the input is a noised vector $\mathbf{z}$, containing F frames $\{\mathbf{x}_i\}$, each a concatenation of human body representation h_i, object transformation o_i, and contact parameters c_i. As condition $\mathbf{C}$, we additionally input encoded object geometry G and text description T. The training process is then supervised with the ground-truth sequence containing $\hat{h_i}, \hat{o_i}, \hat{c_i}$, minimizing a common objective:

$$\mathbf{L} = \lambda_h ||h_i - \hat{h_i}||_1 + \lambda_o ||o_i - \hat{o_i}||_1 + \lambda_c ||c_i - \hat{c_i}||_2, \tag{5.6}$$

with $\lambda_h = 1.0, \lambda_o = 0.9, \lambda_c = 0.9$. We use classifier-free guidance [168] for improved fidelity during inference, thus masking out the conditioning signal with 10% probability.

5.5 Interaction Generation

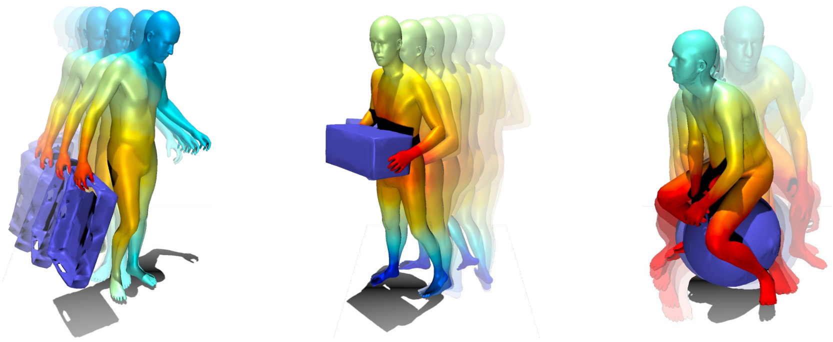


Figure 5.3: An object's trajectory is largely defined by the motion of the region of the body in close contact with the object, e.g. the hand(s) when carrying an object (left, middle) or the lower body when moving with an object while sitting (right). This informs our contact-based approach to generating object motion.

Using our trained network model, we can generate novel human-object interaction sequences for a given object geometry and a short text description using our weighting scheme for generating object transformations, and a custom guidance function on top of classifier-free guidance to generate physically plausible sequences.

Specifically, we use our trained model to reverse the forward diffusion process of Eq. 5.2: Starting with noised sample $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we iteratively use our trained network model $\mathcal{F}$ to estimate cleaned sample $\mathbf{z}_0$:

$$\mathbf{z}_{t-1} = \sqrt{\alpha_t}\tilde{\mathbf{z}} + \sqrt{1-\alpha_t}\epsilon. \tag{5.7}$$

5.5.1 Contact-Based Diffusion Guidance

While our joint human-object-contact training already leads to plausible motions, generated sequences are not explicitly constrained to respect contact estimates during inference, which can lead to inconsistent contact between human and object motion (e.g., floating objects). Thus, we introduce a contact-based guidance during inference to refine predictions, using a cost function $\mathcal{G}(\mathbf{z}_t) = ||c_t - \overline{c}_t||_2^2$ which takes as input the denoised human, object, and contact predictions $\mathbf{z}_t = [h_t, o_t, c_t]$ at diffusion step t and compares predicted c_t and actual contact distances $\overline{c}_t$ for each contact point. Based on this, we then calculate the gradient $\nabla_{\mathbf{z}_t}\mathcal{G}(\mathbf{z}_t)$.

We use this gradient for diffusion guidance, following [170], by re-calculating the mean prediction μ_t at each time t:

$$\hat{\mu}_t = \mu_t + s\sum_t \nabla_{x_t}\mathcal{G}(x_t), \tag{5.8}$$

for a scaling factor s. This guidance is indirect but dense in time, and is able to correct physical contact inconsistencies in the predicted sequences during inference time, without requiring any explicit post-processing steps.

5.5.2 Conditioning on Object Trajectory

While our model has been trained with text and static object geometry as condition, we can also apply the same trained model for conditional generation of a human sequence given an object sequence and text description. Note that this does not require any re-training, as our model has learned a strong correlation between human and object motion. Instead, we use a replacement-based approach, and inject the given object motion O' into the diffusion process during inference at every step. Following Eq. 5.7, we obtain:

$$\mathbf{z}_{t-1} = \sqrt{\alpha_t}\tilde{\mathbf{z}}'_t + \sqrt{1-\alpha_t}\epsilon, \tag{5.9}$$

with $\tilde{\mathbf{z}}' = [h_t, o'_t, c_t]$, concatenating human motion h_t, contact distances c_t, and injected given object motion o'_t.

5.6 Results

We evaluate our approach using two commonly used human-object interaction datasets BEHAVE [164] and CHAIRS [165] on a range of metrics, measuring motion fidelity and diversity. We show that our approach is able to generate realistic and diverse motion on both datasets, across a variety of objects and types of interactions.

		BEHAVE				CHAIRS			
Task	Approach	R-Prec. (top-3) ↑	FID ↓	Diversity →	MModality →	R-Prec. (top-3) ↑	FID ↓	Diversity →	MModality →
	Real (human)	0.73	0.09	4.23	4.55	0.83	0.01	7.34	3.00
Text-Cond. Human Only	MDM [1]	0.52	4.54	5.44	5.12	0.72	5.99	**6.83**	3.45
	InterDiff [2]	0.49	5.36	**3.98**	3.98	0.63	6.76	5.24	2.44
	Ours	**0.60**	**4.26**	4.92	**4.10**	**0.78**	**5.24**	7.90	**3.22**
	Real	0.81	0.17	6.80	6.24	0.87	0.02	9.91	6.12
Motion-Cond. HOI	InterDiff [2]	0.68	3.86	5.62	5.90	0.67	4.83	7.49	4.87
	Ours	**0.71**	**3.52**	**6.89**	**6.43**	**0.79**	**4.01**	**8.42**	**6.29**
Text-Cond. HOI	MDM [1]	0.49	9.21	6.51	8.19	0.53	9.23	6.23	7.44
	InterDiff [2]	0.53	8.70	3.85	4.23	0.69	7.53	5.23	4.63
	Ours	**0.62**	**6.31**	**6.63**	**5.47**	**0.74**	**6.45**	**8.91**	**5.94**

Table 5.1: Quantitative comparison with state-of-the-art approaches MDM [1] and InterDiff [2]. Human Only results are evaluated only on the human pose sequence, and motion-cond. denotes predictions additionally conditioned on past observations of both human and object behavior. For metrics with →, results closer to the real distribution are better. Our approach outperforms these baselines in all three settings, indicating a strong learned correlation between human and object motion.

5.6.1 Experimental Setup

Datasets We conduct our experiments on two datasets containing interactions between whole-body 3D humans and corresponding objects. CHAIRS [165] captures 46 subjects as their SMPL-X [49] bodies interacting with 81 different types of chairs and sofas. We extract sequences in which both human and object are in motion, yielding ≈ 1300 HOI sequences, each labeled with a text description. We use a random 80/10/10 split along object classes, ensuring that test objects are not seen during training. BEHAVE [164] captures 8 participants as their SMPL-H [48] parameters alongside 20 different objects. This yields ≈ 520 sequences with corresponding text descriptions. We use their original train/test split. We sample both datasets at 20 frames per second, and generate 32 frames for CHAIRS and 64 for BEHAVE, leading to generated motion that lasts up to 3 seconds.

Implementation Details We train our model with batch size 64 for 600k steps (≈24 hours), after which we choose the checkpoint that minimizes validation FID, following [2]. Our attention uses 4 heads and a latent dimension of 256. Input text is encoded using a frozen CLIP-ViT-B/32 model. For classifier-free guidance during inference time, we use a guidance scale of 2.5, which empirically provides a good trade-off between diversity and fidelity. For our inference-time contact-based guidance, we use scale $s = 100.0$.

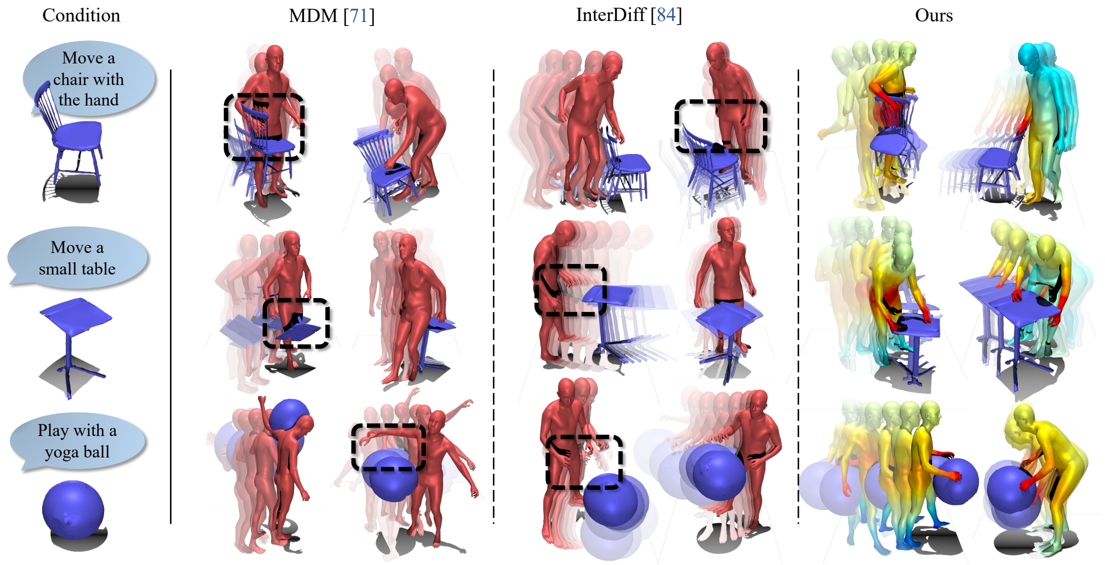


Figure 5.4: Qualitative comparison to state-of-the-art methods MDM [1] and InterDiff [2]. Our approach generates high-quality HOIs by jointly modeling contact (closer contact in red), reducing penetration and floating artifacts (black highlight boxes).

5.6.2 Evaluation Metrics

We measure realism and diversity of combined human and object motion, alongside closeness to the text description, following established practices [1, 46, 219]. We first train a joint human-object motion feature extractor and separate text feature extractor using a contrastive loss to produce geometrically close feature vectors $\{v_i\}$ for matched text-motion pairs, and report the following metrics:

R-Precision measures the closeness of the text condition and generated HOI in latent feature space, and reports whether the correct match falls in the top 3 closest feature vectors.

Frechet Inception Distance (FID) is commonly used to evaluate the similarity between generated and ground-truth distribution in encoded feature space.

Diversity and MultiModality. Diversity measures the motion variance across all text descriptions and is defined as $\frac{1}{N}\sum_{i=1}^{N}||v_i - v_i'||_2$ between two randomly drawn subsets $\{v_i\}$ and $\{v_i'\}$. MultiModality (MModality) measures the average such variance intra-class, for each text description.

Perceptual User Study. The exact perceptual quality of human-object interactions is difficult to capture with any single metric; thus, we additionally conduct a user study with 32 participants to evaluate our method in comparison to baseline approaches. Participants are shown 10 baseline vs. ours pairs each in side-by side views of sequences with the same geometry and text conditioning, and asked to choose 1) Which one follows the given text better and 2) Which one looks more realistic overall.

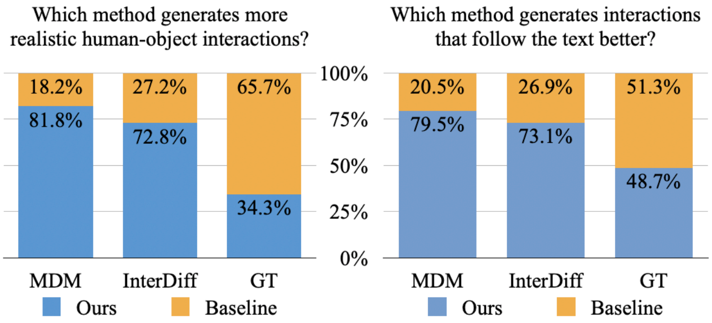


Figure 5.5: Perceptual User Study. Participants significantly favor our method over baselines, for overall realism and text coherence.

	BEHAVE				CHAIRS			
Approach	R-Prec. (top-3) ↑	FID ↓	Diversity →	MModality →	R-Prec. (top-3) ↑	FID ↓	Diversity →	MModality →
Real	0.81	0.17	6.80	6.24	0.87	0.02	9.91	6.12
No cross-attention	0.35	10.44	8.23	7.40	0.49	10.84	12.22	10.64
No contact prediction	0.41	9.64	10.10	**6.89**	0.41	8.53	11.56	9.15
Separate contact pred.	0.47	8.01	5.12	5.12	0.52	9.34	7.65	4.62
No contact weighting	0.55	8.54	6.52	5.29	0.64	7.55	8.56	5.45
No contact guidance	0.59	7.22	7.84	5.30	0.70	7.41	8.05	5.76
Ours	**0.62**	**6.31**	**6.63**	5.47	**0.74**	**6.74**	**8.91**	**5.94**

Table 5.2: Ablation on our design choices. Joint contact prediction with cross-attention encourages the generation of more natural HOIs, and our weighting scheme and inference-time contact guidance together enable the best generation performance.

5.6.3 Comparison to Baselines

As our method is the first to enable generating human and object motion from text, there are no baselines available for direct comparison. InterDiff [2] is closest to our approach, performing forecasting from observed human and object motion as input and predicting a plausible continuation. In Tab. 5.1, we compare to ours first in their setting, using observed motion as condition (motion-cond.), for a fair comparison. Additionally, we modify their approach by replacing observed motion encoders with our text encoder, allowing for a comparison in our setting (text-cond.). We also compare with MDM [1], a state-of-the-art method for human-only sequence generation from text, both in their original setting, only predicting human sequences, and extending theirs to also generate object sequences, by adding additional tokens and geometry conditioning to their transformer encoder formulation. For more details of baseline setup, we refer to the supplemental. We evaluate the quality of generated human-object interactions as well as human-only generation, only evaluating the human sequence for our method, as compared to the generated sequences of MDM.

Both Tab. 5.1 and the user study in Fig. 5.5 show that our approach is able to generate more realistic and physically plausible human-object interaction sequences than baselines. In Fig. 5.4, we see that our approach synthesizes more meaningful human-object interaction with respect to contact and mitigating independent object floating.

5.6.4 Ablation Studies

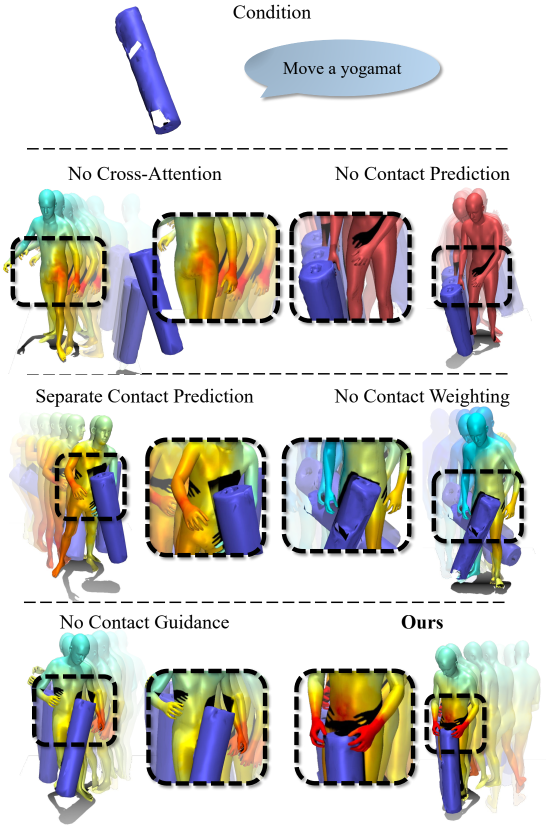


Figure 5.6: Visualization of ablations of our method design: Generation, weighting, and inference-time guidance work together to enable realistic interactions in our method, resolving artifacts such as object floating.

Cross-attention enables learning human-object interdependencies. Tab. 5.2 shows that our human-object-contact cross-attention (Sec. 5.4.2) significantly improves performance by effectively sharing information between human, contact, and object sequence modalities. In Fig. 5.6, we see this encourages realistic contact between human and object.

Contact prediction improves HOI generation performance. Predicting contact (Sec. 5.5) is crucial to generating more realistic human-object sequences, resulting in more realistic interactions between human and object (Fig. 5.6), and improved fidelity (Tab. 5.2). Notably, learning contact jointly with human and object motion improves overall quality, compared to a separately trained contact model used for inference guidance ("Separate contact pred.", Tab. 5.2).

Contact-based object transformation weighting improves generation performance. Weighting predicted object motion hypotheses with predicted contact (Sec. 5.4.3) improves HOI generation over naive object sequence prediction, both quantitatively in Tab. 5.2 ("No contact weighting") and visually as realistic human-object interactions in Fig. 5.6.

Contact-based guidance during inference helps produce physically plausible interactions. As shown in Fig. 5.6 and Tab. 5.2, using our guidance based on predicted contacts leads to a higher degree of fidelity and physical plausibility.

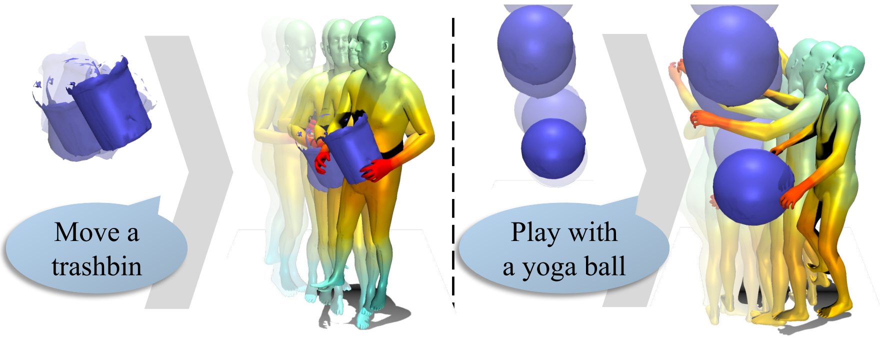


Figure 5.7: Given an object trajectory at inference time, our method can generate corresponding human motion without re-training.

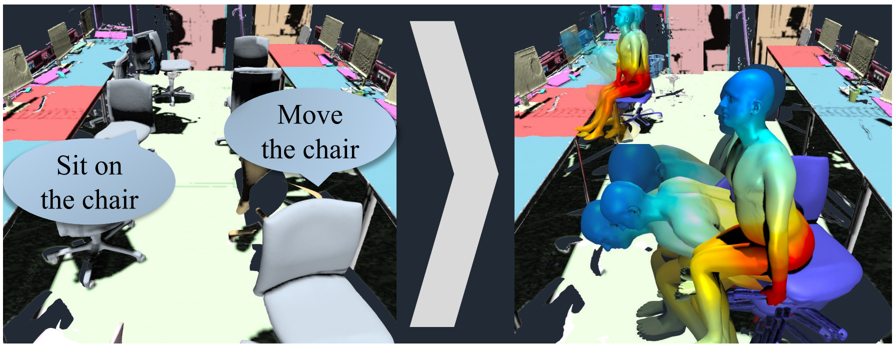


Figure 5.8: Application to static scene scans. Our method can generate HOIs from segmented objects in such environments.

5.6.5 Applications

Human motion generation given object trajectory. Our approach can be directly applied to conditionally generate human sequences given object sequences as condition, as shown in Fig. 5.7. As our model learns a strong correspondence between object and human motion, facilitated by contact distance predictions, we are able to condition without any additional training.

Populating 3D scans. Fig. 5.8 shows that we can also apply our method to generate human-object interactions in static scene scans. Here, we use a scene from the ScanNet++ dataset [220], with their existing semantic object segmentation. This enables the potential to generate realistic human motion sequences only given a static scene environment.

5.6.6 Limitations

While we have demonstrated the usefulness of joint contact prediction in 3D HOI generation, several limitations remain. For instance, our method focuses on realistic interactions with a single object. We show that this can be applied to objects in static 3D scans; however, we do not model multiple objects together, which could have the potential to model more complex long-term human behavior (e.g. cooking sequences). Additionally, our method requires expensive 3D HOI captures for training; a weakly supervised approach leveraging further supervision from 2D action data might be able to represent more diverse scenarios. Similarly, our method depends on manual text annotations; more specific prompts might lead to more control over generated HOIs.

5.7 Conclusion

We propose an approach to generating realistic, dynamic human-object interactions based on contact modeling. Our diffusion model effectively learns interdependencies between human, object, and contact through cross-attention along with our contact-based object transformation weighting. Our predicted contacts further facilitate refinement using custom diffusion guidance, generating diverse, realistic interactions based on text descriptions. Since our model learns a strong correlation between human and object sequences, we can use it to conditionally generate human motion from given object sequences. Extensive experimental evaluation confirms both fidelity and diversity of our generated sequences and shows improved performance compared to baselines.

5.8 Appendix

We show in this appendix additional qualitative and quantitative results, detail our baseline evaluation protocol, elaborate on the metrics used in the main paper, show the architecture used in our approach, and provide additional details regarding the data.

5.8.1 Additional Qualitative Results

Additional Interactions We show additional generated 3D human-object interactions of our method in Fig. 5.10, with object geometry and text condition on the left, and our generated sequence on the right.

Same Prompt, Different Interactions We evaluate the ability of our method to generate diverse interactions for a fixed text condition visually in Fig. 5.9, for text prompt "Move a stool" and "Sit on a stool". In the ground truth training data, move is only done with one or two hands, and feet; moving with the butt sometimes occurs for the text description "Sit on a stool".

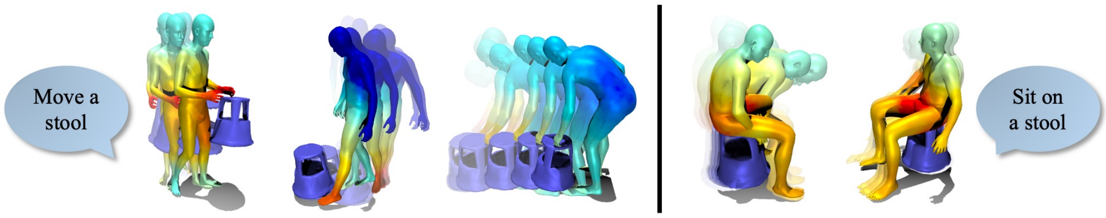


Figure 5.9: Our method is able to generate diverse human-object interactions for the same prompts.

5.8.2 Additional Quantitative Results

Evaluating Penetrations and Floating Our method discourages penetration and floating implicitly, by enforcing correct contact distances as a soft constraint at train and test time. However, the exact fidelity and diversity of our results is hard to capture with any single metric. Thus, we evaluate multiple such metrics in the main paper (R-Precision, FID, Diversity, MultiModality), and conduct a perceptual user study to verify the metrics' expressiveness.

Here, we provide an additional evaluation based on intuitive physics-based metrics: Tab. 5.3 evaluates the mean ratio of frames with some penetration as well as the ratio of penetrating vertices overall, showing that penetrations typically happens with small body parts (e.g., hands, which also occurs in the ground-truth data). We also evaluate the ratio of frames and vertices with human and object not in contact, including floating and stationary objects, which is expected to be close to the dataset ratio.

Results show similar penetration and floating between our generations and ground-truth training data.

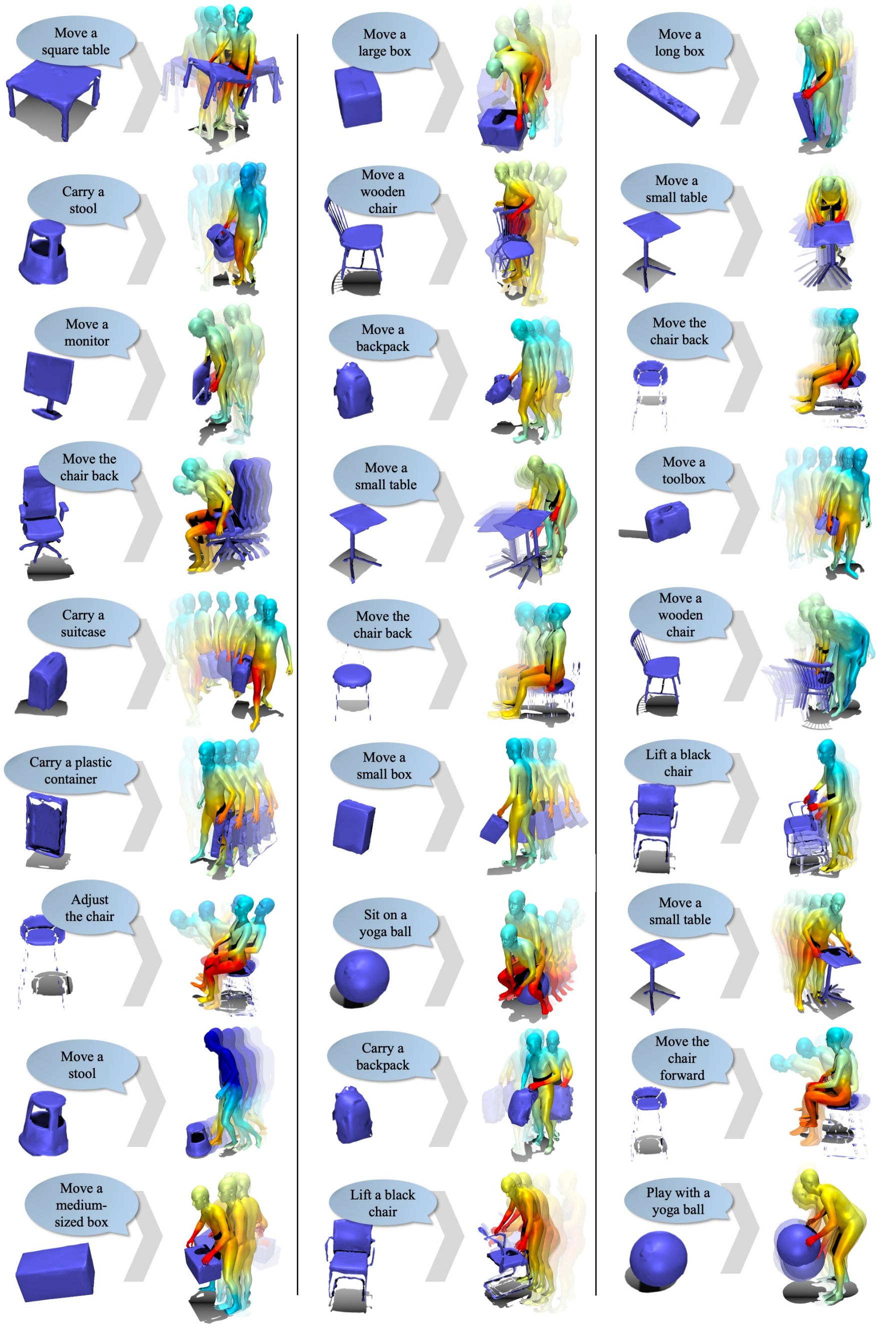


Figure 5.10: Additional qualitative evaluation. Our method produces diverse and realistic 3D human-object interaction sequences, given object geometry and short text description of the action. The sequences depict high-quality human-object interactions by modeling contact, mitigating floating and penetration artifacts.

	BEHAVE				CHAIRS			
	Penetration		Non-Contact		Penetration		Non-Contact	
	Frames	Vertices	Frames	Vertices	Frames	Vertices	Frames	Vertices
Dataset	32.9%	4.1%	21.4%	86.2%	26.9%	1.1%	11.9%	70.4%
Ours	31.3%	3.0%	17.8%	93.3%	35.8%	4.2%	14.1%	74.3%

Table 5.3: Penetration and non-contact (including floating) ratios in terms of frames as well as overall vertices vs ground-truth data.

Evaluating Contact Tab. 5.4 evaluates our contact predictions using precision/recall and distance metrics. We follow [182, 179, 181] to define contact if ≤5cm from object. We also report mean ℓ_1 error in contact distance predictions. All metrics are reported for body parts $\leq 1m$ of the object, to focus on contact scenarios. Better contact prediction corresponds with better HOI generations. Note that none of our baselines predict contact distances.

	BEHAVE			CHAIRS		
Approach	Precision ↑	Recall ↑	Distance ↓	Precision ↑	Recall ↑	Distance ↓
Separate contact pred.	23.4%	25.6%	0.53	58.6%	49.1%	0.24
No contact weighting	29.5%	33.5%	0.34	60.6%	63.4%	0.10
No contact guidance	46.3%	39.2%	0.31	64.2%	70.2%	0.12
Ours	**63.6%**	**59.5%**	**0.07**	**78.3%**	**84.5%**	**0.04**

Table 5.4: Evaluation of predicted contact distances, in terms on precision and recall ($\leq 5cm$ distance), as well as mean contact ℓ_1 error in meters.

Novelty of Generated Interactions We perform an additional interaction novelty analysis to verify that our method does not simply retrieve memorized train sequences but is indeed able to generate novel human-object interactions. To do so, we generate ≈ 500 sequences from both datasets and retrieve the top-3 most similar train sequences, as measured by the l_2 distance in human body and object transformation parameter space.

Fig. 5.11 shows the top-3 closest train sequences, along with a histogram of l_2 distances computed on our test set of ≈ 500 generated sequences. In red, we mark the intra-trainset distance between samples in the train set. We observe that the distance between our generated sequences and the closest train sequence is mostly larger than the intra-train distance. Thus, our method is able to produce samples that are novel and not simply retrieved train sequences.

SMPL Bodies vs. HumanML3D Skeletons We observe slight pose jitter and foot skating in our ground-truth training data (especially BEHAVE, captured with Kinect sensors). As a result, our model reflects some of these effects. Skeleton representations such as HumanML3D [46] could tackle these artifacts, but do not work with contact as effectively as SMPL bodies. Nevertheless, we train ours with HumanML3D parameters for comparison in Tab. 5.5 (fitting SMPL after for comparable evaluation) which leads to degraded performance due to less effective contact guidance.

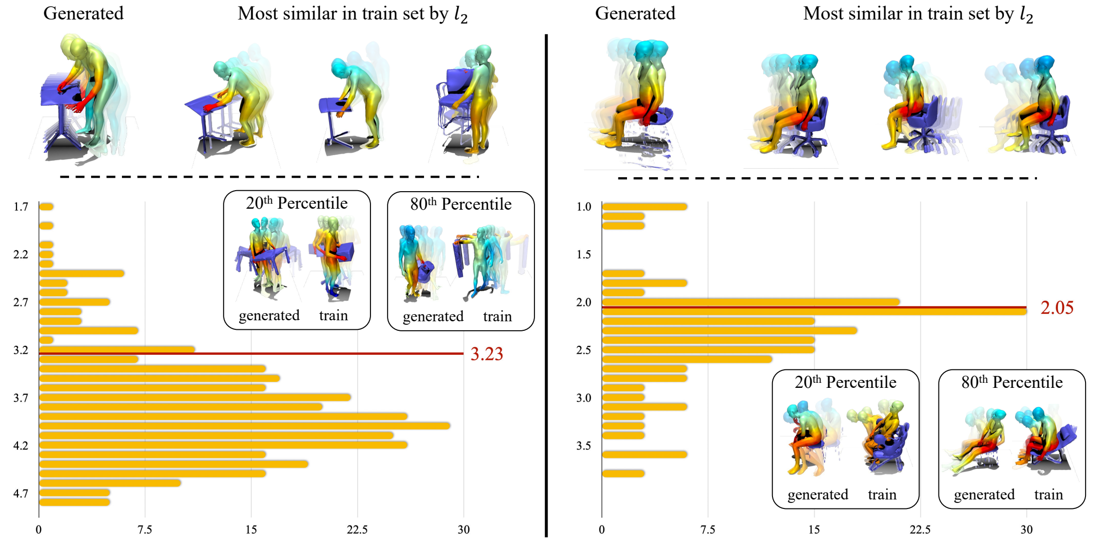


Figure 5.11: Human-Object Interaction Sequence Novelty Analysis. Performed on BEHAVE [164] (left) and CHAIRS [165] (right). We retrieve top-3 most similar sequences from the train set, and plot a histogram of distances to the closest train sample. While sequences at the 20th percentile still resemble the generated interactions, there is a large gap in the 80th percentile. We show the intra-trainset distance in red. Our approach generates novel shapes, not simply retrieving memorized train samples.

5.8.3 Baseline Evaluation Setup

There is no previous approach to modeling 3D human-object interactions from text and object geometry for direct comparison. Thus, we compare to the two closest methods, and compare to them in multiple settings, for a fair comparison.

The most related approach is InterDiff [2]. Their setting is to generate a short sequence of human-object interactions, from an observed such sequence as condition, with geometry but no text input. Their goal is to generate one, the most likely, sequence continuing the observation. We use their full approach, including the main diffusion training together with the post-processing refinement step. We compare in two different settings: First, in their native setup, running their method unchanged and modifying ours to take in geometry and past sequence observation instead of text (Motion-Cond. HOI in Tab. 1 main). Then, we modify their approach to take in geometry and text,

	BEHAVE				CHAIRS			
Representation	R-Prec. (top-3) ↑	FID ↓	Diversity →	MModality →	R-Prec. (top-3) ↑	FID ↓	Diversity →	MModality →
Ours (HumanML3D)	0.33	11.94	2.15	3.75	0.48	12.83	4.39	5.11
Ours	**0.62**	**6.31**	**6.63**	**5.47**	**0.74**	**6.45**	**8.91**	**5.94**

Table 5.5: Ours (using SMPL bodies) vs. using HumanML3D [46] skeletons and fitting SMPL bodies afterwards. While HumanML3D is designed to reduce jitter and foot skating, it leads to degraded performance in our scenario due to less effective contact guidance.

replacing their past motion encoder with our CLIP-based text encoder (Text-Cond. HOI in Tab. 1 main). We observe that our method is able to outperform InterDiff in both scenarios, for both datasets.

We additionally compare to MDM [1], a recent diffusion-based state-of-the-art human motion generation approach. Their approach is based on a transformer encoder formulation, using each human body as a token in the attention. We run their method on SMPL parameters and first compare in their native setting, only predicting human motion. We compare to the human motion generated by our method which is trained to generate full human-object interactions (Text-Cond. Human Only in Tab. 1 main). We also compare to human motion sequences generated by InterDiff in this setting. We see that our method is able to outperform both baselines even in this setting, demonstrating the added benefit of learning interdependencies of human and object motion. For the comparison in our setting, we modify MDM by adding additional tokens for the objects to the attention formulation. Our approach performs more realistic and diverse sequences in both settings which better follow the text condition.

5.8.4 Fidelity and Diversity Metrics

We base our fidelity and diversity metrics R-Precision, FID score, Diversity, and MultiModality on practices established for human motion generation [1, 46, 219], with minor modifications: We use the same networks used by these previous approaches, and adapt the input dimensions to fit our feature lengths, $F = 79$ when evaluating human body motion only, and $F = 79 + 128 + 9 = 216$ (SMPL parameters, contact distances, object transformations) for full evaluation in the human-object interaction scenario.

5.8.5 Architecture Details

Fig. 5.12 shows our detailed network architecture, including encoder, bottleneck, and decoder formulations.

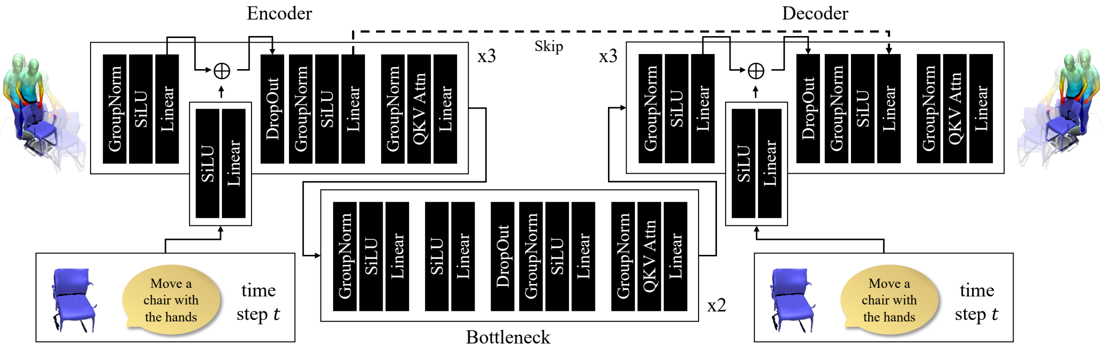


Figure 5.12: Network architecture specification.

5.8.6 Data Details

CHAIRS [165] captures 46 subjects as their SMPL-X [49] parameters using a mocap suit, in various settings interacting with a total of 81 different types of chairs and sofas, from office chairs over simple wooden chairs to more complex models like suspended seating structures. Each captured sequence consists of 6 actions and a given script; the exact separation into corresponding textual descriptions was manually annotated by the authors of this paper. In total, this yields ≈ 1300 sequences of human and object motion, together with a textual description. Every object geometry is provided as their canonical mesh; we additionally generate ground-truth contact and distance labels based on posed human and object meshes per-frame for each sequence. We use a random 80/10/10 split along object types, making sure that test objects are not seen during training.

BEHAVE [164] captures 8 participants as their SMPL-H [48] parameters captured in a multi-Kinect setup, along with the per-frame transformations and canonical geometries of 20 different object with a wide range, including yoga mats and tables. This yields ≈ 130 longer sequences. We use their original train/test split.

Object Geometry Representation We represent object geometry as a point cloud, to be processed by a PointNet [217] encoder. For this, we sample $N = 256$ points uniformly at random on the surface of an object mesh. Each object category is sampled once as a pre-processing step and kept same for train and inference.

Part III

Conclusion & Outlook

6 Conclusion

This dissertation investigates how to approach human behavior understanding from an action and interaction generation perspective. We mainly focus on three problems: Generating long-term goal-based *characteristic* 3D human action poses, forecasting complex long-term behavior based on 2D RGB observations, and generating human-object interaction sequences, modeling motion of both the human and the object interdependently. Each of these problems were introduced in detail in Part II, and we present concluding remarks in the following.

Forecasting Characteristic 3D Poses of Human Actions. In chapter 3, we introduce the new task of predicting future *characteristic 3D poses* of human activities from short sequences of pose observations. We introduce a probabilistic approach to capturing the most likely modes in these characteristic poses, coupled with an autoregressive formulation for pose joint prediction to sample consistent 3D poses from a predicted joint distribution. We trained and evaluated our approach on a new annotated dataset for characteristic 3D pose prediction, outperforming deterministic and multi-modal state-of-the-art approaches. We believe this opens up many possibilities for goal-oriented 3D human pose forecasting and understanding anticipation of human movements.

FutureHuman3D: Forecasting Complex Long-Term 3D Human Behavior from Video Observations. In chapter 4, we propose to forecast future human behavior by jointly predicting future actions alongside characteristic 3D poses. We do not require any 3D annotated action sequences or 3D input data; instead, we learn complex action sequences from 2D action video data and regularize predicted poses with an adversarial formulation against uncorrelated 3D pose data. Experiments demonstrate that our joint forecasting enables complementary feature learning, outperforming each individual task considered separately.

CG-HOI: Contact-Guided 3D Human-Object Interactions. In chapter 5, we propose an approach to generating realistic, dynamic human-object interactions based on contact modeling. Our diffusion model effectively learns interdependencies between human, object, and contact through cross-attention along with our contact-based object transformation weighting. Our predicted contacts further facilitate refinement using custom diffusion guidance, generating diverse, realistic interactions based on text descriptions. Since our model learns a strong correlation between human and object sequences, we can use it to conditionally generate human motion from given object sequences. Extensive experimental evaluation confirms both fidelity and diversity of our generated sequences and shows improved performance compared to baselines.

7 Limitations & Future Work

If you believe too much you'll never notice the flaws; if you doubt too much you won't get started. It requires a lovely balance.
– Richard Hamming, *You and Your Research (1986)*

Forecasting Characteristic 3D Poses of Human Actions. In this work, we introduced the new task of predicting future *characteristic 3D poses*. A probabilistic approach to capturing the most likely modes in these characteristic poses, coupled with an autoregressive formulation for pose joint prediction, enables goal-oriented 3D human pose forecasting. However, several limitations remain for our approach. For instance, while our offset predictions help alleviate the ties to a volumetric heatmap grid, more precise modeling of smaller-scale behavior (e.g., detailed hand movement) would require more efficient representations such as sparse grids. In addition, our method relies on manually annotated characteristic 3D poses for supervision; while characteristic pose annotation is very efficient for new datasets, self-supervised formulations would also be an interesting future direction.

FutureHuman3D: Forecasting Complex Long-Term 3D Human Behavior from Video Observations. In FutureHuman3D, we proposed forecasting future human behavior by jointly predicting future actions alongside characteristic 3D poses. We do not require any 3D annotated action sequences or 3D input data; instead, we learn complex action sequences from 2D action video data and regularize predicted poses with an adversarial formulation against uncorrelated 3D pose data. We demonstrated the potential of this approach, but several limitations remain. For instance, our method leverages a separate 2D pose extraction as input to training, while an end-to-end formulation could potentially better leverage other useful signals in the input frames. Additionally, a more holistic body representation than pose joints would be necessary for finer-grained interactions that involve reasoning over small limbs (e.g., hands) and body surface contact.

CG-HOI: Contact-Guided 3D Human-Object Interactions. CG-HOI is an approach to generating realistic, dynamic human-object interactions based on contact modeling. Our diffusion model effectively learns interdependencies between human, object, and contact through cross-attention along with our contact-based object transformation weighting. Our predicted contacts further facilitate refinement using custom diffusion guidance, generating diverse, realistic interactions based on text descriptions. Since our model learns a strong correlation between human and object sequences, we can use it to conditionally generate human motion from given object sequences.

While we demonstrated the usefulness of this method, several limitations remain. For instance, our method focuses on realistic interactions with a single object. We show that this can be applied to objects in static 3D scans; however, we do not model multiple objects together, which could have the potential to model more complex long-term human behavior (e.g., cooking sequences). Additionally, our method requires expensive 3D HOI captures for training; a weakly supervised approach leveraging further supervision from 2D action data might be able to represent more diverse scenarios. Similarly, our method depends on manual text annotations; more specific prompts might lead to more control over generated HOIs.

Bibliography

[1] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023.* OpenReview.net, 2023. URL: `https://openreview.net/pdf?id=SJ1kSyO2jwu`.

[2] S. Xu, Z. Li, Y. Wang, and L. Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 14882–14894. IEEE, 2023. URL: `https://doi.org/10.1109/ICCV51070.2023.01371`, `doi:10.1109/ICCV51070.2023.01371`.

[3] C. Diller and A. Dai. CG-HOI: contact-guided 3d human-object interaction generation. *CoRR*, abs/2311.16097, 2023. URL: `https://doi.org/10.48550/arXiv.2311.16097`, `arXiv:2311.16097`, `doi:10.48550/ARXIV.2311.16097`.

[4] M. Bar. The proactive brain: memory for predictions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1235–1243, 2009.

[5] J. Hohwy. *The predictive mind.* Oxford University Press, 2013.

[6] A. Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204, 2013.

[7] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert. Activity forecasting. In A. W. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV*, volume 7575 of *Lecture Notes in Computer Science*, pages 201–214. Springer, 2012. URL: `https://doi.org/10.1007/978-3-642-33765-9_15`, `doi:10.1007/978-3-642-33765-9\_15`.

[8] H. Kuehne, A. B. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 780–787. IEEE Computer Society, 2014. URL: `https://doi.org/10.1109/CVPR.2014.105`, `doi:10.1109/CVPR.2014.105`.

[9] N. Rhinehart and K. M. Kitani. First-person activity forecasting with online inverse reinforcement learning. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3716–3725. IEEE

Computer Society, 2017. URL: `https://doi.org/10.1109/ICCV.2017.399`, `doi:10.1109/ICCV.2017.399`.

[10] A. Furnari and G. M. Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(11):4021–4036, 2021. URL: `https://doi.org/10.1109/TPAMI.2020.2992889`, `doi:10.1109/TPAMI.2020.2992889`.

[11] R. Girdhar and K. Grauman. Anticipative video transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 13485–13495. IEEE, 2021. URL: `https://doi.org/10.1109/ICCV48922.2021.01325`, `doi:10.1109/ICCV48922.2021.01325`.

[12] D. Gong, J. Lee, M. Kim, S. J. Ha, and M. Cho. Future transformer for long-term action anticipation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 3042–3051. IEEE, 2022. URL: `https://doi.org/10.1109/CVPR52688.2022.00306`, `doi:10.1109/CVPR52688.2022.00306`.

[13] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik. Recurrent network models for human dynamics. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4346–4354. IEEE Computer Society, 2015. URL: `https://doi.org/10.1109/ICCV.2015.494`, `doi:10.1109/ICCV.2015.494`.

[14] W. Mao, M. Liu, M. Salzmann, and H. Li. Learning trajectory dependencies for human motion prediction. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9488–9496. IEEE, 2019. URL: `https://doi.org/10.1109/ICCV.2019.00958`, `doi:10.1109/ICCV.2019.00958`.

[15] W. Mao, M. Liu, and M. Salzmann. History repeats itself: Human motion prediction via motion attention. In A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIV*, volume 12359 of *Lecture Notes in Computer Science*, pages 474–489. Springer, 2020. URL: `https://doi.org/10.1007/978-3-030-58568-6_28`, `doi:10.1007/978-3-030-58568-6\_28`.

[16] Y. Yuan and K. Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IX*, volume 12354 of *Lecture Notes in Computer Science*, pages 346–364. Springer, 2020. URL: `https://doi.org/10.1007/978-3-030-58545-7_20`, `doi:10.1007/978-3-030-58545-7\_20`.

[17] S. Xu, Y. Wang, and L. Gui. Diverse human motion prediction guided by multi-level spatial-temporal anchors. In S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXII*, volume 13682 of *Lecture Notes in Computer Science*, pages 251–269. Springer, 2022. URL: `https://doi.org/10.1007/978-3-031-20047-2_15`, `doi:10.1007/978-3-031-20047-2\_15`.

[18] S. Huang, Z. Wang, P. Li, B. Jia, T. Liu, Y. Zhu, W. Liang, and S. Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 16750–16761. IEEE, 2023. URL: `https://doi.org/10.1109/CVPR52729.2023.01607`, `doi:10.1109/CVPR52729.2023.01607`.

[19] Z. Xiao, T. Wang, J. Wang, J. Cao, W. Zhang, B. Dai, D. Lin, and J. Pang. Unified human-scene interaction via prompted chain-of-contacts. *CoRR*, abs/2309.07918, 2023. URL: `https://doi.org/10.48550/arXiv.2309.07918`, `arXiv:2309.07918`, `doi:10.48550/ARXIV.2309.07918`.

[20] J. Wang, Y. Rong, J. Liu, S. Yan, D. Lin, and B. Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 20428–20437. IEEE, 2022. URL: `https://doi.org/10.1109/CVPR52688.2022.01981`, `doi:10.1109/CVPR52688.2022.01981`.

[21] Z. Wang, Y. Chen, T. Liu, Y. Zhu, W. Liang, and S. Huang. HUMANISE: language-conditioned human motion generation in 3d scenes. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL: `http://papers.nips.cc/paper_files/paper/2022/hash/6030db5195150ac86d942186f4abdad8-Abstract-Conference.html`.

[22] K. Zhao, S. Wang, Y. Zhang, T. Beeler, and S. Tang. Compositional human-scene interaction synthesis with semantic control. In S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VI*, volume 13666 of *Lecture Notes in Computer Science*, pages 311–327. Springer, 2022. URL: `https://doi.org/10.1007/978-3-031-20068-7_18`, `doi:10.1007/978-3-031-20068-7\_18`.

[23] K. Zhao, Y. Zhang, S. Wang, T. Beeler, and S. Tang. Synthesizing diverse human motions in 3d indoor scenes. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 14692–14703. IEEE, 2023. URL: `https://doi.org/10.1109/ICCV51070.2023.01354`, `doi:10.1109/ICCV51070.2023.01354`.

[24] Y. Zhang and S. Tang. The wanderings of odysseus in 3d scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 20449–20459. IEEE, 2022. URL: https://doi.org/10.1109/CVPR52688.2022.01983, doi:10.1109/CVPR52688.2022.01983.

[25] J. F. M. Jr., D. Kothandaraman, A. Bera, and D. Manocha. Placing human animations into 3d scenes by learning interaction- and geometry-driven keyframes. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pages 300–310. IEEE, 2023. URL: https://doi.org/10.1109/WACV56688.2023.00038, doi:10.1109/WACV56688.2023.00038.

[26] Y. Zheng, Y. Yang, K. Mo, J. Li, T. Yu, Y. Liu, C. K. Liu, and L. J. Guibas. GIMO: gaze-informed human motion prediction in context. In S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIII*, volume 13673 of *Lecture Notes in Computer Science*, pages 676–694. Springer, 2022. URL: https://doi.org/10.1007/978-3-031-19778-9_39, doi:10.1007/978-3-031-19778-9_39.

[27] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5308–5317. IEEE Computer Society, 2016. URL: https://doi.org/10.1109/CVPR.2016.573, doi:10.1109/CVPR.2016.573.

[28] J. Martinez, M. J. Black, and J. Romero. On human motion prediction using recurrent neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4674–4683. IEEE Computer Society, 2017. URL: https://doi.org/10.1109/CVPR.2017.497, doi:10.1109/CVPR.2017.497.

[29] L. Gui, Y. Wang, X. Liang, and J. M. F. Moura. Adversarial geometry-aware human motion prediction. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, volume 11208 of *Lecture Notes in Computer Science*, pages 823–842. Springer, 2018. URL: https://doi.org/10.1007/978-3-030-01225-0_48, doi:10.1007/978-3-030-01225-0_48.

[30] M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, and Q. Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 211–220. Computer Vision Foundation / IEEE, 2020. URL: https://openaccess.thecvf.com/content_CVPR_2020/html/Li_Dynamic_Multiscale_Graph_

Neural_Networks_for_3D_Skeleton_Based_Human_CVPR_2020_paper.html, doi:10.1109/CVPR42600.2020.00029.

[31] Y. Tang, L. Ma, W. Liu, and W. Zheng. Long-term human motion prediction by modeling motion context and enhancing motion dynamics. In J. Lang, editor, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 935–941. ijcai.org, 2018. URL: https://doi.org/10.24963/ijcai.2018/130, doi:10.24963/ijcai.2018/130.

[32] D. Jayaraman, F. Ebert, A. A. Efros, and S. Levine. Time-agnostic prediction: Predicting predictable video frames. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL: https://openreview.net/forum?id=SyzVb3CcFX.

[33] K. Pertsch, O. Rybkin, J. Yang, S. Zhou, K. G. Derpanis, K. Daniilidis, J. J. Lim, and A. Jaegle. Keyframing the future: Keyframe discovery for visual prediction and planning. In A. M. Bayen, A. Jadbabaie, G. J. Pappas, P. A. Parrilo, B. Recht, C. J. Tomlin, and M. N. Zeilinger, editors, *Proceedings of the 2nd Annual Conference on Learning for Dynamics and Control, L4DC 2020, Online Event, Berkeley, CA, USA, 11-12 June 2020*, volume 120 of *Proceedings of Machine Learning Research*, pages 969–979. PMLR, 2020. URL: http://proceedings.mlr.press/v120/pertsch20a.html.

[34] Z. Zhang, R. Liu, K. Aberman, and R. Hanocka. Tedi: Temporally-entangled diffusion for long-term motion synthesis. *CoRR*, abs/2307.15042, 2023. URL: https://doi.org/10.48550/arXiv.2307.15042, arXiv:2307.15042, doi:10.48550/ARXIV.2307.15042.

[35] S. Raab, I. Leibovitch, G. Tevet, M. Arar, A. H. Bermano, and D. Cohen-Or. Single motion diffusion. *CoRR*, abs/2302.05905, 2023. URL: https://doi.org/10.48550/arXiv.2302.05905, arXiv:2302.05905, doi:10.48550/ARXIV.2302.05905.

[36] M. Zhao, M. Liu, B. Ren, S. Dai, and N. Sebe. Modiff: Action-conditioned 3d motion generation with denoising diffusion probabilistic models. *CoRR*, abs/2301.03949, 2023. URL: https://doi.org/10.48550/arXiv.2301.03949, arXiv:2301.03949, doi:10.48550/ARXIV.2301.03949.

[37] R. Dabral, M. H. Mughal, V. Golyanik, and C. Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 9760–9770. IEEE, 2023. URL: https://doi.org/10.1109/CVPR52729.2023.00941, doi:10.1109/CVPR52729.2023.00941.

[38] Z. Ren, Z. Pan, X. Zhou, and L. Kang. Diffusion motion: Generate text-guided 3d human motion by diffusion model. In *IEEE International Conference on Acoustics,*

Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023, pages 1–5. IEEE, 2023. URL: https://doi.org/10.1109/ICASSP49357.2023.10096441, doi:10.1109/ICASSP49357.2023.10096441.

[39] X. Chen, B. Jiang, W. Liu, Z. Huang, B. Fu, T. Chen, and G. Yu. Executing your commands via motion diffusion in latent space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 18000–18010. IEEE, 2023. URL: https://doi.org/10.1109/CVPR52729.2023.01726, doi:10.1109/CVPR52729.2023.01726.

[40] C. Diller, T. A. Funkhouser, and A. Dai. Forecasting characteristic 3d poses of human actions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15893–15902. IEEE, 2022. URL: https://doi.org/10.1109/CVPR52688.2022.01545, doi:10.1109/CVPR52688.2022.01545.

[41] C. Diller, T. A. Funkhouser, and A. Dai. Futurehuman3d: Forecasting complex long-term 3d human behavior from video observations. *CoRR*, abs/2211.14309, 2022. URL: https://doi.org/10.48550/arXiv.2211.14309, arXiv:2211.14309, doi:10.48550/ARXIV.2211.14309.

[42] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In M. C. Stone, editor, *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1987, Anaheim, California, USA, July 27-31, 1987*, pages 163–169. ACM, 1987. URL: https://doi.org/10.1145/37401.37422, doi:10.1145/37401.37422.

[43] J. J. Park, P. R. Florence, J. Straub, R. A. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 165–174. Computer Vision Foundation / IEEE, 2019. URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Park_DeepSDF_Learning_Continuous_Signed_Distance_Functions_for_Shape_Representation_CVPR_2019_paper.html, doi:10.1109/CVPR.2019.00025.

[44] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas. GRAB: A dataset of whole-body human grasping of objects. In A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, volume 12349 of *Lecture Notes in Computer Science*, pages 581–600. Springer, 2020. URL: https://doi.org/10.1007/978-3-030-58548-8_34, doi:10.1007/978-3-030-58548-8_34.

[45] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2014. URL: https://doi.org/10.1109/TPAMI.2013.248, doi:10.1109/TPAMI.2013.248.

[46] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng. Generating diverse and natural 3d human motions from text. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5142–5151. IEEE, 2022. URL: `https://doi.org/10.1109/CVPR52688.2022.00509`, `doi:10.1109/CVPR52688.2022.00509`.

[47] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: a skinned multi-person linear model. volume 34, pages 248:1–248:16, 2015. URL: `https://doi.org/10.1145/2816795.2818013`, `doi:10.1145/2816795.2818013`.

[48] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 36(6):245:1–245:17, 2017. URL: `https://doi.org/10.1145/3130800.3130883`, `doi:10.1145/3130800.3130883`.

[49] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10975–10985. Computer Vision Foundation / IEEE, 2019. URL: `http://openaccess.thecvf.com/content_CVPR_2019/html/Pavlakos_Expressive_Body_Capture_3D_Hands_Face_and_Body_From_a_CVPR_2019_paper.html`, `doi:10.1109/CVPR.2019.01123`.

[50] A. A. A. Osman, T. Bolkart, and M. J. Black. STAR: sparse trained articulated human body regressor. In A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI*, volume 12351 of *Lecture Notes in Computer Science*, pages 598–613. Springer, 2020. URL: `https://doi.org/10.1007/978-3-030-58539-6_36`, `doi:10.1007/978-3-030-58539-6\_36`.

[51] A. A. A. Osman, T. Bolkart, D. Tzionas, and M. J. Black. SUPR: A sparse unified part-based human body model. In *European Conference on Computer Vision (ECCV)*, 2022. URL: `https://supr.is.tue.mpg.de`.

[52] P. R. Palafox, A. Bozic, J. Thies, M. Nießner, and A. Dai. Npms: Neural parametric models for 3d deformable shapes. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 12675–12685. IEEE, 2021. URL: `https://doi.org/10.1109/ICCV48922.2021.01246`, `doi:10.1109/ICCV48922.2021.01246`.

[53] P. R. Palafox, N. Sarafianos, T. Tung, and A. Dai. Spams: Structured implicit parametric models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 12841–12850. IEEE, 2022. URL: `https://doi.org/10.1109/CVPR52688.2022.01251`, `doi:10.1109/CVPR52688.2022.01251`.

[54] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision, Kerkyra, Corfu, Greece, September 20-25, 1999*, pages 1150–1157. IEEE Computer Society, 1999. URL: https://doi.org/10.1109/ICCV.1999.790410, doi:10.1109/ICCV.1999.790410.

[55] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski. ORB: an efficient alternative to SIFT or SURF. In D. N. Metaxas, L. Quan, A. Sanfeliu, and L. V. Gool, editors, *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, pages 2564–2571. IEEE Computer Society, 2011. URL: https://doi.org/10.1109/ICCV.2011.6126544, doi:10.1109/ICCV.2011.6126544.

[56] A. Shahroudy, J. Liu, T. Ng, and G. Wang. NTU RGB+D: A large scale dataset for 3d human activity analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1010–1019. IEEE Computer Society, 2016. URL: https://doi.org/10.1109/CVPR.2016.115, doi:10.1109/CVPR.2016.115.

[57] J. Liu, A. Shahroudy, M. Perez, G. Wang, L. Duan, and A. C. Kot. NTU RGB+D 120: A large-scale benchmark for 3d human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(10):2684–2701, 2020. URL: https://doi.org/10.1109/TPAMI.2019.2916873, doi:10.1109/TPAMI.2019.2916873.

[58] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[59] H. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y. Li, and C. Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(6):7157–7173, 2023. URL: https://doi.org/10.1109/TPAMI.2022.3222784, doi:10.1109/TPAMI.2022.3222784.

[60] H. Joo, H. Liu, L. Tan, L. Gui, B. C. Nabbe, I. A. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 3334–3342. IEEE Computer Society, 2015. URL: https://doi.org/10.1109/ICCV.2015.381, doi:10.1109/ICCV.2015.381.

[61] E. Aksan, M. Kaufmann, P. Cao, and O. Hilliges. A spatio-temporal transformer for 3d human motion prediction. In *International Conference on 3D Vision, 3DV 2021, London, United Kingdom, December 1-3, 2021*, pages 565–574. IEEE, 2021. URL: https://doi.org/10.1109/3DV53792.2021.00066, doi:10.1109/3DV53792.2021.00066.

[62] M. Petrovich, M. J. Black, and G. Varol. Action-conditioned 3d human motion synthesis with transformer VAE. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 10965–10975. IEEE, 2021. URL: `https://doi.org/10.1109/ICCV48922.2021.01080`, `doi:10.1109/ICCV48922.2021.01080`.

[63] A. A. S. Gunawan, A. P. Iman, and D. Suhartono. Automatic music generator using recurrent neural network. *Int. J. Comput. Intell. Syst.*, 13(1):645–654, 2020. URL: `https://doi.org/10.2991/ijcis.d.200519.001`, `doi:10.2991/IJCIS.D.200519.001`.

[64] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997. URL: `https://doi.org/10.1162/neco.1997.9.8.1735`, `doi:10.1162/NECO.1997.9.8.1735`.

[65] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In A. Moschitti, B. Pang, and W. Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734. ACL, 2014. URL: `https://doi.org/10.3115/v1/d14-1179`, `doi:10.3115/V1/D14-1179`.

[66] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL: `https://openreview.net/forum?id=SJU4ayYgl`.

[67] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL: `https://openreview.net/forum?id=rJXMpikCZ`.

[68] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL: `http://papers.nips.cc/paper/7181-attention-is-all-you-need`.

[69] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine*

Learning Research, pages 8821–8831. PMLR, 2021. URL: http://proceedings.mlr.press/v139/ramesh21a.html.

[70] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[71] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. URL: https://doi.org/10.18653/v1/n19-1423, doi:10.18653/V1/N19-1423.

[72] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021.* OpenReview.net, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[73] Á. Martínez-González, M. Villamizar, and J. Odobez. Pose transformers (POTR): human motion prediction with non-autoregressive transformers. In *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11-17, 2021*, pages 2276–2284. IEEE, 2021. URL: https://doi.org/10.1109/ICCVW54120.2021.00257, doi:10.1109/ICCVW54120.2021.00257.

[74] D. Roy and B. Fernando. Action anticipation using pairwise human-object interactions and transformers. *IEEE Trans. Image Process.*, 30:8116–8129, 2021. URL: https://doi.org/10.1109/TIP.2021.3113114, doi:10.1109/TIP.2021.3113114.

[75] S. Tian, M. Zheng, and X. Liang. Transfusion: A practical and effective transformer-based diffusion model for 3d human motion prediction. *CoRR*, abs/2307.16106, 2023. URL: https://doi.org/10.48550/arXiv.2307.16106, arXiv:2307.16106, doi:10.48550/ARXIV.2307.16106.

[76] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In F. R. Bach and D. M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2256–2265. JMLR.org, 2015. URL: `http://proceedings.mlr.press/v37/sohl-dickstein15.html`.

[77] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. 2021. URL: `https://openreview.net/forum?id=St1giarCHLP`.

[78] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. 2020. URL: `https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html`.

[79] Z. Yang, B. Su, and J. Wen. Synthesizing long-term human motions with diffusion models via coherent sampling. In A. El-Saddik, T. Mei, R. Cucchiara, M. Bertini, D. P. T. Vallejo, P. K. Atrey, and M. S. Hossain, editors, *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 3954–3964. ACM, 2023. URL: `https://doi.org/10.1145/3581783.3611887`, `doi:10.1145/3581783.3611887`.

[80] S. Yang, Z. Yang, and Z. Wang. Longdancediff: Long-term dance generation with conditional diffusion model. *CoRR*, abs/2308.11945, 2023. URL: `https://doi.org/10.48550/arXiv.2308.11945`, `arXiv:2308.11945`, `doi:10.48550/ARXIV.2308.11945`.

[81] Y. Shafir, G. Tevet, R. Kapon, and A. H. Bermano. Human motion diffusion as a generative prior. *CoRR*, abs/2303.01418, 2023. URL: `https://doi.org/10.48550/arXiv.2303.01418`, `arXiv:2303.01418`, `doi:10.48550/ARXIV.2303.01418`.

[82] Y. Wang, Z. Leng, F. W. B. Li, S. Wu, and X. Liang. Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 21978–21987. IEEE, 2023. URL: `https://doi.org/10.1109/ICCV51070.2023.02014`, `doi:10.1109/ICCV51070.2023.02014`.

[83] H. Ahn, E. V. Mascaro, and D. Lee. Can we use diffusion probabilistic models for 3d motion prediction? In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*, pages 9837–9843. IEEE, 2023. URL: `https://doi.org/10.1109/ICRA48891.2023.10160722`, `doi:10.1109/ICRA48891.2023.10160722`.

[84] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *CoRR*, abs/2208.15001, 2022. URL: `https://doi.org/10.48550/arXiv.2208.15001`, `arXiv:2208.15001`, `doi:10.48550/ARXIV.2208.15001`.

[85] D. Wei, H. Sun, B. Li, J. Lu, W. Li, X. Sun, and S. Hu. Human joint kinematics diffusion-refinement for stochastic motion prediction. In B. Williams, Y. Chen, and J. Neville, editors, *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 6110–6118. AAAI Press, 2023. URL: https://doi.org/10.1609/aaai.v37i5.25754, doi:10.1609/AAAI.V37I5.25754.

[86] J. Sun and G. Chowdhary. Towards globally consistent stochastic human motion prediction via motion diffusion. *CoRR*, abs/2305.12554, 2023. URL: https://doi.org/10.48550/arXiv.2305.12554, arXiv:2305.12554, doi:10.48550/ARXIV.2305.12554.

[87] D. Wei, X. Sun, H. Sun, B. Li, S. Hu, W. Li, and J. Lu. Understanding text-driven motion synthesis with keyframe collaboration via diffusion models. *CoRR*, abs/2305.13773, 2023. URL: https://doi.org/10.48550/arXiv.2305.13773, arXiv:2305.13773, doi:10.48550/ARXIV.2305.13773.

[88] S. Alexanderson, R. Nagy, J. Beskow, and G. E. Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Trans. Graph.*, 42(4):44:1–44:20, 2023. URL: https://doi.org/10.1145/3592458, doi:10.1145/3592458.

[89] J. Choi, D. Shim, and H. J. Kim. Diffupose: Monocular 3d human pose estimation via denoising diffusion probabilistic model. In *IROS*, pages 3773–3780, 2023. URL: https://doi.org/10.1109/IROS55552.2023.10342204, doi:10.1109/IROS55552.2023.10342204.

[90] L. Chen, J. Zhang, Y. Li, Y. Pang, X. Xia, and T. Liu. Humanmac: Masked motion completion for human motion prediction. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 9510–9521. IEEE, 2023. URL: https://doi.org/10.1109/ICCV51070.2023.00875, doi:10.1109/ICCV51070.2023.00875.

[91] G. Barquero, S. Escalera, and C. Palmero. Belfusion: Latent diffusion for behavior-driven human motion prediction. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 2317–2327. IEEE, 2023. URL: https://doi.org/10.1109/ICCV51070.2023.00220, doi:10.1109/ICCV51070.2023.00220.

[92] M. Zhang, X. Guo, L. Pan, Z. Cai, F. Hong, H. Li, L. Yang, and Z. Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 364–373. IEEE, 2023. URL: https://doi.org/10.1109/ICCV51070.2023.00040, doi:10.1109/ICCV51070.2023.00040.

[93] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014. URL: `https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html`.

[94] E. Aksan, M. Kaufmann, and O. Hilliges. Structured prediction helps 3d human motion modelling. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7143–7152. IEEE, 2019. URL: `https://doi.org/10.1109/ICCV.2019.00724`, `doi:10.1109/ICCV.2019.00724`.

[95] H. Chiu, E. Adeli, B. Wang, D. Huang, and J. C. Niebles. Action-agnostic human pose forecasting. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*, pages 1423–1432. IEEE, 2019. URL: `https://doi.org/10.1109/WACV.2019.00156`, `doi:10.1109/WACV.2019.00156`.

[96] B. Wang, E. Adeli, H. Chiu, D. Huang, and J. C. Niebles. Imitation learning for human pose prediction. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7123–7132. IEEE, 2019. URL: `https://doi.org/10.1109/ICCV.2019.00722`, `doi:10.1109/ICCV.2019.00722`.

[97] A. Gopalakrishnan, A. Mali, D. Kifer, C. L. Giles, and A. G. O. II. A neural temporal model for human motion prediction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12116–12125. Computer Vision Foundation / IEEE, 2019. URL: `http://openaccess.thecvf.com/content_CVPR_2019/html/Gopalakrishnan_A_Neural_Temporal_Model_for_Human_Motion_Prediction_CVPR_2019_paper.html`, `doi:10.1109/CVPR.2019.01239`.

[98] J. Bütepage, M. J. Black, D. Kragic, and H. Kjellström. Deep representation learning for human motion prediction and classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1591–1599. IEEE Computer Society, 2017. URL: `https://doi.org/10.1109/CVPR.2017.173`, `doi:10.1109/CVPR.2017.173`.

[99] C. Li, Z. Zhang, W. S. Lee, and G. H. Lee. Convolutional sequence to sequence model for human dynamics. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5226–5234. Computer Vision Foundation / IEEE Computer Society, 2018. URL: `http://openaccess.thecvf.com/content_cvpr_2018/html/`

Li_Convolutional_Sequence_to_CVPR_2018_paper.html, doi:10.1109/CVPR.2018.00548.

[100] Y. Zhou, Z. Li, S. Xiao, C. He, Z. Huang, and H. Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL: https://openreview.net/forum?id=r11Q2SlRW.

[101] E. Barsoum, J. R. Kender, and Z. Liu. HP-GAN: probabilistic 3d human motion prediction via GAN. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1418–1427. Computer Vision Foundation / IEEE Computer Society, 2018. URL: http://openaccess.thecvf.com/content_cvpr_2018_workshops/w29/html/Barsoum_HP-GAN_Probabilistic_3D_CVPR_2018_paper.html, doi:10.1109/CVPRW.2018.00191.

[102] X. Yan, A. Rastogi, R. Villegas, K. Sunkavalli, E. Shechtman, S. Hadap, E. Yumer, and H. Lee. MT-VAE: learning motion transformations to generate multimodal human dynamics. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V*, volume 11209 of *Lecture Notes in Computer Science*, pages 276–293. Springer, 2018. URL: https://doi.org/10.1007/978-3-030-01228-1_17, doi:10.1007/978-3-030-01228-1_17.

[103] M. S. Aliakbarian, F. S. Saleh, M. Salzmann, L. Petersson, and S. Gould. A stochastic conditioning scheme for diverse human motion prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5222–5231. Computer Vision Foundation / IEEE, 2020. URL: https://openaccess.thecvf.com/content_CVPR_2020/html/Aliakbarian_A_Stochastic_Conditioning_Scheme_for_Diverse_Human_Motion_Prediction_CVPR_2020_paper.html, doi:10.1109/CVPR42600.2020.00527.

[104] Z. Cao, H. Gao, K. Mangalam, Q. Cai, M. Vo, and J. Malik. Long-term human motion prediction with scene context. In A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 387–404. Springer, 2020. URL: https://doi.org/10.1007/978-3-030-58452-8_23, doi:10.1007/978-3-030-58452-8_23.

[105] K. Iskakov, E. Burkov, V. S. Lempitsky, and Y. Malkov. Learnable triangulation of human pose. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7717–

7726. IEEE, 2019. URL: https://doi.org/10.1109/ICCV.2019.00781, doi:10.1109/ICCV.2019.00781.

[106] H. Tu, C. Wang, and W. Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 197–212. Springer, 2020. URL: https://doi.org/10.1007/978-3-030-58452-8_12, doi:10.1007/978-3-030-58452-8_12.

[107] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2226–2234, 2016. URL: https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html.

[108] X. Huang, M. Liu, S. J. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III*, volume 11207 of *Lecture Notes in Computer Science*, pages 179–196. Springer, 2018. URL: https://doi.org/10.1007/978-3-030-01219-9_11, doi:10.1007/978-3-030-01219-9_11.

[109] S. Brahmbhatt, C. Ham, C. C. Kemp, and J. Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8709–8719. Computer Vision Foundation / IEEE, 2019. URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Brahmbhatt_ContactDB_Analyzing_and_Predicting_Grasp_Contact_via_Thermal_Imaging_CVPR_2019_paper.html, doi:10.1109/CVPR.2019.00891.

[110] M. Rohrbach, A. Rohrbach, M. Regneri, S. Amin, M. Andriluka, M. Pinkal, and B. Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *Int. J. Comput. Vis.*, 119(3):346–373, 2016. URL: https://doi.org/10.1007/s11263-015-0851-8, doi:10.1007/S11263-015-0851-8.

[111] Y. Ben-Shabat, X. Yu, F. S. Saleh, D. Campbell, C. R. Opazo, H. Li, and S. Gould. The IKEA ASM dataset: Understanding people assembling furniture through actions, objects and pose. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 846–858. IEEE, 2021. URL: https://doi.org/10.1109/WACV48630.2021.00089, doi:10.1109/WACV48630.2021.00089.

[112] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black. AMASS: archive of motion capture as surface shapes. In *2019 IEEE/CVF International*

Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pages 5441–5450. IEEE, 2019. URL: `https://doi.org/10.1109/ICCV.2019.00554`, `doi:10.1109/ICCV.2019.00554`.

[113] D. Pavllo, D. Grangier, and M. Auli. Quaternet: A quaternion-based recurrent model for human motion. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 299. BMVA Press, 2018. URL: `http://bmvc2018.org/contents/papers/0675.pdf`.

[114] L. Dang, Y. Nie, C. Long, Q. Zhang, and G. Li. MSR-GCN: multi-scale residual graph convolution networks for human motion prediction. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11447–11456. IEEE, 2021. URL: `https://doi.org/10.1109/ICCV48922.2021.01127`, `doi:10.1109/ICCV48922.2021.01127`.

[115] M. Li, S. Chen, Z. Liu, Z. Zhang, L. Xie, Q. Tian, and Y. Zhang. Skeleton graph scattering networks for 3d skeleton-based human motion prediction. In *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11-17, 2021*, pages 854–864. IEEE, 2021. URL: `https://doi.org/10.1109/ICCVW54120.2021.00101`, `doi:10.1109/ICCVW54120.2021.00101`.

[116] T. Sofianos, A. Sampieri, L. Franco, and F. Galasso. Space-time-separable graph convolutional network for pose forecasting. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11189–11198. IEEE, 2021. URL: `https://doi.org/10.1109/ICCV48922.2021.01102`, `doi:10.1109/ICCV48922.2021.01102`.

[117] C. Zhong, L. Hu, Z. Zhang, Y. Ye, and S. Xia. Spatio-temporal gating-adjacency GCN for human motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 6437–6446. IEEE, 2022. URL: `https://doi.org/10.1109/CVPR52688.2022.00634`, `doi:10.1109/CVPR52688.2022.00634`.

[118] Q. Cui, H. Sun, and F. Yang. Learning dynamic relationships for 3d human motion prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6518–6526. Computer Vision Foundation / IEEE, 2020. URL: `https://openaccess.thecvf.com/content_CVPR_2020/html/Cui_Learning_Dynamic_Relationships_for_3D_Human_Motion_Prediction_CVPR_2020_paper.html`, `doi:10.1109/CVPR42600.2020.00655`.

[119] Q. Cui and H. Sun. Towards accurate 3d human motion prediction from incomplete observations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 4801–4810. Computer Vision Foundation / IEEE, 2021. URL: `https://openaccess.`

thecvf.com/content/CVPR2021/html/Cui_Towards_Accurate_3D_Human_Motion_Prediction_From_Incomplete_Observations_CVPR_2021_paper.html, doi:10.1109/CVPR46437.2021.00477.

[120] B. Li, J. Tian, Z. Zhang, H. Feng, and X. Li. Multitask non-autoregressive model for human motion prediction. *IEEE Trans. Image Process.*, 30:2562–2574, 2021. URL: https://doi.org/10.1109/TIP.2020.3038362, doi:10.1109/TIP.2020.3038362.

[121] Y. Cai, L. Huang, Y. Wang, T. Cham, J. Cai, J. Yuan, J. Liu, X. Yang, Y. Zhu, X. Shen, D. Liu, J. Liu, and N. Magnenat-Thalmann. Learning progressive joint propagation for human motion prediction. In A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VII*, volume 12352 of *Lecture Notes in Computer Science*, pages 226–242. Springer, 2020. URL: https://doi.org/10.1007/978-3-030-58571-6_14, doi:10.1007/978-3-030-58571-6_14.

[122] O. Medjaouri and K. Desai. HR-STAN: high-resolution spatio-temporal attention network for 3d human motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*, pages 2539–2548. IEEE, 2022. URL: https://doi.org/10.1109/CVPRW56347.2022.00286, doi:10.1109/CVPRW56347.2022.00286.

[123] W. Guo, Y. Du, X. Shen, V. Lepetit, X. Alameda-Pineda, and F. Moreno-Noguer. Back to MLP: A simple baseline for human motion prediction. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pages 4798–4808. IEEE, 2023. URL: https://doi.org/10.1109/WACV56688.2023.00479, doi:10.1109/WACV56688.2023.00479.

[124] A. Bouazizi, A. Holzbock, U. Kressel, K. Dietmayer, and V. Belagiannis. Motionmixer: Mlp-based 3d human body pose forecasting. In L. D. Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 791–798. ijcai.org, 2022. URL: https://doi.org/10.24963/ijcai.2022/111, doi:10.24963/IJCAI.2022/111.

[125] J. N. Kundu, M. Gor, and R. V. Babu. Bihmp-gan: Bidirectional 3d human motion prediction GAN. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 8553–8560. AAAI Press, 2019. URL: https://doi.org/10.1609/aaai.v33i01.33018553, doi:10.1609/AAAI.V33I01.33018553.

[126] J. Walker, K. Marino, A. Gupta, and M. Hebert. The pose knows: Video forecasting by generating pose futures. In *IEEE International Conference on Computer*

Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pages 3352–3361. IEEE Computer Society, 2017. URL: `https://doi.org/10.1109/ICCV.2017.361`, `doi:10.1109/ICCV.2017.361`.

[127] W. Mao, M. Liu, and M. Salzmann. Generating smooth pose sequences for diverse human motion prediction. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 13289–13298. IEEE, 2021. URL: `https://doi.org/10.1109/ICCV48922.2021.01306`, `doi:10.1109/ICCV48922.2021.01306`.

[128] Y. Cai, Y. Wang, Y. Zhu, T. Cham, J. Cai, J. Yuan, J. Liu, C. Zheng, S. Yan, H. Ding, X. Shen, D. Liu, and N. Magnenat-Thalmann. A unified 3d human motion synthesis model via conditional variational auto-encoder*. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11625–11635. IEEE, 2021. URL: `https://doi.org/10.1109/ICCV48922.2021.01144`, `doi:10.1109/ICCV48922.2021.01144`.

[129] T. Salzmann, M. Pavone, and M. Ryll. Motron: Multimodal probabilistic human motion forecasting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 6447–6456. IEEE, 2022. URL: `https://doi.org/10.1109/CVPR52688.2022.00635`, `doi:10.1109/CVPR52688.2022.00635`.

[130] T. Lucas, F. Baradel, P. Weinzaepfel, and G. Rogez. Posegpt: Quantization-based 3d human motion generation and forecasting. In S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VI*, volume 13666 of *Lecture Notes in Computer Science*, pages 417–435. Springer, 2022. URL: `https://doi.org/10.1007/978-3-031-20068-7_24`, `doi:10.1007/978-3-031-20068-7\_24`.

[131] A. Blattmann, T. Milbich, M. Dorkenwald, and B. Ommer. Behavior-driven synthesis of human dynamics. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12236–12246. Computer Vision Foundation / IEEE, 2021. URL: `https://openaccess.thecvf.com/content/CVPR2021/html/Blattmann_Behavior-Driven_Synthesis_of_Human_Dynamics_CVPR_2021_paper.html`, `doi:10.1109/CVPR46437.2021.01206`.

[132] L. Dang, Y. Nie, C. Long, Q. Zhang, and G. Li. Diverse human motion prediction via gumbel-softmax sampling from an auxiliary space. In J. Magalhães, A. D. Bimbo, S. Satoh, N. Sebe, X. Alameda-Pineda, Q. Jin, V. Oria, and L. Toni, editors, *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 5162–5171. ACM, 2022. URL: `https://doi.org/10.1145/3503161.3547956`, `doi:10.1145/3503161.3547956`.

[133] J. Tanke, L. Zhang, A. Zhao, C. Tang, Y. Cai, L. Wang, P. Wu, J. Gall, and C. Keskin. Social diffusion: Long-term multiple human motion anticipation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 9567–9577. IEEE, 2023. URL: https://doi.org/10.1109/ICCV51070.2023.00880, doi:10.1109/ICCV51070.2023.00880.

[134] C. M. Jiang, A. Cornman, C. Park, B. Sapp, Y. Zhou, and D. Anguelov. Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 9644–9653. IEEE, 2023. URL: https://doi.org/10.1109/CVPR52729.2023.00930, doi:10.1109/CVPR52729.2023.00930.

[135] Z. Zhou and B. Wang. UDE: A unified driving engine for human motion generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 5632–5641. IEEE, 2023. URL: https://doi.org/10.1109/CVPR52729.2023.00545, doi:10.1109/CVPR52729.2023.00545.

[136] Y. A. Farha and J. Gall. Uncertainty-aware anticipation of activities. In *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, pages 1197–1204. IEEE, 2019. URL: https://doi.org/10.1109/ICCVW.2019.00151, doi:10.1109/ICCVW.2019.00151.

[137] Y. A. Farha, A. Richard, and J. Gall. When will you do what? - anticipating temporal occurrences of activities. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5343–5352. Computer Vision Foundation / IEEE Computer Society, 2018. URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Abu_Farha_When_Will_You_CVPR_2018_paper.html, doi:10.1109/CVPR.2018.00560.

[138] Q. Ke, M. Fritz, and B. Schiele. Time-conditioned action anticipation in one shot. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9925–9934. Computer Vision Foundation / IEEE, 2019. URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Ke_Time-Conditioned_Action_Anticipation_in_One_Shot_CVPR_2019_paper.html, doi:10.1109/CVPR.2019.01016.

[139] B. Fernando and S. Herath. Anticipating human actions by correlating past with the future with jaccard similarity measures. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 13224–13233. Computer Vision Foundation / IEEE, 2021. URL: https://openaccess.thecvf.com/content/CVPR2021/html/Fernando_

Anticipating_Human_Actions_by_Correlating_Past_With_the_Future_With_CVPR_2021_paper.html, doi:10.1109/CVPR46437.2021.01302.

[140] Y. A. Farha, Q. Ke, B. Schiele, and J. Gall. Long-term anticipation of activities with cycle consistency. In Z. Akata, A. Geiger, and T. Sattler, editors, *Pattern Recognition - 42nd DAGM German Conference, DAGM GCPR 2020, Tübingen, Germany, September 28 - October 1, 2020, Proceedings*, volume 12544 of *Lecture Notes in Computer Science*, pages 159–173. Springer, 2020. URL: https://doi.org/10.1007/978-3-030-71278-5_12, doi:10.1007/978-3-030-71278-5_12.

[141] F. Sener, D. Singhania, and A. Yao. Temporal aggregate representations for long-range video understanding. In A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVI*, volume 12361 of *Lecture Notes in Computer Science*, pages 154–171. Springer, 2020. URL: https://doi.org/10.1007/978-3-030-58517-4_10, doi:10.1007/978-3-030-58517-4_10.

[142] A. Furnari and G. M. Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6251–6260. IEEE, 2019. URL: https://doi.org/10.1109/ICCV.2019.00635, doi:10.1109/ICCV.2019.00635.

[143] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes. Predicting the future: A jointly learnt model for action anticipation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5561–5570. IEEE, 2019. URL: https://doi.org/10.1109/ICCV.2019.00566, doi:10.1109/ICCV.2019.00566.

[144] A. Miech, I. Laptev, J. Sivic, H. Wang, L. Torresani, and D. Tran. Leveraging the present to anticipate the future in videos. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2915–2922. Computer Vision Foundation / IEEE, 2019. URL: http://openaccess.thecvf.com/content_CVPRW_2019/html/Precognition/Miech_Leveraging_the_Present_to_Anticipate_the_Future_in_Videos_CVPRW_2019_paper.html, doi:10.1109/CVPRW.2019.00351.

[145] F. Sener and A. Yao. Zero-shot anticipation for instructional activities. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 862–871. IEEE, 2019. URL: https://doi.org/10.1109/ICCV.2019.00095, doi:10.1109/ICCV.2019.00095.

[146] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena. Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. In D. Kragic, A. Bicchi, and A. D. Luca, editors, *2016 IEEE International Conference on Robotics and Automation, ICRA 2016, Stockholm, Sweden, May 16-*

21, 2016, pages 3118–3125. IEEE, 2016. URL: https://doi.org/10.1109/ICRA.2016.7487478, doi:10.1109/ICRA.2016.7487478.

[147] Y. Wu, L. Zhu, X. Wang, Y. Yang, and F. Wu. Learning to anticipate egocentric actions by imagination. *IEEE Trans. Image Process.*, 30:1143–1152, 2021. URL: https://doi.org/10.1109/TIP.2020.3040521, doi:10.1109/TIP.2020.3040521.

[148] C. Vondrick, H. Pirsiavash, and A. Torralba. Anticipating visual representations from unlabeled video. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 98–106. IEEE Computer Society, 2016. URL: https://doi.org/10.1109/CVPR.2016.18, doi:10.1109/CVPR.2016.18.

[149] T. Han, W. Xie, and A. Zisserman. Memory-augmented dense predictive coding for video representation learning. In A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III*, volume 12348 of *Lecture Notes in Computer Science*, pages 312–329. Springer, 2020. URL: https://doi.org/10.1007/978-3-030-58580-8_19, doi:10.1007/978-3-030-58580-8_19.

[150] M. Liu, S. Tang, Y. Li, and J. M. Rehg. Forecasting human-object interaction: Joint prediction of motor attention and actions in first person video. In A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 704–721. Springer, 2020. URL: https://doi.org/10.1007/978-3-030-58452-8_41, doi:10.1007/978-3-030-58452-8_41.

[151] Y. Zhu, D. S. Doermann, Y. Zhang, Q. Liu, and A. Girgensohn. What and how? jointly forecasting human action and pose. In *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*, pages 771–778. IEEE, 2020. URL: https://doi.org/10.1109/ICPR48806.2021.9412833, doi:10.1109/ICPR48806.2021.9412833.

[152] A. Bar, R. Herzig, X. Wang, A. Rohrbach, G. Chechik, T. Darrell, and A. Globerson. Compositional video synthesis with action graphs. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 662–673. PMLR, 2021. URL: http://proceedings.mlr.press/v139/bar21a.html.

[153] D. Suris, R. Liu, and C. Vondrick. Learning the predictability of the future. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12607–12617. Computer Vision Foundation / IEEE, 2021. URL: https://openaccess.thecvf.com/content/CVPR2021/html/Suris_

Learning_the_Predictability_of_the_Future_CVPR_2021_paper.html, doi:10.1109/CVPR46437.2021.01242.

[154] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 2017. URL: http://proceedings.mlr.press/v70/arjovsky17a.html.

[155] B. Wandt, H. Ackermann, and B. Rosenhahn. 3d reconstruction of human motion from monocular image sequences. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(8):1505–1516, 2016. URL: https://doi.org/10.1109/TPAMI.2016.2553028, doi:10.1109/TPAMI.2016.2553028.

[156] B. Wandt and B. Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7782–7791. Computer Vision Foundation / IEEE, 2019. URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Wandt_RepNet_Weakly_Supervised_Training_of_an_Adversarial_Reprojection_Network_for_CVPR_2019_paper.html, doi:10.1109/CVPR.2019.00797.

[157] C. Xu, R. T. Tan, Y. Tan, S. Chen, Y. G. Wang, X. Wang, and Y. Wang. Eqmotion: Equivariant multi-agent motion prediction with invariant interaction reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 1410–1420. IEEE, 2023. URL: https://doi.org/10.1109/CVPR52729.2023.00142, doi:10.1109/CVPR52729.2023.00142.

[158] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2252–2261. IEEE, 2019. URL: https://doi.org/10.1109/ICCV.2019.00234, doi:10.1109/ICCV.2019.00234.

[159] J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4724–4733. IEEE Computer Society, 2017. URL: https://doi.org/10.1109/CVPR.2017.502, doi:10.1109/CVPR.2017.502.

[160] J. S. Katrolia, A. El-Sherif, H. Feld, B. Mirbach, J. R. Rambach, and D. Stricker. Ticam: A time-of-flight in-car cabin monitoring dataset. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, page 277. BMVA Press, 2021. URL: https://www.bmvc2021-virtualconference.com/assets/papers/0701.pdf.

[161] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social GAN: socially acceptable trajectories with generative adversarial networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2255–2264. Computer Vision Foundation / IEEE Computer Society, 2018. URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Gupta_Social_GAN_Socially_CVPR_2018_paper.html, doi:10.1109/CVPR.2018.00240.

[162] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[163] W. Susanto, M. Rohrbach, and B. Schiele. 3d object detection with multiple kinects. In A. Fusiello, V. Murino, and R. Cucchiara, editors, *Computer Vision - ECCV 2012. Workshops and Demonstrations - Florence, Italy, October 7-13, 2012, Proceedings, Part II*, volume 7584 of *Lecture Notes in Computer Science*, pages 93–102. Springer, 2012. URL: https://doi.org/10.1007/978-3-642-33868-7_10, doi:10.1007/978-3-642-33868-7_10.

[164] B. L. Bhatnagar, X. Xie, I. A. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll. BEHAVE: dataset and method for tracking human object interactions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15914–15925. IEEE, 2022. URL: https://doi.org/10.1109/CVPR52688.2022.01547, doi:10.1109/CVPR52688.2022.01547.

[165] N. Jiang, T. Liu, Z. Cao, J. Cui, Y. Chen, H. Wang, Y. Zhu, and S. Huang. CHAIRS: towards full-body articulated human-object interaction. *CoRR*, abs/2212.10621, 2022. URL: https://doi.org/10.48550/arXiv.2212.10621, arXiv:2212.10621, doi:10.48550/ARXIV.2212.10621.

[166] A. Bhattacharyya, B. Schiele, and M. Fritz. Accurate and diverse sampling of sequences based on a "best of many" sample objective. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8485–8493. Computer Vision Foundation / IEEE Computer Society, 2018. URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Bhattacharyya_Accurate_and_Diverse_CVPR_2018_paper.html, doi:10.1109/CVPR.2018.00885.

[167] S. Xu, Y. Wang, and L. Gui. Stochastic multi-person 3d motion forecasting. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL: https://openreview.net/pdf?id=_s1N-DnxdyT.

[168] J. Ho and T. Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022. URL: https://doi.org/10.48550/arXiv.2207.12598, arXiv:2207.12598, doi:10.48550/ARXIV.2207.12598.

[169] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. 162:16784–16804, 2022. URL: https://proceedings.mlr.press/v162/nichol22a.html.

[170] K. Karunratanakul, K. Preechakul, S. Suwajanakorn, and S. Tang. GMD: controllable human motion synthesis via guided diffusion models. *CoRR*, abs/2305.12577, 2023. URL: https://doi.org/10.48550/arXiv.2305.12577, arXiv:2305.12577, doi:10.48550/ARXIV.2305.12577.

[171] M. Janner, Y. Du, J. B. Tenenbaum, and S. Levine. Planning with diffusion for flexible behavior synthesis. 162:9902–9915, 2022. URL: https://proceedings.mlr.press/v162/janner22a.html.

[172] D. Rempe, Z. Luo, X. B. Peng, Y. Yuan, K. Kitani, K. Kreis, S. Fidler, and O. Litany. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 13756–13766. IEEE, 2023. URL: https://doi.org/10.1109/CVPR52729.2023.01322, doi:10.1109/CVPR52729.2023.01322.

[173] Y. Yuan, J. Song, U. Iqbal, A. Vahdat, and J. Kautz. Physdiff: Physics-guided human motion diffusion model. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 15964–15975. IEEE, 2023. URL: https://doi.org/10.1109/ICCV51070.2023.01467, doi:10.1109/ICCV51070.2023.01467.

[174] J. Zhang, Y. Zhang, X. Cun, Y. Zhang, H. Zhao, H. Lu, X. Shen, and S. Ying. Generating human motion from textual descriptions with discrete representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14730–14740. IEEE, 2023. URL: https://doi.org/10.1109/CVPR52729.2023.01415, doi:10.1109/CVPR52729.2023.01415.

[175] S. Azadi, A. Shah, T. Hayes, D. Parikh, and S. Gupta. Make-an-animation: Large-scale text-conditional 3d human motion generation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 14993–15002. IEEE, 2023. URL: https://doi.org/10.1109/ICCV51070.2023.01381, doi:10.1109/ICCV51070.2023.01381.

[176] M. Petrovich, M. J. Black, and G. Varol. TEMOS: generating diverse human motions from textual descriptions. In S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXII*, volume 13682 of *Lecture Notes in Computer Science*, pages 480–497. Springer, 2022. URL: https://doi.org/10.1007/978-3-031-20047-2_28, doi:10.1007/978-3-031-20047-2_28.

[177] J. Kim, J. Kim, and S. Choi. FLAME: free-form language-based motion synthesis & editing. In B. Williams, Y. Chen, and J. Neville, editors, *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 8255–8263. AAAI Press, 2023. URL: `https://doi.org/10.1609/aaai.v37i7.25996`, `doi:10.1609/AAAI.V37I7.25996`.

[178] S. S. Kalakonda, S. Maheshwari, and R. K. Sarvadevabhatla. Action-gpt: Leveraging large-scale language models for improved and generalized zero shot action generation. *CoRR*, abs/2211.15603, 2022. URL: `https://doi.org/10.48550/arXiv.2211.15603`, `arXiv:2211.15603`, `doi:10.48550/ARXIV.2211.15603`.

[179] S. Zhang, Y. Zhang, Q. Ma, M. J. Black, and S. Tang. PLACE: proximity learning of articulation and contact in 3d environments. In V. Struc and F. G. Fernández, editors, *8th International Conference on 3D Vision, 3DV 2020, Virtual Event, Japan, November 25-28, 2020*, pages 642–651. IEEE, 2020. URL: `https://doi.org/10.1109/3DV50981.2020.00074`, `doi:10.1109/3DV50981.2020.00074`.

[180] M. Hassan, V. Choutas, D. Tzionas, and M. J. Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2282–2292. IEEE, 2019. URL: `https://doi.org/10.1109/ICCV.2019.00237`, `doi:10.1109/ICCV.2019.00237`.

[181] Y. Zhang, M. Hassan, H. Neumann, M. J. Black, and S. Tang. Generating 3d people in scenes without people. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6193–6203. Computer Vision Foundation / IEEE, 2020. URL: `https://openaccess.thecvf.com/content_CVPR_2020/html/Zhang_Generating_3D_People_in_Scenes_Without_People_CVPR_2020_paper.html`, `doi:10.1109/CVPR42600.2020.00623`.

[182] M. Hassan, P. Ghosh, J. Tesch, D. Tzionas, and M. J. Black. Populating 3d scenes by learning human-scene interaction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14708–14718. Computer Vision Foundation / IEEE, 2021. URL: `https://openaccess.thecvf.com/content/CVPR2021/html/Hassan_Populating_3D_Scenes_by_Learning_Human-Scene_Interaction_CVPR_2021_paper.html`, `doi:10.1109/CVPR46437.2021.01447`.

[183] M. Hassan, Y. Guo, T. Wang, M. J. Black, S. Fidler, and X. B. Peng. Synthesizing physical character-scene interactions. In E. Brunvand, A. Sheffer, and M. Wimmer, editors, *ACM SIGGRAPH 2023 Conference Proceedings, SIGGRAPH 2023, Los Angeles, CA, USA, August 6-10, 2023*, pages 63:1–63:9. ACM, 2023. URL: `https://doi.org/10.1145/3588432.3591525`, `doi:10.1145/3588432.3591525`.

[184] Z. Xie, J. Tseng, S. Starke, M. van de Panne, and C. K. Liu. Hierarchical planning and control for box loco-manipulation. *Proc. ACM Comput. Graph. Interact. Tech.*, 6(3):31:1–31:18, 2023. URL: `https://doi.org/10.1145/3606931`, `doi:10.1145/3606931`.

[185] W. Mao, M. Liu, R. I. Hartley, and M. Salzmann. Contact-aware human motion forecasting. 2022. URL: `http://papers.nips.cc/paper_files/paper/2022/hash/3018804d037cc101b73624f381bed0cb-Abstract-Conference.html`.

[186] J. Wang, H. Xu, J. Xu, S. Liu, and X. Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9401–9411. Computer Vision Foundation / IEEE, 2021. URL: `https://openaccess.thecvf.com/content/CVPR2021/html/Wang_Synthesizing_Long-Term_3D_Human_Motion_and_Interaction_in_3D_Scenes_CVPR_2021_paper.html`, `doi:10.1109/CVPR46437.2021.00928`.

[187] J. Wang, S. Yan, B. Dai, and D. Lin. Scene-aware generative network for human motion synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12206–12215. Computer Vision Foundation / IEEE, 2021. URL: `https://openaccess.thecvf.com/content/CVPR2021/html/Wang_Scene-Aware_Generative_Network_for_Human_Motion_Synthesis_CVPR_2021_paper.html`, `doi:10.1109/CVPR46437.2021.01203`.

[188] M. Hassan, D. Ceylan, R. Villegas, J. Saito, J. Yang, Y. Zhou, and M. J. Black. Stochastic scene-aware motion prediction. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11354–11364. IEEE, 2021. URL: `https://doi.org/10.1109/ICCV48922.2021.01118`, `doi:10.1109/ICCV48922.2021.01118`.

[189] O. Taheri, V. Choutas, M. J. Black, and D. Tzionas. GOAL: generating 4d whole-body motion for hand-object grasping. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 13253–13263. IEEE, 2022. URL: `https://doi.org/10.1109/CVPR52688.2022.01291`, `doi:10.1109/CVPR52688.2022.01291`.

[190] P. Tendulkar, D. Surís, and C. Vondrick. FLEX: full-body grasping without full-body grasps. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 21179–21189. IEEE, 2023. URL: `https://doi.org/10.1109/CVPR52729.2023.02029`, `doi:10.1109/CVPR52729.2023.02029`.

[191] X. Zhang, B. L. Bhatnagar, S. Starke, V. Guzov, and G. Pons-Moll. COUCH: towards controllable human-chair interactions. In S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision - ECCV*

2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part V, volume 13665 of *Lecture Notes in Computer Science*, pages 518–535. Springer, 2022. URL: `https://doi.org/10.1007/978-3-031-20065-6_30`, `doi:10.1007/978-3-031-20065-6\_30`.

[192] Y. Wu, J. Wang, Y. Zhang, S. Zhang, O. Hilliges, F. Yu, and S. Tang. SAGA: stochastic whole-body grasping with contact. In S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VI*, volume 13666 of *Lecture Notes in Computer Science*, pages 257–274. Springer, 2022. URL: `https://doi.org/10.1007/978-3-031-20068-7_15`, `doi:10.1007/978-3-031-20068-7\_15`.

[193] J. Lee and H. Joo. Locomotion-action-manipulation: Synthesizing human-scene interactions in complex 3d environments. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 9629–9640. IEEE, 2023. URL: `https://doi.org/10.1109/ICCV51070.2023.00886`, `doi:10.1109/ICCV51070.2023.00886`.

[194] W. Zhang, R. Dabral, T. Leimkühler, V. Golyanik, M. Habermann, and C. Theobalt. ROAM: robust and object-aware motion generation using neural pose descriptors. *CoRR*, abs/2308.12969, 2023. URL: `https://doi.org/10.48550/arXiv.2308.12969`, `arXiv:2308.12969`, `doi:10.48550/ARXIV.2308.12969`.

[195] N. Kulkarni, D. Rempe, K. Genova, A. Kundu, J. Johnson, D. Fouhey, and L. J. Guibas. NIFTY: neural object interaction fields for guided human motion synthesis. *CoRR*, abs/2307.07511, 2023. URL: `https://doi.org/10.48550/arXiv.2307.07511`, `arXiv:2307.07511`, `doi:10.48550/ARXIV.2307.07511`.

[196] J. Li, J. Wu, and C. K. Liu. Object motion guided human motion synthesis. *ACM Trans. Graph.*, 42(6):197:1–197:11, 2023. URL: `https://doi.org/10.1145/3618333`, `doi:10.1145/3618333`.

[197] W. Wan, L. Yang, L. Liu, Z. Zhang, R. Jia, Y. Choi, J. Pan, C. Theobalt, T. Komura, and W. Wang. Learn to predict how humans manipulate large-sized objects from interactive motions. *IEEE Robotics Autom. Lett.*, 7(2):4702–4709, 2022. URL: `https://doi.org/10.1109/LRA.2022.3151614`, `doi:10.1109/LRA.2022.3151614`.

[198] D. Driess, Z. Huang, Y. Li, R. Tedrake, and M. Toussaint. Learning multi-object dynamics with compositional neural radiance fields. In K. Liu, D. Kulic, and J. Ichnowski, editors, *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, volume 205 of *Proceedings of Machine Learning Research*, pages 1755–1768. PMLR, 2022. URL: `https://proceedings.mlr.press/v205/driess23a.html`.

[199] D. Mrowca, C. Zhuang, E. Wang, N. Haber, L. Fei-Fei, J. Tenenbaum, and D. L. K. Yamins. Flexible neural representation for physics prediction. pages 8813–8824, 2018. URL: `https://proceedings.neurips.cc/paper/2018/hash/fd9dd764a6f1d73f4340d570804eacc4-Abstract.html`.

[200] D. Rempe, S. Sridhar, H. Wang, and L. J. Guibas. Predicting the physical dynamics of unseen 3d objects. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 2823–2832. IEEE, 2020. URL: `https://doi.org/10.1109/WACV45572.2020.9093374`, `doi:10.1109/WACV45572.2020.9093374`.

[201] G. Zhu, Z. Huang, and C. Zhang. Object-oriented dynamics predictor. pages 9826–9837, 2018. URL: `https://proceedings.neurips.cc/paper/2018/hash/713fd63d76c8a57b16fc433fb4ae718a-Abstract.html`.

[202] A. Ghosh, R. Dabral, V. Golyanik, C. Theobalt, and P. Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. volume 42, pages 1–12, 2023. URL: `https://doi.org/10.1111/cgf.14739`, `doi:10.1111/CGF.14739`.

[203] J. Braun, S. J. Christen, M. Kocabas, E. Aksan, and O. Hilliges. Physically plausible full-body hand-object interaction synthesis. *CoRR*, abs/2309.07907, 2023. URL: `https://doi.org/10.48550/arXiv.2309.07907`, `arXiv:2309.07907`, `doi:10.48550/ARXIV.2309.07907`.

[204] Y. Ye, X. Li, A. Gupta, S. D. Mello, S. Birchfield, J. Song, S. Tulsiani, and S. Liu. Affordance diffusion: Synthesizing hand-object interactions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 22479–22489. IEEE, 2023. URL: `https://doi.org/10.1109/CVPR52729.2023.02153`, `doi:10.1109/CVPR52729.2023.02153`.

[205] J. Zheng, Q. Zheng, L. Fang, Y. Liu, and L. Yi. CAMS: canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 585–594. IEEE, 2023. URL: `https://doi.org/10.1109/CVPR52729.2023.00064`, `doi:10.1109/CVPR52729.2023.00064`.

[206] K. Zhou, B. L. Bhatnagar, J. E. Lenssen, and G. Pons-Moll. TOCH: spatio-temporal object-to-hand correspondence for motion refinement. In S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part III*, volume 13663 of *Lecture Notes in Computer Science*, pages 1–19. Springer, 2022. URL: `https://doi.org/10.1007/978-3-031-20062-5_1`, `doi:10.1007/978-3-031-20062-5\_1`.

[207] P. G. Kry and D. K. Pai. Interaction capture and synthesis. *ACM Trans. Graph.*, 25(3):872–880, 2006. URL: `https://doi.org/10.1145/1141911.1141969`, `doi:10.1145/1141911.1141969`.

[208] Q. Li, J. Wang, C. C. Loy, and B. Dai. Task-oriented human-object interactions generation with implicit neural representations. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pages 3023–3032. IEEE, 2024. URL: `https://doi.org/10.1109/WACV57701.2024.00301`, `doi:10.1109/WACV57701.2024.00301`.

[209] P. Grady, C. Tang, C. D. Twigg, M. Vo, S. Brahmbhatt, and C. C. Kemp. Contactopt: Optimizing contact to improve grasps. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 1471–1481. Computer Vision Foundation / IEEE, 2021. URL: `https://openaccess.thecvf.com/content/CVPR2021/html/Grady_ContactOpt_Optimizing_Contact_To_Improve_Grasps_CVPR_2021_paper.html`, `doi:10.1109/CVPR46437.2021.00152`.

[210] H. Jiang, S. Liu, J. Wang, and X. Wang. Hand-object contact consistency reasoning for human grasps generation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11087–11096. IEEE, 2021. URL: `https://doi.org/10.1109/ICCV48922.2021.01092`, `doi:10.1109/ICCV48922.2021.01092`.

[211] T. H. E. Tse, Z. Zhang, K. I. Kim, A. Leonardis, F. Zheng, and H. J. Chang. S^{2}contact: Graph-based network for 3d hand-object contact estimation with semi-supervised learning. In S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part I*, volume 13661 of *Lecture Notes in Computer Science*, pages 568–584. Springer, 2022. URL: `https://doi.org/10.1007/978-3-031-19769-7_33`, `doi:10.1007/978-3-031-19769-7\_33`.

[212] W. Mao, M. Liu, and M. Salzmann. Weakly-supervised action transition learning for stochastic human motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 8141–8150. IEEE, 2022. URL: `https://doi.org/10.1109/CVPR52688.2022.00798`, `doi:10.1109/CVPR52688.2022.00798`.

[213] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, and D. Cohen-Or. Motionclip: Exposing human motion generation to CLIP space. In S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXII*, volume 13682 of *Lecture Notes in Computer Science*, pages 358–374. Springer, 2022. URL: `https://doi.org/10.1007/978-3-031-20047-2_21`, `doi:10.1007/978-3-031-20047-2\_21`.

[214] Y. Zhang, M. J. Black, and S. Tang. We are more than our joints: Predicting how 3d bodies move. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3372–3382. Computer Vision Foundation / IEEE, 2021. URL: `https://openaccess.thecvf.com/content/CVPR2021/html/Zhang_We_Are_More_Than_Our_Joints_Predicting_How_3D_Bodies_CVPR_2021_paper.html`, `doi:10.1109/CVPR46437.2021.00338`.

[215] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL: `https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html`.

[216] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL: `http://proceedings.mlr.press/v139/radford21a.html`.

[217] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 77–85. IEEE Computer Society, 2017. URL: `https://doi.org/10.1109/CVPR.2017.16`, `doi:10.1109/CVPR.2017.16`.

[218] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li. On the continuity of rotation representations in neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5745–5753. Computer Vision Foundation / IEEE, 2019. URL: `http://openaccess.thecvf.com/content_CVPR_2019/html/Zhou_On_the_Continuity_of_Rotation_Representations_in_Neural_Networks_CVPR_2019_paper.html`, `doi:10.1109/CVPR.2019.00589`.

[219] C. Guo, X. Zuo, S. Wang, S. Zou, Q. Sun, A. Deng, M. Gong, and L. Cheng. Action2motion: Conditioned generation of 3d human motions. In C. W. Chen, R. Cucchiara, X. Hua, G. Qi, E. Ricci, Z. Zhang, and R. Zimmermann, editors, *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 2021–2029. ACM, 2020. URL: `https://doi.org/10.1145/3394171.3413635`, `doi:10.1145/3394171.3413635`.

[220] C. Yeshwanth, Y. Liu, M. Nießner, and A. Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *IEEE/CVF International Conference on Com-*

puter Vision, ICCV 2023, Paris, France, October 1-6, 2023, pages 12–22. IEEE, 2023. URL: `https://doi.org/10.1109/ICCV51070.2023.00008`, `doi:10.1109/ICCV51070.2023.00008`.

A Project Pages & Videos

A.1 Forecasting Characteristic 3D Poses of Human Actions

- Project Page: charposes.christian-diller.de
- Video: youtube.com/watch?v=kVhn8OWMgME

A.2 FutureHuman3D: Forecasting Complex Long-Term 3D Human Behavior from Video Observations

- Project Page: future-human-3d.christian-diller.de
- Video: youtube.com/watch?v=18du85YFXL0

A.3 CG-HOI: Contact-Guided 3D Human-Object Interaction Generation

- Project Page: cg-hoi.christian-diller.de
- Video: youtube.com/watch?v=GNyQwTwZ15s

B Authored and Co-authored Publications

Authored

1. **Christian Diller**, Thomas Funkhouser, and Angela Dai. "Forecasting Characteristic 3D Poses of Human Actions" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 15914-15923
2. **Christian Diller**, Thomas Funkhouser, and Angela Dai. "FutureHuman3D: Forecasting Complex Long-Term 3D Human Behavior from Video Observations" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024
3. **Christian Diller** and Angela Dai. "CG-HOI: Contact-Guided 3D Human-Object Interactions" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024

Co-Authored

1. Angela Dai, **Christian Diller**, and Matthias Niessner, "SG-NN: Sparse Generative Neural Networks for Self-Supervised Scene Completion of RGB-D Scans" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 849-858

C Original Publications

C.1 Forecasting Characteristic 3D Poses of Human Actions

Copyright Notice

In accordance with the IEEE Thesis/Dissertation Reuse Permissions, we include the accepted version of the original publication [40] in the following.

Forecasting Characteristic 3D Poses of Human Actions

Conference Proceedings:
2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Author: Christian Diller

Publisher: IEEE

Date: June 2022

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

Forecasting Characteristic 3D Poses of Human Actions

Christian Diller[1] Thomas Funkhouser[2] Angela Dai[1]

[1]Technical University of Munich [2]Google

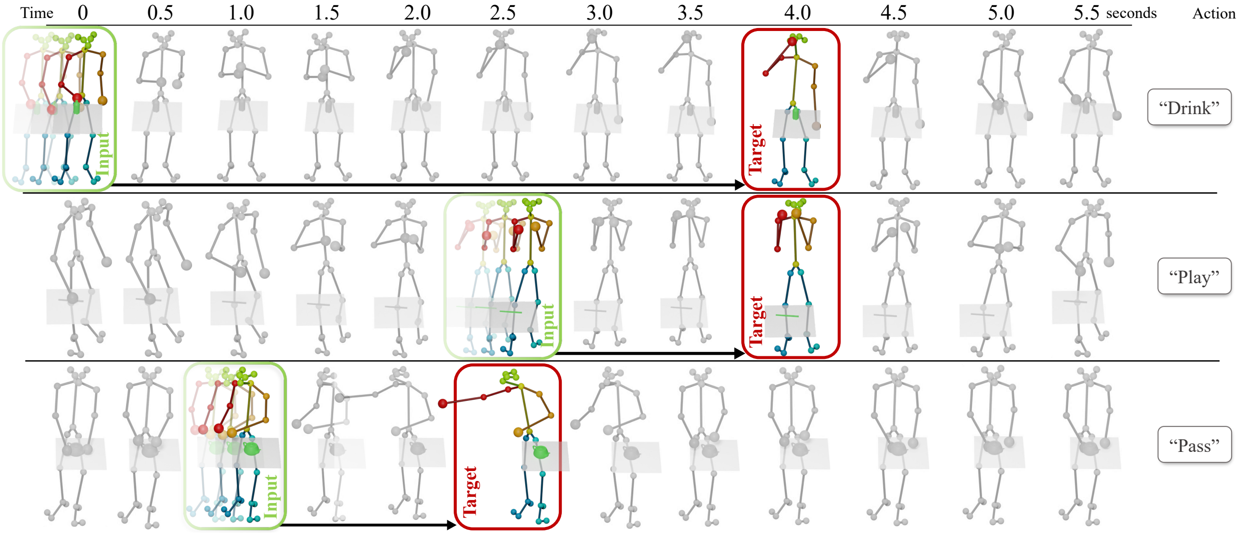


Figure 1. For a real-world 3d skeleton sequence of a human performing an action, we propose to forecast the semantically meaningful *characteristic 3d pose*, representing the action goal for this sequence. As input, we take a short observation of a sequence of consecutive poses leading up to the target characteristic pose. Thus, we propose to take a *goal-oriented* approach, predicting the key moments characterizing future behavior, instead of predicting continuous motion, which can occur at varying speeds with predictions more easily diverging for longer-term (>1s) predictions. We develop an attention-driven probabilistic approach to capture the most likely modes of possible future characteristic poses.

Abstract

We propose the task of forecasting characteristic 3d poses: from a short sequence observation of a person, predict a future 3d pose of that person in a likely action-defining, characteristic pose – for instance, from observing a person picking up an apple, predict the pose of the person eating the apple. Prior work on human motion prediction estimates future poses at fixed time intervals. Although easy to define, this frame-by-frame formulation confounds temporal and intentional aspects of human action. Instead, we define a semantically meaningful pose prediction task that decouples the predicted pose from time, taking inspiration from goal-directed behavior. To predict characteristic poses, we propose a probabilistic approach that models the possible multi-modality in the distribution of likely characteristic poses. We then sample future pose hypotheses from the predicted distribution in an autoregressive fashion to model dependencies between joints. To evaluate our method, we construct a dataset of manually annotated characteristic 3d poses. Our experiments with this dataset suggest that our proposed probabilistic approach outperforms state-of-the-art methods by 26% on average.

1. Introduction

Future human pose forecasting is fundamental towards a comprehensive understanding of human behavior, and consequently towards achieving higher-level perception in machine interactions with humans, such as autonomous robots or vehicles. In fact, prediction is considered to play a foundational part in intelligence [3, 11, 15]. In particular, predicting the 3d pose of a human in the future lays a basis for both structural and semantic understanding of human behavior, and for an agent to take fine-grained anticipatory action towards the forecasted future. For example, a robotic surgical assistant should predict in advance where best to place a tool to assist the surgeon's next action, what sensor viewpoints will be best to observe the surgeon's next manipulation, and how to position itself to be out of the way at critical future moments.

Recently, we have seen notable progress in the task of future 3D human motion prediction – from an initial observation of a person, forecasting the 3D behavior of that person up to ≈ 1 second in the future [12,19,23–25]. Various methods have been developed, leveraging RNNs [12, 14, 19, 25], graph convolutional neural networks [22, 24], and attention [23, 30]. However, these approaches all take a temporal approach towards forecasting future 3D human poses, and predict poses at fixed time intervals to imitate the fixed frame rate of camera capture. This makes it difficult to predict longer-term (several seconds) behavior, which requires predicting both the time-based speed of movement as well as the higher-level goal of the future action.

Thus, we propose to decouple the temporal and intentional behavior, and introduce a new task of forecasting *characteristic 3d poses* of a person's future action: from a short pose sequence observation of a human, the goal is to predict a future pose of the person in a characteristic, action-defining moment. This has many potential applications, including HRI, surveillance, visualization, simulation, and content creation. It could be used to predict the hand-off point when a robot is passing an object to a person; to detect and display future poses worthy of alerts in a safety monitoring system; to coordinate grasps when assisting a person lifting a heavy object; to assist tracking through occlusions; or to predict future keyframes, as is done in video generation [20, 27].

Fig. 2 visualizes the difference between this new task and the traditional, time-based approach: our task is to predict a next characteristic pose at action-defining moments (blue dots) rather than at fixed time-intervals (red dots). As shown in Fig. 1, the characteristic 3d poses are more semantically meaningful and rarely occur at exactly the same times in the future. We believe that predicting possible future characteristic 3d poses takes an important step towards forecasting human action, by understanding the objectives underlying a future action or movement separately from the speed at which they occur.

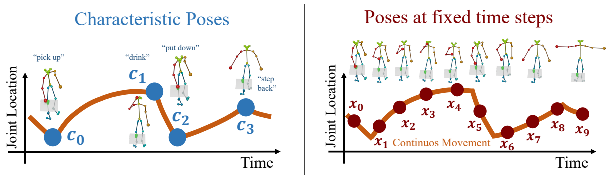


Figure 2. These plots show the salient difference between our new task (left) and the traditional one (right). The orange curve depicts the motion of one joint (e.g., hand position as a person drinks from a glass). It represents a typical piecewise continuous motion, which has discrete action-defining characteristic poses at cusps of the motion curves (e.g., grasping the glass on the table, putting it to ones mouth, etc.) separating smooth trajectories connecting them (e.g., raising or lowering the glass). Our task is to predict future characteristic poses (blue dots on left) rather than in-between poses at regular time intervals (red points on right).

Since future characteristic 3d poses often occur at longer-term intervals (> 1s) in the future, there may be multiple likely modes of the characteristic poses, and we must capture this multi-modality in our forecasting. Rather than deterministic forecasting, as is an approach in many 3D human pose forecasting approaches [22–24], we develop an attention-driven prediction of probability heatmaps representing the likelihood of each human pose joint in its future location. This enables generation of multiple, diverse hypotheses for the future pose. To generate a coherent pose prediction across all pose joints' potentially multi-modal futures, we make autoregressive predictions for the end effectors of the actions (e.g., predicting the right hand, then the left hand conditioned on the predicted right hand location) – this enables a tractable modeling of the joint distribution of the human pose joints.

To demonstrate our proposed approach, we introduce a new benchmark on *characteristic 3D Pose* prediction. We annotate characteristic keyframes in sequences from the GRAB [29] and Human3.6M [17] datasets. Experiments on this benchmark show that our probabilistic approach outperforms time-based state of the art by 26% on average.

In summary, we present the following contributions:

- We propose the task of forecasting *characteristic 3D Poses*: predicting likely next action-defining future moments from a sequence observation of a person, towards goal-oriented understanding of pose forecasting.
- We introduce an attention-driven, probabilistic approach to tackle this problem and model the most likely modes for the next characteristic pose, and show that it outperforms state of the art.
- We autoregressively model the multi-modal distribution of future pose joint locations, casting pose prediction as a product of conditional distributions of end effector locations (e.g., hands), and the rest of the body.
- We introduce a dataset and benchmark on our *characteristic 3D Pose* prediction, comprising 1535 annotated characteristic pose frames from the GRAB [29] and Human3.6M [17] datasets.

2. Related Work

Deterministic Human Motion Forecasting. Many works have focused on human motion forecasting, cast as a sequential task to predict a sequence of human poses according to the fixed frame rate capture of a camera. For this sequential task, recurrent neural networks have been widely used for human motion forecasting [1, 9, 12, 13, 19, 25, 33]. Such approaches have achieved impressive success in

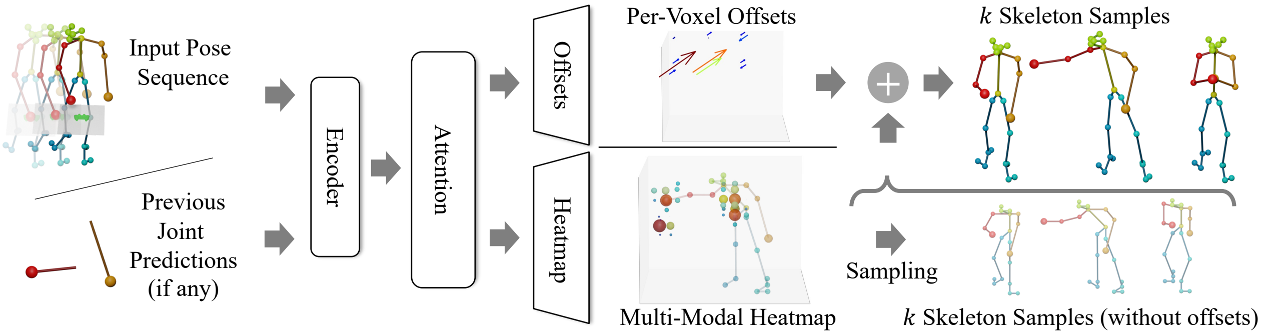


Figure 3. Overview of our approach for characteristic 3d pose prediction. From an input observed pose sequence, as well as any prior joint predictions, we leverage attention to learn inter-joint dependencies, and decode a 3d volumetric heatmap representing the probability distribution for the next joint to be predicted as well as a per-voxel offset field of same size for improved joint placement. This enables autoregressive sampling to obtain final pose hypotheses characterizing likely characteristic 3d poses.

shorter-term prediction (up to $\approx$ 1s, occasionally several seconds for longer term predictions), but the RNN summarization of history into a fixed-size representation struggles to maintain the long-term dependencies needed for forecasting further into the future.

To address some of the drawbacks of RNNs, non-recurrent models have also been adopted, encoding temporal history with convolutional or fully connected networks [6, 21, 24], or attention [23, 30]. Li et al. [36] proposed an auto-conditioned approach enabling synthesizing pose sequences up to 300 seconds of periodic-like motions (walking, dancing). However, these works all focus on frame-by-frame synthesis, with benchmark evaluation of up to 1000 milliseconds. Instead of a frame-by-frame synthesis, we propose a goal-directed task to capture perception of longer-term human action, which not only lends itself towards forecasting more semantically meaningful key moments, but enables a more predictable evaluation: as seen in Fig. 1, there can be significant ambiguity in the number of pose frames to predict towards a key or goal pose, making frame-based evaluation difficult in longer-term forecasting.

Multi-Modal Human Motion Forecasting. While 3d human motion forecasting has typically been addressed in a deterministic fashion, several recent works have introduced multi-modal future pose sequence predictions. These approaches leverage well-studied approaches for multi-modal predictions, such as generative adversarial networks [4] and variational autoencoders [2, 34, 35]. For instance, Aliakbarian et al. [2] stochastically combines random noise with previous pose observations, leading to more diverse sequence predictions. Yuan et al. [35] learns a set of mapping functions which are then used for sampling from a trained VAE, leading to increased diversity in the sequence predictions than simple random sampling. In contrast to these time-based approaches, we consider goal-oriented prediction of characteristic poses, and model multi-modality explicitly as predicted heatmaps for body joints in an autoregressive fashion to capture inter-joint dependencies.

Goal-oriented Forecasting. While a time-based, frame-by-frame prediction is the predominant approach towards future forecasting tasks, several works have proposed to tackle goal-oriented forecasting. Recently, Jayaraman et al. [20] proposed to predict “predictable” future video frames in a time-agnostic fashion, and represent the predictions as subgoals for a robotic tasks. Pertsch et al. [27] predict future keyframes representing a future video sequence of events. Cao et al. [7] plan human trajectories from an image and 2d pose history, first predicting 2d goal locations for a person to walk to in order to synthesize the path. Inspired by such goal-based abstractions, we aim to represent 3d human actions as its key, characteristic poses.

3. Method Overview

Given a sequence of N 3d pose observations $\mathbf{X}_{1:N} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$ of a person, our aim is to estimate a characteristic 3d pose of that person, characterizing the intent of the person's future action. We take J joint locations (represented as their 3d coordinates) for each pose of the input sequence, i.e. $\mathbf{x}_i \in \mathbb{R}^{J\times 3}$. From this input sequence, we predict a joint distribution of J probability heatmaps $\mathbf{H}_j$ and finally, sample K output pose hypotheses $\mathbf{Y}_{1:K}$, characterized by their J 3d joints: $\mathbf{y}_i \in \mathbb{R}^{J\times 3}$. By representing probability heatmaps for the joint predictions, we can capture multiple different modes in likely characteristic poses, enabling more diverse future pose prediction. We note that we are the first to propose using volumetric heatmaps for future human pose forecasting, to the best of our knowledge, while previous work used them for the more deterministic task of pose estimation from multiple images [18, 31].

From the input sequence, we develop a neural network architecture to predict a probability heatmap over a volumetric 3d grid for each joint, corresponding to likely future positions of that joint. This enables effective modeling of multi-modality, but remains tied to a discrete grid, so we

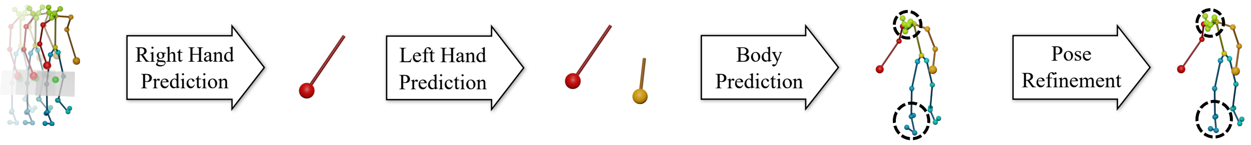


Figure 4. To model joint dependencies within the human skeleton, we sample joints in an autoregressive manner by first predicting the end-effectors (right and left hand), then the rest of the body; pose refinement then improves skeleton consistency.

also regress a corresponding volume of per-voxel offsets, allowing for precise locations to be sampled. Fig. 3 shows an overview of our learned probabilistic predictions.

We model these predictions conditionally in an autoregressive fashion in order to tractably model the joint distribution over all pose joint locations. This enables a consistent pose prediction over the set of pose joints, as a set of joints may have likely modes that are unlikely to be seen all together (e.g., right hand moving forward while the right elbow moves to the side – both are valid independently but not together). To sequentialize the pose joint prediction autoregressively, we first predict probability heatmaps for the end effectors in our dataset – right hand first, then left hand conditioned on the right hand prediction, followed by the rest of the body joints.

4. Capturing Multi-Modality with Heatmap Predictions

We aim to learn to predict likely future locations for an output pose joint j, characterized by a probability heatmap $\mathbf{H}_j$ over a volumetric grid of possible pose joint locations. From the input sequence of N pose observations of J joints, and conditioned on any already predicted joints, we construct an attention-driven neural network to learn the different dependencies between human skeleton joints to inform the final heatmap prediction.

Attention-Driven Sequence Encoding. We represent the body joints of the input sequence $\mathbf{X}_{1:N} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$ as an $N \times J \times 3$ ($N = 10$ as well as $J = 25$ for the GRAB dataset and $J = 17$ for Human 3.6M, respectively) concatenation of the joint locations over time. Features are first extracted with a single-layer GRU [10]. We then compute an attention map from these features, representing dependencies to the input set of pose joints. This way, the network learns not only how different joints in the skeleton affect each other directly (e.g., kinematic relationships) but also learns to exploit more subtle correlations such as likely positions of one hand with respect to the other. Following the formalism of Scaled Dot-Product Attention [32], popularized in natural language processing, our attention maps are computed from a query $\boldsymbol{Q}$ and a set of key-value pairs $\boldsymbol{K}$ and $\boldsymbol{V}$. During training, representations for $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are learned which are shared between all joints. This allows us to project all joints into the same embedding space where we can then compare the joint of interest (represented by $\boldsymbol{Q}$) with all other joints ($\boldsymbol{K}$) to inform which parts of $\boldsymbol{V}$ (the learned latent representation for all joints which will be passed to the decoder) are relevant for this joint of interest.

$$\mathrm{Attn}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{D}}\right)\boldsymbol{V} = \boldsymbol{A}\boldsymbol{V}, \quad (1)$$

Intuitively, the similarity between key and query defines which parts of a learned pose skeleton representation are important for the desired prediction. Formally, this is defined in Eq. 1: The value representation $\boldsymbol{V}$ is weighed per-element by the result of the dot-product between $\boldsymbol{Q}$ and $\boldsymbol{K}$ (scaled by the dimension of the embedding vector D and a softmax operation). In our case, the attention map $\boldsymbol{A}$ has a dimensionality of $J' \times N$ with J' indicating the number of joints to be predicted. Any prior joint predictions for autoregressive prediction are considered as an additional node to our attention map, giving the attention map dimension $J' \times (N + n_p)$ for n_p prior joints.

Heatmap Prediction. Based on the attention scoring, we then use a series of nine 3D convolutions to decode an output probability heatmap $\mathbf{H}_j$ for each body joint j. The grids are centered at the skeleton's hip joint; we use a grid size of 16^3 voxels, spanning $2m^3$. A value in the grid of $\mathbf{H}_j$ at location $\mathbf{H}_j(x, y, z)$ corresponds to a probability of joint j being at location (x, y, z) in the future characteristic pose. Instead of directly regressing the probability values, we predict $\mathbf{H}_j(x, y, z)$ as a classification problem by discretizing the output values into $n_{discr} = 10$ bins in the $[0, 1]$ space. We then use a cross entropy loss with the discretized target heatmap to train our heatmap predictions. In our experiments, we found that this classification formulation for $\mathbf{H}_j$ produced better results than an ℓ_2 or ℓ_1 regression loss, as it mitigated tending towards the average or median.

Offset Prediction. Since predicting joint locations in a discrete grid inherently leads to grid artifacts in sampled output poses, we additionally learn an offset field $\mathbf{O}_j$ over the same volumetric grid. Here, each voxel $\mathbf{O}_j(x, y, z) \in \mathbb{R}^3$ represents the shift to be added after sampling a joint from the heatmap at $\mathbf{H}_j(x, y, z)$. We predict these offsets similarly to the heatmap volume, with a series of nine 3D convolutions, and clamp each offset vector $\mathbf{O}_j(x, y, z)$ to move the joint at most one voxel length. Output poses are then estimated by sampling the heatmap, followed by refinement using the corresponding predicted offset.

4.1. Training Details

Note that for real-world data captured of human movement, we do not have a full ground truth probability distribution for the future characteristic pose, but rather a set of paired observations of input pose to the target pose. Thus, we generate target heatmap data from a single future observation in the training data by applying a Gaussian kernel (size 5, $\sigma = 2$) over the target joint location. At test time, we apply softmax scaling to the predicted heatmaps with a temperature of 0.025 and from there, sample our final joint locations. We learn multi-modality by generalizing across train set observations which results in seeing multiple possibilities for similar inputs (e.g., right vs. forward pass), encouraging learned heatmaps to represent multiple modes. We show that our formulation can effectively model multi-modal heatmaps in Section 7.

We train our models on a single NVIDIA GeForce RTX 2080Ti. We use an ADAM optimizer with a weight decay of 0.001 and a linear warmup schedule for 1000 steps; learning rate is then kept at 0.001. We use a batch size of 100, as a larger batch size helps with training our attention mechanism. Our model trains for up to 8 hours until convergence. During training, we apply teacher forcing, i.e. pose joint predictions conditioned on prior joint predictions are trained using the ground truth locations of the prior joints. For a detailed specification of our network architecture, please refer to the appendix.

5. Autoregressive Joint Prediction

Given a set of heatmaps for each pose joint location, the next step is to predict specific joint locations. Since they are not independent of one another, we cannot simply sample joint locations from each heatmap independently. Instead, we must model the interdependencies between pose joints.

To do this, we model the joint distribution of pose joints autoregressively, as visualized in Fig. 4: we first predict end effector joints, followed by other body joints. For our experiments, we find that the right and left hands tend to have a large variability, so we first predict the right hand, then the left hand conditioned on the right hand location, followed by the rest of the body joints. Empirically, we found that the hands tended to define the body pose, while the order of the rest has little impact. To sample from a joint heatmap, we use temperature scaling to concentrate the heatmap near its local maxima, followed by random sampling.

Pose Refinement. While our autoregressive pose joint prediction encourages a coherent pose prediction with respect to coarse global structure, pose joints may still be slightly offset from natural skeleton structures. Thus, we employ a pose refinement optimization to encourage the predicted pose to follow inherent skeleton bone length and angle constraints while keeping all joints in areas of high probability and the end-effectors close to their original prediction, as formulated in the objective function:

$$\begin{aligned}\mathrm{E}_R(\mathbf{x}, \mathbf{e}, \mathbf{b}, \mathbf{x_0}, \theta, H) = \\ w_e\|\mathbf{x}_e - \mathbf{e}\|_2 + w_b\|\text{bonelengths}(x) - \mathbf{b}\|_1 \\ + w_a\|\text{angles}(x) - \theta\|_1 + w_c\|x - x_0\|_1 \\ + w_h\Sigma_j(1 - H_j)\end{aligned} \tag{2}$$

where $\mathbf{x}$ the raw predicted pose skeleton as a vector of N 3D joint locations; $\mathbf{b}$ and θ the bone lengths and joint angles, respectively, of the initially observed pose skeleton; x_0 the joint locations of the last skeleton in the input sequence; H_j the heatmap probability for each joint; $\mathbf{e}$ the sampled end effector locations; and w_e, w_b, w_a, w_h, w_c weighting parameters (in all our experiments, we use $w_e = 0.2, w_b = 1.0, w_a = 0.4, w_h = 0.1, w_c = 0.1$). We then optimize for $\mathbf{x}$ under this objective to obtain our final pose prediction.

6. Characteristic 3D Pose Dataset

To train and evaluate the task of characteristic 3d pose forecasting, we introduce a dataset of annotated characteristic poses, built on GRAB [29] and Human3.6M [17].

- **Human3.6M** is a commonly used dataset for human pose forecasting, comprising 210 actions performed by 11 professional actors in 17 scenarios for a total of 3.6 million frames. 3d locations are obtained for 32 joints via a high-speed motion capture system; we use a reduced 17-joint layout in our method, removing redundant and unused joints, following [35].

- **GRAB** is a recent dataset with over 1 million frames in 1334 sequences of 10 different actors performing a total of 29 actions with various objects. Each actor

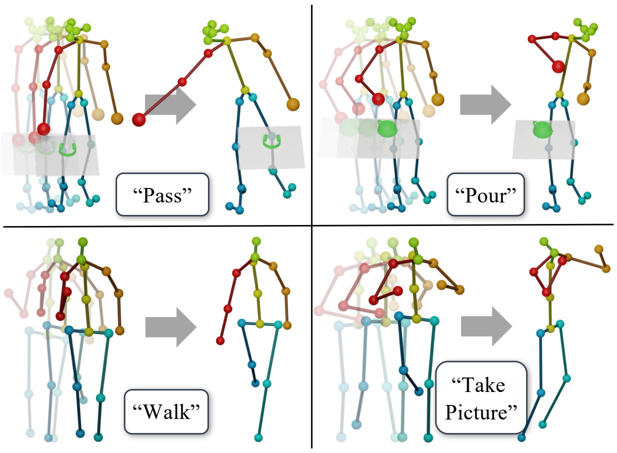


Figure 5. Example input observations and target characteristic 3d poses from our annotated datasets, based on GRAB (top) and Human3.6M (bottom).

starts in a T-Pose, moves towards a table with an object, performs an action with the object, and then steps back to the T-Pose. The human motions are captured using modern motion capture techniques, with an accuracy in the range of a few millimeters. GRAB provides SMPL-X [26] parameters from which we extract the 25 most defining body joints. For more details, we refer to the appendix.

We then annotate the timesteps of the captured sequences corresponding to characteristic poses. Input sequence start frames are randomly sampled, up until the characteristic pose frame. Several example input sequence-characteristic pose pairs are visualized in Fig. 5. Annotations were performed by the authors, within a time span of one day. This is the total time for annotating more that 1000 sequences across two datasets, with each annotation taking 10-30 seconds; this annotation efficiency enables quick and easy adoption of new datasets in the future. We define a characteristic pose as the point in time when the action is most articulated, i.e. right before the actor starts returning back to another pose (e.g., when the hand is furthest from the person when passing, most tilted when pouring, etc.). For sequences containing multiple occurrences of the same action, like lifting, we chose the repetition with most articulation, e.g. when the object is lifted highest. In the case of Human3.6M, where there are sometimes multiple possible options for characteristic poses, we pick the first one that is representative of the action, e.g., the first sitting pose.

Characteristic 3D Pose Prediction. For the task of characteristic 3d pose prediction, we consider an input sequence of $N = 10$ 3d pose observations of a person, represented as $J = 25$ 3d joint locations for the GRAB dataset and $J = 17$ for the Human3.6M dataset (in their native joint layouts; for more details we refer to the appendix). From this observation, the next characteristic pose is predicted as J 3d joint locations. All poses are considered in their hip-centered coordinate systems. Note that while we have action labels in the annotated dataset, we do not use them for this task.

The N input pose observations can occur at any time, so methods are trained with random input sequences up to the characteristic 3d pose. At test time, five input points are evaluated for each method, with the five input points selected to evenly distribute between the beginning of the sequence to N frames before the characteristic pose.

Evaluation. We use a train/val/test split by actor in each dataset. For GRAB we have 8/1/1 train/val/test actors, resulting in 992/197/136 train/val/test sequences. For Human3.6M, we follow the split of [23]: 5/1/1 and 150/30/30 train/val/test actors and sequences, respectively.

To evaluate our task of characteristic 3d pose prediction, we aim to consider the multi-modal nature of the task. Since we do not have ground truth probability distributions available, and only a single observed characteristic pose for each input pose observation, we follow previous work on multi-modal human pose sequence predictions [2,4,34,35]: At test time, we consider $k = 10$ hypotheses from each method. To characterize these hypotheses holistically, we consider several metrics to assess accuracy, diversity, and quality of predictions.

Accuracy. First, we evaluate the sampling error using the mean per-joint position error (MPJPE) [17] by comparing the most similar prediction p' to the ground-truth pose p:

$$\mathrm{E}_{\text{MPJPE}} = \frac{1}{N}\sum_{j=1}^{N} ||p'_j - p_j||_2^2 \tag{3}$$

This evaluates whether the predicted hypotheses capture the target well and allows for comparison with deterministic baselines (where all hypotheses are identical).

Diversity. We evaluate the diversity as the MPJPE between all sampled poses for the same sequence. This evaluates the multi-modality of predicted distributions.

Quality. Finally, we evaluate quality of our multi-modal predictions with the Inception Score [28] (IS) over the set of predicted hypotheses for all test sequences. The Inception Score is widely used to measure the quality generative model outputs. More specifically, we use the conditional formulation first introduced in [16]. Similar to [2], we adapt this idea to our use case by training a simple skeleton-based action classifier on ground-truth samples from our datasets. Overall, this metric estimates how well the predictions capture an action while still producing diverse poses.

7. Experimental Evaluation

We evaluate the task of characteristic 3d pose prediction, using our annotated dataset built from the real-world GRAB [29] and Human3.6M [17] datasets.

Comparison to time-based state-of-the-art forecasting. In Tab. 1, we compare to state-of-the-art multi-modal sequence forecasting approach DLow [35], which is based on a conditional VAE, as well as to recent deterministic approaches for frame-based future human motion prediction, Learning Trajectory Dependencies [24] and History Repeats Itself [23], which use a graph neural network and an attention-based model, respectively, to predict human pose sequences. We train all of these sequential approaches on our datasets, given the input sequence of N frames, to predict an output N_o-frame pose sequence, with $N_o = 100$ frames to ensure that the characteristic pose falls within each target sequence. Since these sequence-based approaches each predict output sequences, we additionally allow them to predict the time step of the characteristic pose with an MLP to obtain the final characteristic pose prediction (see the appendix for additional detail).

	Method	GRAB			Human3.6m		
		MPJPE ↓	Diversity ↑	IS ↑	MPJPE ↓	Diversity ↑	IS ↑
Statistical	Random Sampling	1.018	-	-	1.159	-	-
	Average Train Pose	0.146	-	-	0.179	-	-
	Zero Velocity	0.063	-	-	0.166	-	-
Algorithmic	Learning Trajectory Dependencies [24]	0.077	-	-	0.165	-	-
	History Repeats Itself [23]	0.071	-	-	0.116	-	-
	DLow [35]	0.071	0.089	1.257 ±0.02	0.119	0.104	1.623 ±0.08
	Ours	**0.054**	**0.105**	**4.153** ±0.87	**0.092**	**0.189**	**3.139** ±0.32

Table 1. Characteristic 3d pose performance, in comparison with state of the art and statistical baselines. We evaluate MPJPE for all methods and additionally, the diversity of multi-modal methods in terms of MPJPE between samples as well as their quality with the Inception Score, similar to [2].

Since we aim to predict a characteristic 3d pose given an arbitrary sequence observation, we sample different start points for the input sequence, and analyze performance across varying distance from the goal pose.

We report the MPJPE, Diversity, and IS metrics in Tab. 1; we first measure the performance for each of the five input sequence start times mentioned above and average over those for the final result. Our approach more accurately characterizes the future characteristic poses while also producing improved diversity and quality. For comparison, we also report baseline performance when given an oracle providing the ground-truth characteristic time step in Tab. 2. Even with this additional information, our characteristic pose formulation achieves improved results. Qualitative results are shown in Fig. 6; our probabilistic approach more effectively captures a realistic set of characteristic modes.

In Fig. 7, we visualize the diversity of our predictions in comparison with multi-modal baselines. Our predicted pose hypotheses show more diversity in both joint placement and action representation, while still capturing the target pose.

Comparison to statistical baselines. We also compare with three statistical baselines: full random sampling from an evenly distributed heatmap, the average target train pose over the entire dataset, and a zero-velocity baseline (i.e., the error of simply using the last input pose as prediction), which was shown by Martinez et al. [25] to be competitive with and sometimes outperform state of the art. Our approach outperforms these statistical baselines, indicating learning of strong characteristic pose patterns.

Method	GRAB		Human3.6m	
	MPJPE ↓	IS ↑	MPJPE ↓	IS ↑
L. T. D. [24]	0.075	-	0.156	-
H. R. I. [23]	0.066	-	0.116	-
DLow [35]	0.059	1.567 ±0.02	0.108	1.418 ±0.14
Ours	**0.054**	**4.153** ±0.87	**0.092**	**3.139** ±0.32

Table 2. Characteristic 3d pose performance comparison. In contrast to Tab 1, baselines are provided with ground-truth characteristic time step information.

8. Ablation Studies

Does a probabilistic prediction help? In addition to comparing to state-of-the-art alternative approaches which make deterministic predictions, we compare in Tab. 3 with our model backbone with a deterministic output head (an MLP) replacing the volumetric heatmap decoder which re-

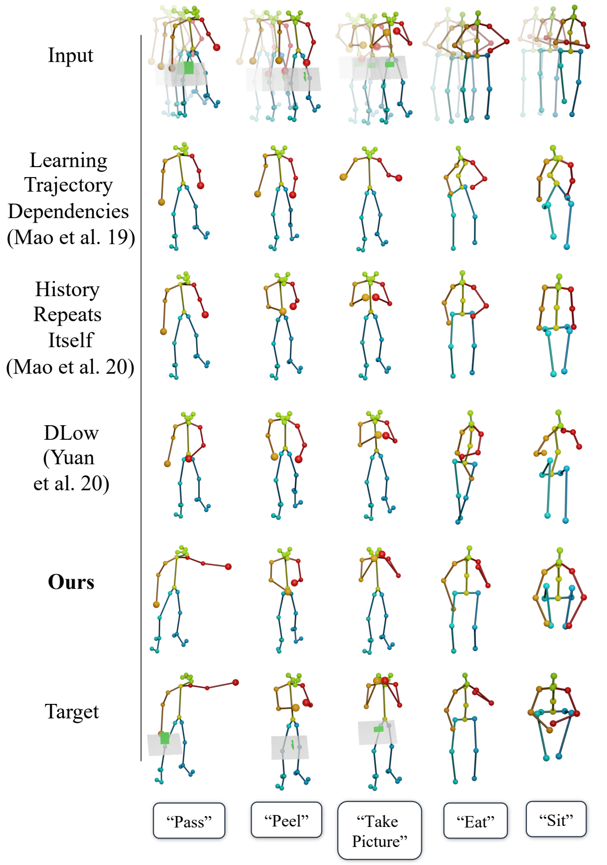

Input
Learning Trajectory Dependencies (Mao et al. 19)
History Repeats Itself (Mao et al. 20)
DLow (Yuan et al. 20)
Ours
Target
"Pass"
"Peel"
"Take Picture"
"Eat"
"Sit"

Figure 6. Qualitative results on characteristic 3d pose prediction. In comparison to deterministic [23, 24] (rows 2 and 3) and probabilistic [35] (row 4) approaches, our method more effectively predicts likely intended action poses. Note that action labels are only shown for visualization purposes.

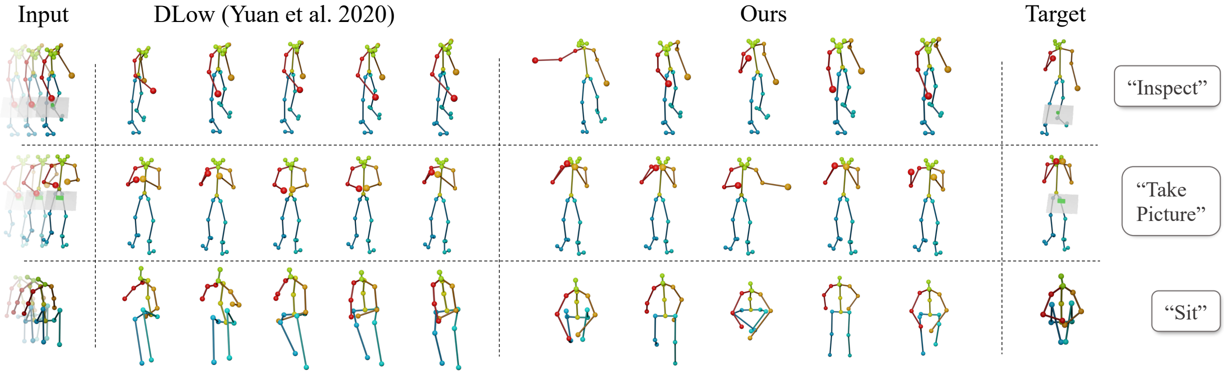


Figure 7. Qualitative results on characteristic 3d pose prediction, showing the diversity of our predictions in comparison with DLow [35].

gresses offset positions for each pose joint relative to the input positions. Removing our heatmap predictions similarly fails to effectively capture the characteristic modes; our probabilistic, heatmap-based predictions notably improve performance.

Does per-voxel offset prediction help? We analyze the effect of per-voxel offset prediction in Tab. 3, showing that they notably improve pose predictions. Applying pose refinement without offset prediction fails to achieve the same level of improvement.

Does autoregressive pose joint sampling help? We analyze the effect of our autoregressive pose joint sampling in Tab. 3. We compare against a version of our model trained to predict each pose joint heatmap independently, with pose joints sampled independently, which often results in valid individual pose joint predictions that are globally inconsistent with the other pose joints. In contrast, our autoregressive sampling helps to generate a likely, consistent pose.

How diverse are the sampled poses? We show qualitative examples of our multi-modal predictions in Fig. 7, outlining the diversity of both heatmap predictions and sampled skeletons. We also evaluate our prediction diversity as MPJPE between our sampled outputs as part of Tab. 1.

		GRAB		Human3.6m	
	Ablation	MPJPE ↓	IS ↑	MPJPE ↓	IS ↑
Loss	ℓ_1 loss	0.132	1.132 ±0.01	0.198	2.246 ±0.24
	ℓ_2 loss	0.130	1.146 ±0.01	0.206	1.976 ±0.08
Model	Deterministic	0.064	-	0.108	-
	Not autoreg.	0.077	1.583 ±0.15	0.109	1.929 ±0.09
Sampling	No offsets	0.132	1.328 ±0.02	0.172	2.537 ±0.07
	$\hookrightarrow$ refined	0.127	1.509 ±0.03	0.163	2.978 ±0.14
	$k = 50$	0.049	1.222 ±0.02	0.082	1.845 ±0.19
	Not refined	0.057	3.989 ±0.95	0.098	2.418 ±0.11
	Ours	**0.054**	**4.153** ±0.87	**0.092**	**3.139** ±0.32

Table 3. Ablation study over varying heatmap losses, deterministic and non-autoregressive pose sampling, no offset prediction (with and without pose refinement), number of samples taken for the evaluation, and without pose refinement.

What is the effect of the number of pose samples? If we take more pose samples from our predicted joint distribution (from 10 to 50), we can, as expected, better predict the potential target characteristic pose, as seen in Tab. 1.

Do different heatmap losses matter? We evaluate our formulation for heatmap prediction as a discretized heatmap with a cross entropy loss against regressing heatmaps with an ℓ_1 or ℓ_2 loss, and find that our discretized formulation much more effectively models the relevant modes.

Limitations. Several limitations remain for our approach of characteristic 3d action pose forecasting. For instance, while our offset predictions help alleviate the ties to a volumetric heatmap grid, more precise modeling of smaller-scale behavior (e.g., detailed hand movement) would require more efficient representations such as sparse grids. In addition, our method relies on manually annotated characteristic 3d poses for supervision; while characteristic pose annotation is very efficient for new datasets, self-supervised formulations would also be an interesting future direction.

9. Conclusion

In this paper, we introduced a new task: predicting future *characteristic 3d poses* of human activities from short sequences of pose observations. We introduce a probabilistic approach to capturing the most likely modes in these characteristic poses, coupled with an autoregressive formulation for pose joint prediction to sample consistent 3d poses from a predicted joint distribution. We trained and evaluated our approach on a new annotated dataset for characteristic 3d pose prediction, outperforming deterministic and multi-modal state-of-the-art approaches. We believe that this opens up many possibilities towards goal-oriented 3d human pose forecasting and understanding anticipation of human movements.

Acknowledgements

This project is funded by the Bavarian State Ministry of Science and the Arts and coordinated by the Bavarian Research Institute for Digital Transformation (bidt).

References

[1] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7143–7152. IEEE, 2019. 2

[2] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5223–5232, 2020. 3, 6, 7, 15

[3] Moshe Bar. The proactive brain: memory for predictions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1235–1243, 2009. 1

[4] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1418–1427, 2018. 3, 6

[5] Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. ContactDB: Analyzing and predicting grasp contact via thermal imaging. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 13

[6] Judith Bütepage, Michael J. Black, Danica Kragic, and Hedvig Kjellström. Deep representation learning for human motion prediction and classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1591–1599. IEEE Computer Society, 2017. 3

[7] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 387–404. Springer, 2020. 3

[8] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 13, 14

[9] Hsu-Kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*, pages 1423–1432. IEEE, 2019. 2

[10] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 4

[11] Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204, 2013. 1

[12] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4346–4354. IEEE Computer Society, 2015. 2

[13] Anand Gopalakrishnan, Ankur Mali, Dan Kifer, C. Lee Giles, and Alexander G. Ororbia II. A neural temporal model for human motion prediction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12116–12125. Computer Vision Foundation / IEEE, 2019. 2

[14] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José M. F. Moura. Adversarial geometry-aware human motion prediction. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, volume 11208 of *Lecture Notes in Computer Science*, pages 823–842. Springer, 2018. 2

[15] Jakob Hohwy. *The predictive mind*. Oxford University Press, 2013. 1

[16] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018. 6

[17] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2014. 2, 5, 6, 13

[18] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7718–7727, 2019. 3

[19] Ashesh Jain, Amir Roshan Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5308–5317. IEEE Computer Society, 2016. 2

[20] Dinesh Jayaraman, Frederik Ebert, Alexei A. Efros, and Sergey Levine. Time-agnostic prediction: Predicting predictable video frames. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 2, 3

[21] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5226–5234. IEEE Computer Society, 2018. 3

[22] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 211–220. IEEE, 2020. 2

[23] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and

Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIV*, volume 12359 of *Lecture Notes in Computer Science*, pages 474–489. Springer, 2020. 2, 3, 6, 7, 15

[24] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9488–9496. IEEE, 2019. 2, 3, 6, 7, 15

[25] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4674–4683. IEEE Computer Society, 2017. 2, 7

[26] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10975–10985. Computer Vision Foundation / IEEE, 2019. 6, 14

[27] Karl Pertsch, Oleh Rybkin, Jingyun Yang, Shenghao Zhou, Konstantinos Derpanis, Kostas Daniilidis, Joseph Lim, and Andrew Jaegle. Keyframing the future: Keyframe discovery for visual prediction and planning. In *Learning for Dynamics and Control*, pages 969–979. PMLR, 2020. 2, 3

[28] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242, 2016. 6

[29] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, volume 12349 of *Lecture Notes in Computer Science*, pages 581–600. Springer, 2020. 2, 5, 6, 13

[30] Yongyi Tang, Lin Ma, Wei Liu, and Wei-Shi Zheng. Long-term human motion prediction by modeling motion context and enhancing motion dynamics. In Jérôme Lang, editor, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 935–941. ijcai.org, 2018. 2, 3

[31] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *European Conference on Computer Vision*, pages 197–212. Springer, 2020. 3

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 4

[33] Borui Wang, Ehsan Adeli, Hsu-Kuang Chiu, De-An Huang, and Juan Carlos Niebles. Imitation learning for human pose prediction. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7123–7132. IEEE, 2019. 2

[34] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 265–281, 2018. 3, 6

[35] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*, pages 346–364. Springer, 2020. 3, 5, 6, 7, 8, 15

[36] Yi Zhou, Zimo Li, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 3

Appendix

In this appendix, we show additional qualitative results (Sec. A), additional quantitative analysis (Sec. B), detail our network architecture specification (Sec. C), provide additional details regarding the dataset (Sec. D) as well as our training setup (Sec. E), and discuss potential negative societal impacts of our method (Sec. F).

A. Additional Qualitative Results.

We show additional qualitative results of our method in Fig. 9, which demonstrate the diversity of our characteristic pose predictions for a given input sequence. Our approach not only effectively models the multi-modal nature of characteristic poses, but also captures the final target action pose (highlighted pose prediction).

In cases where the time between input sequence and target pose is longer, such as in ‘sit’ or ‘greet’, our approach produces a more diverse set of action poses, capturing the ambiguity in the future characteristic pose. When the input sequence is close to the target pose, our approach converges to a small set of probable poses (for example, in ‘drink’), reflecting the reduced ambiguity.

B. Additional Quantitative Results.

MPJPE baseline comparison, by goal-normalized input time Fig. 8 shows MPJPE for varying input sequence start times in comparison with state of the art, goal-normalized from the start of each sequence (0) to N frames before the characteristic pose (1), with three steps inbetween.

Autoregressive Joint Order. We determined the order of the joints for the autoregressive prediction empirically; most ambiguity occurred in active end-effectors (i.e. right and left hands), whereas the rest of the body tended to have lower variability. In Tab. 4, we compare our original approach of (right hand, left hand, rest) with two alternatives: (left hand, right hand, rest), and (full autoregressive from human kinematic chain following left/right hands). Our method is robust to these orderings (though diversity of the rest of the body except hands decreases with autoregression through the kinematic chain).

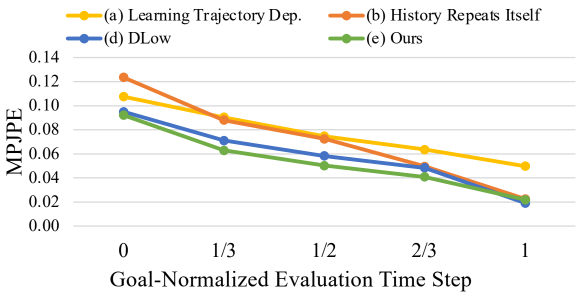


Figure 8. MPJPE comparison to baselines, evaluating with the input sequence at different points in time: from the start of the sequence (0) to N frames before the target characteristic pose (1).

Order	MPJPE ↓	Div. ↑	IS ↑
right hand → left hand → rest	**0.054**	**0.105**	**4.15** ±0.9
left hand → right hand → rest	0.057	0.049	4.09 ±1.6
following the kinematic chain	0.058	0.018	4.02 ±0.9

Table 4. Ablation analysis on autoregressive order on GRAB data.

Grid Resolution and Offset Prediction. We show additional ablations on the effect of grid resolution and offset prediction in Tab 5 on GRAB data; A resolution of 16^3 performs better than 8^3 or 32^3. Our offset prediction helps mitigate grid artifacts even at 32^3.

Resolution	Offsets	MPJPE ↓	Diversity ↑	IS ↑
8^3	×	0.242	**0.189**	1.40 ±0.3
8^3	✓	0.092	0.068	1.71 ±0.1
16^3	×	0.127	0.081	1.51 ±0.1
$\mathbf{16^3}$	✓	**0.054**	0.105	**4.15** ±0.9
32^3	×	0.118	0.122	2.39 ±0.2
32^3	✓	0.066	0.058	1.91 ±0.2

Table 5. Ablation analysis on heatmap grid size and offset prediction on GRAB data.

Per-Bodypart MPJPE. In Tab. 9, we show our final pose prediction performance in MPJPE, broken down per bodypart, as compared to sequential baselines.

Characteristic Pose Forecasting with Ground Truth Action Labels. In Tab. 6, we additionally evaluate our approach using ground truth action labels as input to provide additional contextual information.

The ground truth action label is processed as an additional attention node alongside input and previously predicted joint locations. This action label information reduces ambiguity in the possible set of output poses, resulting in reduced diversity, as is reflected in the diversity metric and inception score (as this directly considers diversity).

In our original action-agnostic scenario, our approach predicts plausible and diverse characteristic poses across all actions.

	GRAB			Human3.6M		
	MPJPE ↓	Div. ↑	IS ↑	MPJPE ↓	Div. ↑	IS ↑
×	0.054	0.105	4.153 ±0.87	0.092	0.189	3.139 ±0.32
✓	0.051	0.026	1.085 ±0.02	0.094	0.044	1.700 ±0.06

Table 6. Comparison of ours to an ablation with ground truth action labels as additional input.

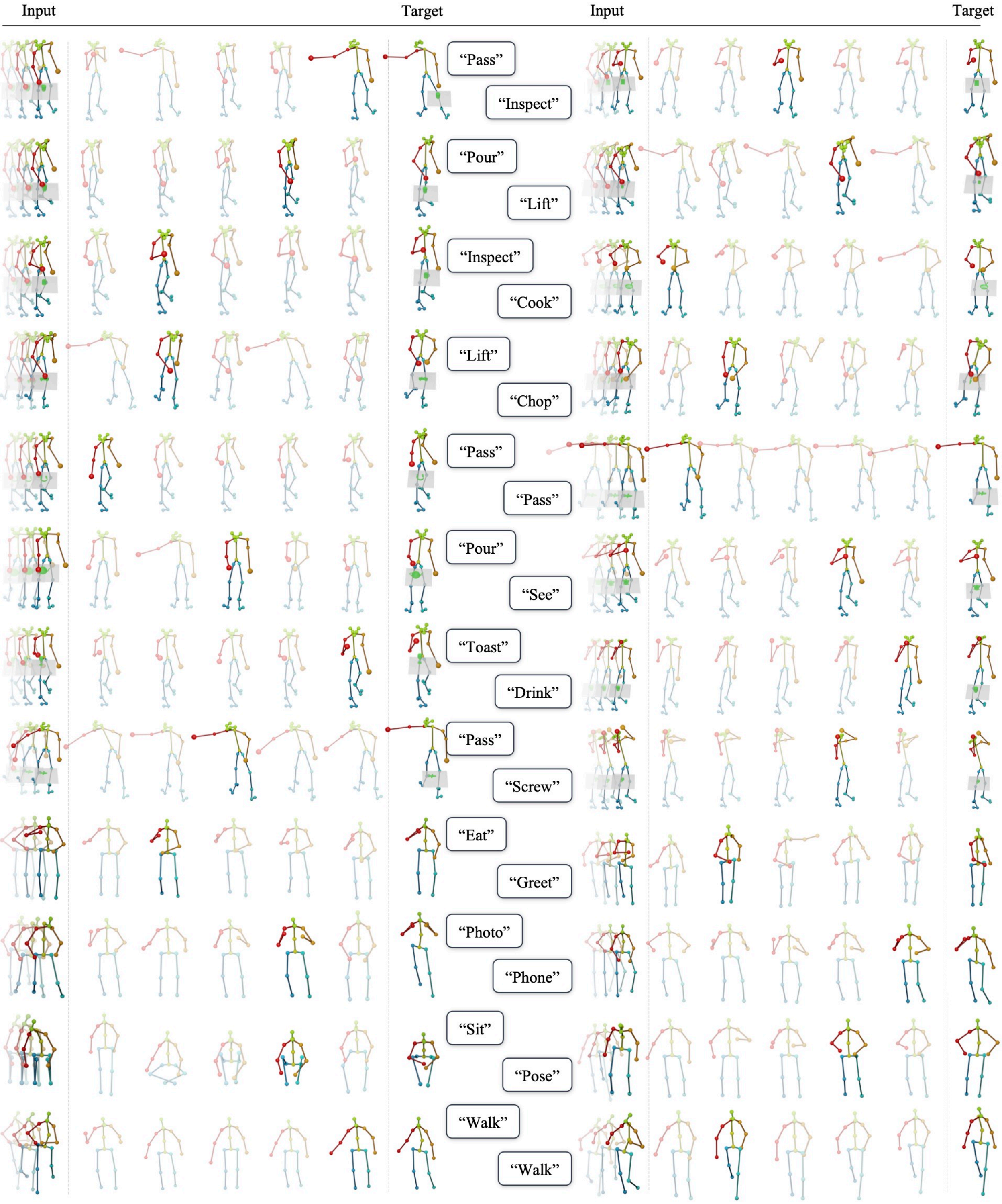


Figure 9. Additional qualitative results, showing the for each action sequence the inputs (left), our diverse set of predictions (middle) and the target action pose (right). Our final pose prediction is highlighted for each action sequence.

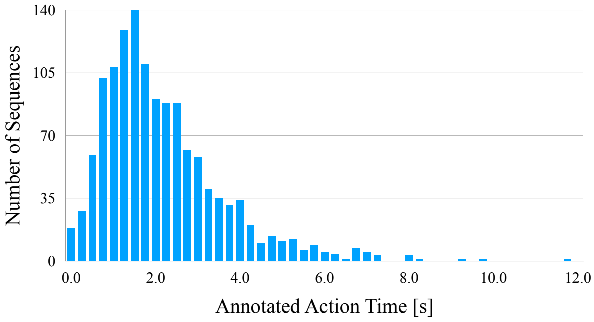


Figure 10. Times at which characteristic poses occur for GRAB.

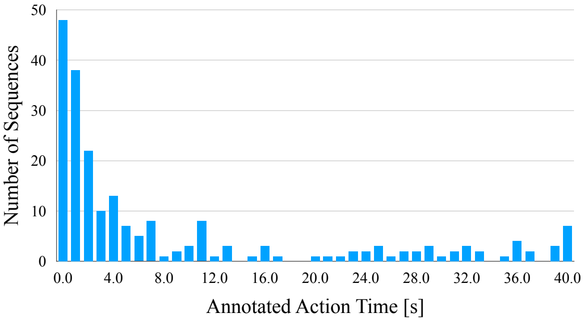


Figure 11. Times at which char. poses occur for Human3.6M.

C. Architecture Details

Fig. 13 details our network specification from input (left) to heatmap and offsets output (right). For each GRU layer, we provide the hidden dimension and number of layers in parentheses, for normalization layers the dimension to be normalized over, for dropout layers the dropout probability p, and for convolutions the number of input and output channels as well as kernel size (ks), stride (str), and padding (pad). We apply cross-entropy (CE) losses at a heatmap resolution of 8^3 and at the final resolution of 16^3; for the offsets prediction, we concatenate the offsets volume generated from the last input skeleton after 5 convolution blocks and supervise the final predictions with an ℓ_1 loss.

We take as input 25 joints in the case of GRAB and 17 joints for Human3.6M (#in_joints). The number of output joints (#out_joints) depends on whether the right or left hand is being predicted (#out_joints=1) or the rest of the body (#out_joints=23 for GRAB, #out_joints=15 for Human3.6M). In all our experiments, we use 10 as the number of probability bins.

D. Dataset

GRAB Pose Layout. Since GRAB [29] not only provides a human skeleton representation but full body shape parameters, we preprocess all pose sequences by first extracting relevant joints for our approach. For this, we chose the 3d OpenPose [8] layout as it describes the prevalent body joints and is widely used for representing 3d poses. Note that we do not apply the OpenPose method on 2d data; we only use their joint definitions in 3d. We extract 25 body joints from the SMPL-X body given by the GRAB dataset [29] using the correspondences shown in Tab. 8. Additionally, we denote in Tab. 8 the correspondences of joints to body parts, for the body part analysis in Tab. 9. Fig. 12 (left) visualizes our joint selection, overlaying the body shape given in GRAB as a point cloud over the 25-joint skeleton.

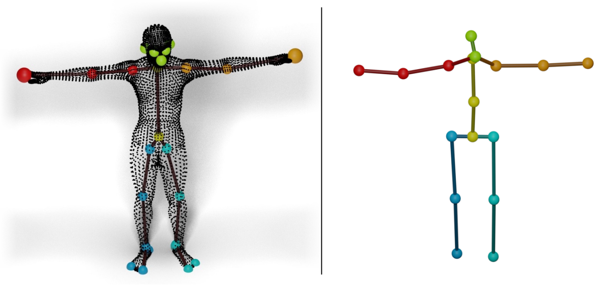


Figure 12. GRAB [29] body and our extracted skeleton joints overlaid (left); 17-joint skeleton based on Human3.6M [17] (right).

Human3.6M Pose Layout. For all our experiments on Human3.6M [17], we use 17 pose joints, visualized in Fig. 12 (right). Tab. 7 describes the exact joints used as well as the correspondences of joints to body parts, as used in Tab. 9.

Visualization Details. While our approach is agnostic to context or action, we visualize the context provided by GRAB [5, 29] (of the table and object) and action label provided by both GRAB and Human3.6M to help contextualize the pose visualizations. The context and action labels are not taken into account by the network or the evaluation, meaning that our approach infers plausible human action poses while being agnostic towards action and context.

Additional Characteristic 3D Pose Details. We show additional characteristic 3d poses in their original sequences in Fig. 14, and note the strong time differences at which the characteristic poses occur. Furthermore, Fig. 10 and Fig. 11 show the times during the sequences at which the characteristic 3d poses are annotated for GRAB and Human3.6M; these characteristic poses are distributed across a wide range (0-12 seconds and 0-40 seconds, respectively) of time.

E. Additional Training Details

Cross Entropy Loss. Since our approach learns to predict the probabilities of a Gaussian-smoothed target point during training, we observe a very large class imbalance between the no-probability bin (bin 0) and the rest of the bins. We thus weigh the classes in the cross entropy loss to account for the class imbalances, by the inverse of their log-scaled occurrence, and a weight of 0.1 for the no-probability bin.

	Ours (17-Joint)		Base (Human3.6M)	
	Idx	Label	Label	Idx
R. Leg	1	R. Hip	R. Hip	1
	2	R. Knee	R. Knee	2
	3	R. Foot	R. Heel	3
L. Leg	4	L. Hip	L. Hip	6
	5	L. Knee	L. Knee	7
	6	L. Foot	L. Heel	8
R. Arm	14	R. Shoulder	R. Shoulder	25
	15	R. Elbow	R. Elbow	26
	16	R. Hand	R. Hand	27
L. Arm	11	L. Shoulder	L. Shoulder	17
	12	L. Elbow	L. Elbow	18
	13	L. Hand	L. Hand	19
Spine	7	Spine	Spine	12
	0	Hip	Hip	0
Head	9	Nose	Nose	14
	10	Head	Head	15
	8	Thorax	Thorax	13

Table 7. Joint Correspondences for Human3.6M

	Ours (OpenPose [8])		Base (SMPL-X [26])	
	Idx	Label	Label	Idx
R. Arm	2	Right Shoulder	Right Shoulder	17
	3	Right Elbow	Right Elbow	19
	4	Right Finger	Right Index 3	42
L. Arm	5	Left Shoulder	Left Shoulder	16
	6	Left Elbow	Left Elbow	18
	7	Left Finger	Left Index 3	27
Right Leg	9	Right Hip	Right Hip	2
	10	Right Knee	Right Knee	5
	11	Right Ankle	Right Ankle	8
	22	Right Big Toe	Right Big Toe	63
	23	Right Small Toe	Right Small Toe	64
	24	Right Heel	Right Heel	65
Left Leg	12	Left Hip	Left Hip	1
	13	Left Knee	Left Knee	4
	14	Left Ankle	Left Ankle	7
	19	Left Big Toe	Left Big Toe	60
	20	Left Small Toe	Left Small Toe	61
	21	Left Heel	Left Heel	62
Head	0	Nose	Nose	55
	1	Neck	Neck	12
	15	Right Eye	Right Eye	24
	16	Left Eye	Left Eye	23
	17	Right Ear	Right Ear	58
	18	Left Ear	Left Ear	59
	8	Mid-Hip	Pelvis	0

Table 8. Joint Correspondences for GRAB

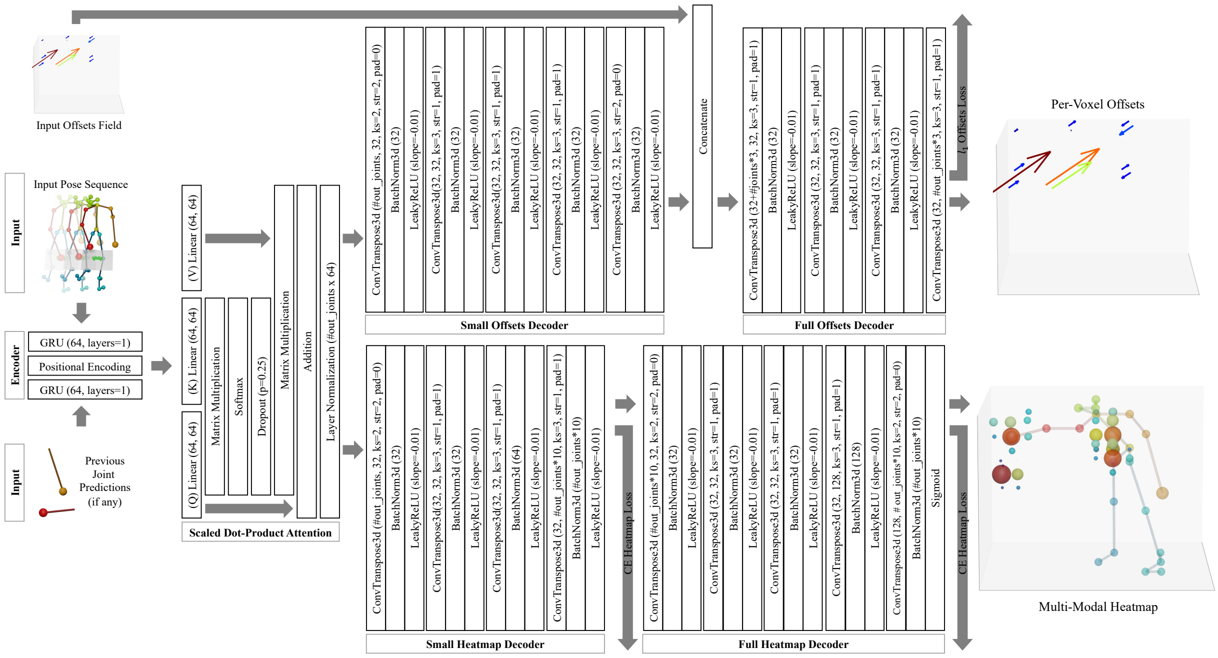


Figure 13. Our network architecture with details for encoder, scaled dot-product attention, as well as heatmap and offsets decoders.

	GRAB						H3.6M					
Method	R. Arm ↓	L. Arm ↓	R. Leg ↓	L. Leg ↓	Spine ↓	Head ↓	R. Arm ↓	L. Arm ↓	R. Leg ↓	L. Leg ↓	Spine ↓	Head ↓
L. T. D. [24]	0.165	0.115	0.058	0.057	0.028	0.085	0.225	0.225	0.135	0.146	0.108	0.123
H. R. I. [23]	0.160	0.113	0.056	0.055	0.026	0.079	0.199	0.191	**0.079**	0.088	0.040	0.089
DLow [35]	0.146	0.109	0.052	0.050	0.024	0.068	0.174	0.169	0.108	0.112	0.044	0.096
Ours	**0.105**	**0.084**	**0.045**	**0.045**	**0.020**	**0.057**	**0.147**	**0.122**	0.091	**0.085**	**0.033**	**0.066**

Table 9. Characteristic 3d pose prediction performance comparison to baselines, broken down by body part MPJPE.

State-of-the-art comparisons. We use the official code with default settings of the methods we compare to ([24], [23], and [35]). We train all methods from scratch on our characteristic 3d pose dataset, setting the number of input frames to 10 and the number of output frames to 100. From the predicted sequence, we evaluate the pose at a timestep predicted by the baselines themselves as characteristic pose and compare it to the target. This scenario is the closest to our approach, as predicting characteristic 3d poses involves which pose is the characteristic pose.

Therefore, we modified each baseline with a small prediction head to predict the characteristic pose frame within all 100 frames of the predicted sequence. In all cases, we supervise this prediction as a classification problem with a cross entropy loss and train the additional head together with the rest of the model.

For DLow [35], we add one linear layer to the final feature output of each of the 100 steps, followed by a ReLU, reducing each step's output dimension to 10. Then, one additional linear layer summarizes the combined output of all steps ($100 * 10$) down to a vector of size 100.

In the case of History Repeats Itself [23], we add a classification head consisting of one linear layer, a 1d batch norm, a ReLU, and one additional linear layer to the output of their last Graph Convolution Block (GCN). While the first linear layer keeps the original dimensionality of 100, the second linear layer reduces the dimension from $\#graph\_nodes * 100$ down to 100.

Finally, for Learning Trajectory Dependencies [24], we apply the same architecture and add a linear layer, a 1d batch norm, a ReLU, and a second linear layer after the final GCN. Here, we first reduce the per-node feature dimension from 256 to 100 and combine the features of all nodes with the second linear layer, going from $\#graph\_nodes * 100$ down to 100.

In the main paper, we additionally evaluated against these baseline approaches when given ground-truth time steps instead; in this scenario, our predictions also outperform the baselines given ground truth times for characteristic poses.

To evaluate the diversity and quality of multi-modal outputs, 10 samples are taken from a probabilistic method for each input sequence, and we report diversity in terms of MPJPE between samples as well as the Inception Score, following [2].

F. Potential Negative Societal Impacts

As we aim to study human pose behavior, we must take care to ensure that datasets used represent notable diversity in those represented. Our approach currently operates on skeleton abstractions that do not characterize finer-scale appearance differences; in possible future studies that may aim to characterize fine-scale interactions, diversity in body shape representations which must be taken into account for data collection and analysis.

In particular, in our scenario of forecasting probable future human behavior, we must also ensure that this possibility cannot be easily used for generating fraudulent motion video of a person. Such usage is currently severely limited in our proposed approach, as it does not target individual people, and does not model photo-realistic characteristics of people.

Another concern might arise with the possibility of surveillance, in the context of predicting specific actions from only a short and possibly ambiguous observation of a person. The types of actions are currently limited by the training data to everyday activities such as eating or walking. With modified datasets, the prediction of various specific action sub-categories might be possible (e.g., forecasting possible malicious actions). While simpler methods may be more suitable for this kind of task, here we look to efforts in data transparency; we will provide our annotations and various statistics to characterize the everyday activities in our considered data.

Another axis to consider is that of environmental impact, in the cost of training deep neural networks. Our training time is relatively short with only a few hours until convergence and a moderately sized neural network. Additionally, adversarial attacks are a possibility to disrupt future predictions, but do not induce security concerns for our approach directly.

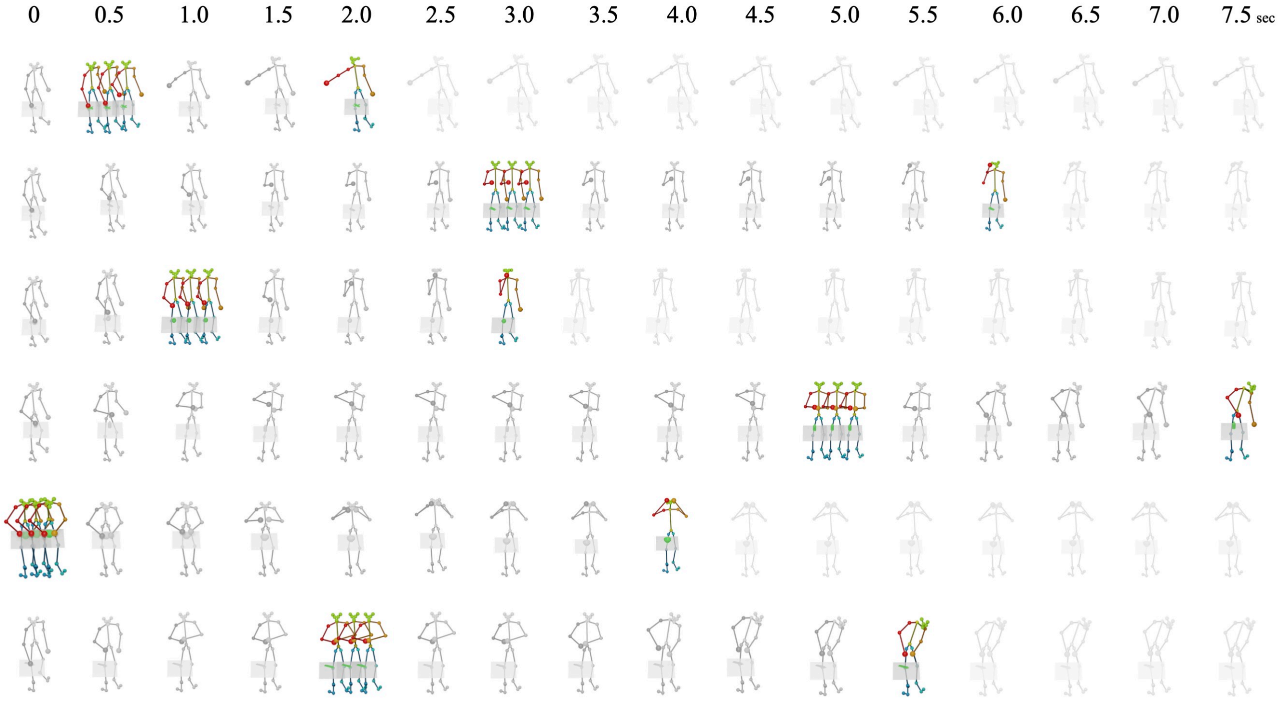


Figure 14. Sample input-target pairs (colored) for our characteristic 3d pose forecasting task, with temporal snapshots along the sequence (grayscale). Each snapshot is half a second apart. Depicted as input is the last frame of the respective input sequence.

C.2 FutureHuman3D: Forecasting Complex Long-Term 3D Human Behavior from Video Observations

Copyright Notice

In accordance with the IEEE Thesis/Dissertation Reuse Permissions, we include the accepted version of the original publication [41] in the following.

IEEE
Requesting permission to reuse content from an IEEE publication

FutureHuman3D: Forecasting Complex Long-Term 3D Human Behavior from Video Observations

Conference Proceedings:
2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Author: Christian Diller

Publisher: IEEE

Date: 16 June 2024

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

FutureHuman3D: Forecasting Complex Long-Term 3D Human Behavior from Video Observations

Christian Diller
Technical University of Munich
christian.diller@tum.de

Thomas Funkhouser
Google
tfunkhouser@google.com

Angela Dai
Technical University of Munich
angela.dai@tum.de

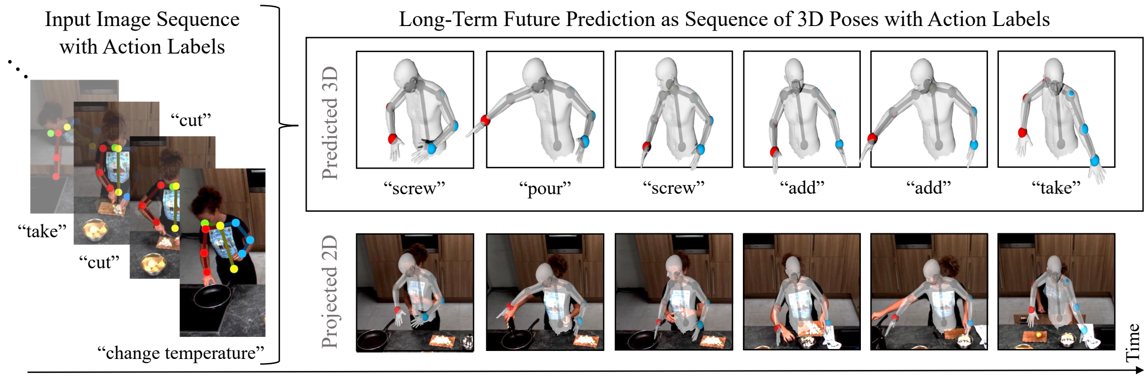

Input Image Sequence with Action Labels
"cut"
"take"
"cut"
"change temperature"
Long-Term Future Prediction as Sequence of 3D Poses with Action Labels
Predicted 3D
"screw"
"pour"
"screw"
"add"
"add"
"take"
Projected 2D
Time

Figure 1. We propose a novel generative approach to model long-term future human behavior by jointly forecasting a sequence of coarse action labels and their concrete realizations as 3D body poses. For broad applicability, our autoregressive method only requires weak supervision and past observations in the form of 2D RGB video data, together with a database of uncorrelated 3D human poses.

Abstract

We present a generative approach to forecast long-term future human behavior in 3D, requiring only weak supervision from readily available 2D human action data. This is a fundamental task enabling many downstream applications. The required ground-truth data is hard to capture in 3D (mocap suits, expensive setups) but easy to acquire in 2D (simple RGB cameras). Thus, we design our method to only require 2D RGB data at inference time while being able to generate 3D human motion sequences. We use a differentiable 2D projection scheme in an autoregressive manner for weak supervision, and an adversarial loss for 3D regularization. Our method predicts long and complex human behavior sequences (e.g., cooking, assembly) consisting of multiple sub-actions. We tackle this in a semantically hierarchical manner, jointly predicting high-level coarse action labels together with their low-level fine-grained realizations as characteristic 3D human poses. We observe that these two action representations are coupled in nature, and joint prediction benefits both action and pose forecasting. Our experiments demonstrate the complementary nature of joint action and 3D pose prediction: our joint approach outperforms each task treated individually, enables robust longer-term sequence prediction, and improves over alternative approaches to forecast actions and characteristic 3D poses.

1. Introduction

Predicting future human behavior is fundamental to machine intelligence, with many applications in content creation, robotics, mixed reality, and more. For instance, a monitoring system might issue early warnings of potentially dangerous behaviour, and a robotic assistant can use forecasting to place tools at the right place and time they will be needed in the future. Consider the specific scenario of an assembly line monitoring system deployed to issue early warnings of behavior that could be harmful in the near future: The system needs to have a long-term understanding of the future, enabling it to forecast multiple action steps ahead so that it can act in time before a harmful action occurs. However, simply understanding the next action steps on a high level is not sufficient: it must also reason about *where* the action will occur. Actions such as "grab a tool" are likely harmless when performed in a toolbox; they become dangerous when done next to an active table saw or moving robot arm. The monitoring system thus also needs to be able to reason about spatial relations in 3D – for both the location and body pose of involved humans.

To support these types of applications, we must address two tasks: 1) forecasting long-term action sequences, and 2) predicting future 3D human poses. Prior work has focused on each of these tasks separately: activity forecasting predicts future action labels without considering the 3D poses [33, 35, 36, 49, 51, 71], while 3D pose forecasting focuses on fixed frame rate sequence prediction limited to single actions in short-term time frames without considering longer-term action sequences [31, 61, 62, 94, 97].

We propose that these two tasks are coupled in nature: predicting action labels with realized 3D poses helps to encourage richer feature learning and can materialize subcategory level differences in actions for predicting future activities, and grounding 3D poses with actions provides global structure for longer-term forecasting.

Leveraging this insight, we design a method that takes in a sequence of recent RGB image observations and their action labels, and jointly predicts a sequence of future 3D characteristic poses and action labels (Fig. 1). In our design, we had to address two significant research challenges: 1) forecasting 3D poses from 2D images without any paired 3D training data, and 2) forecasting long sequences of actions comprising several discrete action steps.

The first challenge arises from limited training data. It would be ideal to have a dataset with ground truth 3D pose and action annotations for complex sequences of actions. Unfortunately, no such dataset exists. There are RGB video datasets with tracked 3D poses for limited types of actions (e.g., walking or waving); and there are video datasets with action labels for complex sequences of actions (e.g., cooking or assembly). However, there is no single dataset that has both types of annotations, and capturing one would be difficult due to the challenges of setting up 3D pose trackers in settings where people typically perform complex sequences of actions (e.g., cooking in a kitchen). Instead, we have to learn to use 2D video observations for 3D pose and action label forecasting without paired data. We achieve this by weakly supervising our pose forecasting in 2D using readily available 2D action datasets [8, 72] and formulate an adversarial loss encouraging likely 3D characteristic poses with respect to a distribution learned from 3D pose datasets [42, 60, 82]. Crucially, this does not require any correspondence between the 2D video and 3D pose data.

The second challenge arises from the difficulties of predicting long sequences of discrete events. One option would be to train a model to output a multi-step sequence of actions and poses all at once – however, this is impossible given the exponential growth of multi-step sequences and the limited amount of available training data. Another option would be to train a model that predicts the next future poses and actions at fixed time points in the future (e.g., 1s in advance) and then recurrently make long-term predictions – however, this time-based forecasting approach produces sequences that tend to "drift" over the long-term, since the intermediate poses at fixed time steps are usually "in between" semantically meaningful actions and thus do not provide a distinctive input representation for the next prediction. To address this issue, we train our autoregressive approach to iteratively generate the next discrete action label along with the *3D characteristic pose* for that action. A 3D characteristic pose [22] is the set of 3D joint positions corresponding to the most distinctive moment a semantic action is being performed (e.g., when a hand grasps an object, when two objects are first brought together, etc.). By training our method to produce these poses as intermediate outputs (and inputs to the next step), we are able to generate more semantically plausible forecasts over longer action sequences.

Our experiments with two RGB video datasets demonstrate that our approach for joint prediction of action behaviors and 3D poses outperforms state-of-the-art methods applied separately to each task. Additionally, we find that predicting actions and their 3D characteristic poses enables more robust autoregressive prediction for longer-term forecasting. Overall, our contributions are:

- The first method to learn forecasting of future 3D poses from datasets with only 2D RGB video and action label data (i.e., without any paired 3D data).
- The first method to forecast future 3D poses jointly with action labels from commonly available video input.
- The first method to forecast future characteristic 3D poses and action labels for long-term and complex behaviors.

2. Related Work

3D Human Pose Forecasting. Forecasting 3D human poses has been studied in many previous works and is commonly formulated as a 3D sequential motion prediction task, taking an input 3D sequence of poses and generating an output 3D sequence of poses. For short-term future prediction (up to ≈ 1 second), RNN-based approaches have achieved impressive performance [1, 16, 30, 37, 38, 44, 64, 69, 90]. As RNNs tend to struggle to capture longer-term dependencies with a fixed-size history, graph-based [17, 18, 20, 53, 55, 56, 61, 77, 94, 99] and attention-based [2, 11, 62, 65, 83] approaches have been proposed to encode temporal history. Some methods also explored the applicability of temporal convolutions [54, 66] and MLP-only architectures [10, 39] to the task of human motion forecasting. Additionally, various approaches have been proposed to model future human motion stochastically to produce diverse future sequence predictions, either with adversarial GAN formulations [7, 52], conditional variational autoencoders (VAEs) [3, 9, 12, 59, 63, 74, 87, 94, 96], or diverse sampling [21, 97]. More recently, diffusion methods [78, 79] have been used for human motion generation and forecasting [6, 19, 46, 84, 85, 95, 98, 100]. These methods require 3D ground truth sequences for training, limiting applicability to scenarios where 3D inputs and ground-truth are not available. Ours requires only 2D training data for the action sequences, which is far more plentiful and easier to obtain. We generate valid 3D poses by leveraging an adversarial loss formulation, operating on a database of uncorrelated 3D poses.

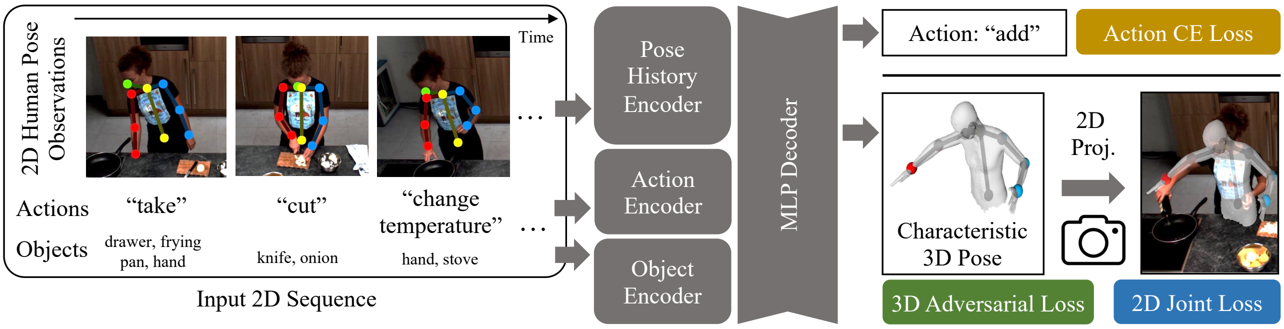


Figure 2. Our approach takes as input a sequence of RGB images, from which 2D poses are extracted, as well as their corresponding action label and initial set of objects. Each input is encoded into a joint latent space to jointly predict the next action label and characteristic 3D pose. While action labels are directly supervised, the 3D pose decoder is trained to match the next 2D action pose using differentiable projection, and an adversarial 3D loss encourages valid 3D pose prediction.

Human Action Forecasting. Action forecasting has been studied by many approaches to predict future actions from a sequence of observed actions [25, 26, 29, 48] or directly from an input video sequence [28, 32, 34, 35, 67, 75, 76, 76]. Various methods have been developed to learn effective representations, including Hidden Markov Models [51], RNNs [25–28, 32, 43, 76, 92], transformer-based networks [35, 36, 73], and self-supervised feature learning [41, 86]. There are approaches that focus on the short-term future [29, 32, 34, 35, 67, 73, 75, 76] or on longer-term actions [25, 26, 28, 29, 32, 34–36, 48, 67, 75, 76, 76]. Such method focus on characterizing anticipation with action labels only, while we aim to predict a richer characterization of the anticipated future by leveraging characteristic 3D poses, representative of future action goals in a sequence of action-pose predictions. Forecasting actions alongside human poses in 2D only has been studied in a few works, for 2D hand placement [57] or full-body 2D human poses at most 1 second into the future [101]. Our approach addresses the benefits of 3D reasoning in human motion forecasting, without requiring full 3D sequences for supervision.

Goal-Driven Future Prediction. Goal-driven forecasting has previously been explored beyond action label forecasting, and has been leveraged to predict goal locations for future human walking trajectories [14] and for future video sequences by predicting keyframes [5, 45, 70, 80]. Diller et al. [22] introduced the task of forecasting *characteristic 3D poses*, salient keyframe poses representing the next action. These goal-based poses are more semantically meaningful and consistent across different action sequences than time-based ones, and thus are better suited for long-term forecasting. We build upon these ideas by introducing a new goal-driven method for joint action anticipation and characteristic 3D pose forecasting in an auto-regressive system that can predict complex, long-term behavior sequences.

3. Method Overview

Our method aims to learn to jointly model future human actions along with the characteristic 3D poses representative of those actions. From a sequence of RGB image observations of a person performing a series of actions and the corresponding action labels, we predict a sequence of future action labels and 3D poses characteristic of these actions. This enables joint reasoning of not only global semantic behavior but also the physical manifestation thereof.

In the absence of 3D pose data of complex human actions, we weakly supervise forecasted 3D poses to align to future poses in 2D, and constrain the poses to be valid in 3D using an adversarial loss with a database of 3D poses. This does not require any correspondence between 3D pose data and 2D video, enabling action sequence supervision on commonly available 2D human action data together with carefully captured but unrelated human poses in 3D.

An overview of this approach is shown in Fig. 2. For an input sequence $S = \{(I_i, a_i, o_i)\}$ of N RGB images $\{I_i\}$ with corresponding actions $\{a_i\}$ and initially involved objects $\{o_i\}$, we aim to predict the future M actions $\{\hat{a}_k\}$ that will be taken along with their characteristic poses in 3D $\{\hat{Y}_k\}$. We define the human pose as a collection of J body joints at salient locations, so each output pose $\hat{Y}_k$ is predicted as a set of J 3D coordinates. We first extract information about the observed 2D pose movement by detecting 2D poses $\{X_i\}$, each with J 2D joints, with a state-of-the-art 2D pose estimator that seamlessly integrates into our pipeline in a pre-trained and frozen form.

Next, we encode this information along with previously observed action and object labels to predict the next future action label $\hat{a}_k$ and characteristic 3D pose $\hat{Y}_k$. We can then forecast a future sequence by autoregressively predicting a series, considering the 2D projections of the previously predicted 3D poses along with previously predicted actions as input to a new prediction.

4. Joint Forecasting of Actions and Characteristic 3D Poses

Our network takes as input the previous 2D observations $\{X_i\}$ extracted from the $\{I_i\}$ images, as well as action and object labels $\{a_i\}$ and $\{o_i\}$ as one-hot vectors. Since we only predict action labels, object labels are given from the objects seen at the beginning of the sequence, and subsequently reused for the entire sequence. Each of these are encoded in parallel with three separate encoders; the actions and objects with MLPs while the poses are projected into latent space with a single linear layer and then processed with a stack of three residual blocks. These encoded features are then all concatenated together in latent space, and processed jointly with an MLP to produce a common latent code z. Finally, we decode both poses and actions in parallel based on z using an MLP decoder each, yielding the next action label class as a vector $\hat{a}_k \in \mathbb{R}^{N_a}$ and 3D characteristic pose $\hat{Y}_k \in \mathbb{R}^{J \times 3}$, with N_a the number of action classes. For a more detailed architecture specification, we refer to the appendix.

We jointly learn future action labels and characteristic 3D poses by supervising $\hat{a}_k$ and $\hat{Y}_k$ to match the observed future 2D video, and constrain $\hat{Y}_k$ to form a valid 3D pose by an adversarial loss, optimizing for the overall loss:

$$\mathcal{L} = \lambda_{action}\mathcal{L}_{action} + \lambda_{pose2d}\mathcal{L}_{pose2d} + \lambda_{adv3d}\mathcal{L}_{adv3d} \quad (1)$$

where $\mathcal{L}_{action}$ denotes the action loss, as described in Sec. 4.1, $\mathcal{L}_{pose2d}$ and $\mathcal{L}_{adv3d}$ constraining the predicted pose, as described in Sec. 4.2, and the λ weighting each loss.

4.1. Action Forecasting

Predicted future actions are decoded from the latent code z by an MLP decoder to predict the action class $\hat{a}_k$, supervised by cross entropy with the ground truth future action: $\mathcal{L}_{action} = \mathrm{CE}(\hat{a}_k, a_k^{\mathrm{gt}})$.

4.2. Characteristic Pose Forecasting

Our goal is to forecast complex action behavior not only in terms of action labels, but also manifested as a sequence of characteristic poses in 3D. Since we only have 2D pose annotations available, we first constrain these poses to represent future actions in 2D and make use of an adversarial regularization in 3D. This does not require any correspondence between 2D and 3D data, only a collection of valid 3D poses, which are readily available.

Differentiable 2D Projection Our generator network predicts the next characteristic action pose $\hat{Y}_k$ as a set of 3D joints. To constrain $\hat{Y}_k$ based on the target future 2D pose X^{gt} extracted from the ground truth future image, we differentiably project $\hat{Y}_k$ into the 2D image with intrinsic parameters K and extrinsic rotation and translation R, t:

$$\hat{X} = K(R\hat{Y}_k + t) \quad (2)$$

Since we learn from third-person video with a fixed camera, we can use the same camera parameters for all sequences used for training. We can then define the 2D pose loss as the mean squared error between the projected pose prediction and the ground truth:

$$\mathcal{L}_{pose2d} = ||X^{\mathrm{gt}} - X_k||_2^2 \quad (3)$$

Note that we only predict the J joints that have been observed in the video data (excluding any joints that remain occluded in the observed video data), so this loss can be applied to all predicted joints.

Adversarial 3D Pose Regularization. While the action and pose prediction losses provide effective predictions when considered in the 2D projections, the $\{\hat{Y}_k\}$ remain underconstrained in 3D and thus tend to exhibit large distortions and implausible bone lengths and angles, when trained with only 2D supervision. We thus constrain the predicted poses to form valid 3D poses by formulating an adversarial 3D loss from a critic network which is simultaneously trained to distinguish predicted poses from a database of real 3D skeleton samples. Note that there is no correspondence between these skeletons and the 2D poses extracted from the action video sequences – any database of 3D skeletons can be used. We can thus train our approach with an entirely uncorrelated 3D pose dataset without requiring 3D action pose correlations.

We then formulate $\mathcal{L}_{adv3d}$ as a Wasserstein loss [4], training the critic network in an alternating fashion with the generator. This enables effective forecasting of future 3D characteristic poses for predicted future action labels, without requiring any 3D observations as input.

In order to enable the critic network to learn effectively about likely intrinsic pose constraints (e.g., lengths, kinematic chains, or valid joint angles), the critic takes as input not only the 3D joint locations of $\hat{Y}_k$ but also their kinematic statistics as a matrix Ψ, following [88, 89].

Ψ encodes joint angles and bone lengths as $\Psi = B^T B$, where $B = (b_1, b_2, \ldots, b_b)$ is a matrix with columns $b_i = j_k - j_l$ representing the vectors between each joint j_k and j_l. Ψ then contains bone lengths l_i^2 on its diagonal, and angular representations on the off-diagonal entries.

4.3. Sequence Prediction

In order to forecast longer-term future behavior, our 3D pose predictions enable a natural autoregressive sequence prediction by taking the predictions $\hat{X}_t, \hat{a}_t$ at time step t as part of the input for time step $t+1$. We can thus predict a sequence of M future action labels $\{\hat{a}_k\}$ and characteristic 3D poses $\{\hat{Y}_k\}$; we use $M = 10$ for MPII Cooking II [72] and $M = 5$ for IKEA-ASM [8], respectively.

4.4. Training Details

We train our approach for the $J = 9$ joints commonly seen across the input observed video data, characterizing the up-

	MPII Cooking II				IKEA ASM			
	2d	3d	Action Accuracy		2d	3d	Action Accuracy	
Approach	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
Zero Velocity	118	–	–	–	74	–	–	–
Train Average	166	–	–	–	91	–	–	–
AVT [35] RGB	–	–	19%	42%	–	–	22%	49%
AVT [35] RGB+Skeleton	–	–	20%	40%	–	–	23%	47%
FUTR [36] RGB	–	–	27%	48%	–	–	19%	45%
FUTR [36] RGB+Skeleton	–	–	27%	49%	–	–	20%	46%
RepNet [88] + DLow (min-10) [97]	72	**0.72**	–	–	45	0.31	–	–
RepNet [88] + GSPS (min-10) [63]	59	0.66	–	–	51	0.15	–	–
RepNet [88] + STARS (det.) [94]	70	0.62	–	–	54	0.27	–	–
RepNet [88] + EqMotion [93]	68	0.66	–	–	55	0.23	–	–
Joint 2D Pose & Action [101]	55	-	27%	43%	44	-	22%	46%
Ours	**50**	0.55	**29%**	**51%**	**40**	**0.31**	**29%**	**50%**

Table 1. Quantitative comparison with state-of-the-art action label and 3D pose forecasting. Our joint approach enables more accurate future action and pose predictions, compared to approaching both tasks separately, and outperforms joint action and 2D pose forecasting.

per body in MPII Cooking II [72] and IKEA-ASM [8].

Additionally, we use loss weights $\lambda_{action} = 1e^6$, $\lambda_{pose} = 1$, and $\lambda_{adv3d} = 1$, empirically chosen to numerically balance each individual loss with the others.

We train our approach on a single NVIDIA GeForce RTX 2080TI for $\approx$ 12 hours until convergence. We use ADAM with batch size 4096, weight decay 0.001, and a constant learning rate of 0.0001 for both generator and discriminator.

4.5. Datasets

We train and evaluate our approach on two datasets: MPII Cooking II [72] and IKEA-ASM [8]. Both datasets contain sequences of human actors performing complex, unscripted actions, and provide annotations of fine-grained sub-action steps. MPII Cooking II [72] is an action recognition dataset with 272 complex cooking sequences and an average sequence time of 182s (35 annotated sub-actions, each 5.2s on average). IKEA-ASM contains 370 sequences of actors assembling IKEA furniture, with an average of 74s per sequence (15 annotated sub-actions, each 4.9s on average).

In both datasets, each action sequence has been filmed from a fixed camera setup; the third-person point of view enables extraction of 2D poses with an off-the-shelf 2D pose estimator. We use OpenPose [13] in our experiments and note that our approach is agnostic to the concrete method of 2D pose detection. We provide more in-depth discussion and additional experiments in the appendix.

We consider the 9 upper-body joints of the OpenPose skeletons, as the other joints are almost always occluded in the video observations, and remove global translation by centering each 2D pose at the neck joint.

Characteristic poses, in contrast to an arbitrary pose within a labeled action range, are the most representative pose of that action, and are annotated for all sub-actions in each sequence as the most articulated pose of that sub-action, following the annotation protocol of [22]. Annotation can be done efficiently and was performed by the authors within just 32 hours, yielding a total of $\approx$18,000 characteristic poses ($\approx$12,000 for MPII Cooking II and $\approx$6,000 for IKEA-ASM). These poses are indicative of the action they represent as demonstrated in Tab. 2: Using such poses significantly improves performance, validating our annotation protocol.

For the 3D adversarial loss, we use $\approx$800,000 human poses from popular 3D pose datasets: Human3.6m [42], AMASS [60], and GRAB [82]. Note that none of these 3D poses have any correspondence with the 2D posed actions from the MPII Cooking II dataset, instead depicting various human skeletons in natural and diverse poses.

5. Results

We evaluate sequence forecasting of action labels and characteristic 3D poses on the MPI Cooking II [72] and IKEA-ASM [8] datasets, and 3D pose quality by comparing to our database of high-fidelity 3D poses.

5.1. Evaluation Metrics

2D Pose Error. Since we only have 2D ground-truth data available for complex action sequences, we first project predicted 3D poses back into 2D, and evaluate the 2D mean per-joint position error (MPJPE) [42], in comparison with 2D poses extracted from ground-truth future frames using [13]: $\text{E}_{\text{MPJPE}} = \frac{1}{M}\sum_{j=1}^{M} ||\hat{X} - X^{\text{gt}}||_2$.

3D Pose Quality. In the absence of annotated ground truth 3D poses for the action video sequences, we measure the quality of predicted 3D poses as how distinguishable they are in comparison to a set of real 3D poses. We follow [3] and evaluate quality by training a binary classifier on 50,000 human poses generated at different training steps (representing examples of unrealistic 3D poses) and 50,000 real 3D pose samples. For classification accuracy a of this classifier, quality is measured as $1 - a$, with a quality of 1 indicating full indistinguishability from real poses. We refer to the appendix for more details on this quality metric.

Action Accuracy. We report the action accuracy of the predicted sequences, as the mean over all sequences in the test

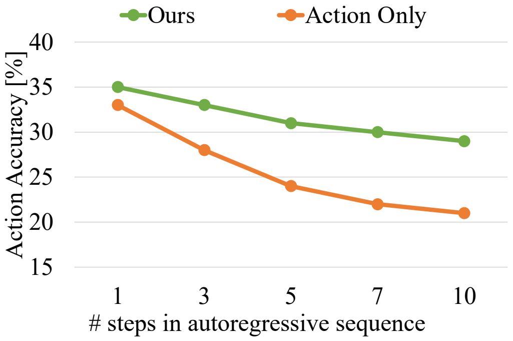


Figure 3. Action accuracy over time. Our joint action-characteristic pose forecasting enables more robust autoregressive action forecasting than action prediction without considering pose.

set. We evaluate the top-n accuracy based on whether the ground truth action is among the n highest scoring predictions, for $n = 1$ and $n = 3$.

5.2. Comparison to Human Pose Forecasting

Tab. 1 compares our method to state-of-the-art 3D pose forecasting methods DLow [97], GSPS [63], STARS [94], and EqMotion [93]. These methods expect sequences of observed 3D human poses as input; we thus first apply a state-of-the-art weakly supervised 3D pose estimator [88] on our 2D input poses, producing inputs and supervision in 3D. This method estimates 3D poses using an adversarial formulation, requiring a database of 3D poses not correlated with the 2D pose inputs. To ensure a fair comparison, this database is exactly the same as the one our method uses.

We chose the 3D pose estimator of [88] since its weakly supervised formulation is most comparable to our approach. An additional comparison to a fully supervised approach for 3D pose lifting (SPIN [50]) is provided in the appendix.

We then train the 3D pose prediction methods from scratch on this generated data, using their original parameter settings. Stochastic methods DLow and GSPS are set to predict 10 possible future sequences; we report the minimum error across these. We use STARS in the method's deterministic mode. Each method takes as input a pose history of M poses and outputs a sequence of M poses, analogous to our setup where each pose is a characteristic pose corresponding to an action step ($M = 10$ for MPII Cooking II and $M = 5$ for IKEA-ASM). Our approach to lift 2D to future 3D poses and actions in an end-to-end fashion enables more effective pose forecasting than these state-of-the-art 3D pose forecasting approaches on both datasets.

In addition, we compare to the joint 2D action and pose forecasting approach of Zhu et al. [101]. Our approach of forecasting long-term sequences of 3D poses alongside actions is able to outperform their 2D MPJPE pose prediction and action accuracy performance, due to improved spatial reasoning when forecasting 3D poses.

Statistical 2D Baselines. We additionally compare with two statistical baselines in 2D, following [22]: the average target train pose, and a zero-velocity baseline which was introduced by Martinez et al. [64] as competitive with state of the art. We outperform both baselines, indicating that our method learns a strong action pose representation.

5.3. Comparison to Action Label Forecasting

We compare the action accuracy of our joint action-pose forecasting to AVT [35] and FUTR [36], two state-of-the-art action anticipation methods, in Tab. 1. We train and evaluate both AVT and FUTR on input RGB frames and their action and object labels, equal to our training setup, and use their original training settings initialized with a pre-trained vision transformer [23] for AVT and extracted I3D features [15] from our datasets for FUTR. Additionally, as we consider extracted 2D poses from the input RGB images, we also evaluate a variant that is trained and evaluated on RGB images overlaid with 2D poses ("+Skeleton"). Our approach outperforms these baselines in both scenarios, by jointly predicting future actions and characteristic 3D poses.

5.4. Ablation Studies

What is the effect of pose forecasting on long-term action understanding? Tab. 3 shows that there is a notable improvement in action accuracy between training only with an action loss vs. training action and 2D pose loss jointly. This becomes more apparent when training action only vs action and full pose prediction (2D and 3D losses). In addition, Fig. 3 shows the correspondence between autoregressive prediction length and action accuracy: jointly forecasting poses and actions enables more robust autoregressive forecasting over time. We conclude that pose forecasting is beneficial for long-term action understanding.

How does action forecasting affect pose prediction performance? Tab. 3 demonstrates that pose forecasting trained jointly with action prediction is complementary and enables more accurate pose prediction.

What is the effect of characteristic pose forecasting? Since state-of-the-art pose forecasting focuses on fixed frame rate predictions independent of actions, we compare with

Poses		2D	3D	Action Accuracy	
Train	Test	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
Uncoupled	Uncoupled	75	0.29	28%	48%
Middle	Middle	58	0.45	26%	43%
Random	Random	67	0.37	22%	42%
Characteristic	**Characteristic**	**50**	**0.55**	**29%**	**51%**

Table 2. Ablation on pose forecasting on MPII Cooking II [72]. Our characteristic pose representation maximizes MPJPE and action performance: We consider pose prediction following state-of-the-art pose forecasting as decoupled from actions (uncoupled), as well as poses coupled to actions but in the middle of an action range, or at a random time therein, and our characteristic pose prediction. The same pose type is used for both train and evaluation.

Losses During Training			MPII Cooking II				IKEA ASM			
			2D	3D	Action Accuracy		2D	3D	Action Accuracy	
Action	2D Proj.	3D Adv.	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
✓	×	×	–	–	21%	41%	–	–	24%	45%
✓	✓	×	62	0.10	26%	49%	46	0.05	27%	49%
×	✓	×	54	0.21	–	–	44	0.09	–	–
×	✓	✓	58	0.53	–	–	43	0.29	–	–
✓	✓	✓	**50**	**0.55**	**29%**	**51%**	**40**	**0.31**	**29%**	**50%**

Table 3. Ablation on the effect of the action, 2D projection, and 3D adversarial losses. Combining all together for joint forecasting enables complementary learning to produce the best performance.

such joint forecasting of action and pose where predicted poses are sampled at equally spaced points in time in Tab. 2 (uncoupled). Additionally, we consider alternative poses to forecast for each action rather than a characteristic 3D pose (middle of the annotated action range, and randomly selected within the action range). We keep the same pose representation for training and testing (i.e., evaluate on middle poses when trained on them, etc.), for a fair comparison. We observe the best performance when forecasting characteristic 3D poses along with action labels, showing their usefulness for forecasting long sequences of 3D poses and actions.

5.5. Qualitative Results

Qualitative evaluations for the predicted poses are shown in Fig. 5 on data from MPII Cooking II [72] and in Fig. 4 on data from IKEA-ASM [8]. We compare our approach with state-of-the-art 3D pose forecasting of DLow [97], GSPS [63], and STARS [94]. For each method, we show a 3D body mesh in addition to the predicted 3D pose joints, to more comprehensively show the 3D structure of the forecasting results; we obtain body meshes by fitting SMPL [58] to each methods' predicted 3D body joints.

As there is no 3D ground truth available, we show the camera perspective with background for context as well as without background for a 3D pose only version. The two views demonstrate the fit to the ground truth 2D along with the quality of the 3D pose, respectively. Our approach leads to poses that better follow the ground-truth action poses in 2D compared to both previous methods while still maintaining a valid pose structure in 3D. Notably, this is true for both datasets, as our approach effectively forecasts the different data characteristics of both cooking as well as furniture assembly. In particular, our joint action-3D pose forecasting enables more accurate forecasting with diverse and accurate 3D pose structures.

5.6. Limitations

While we have demonstrated the potential of joint action and 3D pose forecasting, several limitations remain. For instance, our method leverages a separate 2D pose extraction as input to training, while an end-to-end formulation could potentially better leverage other useful signal in the input

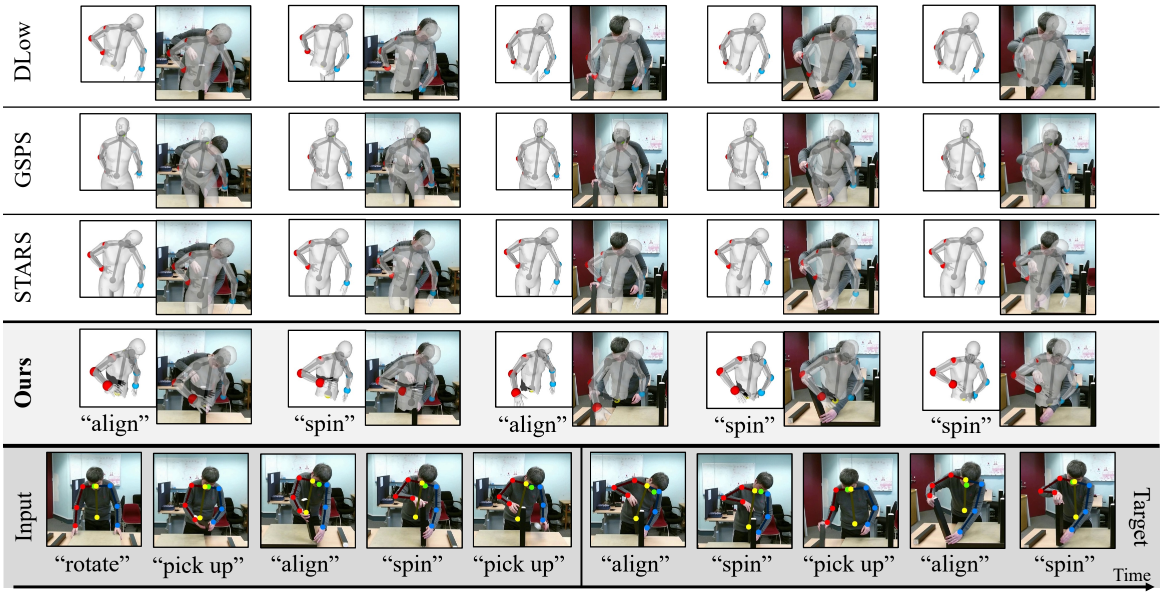


Figure 4. Qualitative comparison between DLow [97], GSPS [63], STARS [94], and our method on IKEA-ASM [8] data. For each method, we show the 3D predicted pose projected into the 2D target view, without background (small) and with background for context (full size). Our joint reasoning captures the individual characteristic action poses more faithfully while producing spatially plausible 3D poses.

frames. Additionally, a more holistic body representation than pose joints would be important for finer-grained interactions that involve reasoning over small limbs (e.g., hands) and body surface contact.

6. Conclusion

In this paper, we proposed to forecast future human behavior by jointly predicting future actions alongside characteristic 3D poses. We do not require any 3D annotated action sequences, or 3D input data; instead, we learn complex action sequences from 2D action video data, and regularize predicted poses with an adversarial formulation against uncorrelated 3D pose data. Experiments demonstrate that our joint forecasting enables complementary feature learning, outperforming each individual task considered separately.

Acknowledgements

This project is funded by the Bavarian State Ministry of Science and the Arts and coordinated by the Bavarian Research Institute for Digital Transformation (bidt), and the German Research Foundation (DFG) Grant "Learning How to Interact with Scenes through Part-Based Understanding".

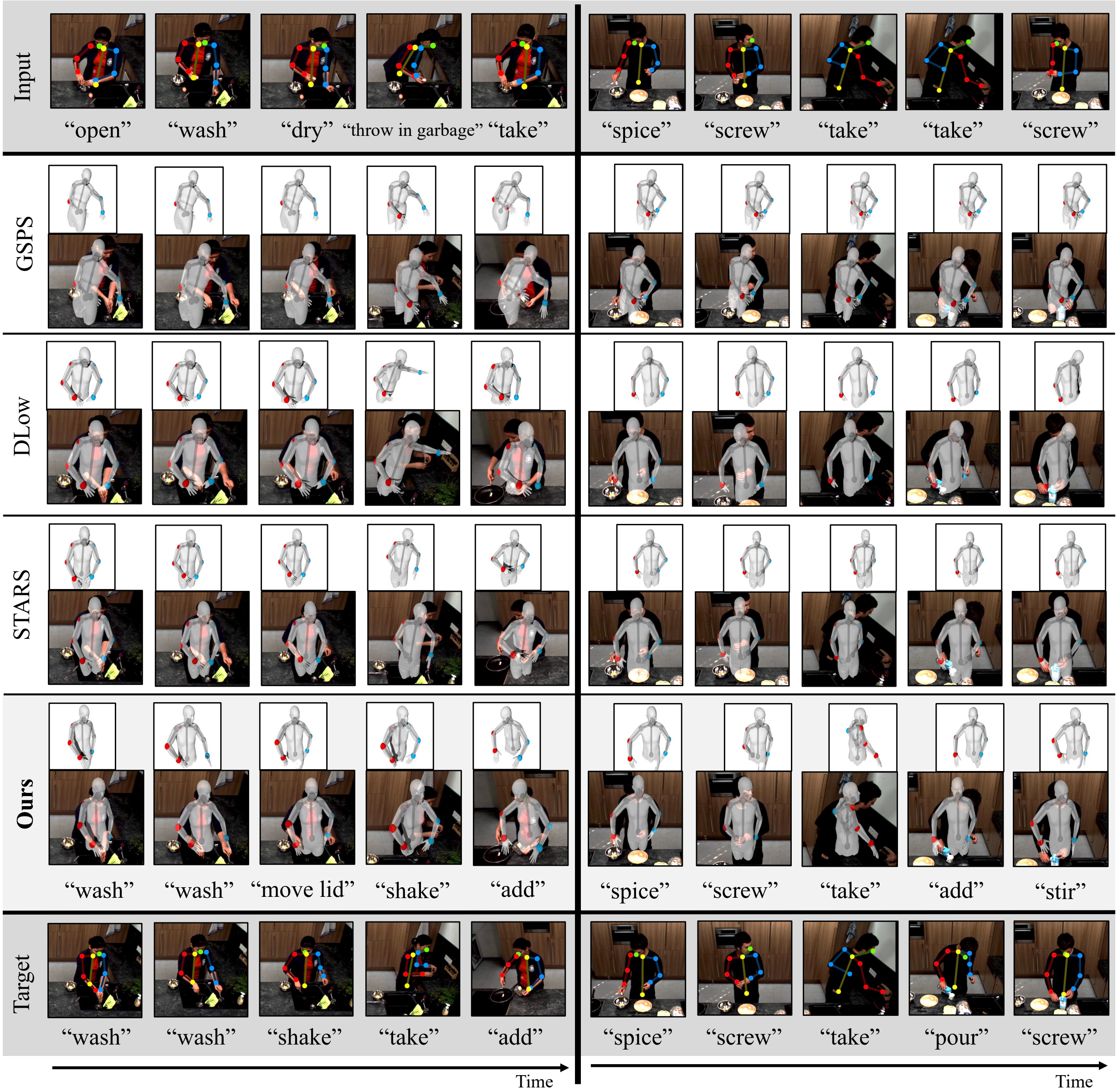


Figure 5. Qualitative comparison between DLow [97], GSPS [63], STARS [94], and our method on two sequences (left and right) from MPII Cooking II [72]. For each method, we show the 3D predicted pose projected into 2D, without background (small) and with background for context (full size). By considering both 3D pose and action forecasting together, we more effectively forecast the longer-term behavior.

References

[1] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7143–7152. IEEE, 2019. 2

[2] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. In *International Conference on 3D Vision, 3DV 2021, London, United Kingdom, December 1-3, 2021*, pages 565–574. IEEE, 2021. 2

[3] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5223–5232, 2020. 2, 5

[4] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 214–223. PMLR, 2017. 4

[5] Amir Bar, Roei Herzig, Xiaolong Wang, Anna Rohrbach, Gal Chechik, Trevor Darrell, and Amir Globerson. Compositional video synthesis with action graphs. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 662–673. PMLR, 2021. 3

[6] Germán Barquero, Sergio Escalera, and Cristina Palmero. Belfusion: Latent diffusion for behavior-driven human motion prediction. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 2317–2327. IEEE, 2023. 2

[7] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1418–1427, 2018. 2

[8] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 847–859, 2021. 2, 4, 5, 7, 14, 15, 16, 17

[9] Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Björn Ommer. Behavior-driven synthesis of human dynamics. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12236–12246. Computer Vision Foundation / IEEE, 2021. 2

[10] Arij Bouazizi, Adrian Holzbock, Ulrich Kressel, Klaus Dietmayer, and Vasileios Belagiannis. Motionmixer: Mlp-based 3d human body pose forecasting. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 791–798. ijcai.org, 2022. 2

[11] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, Ding Liu, Jing Liu, and Nadia Magnenat-Thalmann. Learning progressive joint propagation for human motion prediction. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VII*, pages 226–242. Springer, 2020. 2

[12] Yujun Cai, Yiwei Wang, Yiheng Zhu, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Chuanxia Zheng, Sijie Yan, Henghui Ding, Xiaohui Shen, Ding Liu, and Nadia Magnenat-Thalmann. A unified 3d human motion synthesis model via conditional variational auto-encoder∗. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11625–11635. IEEE, 2021. 2

[13] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 5, 16, 17

[14] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision*, pages 387–404. Springer, 2020. 3

[15] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4724–4733. IEEE Computer Society, 2017. 6, 16

[16] Hsu-Kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*, pages 1423–1432. IEEE, 2019. 2

[17] Qiongjie Cui and Huaijiang Sun. Towards accurate 3d human motion prediction from incomplete observations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 4801–4810. Computer Vision Foundation / IEEE, 2021. 2

[18] Qiongjie Cui, Huaijiang Sun, and Fei Yang. Learning dynamic relationships for 3d human motion prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6518–6526. Computer Vision Foundation / IEEE, 2020. 2

[19] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 9760–9770. IEEE, 2023. 2

[20] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. MSR-GCN: multi-scale residual graph convolution networks for human motion prediction. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11447–11456. IEEE, 2021. 2

[21] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Diverse human motion prediction via gumbel-softmax sampling from an auxiliary space. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 5162–5171. ACM, 2022. 2

[22] Christian Diller, Thomas Funkhouser, and Angela Dai. Forecasting characteristic 3d poses of human actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15914–15923, 2022. 2, 3, 5, 6

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 6

[24] Haoshu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45 (6):7157–7173, 2023. 17

[25] Yazan Abu Farha and Juergen Gall. Uncertainty-aware anticipation of activities. In *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, pages 1197–1204. IEEE, 2019. 3

[26] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what? - anticipating temporal occurrences of activities. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5343–5352. Computer Vision Foundation / IEEE Computer Society, 2018. 3

[27] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what? - anticipating temporal occurrences of activities. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5343–5352. Computer Vision Foundation / IEEE Computer Society, 2018.

[28] Yazan Abu Farha, Qiuhong Ke, Bernt Schiele, and Juergen Gall. Long-term anticipation of activities with cycle consistency. In *Pattern Recognition - 42nd DAGM German Conference, DAGM GCPR 2020, Tübingen, Germany, September 28 - October 1, 2020, Proceedings*, pages 159–173. Springer, 2020. 3

[29] Basura Fernando and Samitha Herath. Anticipating human actions by correlating past with the future with jaccard similarity measures. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 13224–13233. Computer Vision Foundation / IEEE, 2021. 3

[30] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4346–4354. IEEE Computer Society, 2015. 2

[31] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision*, pages 4346–4354, 2015. 2

[32] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6251–6260. IEEE, 2019. 3

[33] Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4021–4036, 2020. 1

[34] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Predicting the future: A jointly learnt model for action anticipation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5561–5570. IEEE, 2019. 3

[35] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 13485–13495. IEEE, 2021. 2, 3, 5, 6, 15

[36] Dayoung Gong, Joonseok Lee, Manjin Kim, Seong Jong Ha, and Minsu Cho. Future transformer for long-term action anticipation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 3042–3051. IEEE, 2022. 2, 3, 5, 6, 15

[37] Anand Gopalakrishnan, Ankur Mali, Dan Kifer, C. Lee Giles, and Alexander G. Ororbia II. A neural temporal model for human motion prediction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12116–12125. Computer Vision Foundation / IEEE, 2019. 2

[38] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José M. F. Moura. Adversarial geometry-aware human motion prediction. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, pages 823–842. Springer, 2018. 2

[39] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Back to MLP: A simple baseline for human motion prediction. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pages 4798–4808. IEEE, 2023. 2

[40] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social GAN: socially acceptable trajectories with generative adversarial networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2255–2264. Computer Vision Foundation / IEEE Computer Society, 2018. 16

[41] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation

learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III*, pages 312–329. Springer, 2020. 3

[42] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2014. 2, 5, 17

[43] Ashesh Jain, Avi Singh, Hema S Koppula, Shane Soh, and Ashutosh Saxena. Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3118–3125. IEEE, 2016. 3

[44] Ashesh Jain, Amir Roshan Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5308–5317. IEEE Computer Society, 2016. 2

[45] Dinesh Jayaraman, Frederik Ebert, Alexei A. Efros, and Sergey Levine. Time-agnostic prediction: Predicting predictable video frames. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 3

[46] Chiyu ”Max” Jiang, Andre Cornman, Cheolho Park, Benjamin Sapp, Yin Zhou, and Dragomir Anguelov. Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 9644–9653. IEEE, 2023. 2

[47] Jigyasa Singh Katrolia, Ahmed El-Sherif, Hartmut Feld, Bruno Mirbach, Jason R. Rambach, and Didier Stricker. Ticam: A time-of-flight in-car cabin monitoring dataset. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, page 277. BMVA Press, 2021. 16

[48] Qiuhong Ke, Mario Fritz, and Bernt Schiele. Time-conditioned action anticipation in one shot. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9925–9934. Computer Vision Foundation / IEEE, 2019. 3

[49] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *European conference on computer vision*, pages 201–214. Springer, 2012. 2

[50] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2252–2261. IEEE, 2019. 6, 16

[51] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014. 2, 3

[52] Jogendra Nath Kundu, Maharshi Gor, and R. Venkatesh Babu. Bihmp-gan: Bidirectional 3d human motion prediction GAN. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 8553–8560. AAAI Press, 2019. 2

[53] Bin Li, Jian Tian, Zhongfei Zhang, Hailin Feng, and Xi Li. Multitask non-autoregressive model for human motion prediction. *IEEE Trans. Image Process.*, 30:2562–2574, 2021. 2

[54] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5226–5234. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[55] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 211–220. Computer Vision Foundation / IEEE, 2020. 2

[56] Maosen Li, Siheng Chen, Zihui Liu, Zijing Zhang, Lingxi Xie, Qi Tian, and Ya Zhang. Skeleton graph scattering networks for 3d skeleton-based human motion prediction. In *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11-17, 2021*, pages 854–864. IEEE, 2021. 2

[57] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *European Conference on Computer Vision*, pages 704–721. Springer, 2020. 3

[58] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 7

[59] Thomas Lucas, Fabien Baradel, Philippe Weinzaepfel, and Grégory Rogez. Posegpt: Quantization-based 3d human motion generation and forecasting. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VI*, pages 417–435. Springer, 2022. 2

[60] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, 2019. 2, 5, 17

[61] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9489–9497, 2019. 2

[62] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*, pages 474–489. Springer, 2020. 2

[63] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Generating smooth pose sequences for diverse human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13309–13318, 2021. 2, 5, 6, 7, 8, 14, 15, 16

[64] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4674–4683. IEEE Computer Society, 2017. 2, 6

[65] Ángel Martínez-González, Michael Villamizar, and Jean-Marc Odobez. Pose transformers (POTR): human motion prediction with non-autoregressive transformers. In *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11-17, 2021*, pages 2276–2284. IEEE, 2021. 2

[66] Omar Medjaouri and Kevin Desai. HR-STAN: high-resolution spatio-temporal attention network for 3d human motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*, pages 2539–2548. IEEE, 2022. 2

[67] Antoine Miech, Ivan Laptev, Josef Sivic, Heng Wang, Lorenzo Torresani, and Du Tran. Leveraging the present to anticipate the future in videos. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2915–2922. Computer Vision Foundation / IEEE, 2019. 3

[68] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 17

[69] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 299. BMVA Press, 2018. 2

[70] Karl Pertsch, Oleh Rybkin, Jingyun Yang, Shenghao Zhou, Konstantinos Derpanis, Kostas Daniilidis, Joseph Lim, and Andrew Jaegle. Keyframing the future: Keyframe discovery for visual prediction and planning. In *Learning for Dynamics and Control*, pages 969–979. PMLR, 2020. 3

[71] Nicholas Rhinehart and Kris M Kitani. First-person activity forecasting with online inverse reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3696–3705, 2017. 2

[72] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision*, pages 1–28, 2015. 2, 4, 5, 6, 7, 8, 14, 15, 17

[73] Debaditya Roy and Basura Fernando. Action anticipation using pairwise human-object interactions and transformers. *IEEE Trans. Image Process.*, 30:8116–8129, 2021. 3

[74] Tim Salzmann, Marco Pavone, and Markus Ryll. Motron: Multimodal probabilistic human motion forecasting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 6447–6456. IEEE, 2022. 2

[75] Fadime Sener and Angela Yao. Zero-shot anticipation for instructional activities. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 862–871. IEEE, 2019. 3

[76] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVI*, pages 154–171. Springer, 2020. 3

[77] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-time-separable graph convolutional network for pose forecasting. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11189–11198. IEEE, 2021. 2

[78] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 2256–2265. JMLR.org, 2015. 2

[79] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2

[80] Didac Suris, Ruoshi Liu, and Carl Vondrick. Learning the predictability of the future. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12607–12617. Computer Vision Foundation / IEEE, 2021. 3

[81] Wandi Susanto, Marcus Rohrbach, and Bernt Schiele. 3d object detection with multiple kinects. In *European Conference on Computer Vision*, pages 93–102. Springer, 2012. 17

[82] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision – ECCV 2020*, pages 581–600, Cham, 2020. Springer International Publishing. 2, 5, 17

[83] Yongyi Tang, Lin Ma, Wei Liu, and Wei-Shi Zheng. Long-term human motion prediction by modeling motion context and enhancing motion dynamics. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 935–941. ijcai.org, 2018. 2

[84] Julian Tanke, Linguang Zhang, Amy Zhao, Chengcheng Tang, Yujun Cai, Lezi Wang, Po-Chen Wu, Juergen Gall, and Cem Keskin. Social diffusion: Long-term multiple human motion anticipation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9601–9611, 2023. 2

[85] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 2

[86] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 98–106. IEEE Computer Society, 2016. 3

[87] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3352–3361. IEEE Computer Society, 2017. 2

[88] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7782–7791, 2019. 4, 5, 6, 14, 15, 16, 17

[89] Bastian Wandt, Hanno Ackermann, and Bodo Rosenhahn. 3d reconstruction of human motion from monocular image sequences. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1505–1516, 2016. 4

[90] Borui Wang, Ehsan Adeli, Hsu-Kuang Chiu, De-An Huang, and Juan Carlos Niebles. Imitation learning for human pose prediction. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7123–7132. IEEE, 2019. 2

[91] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 17

[92] Yu Wu, Linchao Zhu, Xiaohan Wang, Yi Yang, and Fei Wu. Learning to anticipate egocentric actions by imagination. *IEEE Transactions on Image Processing*, 30:1143–1152, 2020. 3

[93] Chenxin Xu, Robby T Tan, Yuhong Tan, Siheng Chen, Yu Guang Wang, Xinchao Wang, and Yanfeng Wang. Eqmotion: Equivariant multi-agent motion prediction with invariant interaction reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1410–1420, 2023. 5, 6

[94] Sirui Xu, Yu-Xiong Wang, and Liang-Yan Gui. Diverse human motion prediction guided by multi-level spatial-temporal anchors. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXII*, pages 251–269. Springer, 2022. 2, 5, 6, 7, 8, 14, 15, 16

[95] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14928–14940, 2023. 2

[96] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 265–281, 2018. 2

[97] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*, pages 346–364. Springer, 2020. 2, 5, 6, 7, 8, 14, 15, 16

[98] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *CoRR*, abs/2208.15001, 2022. 2

[99] Chongyang Zhong, Lei Hu, Zihao Zhang, Yongjing Ye, and Shihong Xia. Spatio-temporal gating-adjacency GCN for human motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 6437–6446. IEEE, 2022. 2

[100] Zixiang Zhou and Baoyuan Wang. UDE: A unified driving engine for human motion generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 5632–5641. IEEE, 2023. 2

[101] Yanjun Zhu, David Doermann, Yanxia Zhang, Qiong Liu, and Andreas Girgensohn. What and how? jointly forecasting human action and pose. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 771–778. IEEE, 2021. 3, 5, 6, 14

Appendix

We show in this appendix additional qualitative (Sec. A) and quantitative (Sec. B) results, detail our baseline evaluation protocol (Sec. C), elaborate on the 3D quality metric (Sec. D), demonstrate the ability of our method to generalize to multi-actor scenarios (Sec. E), verify our method's robustness to 2D detection results (Sec. F), show the architecture used in our approach (Sec. G), and provide additional details regarding the data (Sec. H).

A. Additional Qualitative Results

Fig. 6 shows additional qualitative results of our method, on both MPII Cooking 2 [72] (left column) and IKEA-ASM [8] (right column), as compared to pose baselines DLow [97], GSPS [63], and STARS [94].

B. Additional Quantitative Results

B.1. Characteristic Poses

Analogous to Tab. 2 in the main paper, Tab. 8 shows an ablation on pose timings and compares our approach of using characteristic poses to poses taken at regular time intervals ("uncoupled") as well as in the middle or at a random time of an action, on IKEA-ASM [8] data. To further illustrate this point, Tab. 4 shows additional ablations: Poses predicted at random points in the sequence, but at most 1s from the closest characteristic pose ("centered on the characteristic pose") and predicting characteristic poses but evaluating interpolated regularly spaced poses. Both demonstrate that the usage of characteristic poses improves performance compared to other approaches while still being outperformed by directly predicting characteristic poses.

	2D	3D	Action Accuracy	
Poses	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
Uncoupled	75	0.29	28%	48%
Middle	58	0.45	26%	43%
Random	67	0.37	22%	42%
Centered on Char. Poses	69	0.33	28%	50%
Interp. from Char. Poses	62	0.13	29%	51%
Characteristic	**50**	**0.55**	**29%**	**51%**

Table 4. Ablation on pose forecasting on MPII Cooking II [72]. We consider pose prediction following state-of-the-art pose forecasting as decoupled from actions (uncoupled), as well as poses coupled to actions in various fashions: middle (the middle pose of an action range), random (a random pose of the action), random but at most 1s from the closest characteristic pose (centered), regularly spaced poses interpolated from characteristic pose prediction, and our characteristic pose prediction.

B.2. Lifting 2D Predictions to 3D

In Tab. 1 in the main paper, we compare to first lifting input poses into 3D, then performing 3D motion prediction. Tab. 5 evaluates the other way around: Predicting 2D poses and action labels jointly with [101], then lifting the predicted 2D poses into 3D with RepNet [88] for evaluation. Our method outperforms both approaches.

	MPII Cooking II				IKEA ASM			
	2d	3d	Action Accuracy		2d	3d	Action Accuracy	
Approach	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
[101] + [88]	63	0.50	27%	43%	53	0.21	22%	46%
Ours	**50**	**0.55**	**29%**	**51%**	**40**	**0.31**	**29%**	**50%**

Table 5. Our approach of jointly forecasting 3D poses and actions achieves better performance compared to 2D pose + action forecasting [101] and then lifting forecasted 2D poses into 3D using [88].

B.3. Input Noise Ablation

Tab. 6 shows the effect using a noise vector as additional input to our method. It encourages more diversity in predictions, which benefits pose and action forecasting.

B.4. Input Objects Ablation

Inputting initially observed objects slightly improves results (Tab. 6), due to added context for broad actions like "add," e.g."add ingredient" vs. "add water to pot.".

	MPII Cooking II				IKEA ASM			
	2d	3d	Action Accuracy		2d	3d	Action Accuracy	
Approach	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
No Objects	61	0.52	28%	51%	42	0.30	29%	50%
No Noise	55	0.49	29%	50%	48	0.29	**30%**	**51%**
Ours	**50**	**0.55**	**29%**	**51%**	**40**	**0.31**	29%	50%

Table 6. Ablations studies with no object input and no noise input.

B.5. Statistical Action Baselines

We additionally evaluate "Zero Velocity" and "Train Average" for action labels, analogous to forecasted poses, i.e. repeating the last action label and repeating the most frequent train action label, in Tab. 7. These baselines perform particularly poorly since actions are rarely repeated or fixed for entire sequences.

	MPII Cooking II		IKEA ASM	
Approach	top-1 ↑	top-3 ↑	top-1 ↑	top-3 ↑
Repeat Last Input	9%	43%	8%	35%
Most Common in Train	6%	10%	7%	26%
Ours	**29%**	**51%**	**29%**	**50%**

Table 7. Statistical action baselines: (1) Repeat the last input action label (2) Using the most common action label of the train set.

C. Baseline Evaluation Details

C.1. State-of-the-Art Pose Forecasting

We evaluate the performance of our baselines using the same input data that is available to our method. Pose forecast-

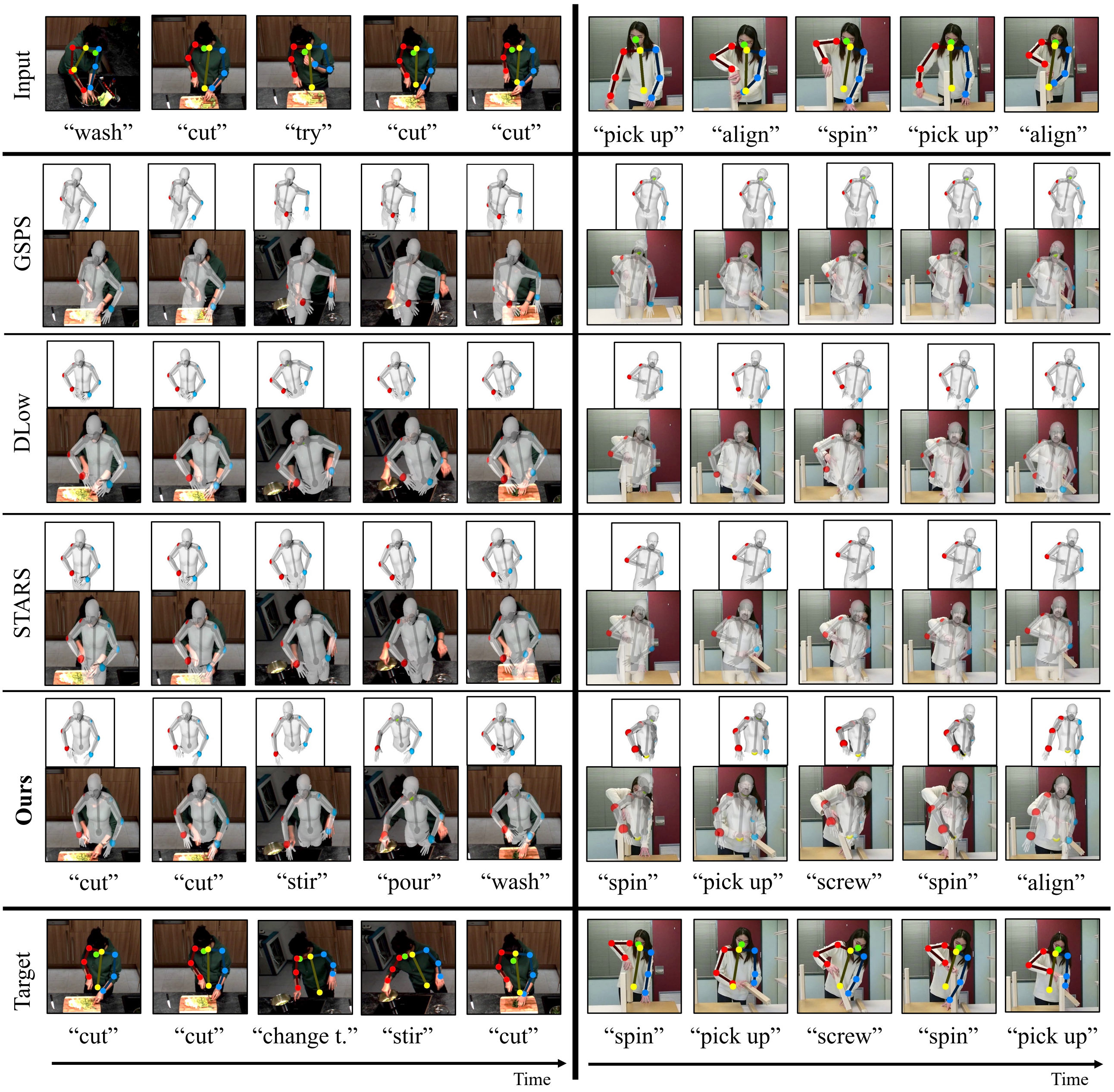


Figure 6. Additional qualitative comparison between DLow [97], GSPS [63], STARS [94], and our method on two sequences (left on MPII Cooking 2 [72], right on IKEA-ASM [8]). For each method, we show a the 3D predicted pose projected into the 2D target view, without background for a pose only version (small) as well as with background for context (full size).

ing baselines DLow [97], GSPS [63], and STARS [94] are trained and evaluated on sequences of our manually annotated characteristic poses. Since there is no ground-truth 3D pose data available, we first use RepNet [88], a state-of-the-art 3D pose estimation method, to retrieve 3D skeletons from our 2D characteristic poses. We train this method from scratch using the same database of valid 3D poses that is available to our method, allowing for a fair comparison.

C.2. State-of-the-Art Action Label Forecasting

We train action baselines AVT [35] and FUTR [36] using sequences of our characteristic pose frames together with the corresponding action labels as input. For AVT, we use their default parameters used by the original authors for

	2D	3D	Action Accuracy	
Poses	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
Uncoupled	64	0.30	28%	48%
Middle	47	0.35	28%	47%
Random	49	0.24	28%	49%
Characteristic	**41**	**0.35**	**29%**	**50%**

Table 8. Ablation on pose forecasting, on the IKEA-ASM [8] dataset. We consider predicting poses following state-of-the-art pose forecasting in a decoupled fashion from actions (uncoupled), as well as poses coupled to actions in various fashions: middle (the middle pose of an action range), random (a random pose of the action), and our characteristic pose prediction, which benefits action prediction the most.

their ablation on third-person dataset 50Salads/Breakfast, inputting our RGB frames instead. For a fair comparison, we also supply the action and object history for each step by encoding both label sequences with a small encoder (a single linear layer) each and fuse these features with the image features generated by the AVT encoder. For FUTR, we first generate I3D features [15] from our RGB frames and concatenate them with action and object history after encoding these in the same way as for AVT.

We then train two variants of both methods: One with the raw RGB frames, action history, and object history as input ("AVT RGB" and "FUTR RGB" in the main results figure), and one with additional 2D skeleton input (skeletons rendered on top of the RGB frames) from the skeletons that we extract with OpenPose [13] ("AVT RGB+Skeleton" and "FUTR RGB+Skeleton").

C.3. Supervised 3D Pose Lifting

For better comparability, we used weakly supervised approach [88] for pose lifting. This is important, since there is no ground-truth coupling between 2D and corresponding 3D action poses in our setting. Nevertheless, we compare to baselines [63, 94, 97] in Tab. 9 with poses lifted using fully supervised pre-trained SPIN [50]; our approach outperforms even these improved baselines in terms of 2D MPJPE.

	MPII Cooking II		IKEA ASM	
	2d	3d	2d	3d
Approach	MPJPE [px] ↓	Quality ↑	MPJPE [px] ↓	Quality ↑
SPIN [50] + DLow [97]	81	**0.89**	43	**0.43**
SPIN [50] + GSPS [63]	74	0.66	45	0.29
SPIN [50] + STARS [94]	66	0.80	41	0.40
Ours	**50**	0.55	**40**	0.31

Table 9. Comparison to pose baselines using fully-supervised pre-trained 3D pose estimation method SPIN [50]. In our main experiments, we instead compare to weakly supervised baseline RepNet [88] for a fair comparison.

D. 3D Quality Metric Details

For our pose quality metric, we use a 3-layer MLP binary classifier of 3D poses. Training poses are randomly sampled from ground-truth (real) and predicted (fake) collected during the training process of our method and all baselines, producing a total of 100k real and fake poses each. Fake poses exhibit a range of small to large unrealistic deformations, depending on when they were sampled, ranging from random joint placements to widely inconsistent bone lengths to unnatural joint angles to only minor inconsistencies in the bone lengths. The classifier is trained once and then used to evaluate all methods, to ensure a fair comparison.

As an additional intuitive metric we show the mean absolute bone length difference of right and left body in 3D in Tab. 10. We observe that this metric correlates with our classifier-based quality.

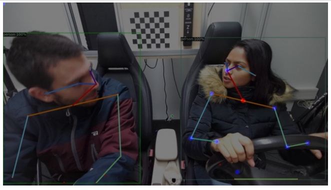
	MPII Cooking II		IKEA ASM	
Approach	Symm. [mm] ↓	Quality ↑	Symm. [mm] ↓	Quality ↑
RepNet [88] + DLow [97]	**13**	**0.72**	45	0.31
RepNet [88] + GSPS [63]	18	0.66	56	0.15
RepNet [88] + STARS [94]	16	0.62	46	0.27
No 3D Adversarial Loss	75	0.10	66	0.05
2D Projection Loss Only	57	0.21	61	0.09
No Action Loss	22	0.53	39	0.29
Ours	22	0.55	**39**	**0.31**

Table 10. Additional quality metric and its correlation to our classifier-based metric: Absolute bone length difference between right and left body, compared to pose baselines and ablations.

E. Multi-Actor Interaction Scenario

In addition to our experiments with single human actors in the main paper, we show here that our approach is able to generalize to multi-actor scenarios, with minor modifications. We show this in Tab. 11 with additional dataset TICaM [47] where driver and passenger are interacting in an in-car driving scenario (actions include "talking", various handoffs). Our modifications are: **(1)** Additional encoder and decoder for the second person **(2)** Interaction pooling introduced in Social GAN [40]. Our modified method outperforms simple combinations of previous works, with and without interaction modelling, demonstrating the wide applicability of our method.

	2d	3d	Action Accuracy	
Approach	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
FUTR RGB + Skeleton	-	-	38%	64%
RepNet + STARS	89	0.34	-	-
Ours (No Interactions)	68	0.40	40%	67%
Ours (Interaction Modeling)	**58**	**0.41**	**48%**	**73%**



Setting

Table 11. Our approach can also be applied to multi-actor scenarios: We demonstrate improved performance on suitable dataset TICaM [47], with and without explicit interaction modeling.

F. 2D Input Pose Quality

In Fig. 12, we replace OpenPose with AlphaPose [24] and Detectron2 [91], both only slightly changing the final results, indicating that our method does not depend on a specific 2D pose detector. We also experiment with added random noise to OpenPose: our method remains relatively robust. The coupled changes in pose and action accuracy further demonstrate the effectiveness of our joint feature learning.

MPII Cooking II	2d	3d	Action Accuracy	
Approach	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
OpenPose + max. 20px noise	59	0.45	26%	47%
OpenPose + max. 10px noise	57	0.47	26%	46%
Ours (using Detectron2)	47	**0.54**	28%	55%
Ours (using AlphaPose)	**46**	0.57	28%	**56%**
Ours (using OpenPose)	50	0.55	**29%**	51%

Table 12. Robustness of our method to different 2D pose detectors Detectron2 [91] and AlphaPose [24] as well as randomly added 2D noise. This only slightly affects our pose and action accuracy, further demonstrating the effectiveness of our joint feature learning.

G. Architecture Details

Generator Network Fig. 7 shows our generator architecture in detail with input and output dimensions for linear layers, and the slope for leaky ReLU layers.

Critic Network Our adversarial critic network processes generator outputs with 4 linear layers and 3 kinematic chain layers which are designed to encourage correct bone lengths (as shown in [88]), in parallel. 2 linear layers then combine both outputs and produce the final critic score.

H. Data Details

H.1. Camera Parameters

While intrinsic camera parameters are often available in captured image data, the camera parameters for captured video were not available from the MPII Cooking 2 [72] dataset to use for pose projection. We thus optimized for intrinsic camera parameters from the video sequence data in correspondence with the 3D scene reconstruction of the empty kitchen environment, as given by [81]. For IKEA-ASM [8], we use the provided intrinsic camera parameters directly. Note that camera parameters are only required to be fixed within a sequence (i.e. no moving camera) but can change between sequences.

H.2. 3D Pose Database Alignment

We use popular 3D pose datasets Human3.6m [42], AMASS [60], and GRAB [82] for our database of uncorrelated valid 3D poses. All poses are pre-processed to follow the OpenGL coordinate system and aligned with respect to the neck joint.

Ours		OpenPose		Human3.6m		SMPL-X	
Idx	Name	Idx	Name	Idx	Name	Idx	Name
0	head	0	nose	15	head	15	head
1	neck	1	neck	13	thorax	12	neck
2	right shoulder	2	right shoulder	25	right shoulder	17	right shoulder
3	right elbow	3	right elbow	26	right elbow	19	right elbow
4	right hand	4	right hand	27	right wrist	42	right index 3
5	left shoulder	5	left shoulder	17	left shoulder	16	left shoulder
6	left elbow	6	left elbow	18	left elbow	18	left elbow
7	left hand	7	left wrist	19	left wrist	27	left index 3
8	hip	8	mid-hip	0	hip	0	pelvis

Table 13. Human skeleton joint layout used in our experiments, for both 2D and 3D skeletons.

H.3. Pose Joint Layout

We use the 9 upper-body joints of the native OpenPose [13] joint layout for skeletons in 2D, and adapt skeletons in our 3D database to use the same format. Tab. 13 shows the correspondence between our joint layout, OpenPose [13], Human3.6m [42], and SMPL-X [68]. 3D datasets AMASS [60] and GRAB [82] provide human bodies in SMPL-X format; we first extract their skeleton joints using their publicly available code and then convert it into our layout using the correspondences in Tab. 13.

H.4. MPII Cooking 2 Details

We use action labels as annotated in the 2D cooking action dataset MPII Cooking 2 [72]. These annotations provide action labels (87 classes) for frame ranges in each sequence as well as the involved objects (187 classes). We first cluster similar actions together, yielding a total of 37 action clusters, which we then use as action classes in our experiments.

In addition, since our goal is to forecast upper-body actions with objects in the foreground, we remove instances of poses and corresponding actions that occur in the background - e.g., when taking out objects from the cupboard, or from the fridge.

In total, there are 272 cooking action sequences; we create a random train/val/test split along sequences with a ratio of 70% / 15% / 15%, yielding 190, 40, 40 sequences for each set.

H.5. IKEA-ASM Details

We use action labels as annotated in the IKEA furniture assembly dataset IKEA-ASM [8]. These annotations provide action labels (31 classes) for frame ranges in each sequence; we use them without explicit object information since each action already encodes its associated object.

In total, there are 370 furniture assembly action sequences; we create a random train/val/test split along sequences with a ratio of 70% / 15% / 15%, yielding 227, 48, 48 sequences for each set.

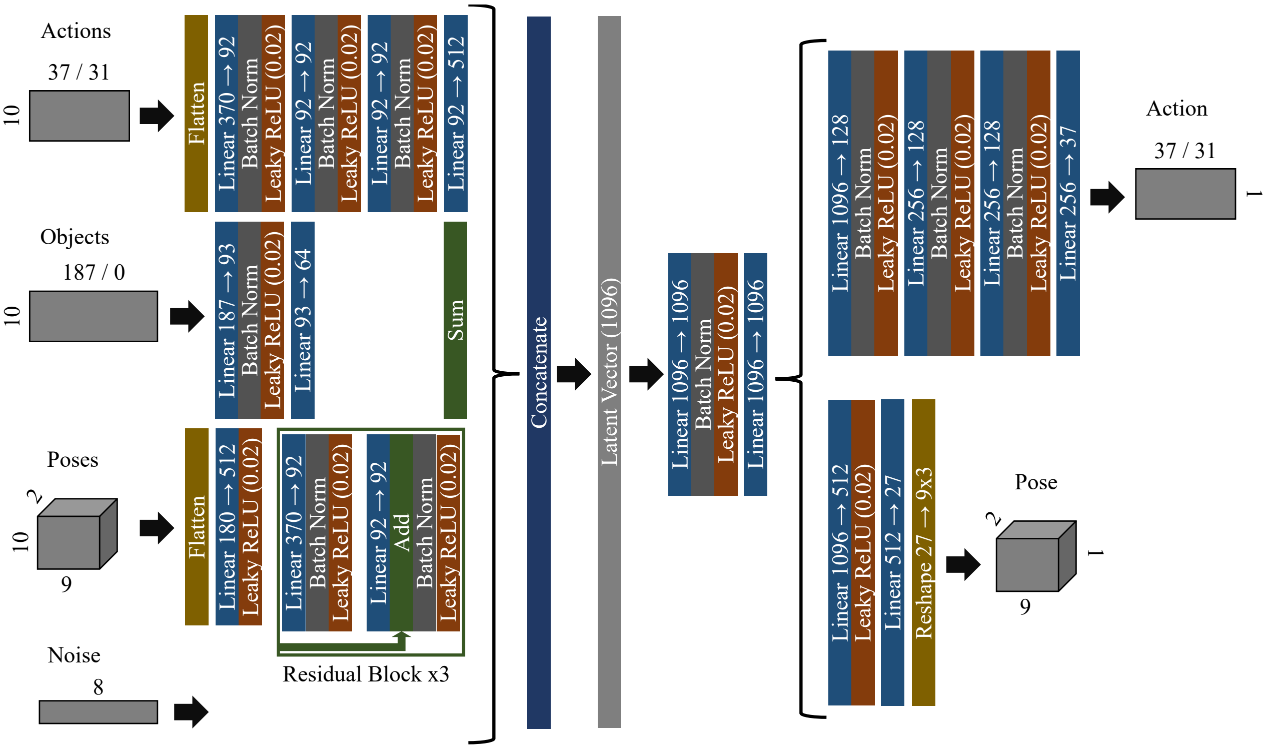


Figure 7. Network architecture specification.

C.3 CG-HOI: Contact-Guided 3D Human-Object Interaction Generation

Copyright Notice

In accordance with the IEEE Thesis/Dissertation Reuse Permissions, we include the accepted version of the original publication [3] in the following.

CG-HOI: Contact-Guided 3D Human-Object Interaction Generation

Conference Proceedings:
2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Author: Christian Diller

Publisher: IEEE

Date: 16 June 2024

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis online.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

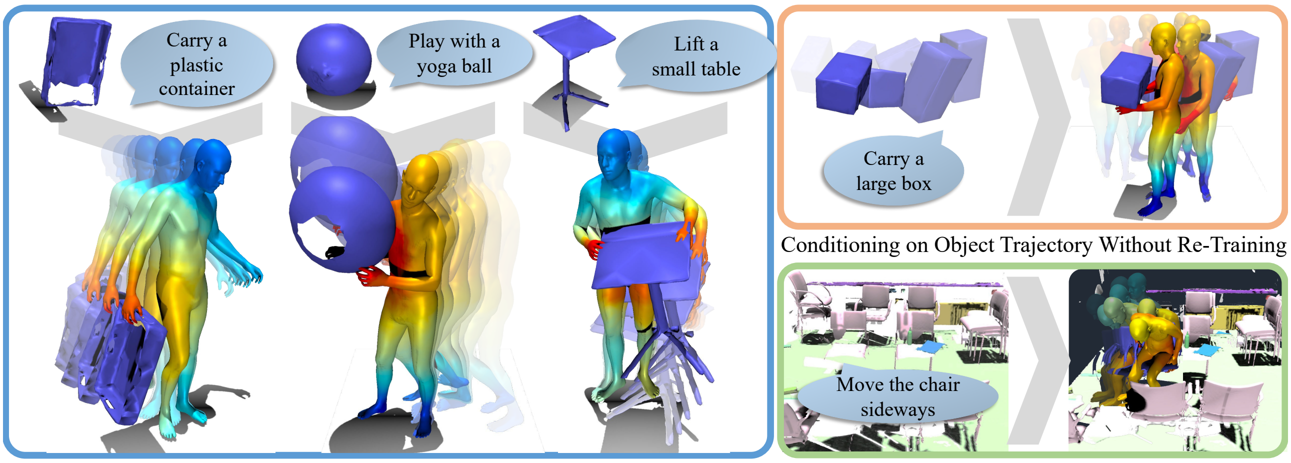
If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

CG-HOI: Contact-Guided 3D Human-Object Interaction Generation

Christian Diller
Technical University of Munich
christian.diller@tum.de

Angela Dai
Technical University of Munich
angela.dai@tum.de



Conditioning on Object Trajectory Without Re-Training

Contact-Guided 3D Human-Object Interaction Synthesis from Text

Application to Objects in Static 3D Scene Scans

Figure 1. We present an approach to generate realistic 3D human-object interactions (HOIs), from a text description and given static object geometry to be interacted with (left). Our main insight is to explicitly model contact (visualized as colors on the body mesh, closer contact in red), in tandem with human and object sequences, in a joint diffusion process. In addition to synthesizing HOIs from text, we can also synthesize human motions conditioned on given object trajectories (top right), and generate interactions in static scene scans (bottom right).

Abstract

We propose CG-HOI, the first method to address the task of generating dynamic 3D human-object interactions (HOIs) from text. We model the motion of both human and object in an interdependent fashion, as semantically rich human motion rarely happens in isolation without any interactions. Our key insight is that explicitly modeling contact between the human body surface and object geometry can be used as strong proxy guidance, both during training and inference. Using this guidance to bridge human and object motion enables generating more realistic and physically plausible interaction sequences, where the human body and corresponding object move in a coherent manner. Our method first learns to model human motion, object motion, and contact in a joint diffusion process, inter-correlated through cross-attention. We then leverage this learned contact for guidance during inference to synthesize realistic and coherent HOIs. Extensive evaluation shows that our joint contact-based human-object interaction approach generates realistic and physically plausible sequences, and we show two applications highlighting the capabilities of our method. Conditioned on a given object trajectory, we can generate the corresponding human motion without re-training, demonstrating strong human-object interdependency learning. Our approach is also flexible, and can be applied to static real-world 3D scene scans.

1. Introduction

Generating human motion sequences in 3D is important for many real-world applications, e.g. efficient realistic character animation, assistive robotic systems, room layout planning, or human behavior simulation. Crucially, human interaction is interdependent with the object(s) being interacted with; the object structure of a chair or ball, for instance, constrains the possible human motions with the object (e.g., sitting, lifting), and the human action often impacts the object motion (e.g., sitting on a swivel chair, carrying a backpack).

Existing works typically focus solely on generating dynamic humans, and thereby disregarding their surroundings [14, 17, 59, 63, 106, 109], or grounding such motion generations in a static environment that remains unchanged throughout the entire sequence [32, 37, 80, 82, 86, 103, 107, 108, 111]. However, real-world human interactions affect the environment. For instance, even when simply sitting down on a chair, the chair is typically moved: to adjust it to the needs of the interacting human, or to move it away from other objects such as a table. Thus, for realistic modeling of human-object interactions, we must consider the interdependency of object and human motions.

We present CG-HOI, the first approach to address the task of generating realistic 3D human-object interactions from text descriptions, by jointly predicting a sequence of 3D human body motion along with the object motion. Key to our approach is to not only model human and object motion, but to also explicitly model contact as a bridge between human and object. In particular, we model contact by predicting contact distances from the human body surface to the closest point on the surface of the object being interacted with. This explicit modeling of contact helps to encourage human and object motion to be semantically coherent, as well as to provide a constraint indicating physical plausibility (e.g., discouraging objects to float without support).

CG-HOI jointly models human, object, and contact together in a denoising diffusion process. Our joint diffusion model is designed to encourage information exchange between all three modalities through cross-attention blocks. Additionally, we employ a contact weighting scheme, based on the insight that object motion, when being manipulated by a human, is most defined by the motion of the body part in closest contact (Fig. 3). We make use of this by generating separate object motion hypotheses for multiple parts of the human body and aggregating them based on that part's predicted contact. During inference, we leverage the predicted contact distances to refine synthesized sequences through our contact-based diffusion guidance, which penalizes synthesizing sequences with human-object contact far from the predicted contact distances.

Our method is able to generate realistic and physically plausible human-object interactions, and we evaluate our approach on two widely-used interaction datasets, BEHAVE [9] and CHAIRS [36]. In addition, we also demonstrate the usefulness of our model with two related applications: First, generating human motion given a specific object trajectory without any retraining, which demonstrates our learned human-object motion interdependencies. Second, populating a static 3D scene scan with human-object interactions of segmented object instances, showing the applicability of our method to general real-world 3D scans.

In summary, our contributions are three-fold:

- We propose an approach to generate realistic, diverse, and physically plausible human-object interaction sequences by jointly modeling human motion, object motion, and contact through cross-attention in a diffusion process.
- We formulate a holistic contact representation: Object motion hypotheses are generated for multiple pre-defined points on the surface of the human body and aggregated based on predicted contact distances, enabling comprehensive body influence on contact while focusing on the body parts in closer contact to the object.
- We propose a contact-based guidance during synthesis of human-object interactions, leveraging predicted contacts to refine generated interactions, leading to more physically plausible results.

2. Related Work

3D Human Motion Generation. Generating sequences of 3D humans in motion is a task which evolved noticeably over the last few years. Traditionally, many methods used recurrent approaches [2, 15, 21, 23, 33, 52] and, improving both fidelity and predicted sequence length, graph- and attention-based frameworks [47, 48, 70]. Notably, generation can either happen deterministically, predicting one likely future human pose sequence [19, 21, 47, 48, 52], or stochastically, thereby also modelling the uncertainty inherent to future human motion [4, 7, 10, 18, 49, 89, 90, 95].

Recently, denoising diffusion models [66, 67] showed impressive results in 2D image generation, producing high fidelity and diverse images [31, 67]. Diffusion models allow for guidance during inference, with classifier-free guidance [8, 54] widely used to trade off between generation quality and diversity. Inspired by these advances, various methods have been proposed to model 3D human motion through diffusion, using U-Nets [14, 17, 59, 63, 106, 109], transformers [1, 59, 65, 68, 73, 74, 81, 83, 84, 91, 92, 98], or custom architectures [3, 6, 13, 16, 99]. Custom diffusion guidance has also been shown to aid controllability [34, 39, 62] and physical plausibility [96].

In addition to unconditional motion generation, conditioning on text descriptions allows for more control over the generation result [63, 73, 81, 84, 98, 109]. In fact, generating plausible and corresponding motion from textual descriptions has been an area of interest well before the popularity of diffusion models [5, 14, 26, 38, 40, 57, 97].

These methods show strong potential for 3D human motion generation, but focus on a skeleton representation of the human body, and only consider human motion in isolation, without naturally occurring interactions. To generate realistic human-object interactions, we must consider the surface of the human body and its motion with respect to object motion, which we characterize as contact.

3D Human Motion in Scenes. As human motion typically occurs not in isolation but in the context of an object or surrounding environment, various methods have explored learning plausible placement of humans into scenes, both physically and semantically, [27, 29, 30, 87, 100, 104], forecasting future motion given context [12, 50], or generating

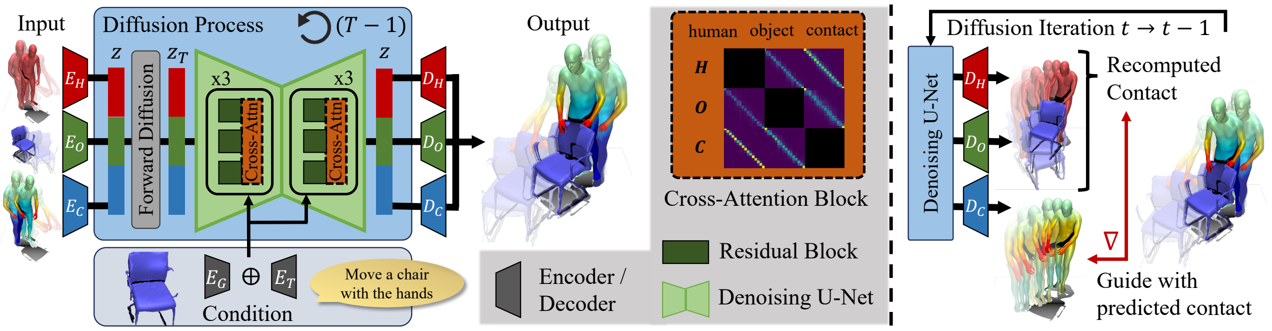


Figure 2. Method Overview. Given a text description and object geometry, CG-HOI produces a human-object interaction (HOI) sequence, modeling both human and object motion. To produce realistic HOIs, we additionally model contact to bridge the interdependent motions. Our method jointly generates all three during training (left), using a U-Net-based diffusion with cross-attention across human, object, and contact. During inference (right), we drive synthesis under guidance of estimated contact to sample more physically plausible interactions.

plausible walking and sitting animations [28, 32, 37, 78–80, 82, 86, 103, 107, 108, 111]. This enables more natural modeling of human reactions to their environment; however, the generated interactions remain limited due to the assumption of a static scene environment, resulting in a focus on walking or sitting movements.

Recent methods have also focused on more fine-grained interactions by generating human motion given a single static object [42, 43, 69, 71, 85, 101, 102]. While these methods only focus on human motion generation for a static object, [44] generates human motion conditioned on object motion and [77, 88] generate full human-object interaction sequences directly from an initial sequence observation. Our approach also models both human and object motion, but we formulate a flexible text-conditioned generative model for dynamic human and object motion, modeling the interdependency between human, object, and contact to synthesize more realistic interactions under various application settings.

Contact Prediction for Human-Object Interactions. While there is a large corpus of related work for human motion prediction, only few works focus on object motion generation [20, 53, 61, 114]. Notably, these methods predict object movement in isolation, making interactions limited, as they typically involve interdependency with human motion.

Contact prediction has been most studied in recent years for the task of fine-grained hand-object interaction [11, 22, 41, 45, 93, 110, 112]. It is defined either as binary labels on the surface [11, 22, 41, 45, 93, 110] or as the signed distance to a corresponding geometry point [112]. In these works, predicting object and hand states without correct contact leads to noticeable artifacts. Contact prediction itself has also been the focus of several works [24, 35, 75, 85], either predicting contact areas or optimizing for them.

Applied to the task of generating whole-body human-object interactions, this requires access to the full surface geometry of both object and human. Only few recent motion generation works focus on generating full-body geometric representations of humans [51, 56, 57, 72, 89, 103, 105] instead of simplified skeletons which is a first step towards physically correct interaction generation. However, while several of these works acknowledge that contact modeling would be essential for more plausible interactions [56, 57, 103], they do not model full-body contact.

We approach the task of generating plausible human-object motion from only the object geometry and a textual description as a joint task and show that considering the joint behavior of full-body human, object, and contact between the two benefits output synthesis to generate realistic human-object interaction sequences.

3. Method Overview

CG-HOI jointly generates sequences of human body and object representations, alongside contact on the human body surface. Reasoning jointly about all three modalities in both training and inference enables generation of semantically meaningful human-object interaction sequences.

Fig. 2 shows a high-level overview of our approach: We consider as condition a brief text description T of the action to be performed, along with the static geometry G of the object to be interacted with, and generate a sequence of F frames $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_F]$ where each frame $\mathbf{x}_i$ consists of representations for the object transformation o_i, for the human body surface h_i, and for the contact c_i between human and object geometry. We denote as $H = \{h_i\}$ the human body representations, $O = \{o_i\}$ the object transformations, and $C = \{c_i\}$ the contact representations.

We first train a denoising diffusion process to generate H, O, and C, using a U-Net architecture with per-modality residual blocks and cross-attention modules. Using cross-attention between human, object motion, and contact allows for effectively learning interdependencies and and feature sharing (Sec. 4). We use the generated contact to guide both training and inference: Instead of predicting one object motion hypothesis per sequence, we generate multiple, and aggregate them based on predicted contacts, such that body parts in closer contact with the object have a stronger correlation with the final object motion (Sec. 4.3). During inference, the trained model generates H, O, and C. For each step of the diffusion inference, we use predicted contact C to guide the generation of H and O, by encouraging closeness of recomputed contact and predicted contact, producing more refined and realistic interactions overall (Sec. 5).

4. Human-Object Interaction Diffusion

4.1. Probabilistic Denoising Diffusion

Our approach uses a diffusion process to jointly generate a sequence of human poses, object transformations, and contact distances in a motion sequence from isotropic Gaussian noise in an iterative process, removing more noise at each step. More specifically, during training we add noise depending on the time step ("forward process") and train a neural network to reverse this process, by directly predicting the clean sample from noisy input. Mathematically, the forward process follows a Markov chain with T steps, yielding a series of time-dependent distributions $q(\mathbf{z}_t|\mathbf{z}_{t-1})$ with noise being injected at each time step until the final distribution $\mathbf{z}_T$ is close to $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Formally,

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\sqrt{\beta_t}\mathbf{z}_{t-1} + (1-\beta_t)\mathbf{I}) \qquad (1)$$

with the variance of the Gaussian noise at time t denoted as β_t, and $\beta_0 = 0$.

Since we adopt the Denoising Diffusion Probabilistic Model [31], we can sample $\mathbf{z}_t$ directly from $\mathbf{z}_0$ as

$$\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{z}_0 + \sqrt{1-\alpha_t}\epsilon \qquad (2)$$

with $\alpha_t = \prod_{t'=0}^{t}(1-\beta_t)$, and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

For the reverse process, we follow [59, 73, 88], directly recovering the original signal $\tilde{\mathbf{z}}$ instead of the added noise.

Human-Object Interactions To model human-object interactions with diffusion, we employ our neural network formulation $\mathcal{F}$. $\mathcal{F}$ operates on the noised vector of concatenated human, object, and contact representations, together with the current time step t, and a condition consisting of object point cloud G, encoded by an encoder E_G, and text information T, encoded by encoder E_T. Formally,

$$\tilde{\mathbf{z}} = \mathcal{F}(\mathbf{z}_t, t, E_G(G) \oplus E_T(T)) \qquad (3)$$

More specifically, in our scenario E_T extracts text features with a pre-trained CLIP [60] encoder. Encoder E_G processes object geometry G as a uniformly sampled point cloud in world coordinate space with a PointNet [58] pre-trained on object parts segmentation.

Object transformations o_i are represented as global translation and rotation using continuous 6D rotation representation [113]. In contrast to prior work [18, 42, 73, 81, 95, 98, 101] which focused on representing human motion in a simplified manner as a collection of J human joints, disregarding both identity-specific and pose-specific body shape, we model physically plausible human-object contacts between body surface and geometry. Thus, we represent the human body h_i in SMPL [46] parameters: $h_i = \{h_i^p, h_i^b, h_i^r, h_i^t\}$ with pose parameters $h_i^p \in \mathbb{R}^{63}$, shape parameters $h_i^b \in \mathbb{R}^{10}$, as well as global rotation $h_i^r \in \mathbb{R}^3$ and translation $h_i^t \in \mathbb{R}^3$. These body parameters can then be converted back into a valid human body surface mesh in a differentiable manner using the SMPL [46] model. This allows us to reason about the contact between human body surface and object geometry. We represent contact c_i on the human body as the distance between a set of $M = 128$ uniformly distributed motion markers on the body surface to the closest point of the object geometry, for each marker. Specifically, we represent contact for frame $\mathbf{x}_i$ and j-th contact marker ($j \in \{0..M-1\}$) c_i^j as its distance from the human body surface to the closest point on the same frame's object geometry surface.

4.2. Human-Object-Contact Cross-Attention

We jointly predict human body sequences $H = \{h_i\}$, object transformations $O = \{o_i\}$, and corresponding contact distances $C = \{c_i\}$ in our diffusion approach. We employ a U-Net backbone for diffusion across these outputs, with separate residual blocks for human, object, and contact representations, building modality-specific latent feature representations. As we aim to model the inter-dependency across human, object, and contact, we introduce custom human-object-contact cross-attention modules after every residual block where each modality attends to the other two.

We follow the formulation of Scaled Dot-Product Attention [76], computing the updated latent human body feature:

$$h_i = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}}\right)\boldsymbol{V}, \qquad (4)$$

with query $\boldsymbol{Q} = H$, and key and value $\boldsymbol{K} = \boldsymbol{V} = O \odot C$ ($\odot$ denotes concatenation), i.e. $\boldsymbol{Q} \in \mathbb{R}^{F \times d}$ and $\boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{2F \times d}$. As in [76], d denotes the dimensionality of query and key. Applying this similarly to O and C yields the final features after each cross-attention module.

4.3. Contact-Based Object Transform Weighting

As visualized in Fig. 3, object motion is naturally most influenced by parts of the human body in very close contact to the object (as they are often the cause of that motion), and less impacted (if at all) by body parts further away. For instance, if a person moves an object with their hands, the object follows the hands but not necessarily other body parts (e.g., body and feet may remain static or walk in a different direction). Thus, instead of directly generating one object motion hypothesis o_i alongside the corresponding

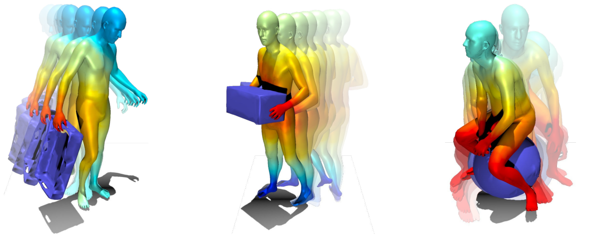


Figure 3. An object's trajectory is largely defined by the motion of the region of the body in close contact with the object, e.g. the hand(s) when carrying an object (left, middle) or the lower body when moving with an object while sitting (right). This informs our contact-based approach to generating object motion.

human motion h_i, we couple o_i to the M body contact points $j \in \{0..M-1\}$ and their predicted distances $\{c_i^j\}$ between human body surface and object geometry.

Formally, we predict object transformation hypotheses o_i^j for each contact point on the human body, and weigh them with the inverse of their predicted contact distance c_i^j:

$$o_i = \frac{1}{\sum_j \max(|c_i|) - |c_i^j|} \sum_{j=0}^{M-1} (\max(|c_i|) - |c_i^j|) o_i^j \quad (5)$$

4.4. Loss Formulation

During training, the input is a noised vector $\mathbf{z}$, containing F frames $\{\mathbf{x}_i\}$, each a concatenation of human body representation h_i, object transformation o_i, and contact parameters c_i. As condition $\mathbf{C}$, we additionally input encoded object geometry G and text description T. The training process is then supervised with the ground-truth sequence containing $\hat{h}_i, \hat{o}_i, \hat{c}_i$, minimizing a common objective:

$$\mathbf{L} = \lambda_h ||h_i - \hat{h}_i||_1 + \lambda_o ||o_i - \hat{o}_i||_1 + \lambda_c ||c_i - \hat{c}_i||_2, \quad (6)$$

with $\lambda_h = 1.0, \lambda_o = 0.9, \lambda_c = 0.9$. We use classifier-free guidance [8] for improved fidelity during inference, thus masking out the conditioning signal with 10% probability.

5. Interaction Generation

Using our trained network model, we can generate novel human-object interaction sequences for a given object geometry and a short text description using our weighting scheme for generating object transformations, and a custom guidance function on top of classifier-free guidance to generate physically plausible sequences.

Specifically, we use our trained model to reverse the forward diffusion process of Eq. 2: Starting with noised sample $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we iteratively use our trained network model $\mathcal{F}$ to estimate cleaned sample $\mathbf{z}_0$:

$$\mathbf{z}_{t-1} = \sqrt{\alpha_t}\tilde{\mathbf{z}} + \sqrt{1-\alpha_t}\epsilon. \quad (7)$$

5.1. Contact-Based Diffusion Guidance

While our joint human-object-contact training already leads to plausible motions, generated sequences are not explicitly constrained to respect contact estimates during inference, which can lead to inconsistent contact between human and object motion (e.g., floating objects). Thus, we introduce a contact-based guidance during inference to refine predictions, using a cost function $\mathcal{G}(\mathbf{z}_t) = ||c_t - \overline{c}_t||_2^2$ which takes as input the denoised human, object, and contact predictions $\mathbf{z}_t = [h_t, o_t, c_t]$ at diffusion step t and compares predicted c_t and actual contact distances $\overline{c}_t$ for each contact point. Based on this, we then calculate the gradient $\nabla_{\mathbf{z}_t}\mathcal{G}(\mathbf{z}_t)$.

We use this gradient for diffusion guidance, following [39], by re-calculating the mean prediction μ_t at each time t:

$$\hat{\mu}_t = \mu_t + s\sum_t \nabla_{x_t}\mathcal{G}(x_t), \quad (8)$$

for a scaling factor s. This guidance is indirect but dense in time, and is able to correct physical contact inconsistencies in the predicted sequences during inference time, without requiring any explicit post-processing steps.

5.2. Conditioning on Object Trajectory

While our model has been trained with text and static object geometry as condition, we can also apply the same trained model for conditional generation of a human sequence given an object sequence and text description. Note that this does not require any re-training, as our model has learned a strong correlation between human and object motion. Instead, we use a replacement-based approach, and inject the given object motion O' into the diffusion process during inference at every step. Following Eq. 7, we obtain:

$$\mathbf{z}_{t-1} = \sqrt{\alpha_t}\tilde{\mathbf{z}}_t' + \sqrt{1-\alpha_t}\epsilon, \quad (9)$$

with $\tilde{\mathbf{z}}' = [h_t, o_t', c_t]$, concatenating human motion h_t, contact distances c_t, and injected given object motion o_t'.

6. Results

We evaluate our approach using two commonly used human-object interaction datasets BEHAVE [9] and CHAIRS [36] on a range of metrics, measuring motion fidelity and diversity. We show that our approach is able to generate realistic and diverse motion on both datasets, across a variety of objects and types of interactions.

6.1. Experimental Setup

Datasets We conduct our experiments on two datasets containing interactions between whole-body 3D humans and corresponding objects. CHAIRS [36] captures 46 subjects as their SMPL-X [55] bodies interacting with 81 different types of chairs and sofas. We extract sequences in which

		BEHAVE				CHAIRS			
Task	Approach	R-Prec. (top-3) ↑	FID ↓	Diversity →	MModality →	R-Prec. (top-3) ↑	FID ↓	Diversity →	MModality →
	Real (human)	0.73	0.09	4.23	4.55	0.83	0.01	7.34	3.00
Text-Cond. Human Only	MDM [73]	0.52	4.54	5.44	5.12	0.72	5.99	**6.83**	3.45
	InterDiff [88]	0.49	5.36	**3.98**	3.98	0.63	6.76	5.24	2.44
	Ours	**0.60**	**4.26**	4.92	**4.10**	**0.78**	**5.24**	7.90	**3.22**
	Real	0.81	0.17	6.80	6.24	0.87	0.02	9.91	6.12
Motion-Cond. HOI	InterDiff [88]	0.68	3.86	5.62	5.90	0.67	4.83	7.49	4.87
	Ours	**0.71**	**3.52**	**6.89**	**6.43**	**0.79**	**4.01**	**8.42**	**6.29**
Text-Cond. HOI	MDM [73]	0.49	9.21	6.51	8.19	0.53	9.23	6.23	7.44
	InterDiff [88]	0.53	8.70	3.85	4.23	0.69	7.53	5.23	4.63
	Ours	**0.62**	**6.31**	**6.63**	**5.47**	**0.74**	**6.45**	**8.91**	**5.94**

Table 1. Quantitative comparison with state-of-the-art approaches MDM [73] and InterDiff [88]. Human Only results are evaluated only on the human pose sequence, and motion-cond. denotes predictions additionally conditioned on past observations of both human and object behavior. For metrics with →, results closer to the real distribution are better. Our approach outperforms these baselines in all three settings, indicating a strong learned correlation between human and object motion.

both human and object are in motion, yielding ≈ 1300 HOI sequences, each labeled with a text description. We use a random 80/10/10 split along object classes, ensuring that test objects are not seen during training. BEHAVE [9] captures 8 participants as their SMPL-H [64] parameters alongside 20 different objects. This yields ≈ 520 sequences with corresponding text descriptions. We use their original train/test split. We sample both datasets at 20 frames per second, and generate 32 frames for CHAIRS and 64 for BEHAVE, leading to generated motion that lasts up to 3 seconds.

Implementation Details We train our model with batch size 64 for 600k steps (≈24 hours), after which we choose the checkpoint that minimizes validation FID, following [88]. Our attention uses 4 heads and a latent dimension of 256. Input text is encoded using a frozen CLIP-ViT-B/32 model. For classifier-free guidance during inference time, we use a guidance scale of 2.5, which empirically provides a good trade-off between diversity and fidelity. For our inference-time contact-based guidance, we use scale $s = 100.0$.

6.2. Evaluation Metrics

We measure realism and diversity of combined human and object motion, alongside closeness to the text description, following established practices [25, 26, 73]. We first train a joint human-object motion feature extractor and separate text feature extractor using a contrastive loss to produce geometrically close feature vectors $\{v_i\}$ for matched text-motion pairs, and report the following metrics:

R-Precision measures the closeness of the text condition and generated HOI in latent feature space, and reports whether the correct match falls in the top 3 closest feature vectors.

Frechet Inception Distance (FID) is commonly used to evaluate the similarity between generated and ground-truth distribution in encoded feature space.

Diversity and MultiModality. Diversity measures the motion variance across all text descriptions and is defined as $\frac{1}{N}\sum_{i=1}^{N}||v_i - v_i'||_2$ between two randomly drawn subsets $\{v_i\}$ and $\{v_i'\}$. MultiModality (MModality) measures the average such variance intra-class, for each text description.

Perceptual User Study. The exact perceptual quality of human-object interactions is difficult to capture with any single metric; thus, we additionally conduct a user study with 32 participants to evaluate our method in comparison to baseline approaches. Participants are shown 10 baseline vs. ours pairs each in side-by side views of sequences with the same geometry and text conditioning, and asked to choose 1) Which one follows the given text better and 2) Which one looks more realistic overall.

6.3. Comparison to Baselines

As our method is the first to enable generating human and object motion from text, there are no baselines available for direct comparison. InterDiff [88] is closest to our approach, performing forecasting from observed human and object motion as input and predicting a plausible continuation. In Tab. 1, we compare to ours first in their setting, using observed motion as condition (motion-cond.), for a fair comparison. Additionally, we modify their approach by replacing observed motion encoders with our text encoder, allowing for a comparison in our setting (text-cond.). We also compare with MDM [73], a state-of-the-art method for human-only sequence generation from text, both in their original setting, only predicting human sequences, and extending theirs to also generate object sequences, by adding additional tokens and geometry conditioning to their transformer encoder formulation. For more details of baseline setup, we refer to the appendix. We evaluate the quality of generated human-object interactions as well as human-only generation, only evaluating the human sequence for our method, as compared to the generated sequences of MDM.

Both Tab. 1 and the user study in Fig. 5 show that our approach is able to generate more realistic and physically plausible human-object interaction sequences than baselines. In Fig. 4, we see that our approach synthesizes more meaningful human-object interaction with respect to contact and mitigating independent object floating.

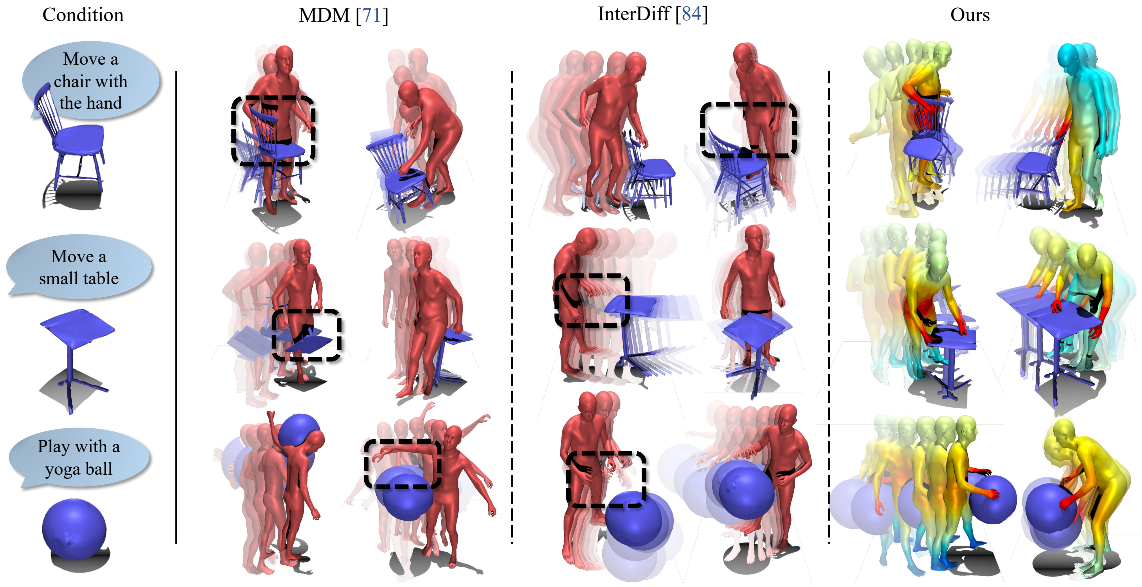


Figure 4. Qualitative comparison to state-of-the-art methods MDM [73] and InterDiff [88]. Our approach generates high-quality HOIs by jointly modeling contact (closer contact in red), reducing penetration and floating artifacts (black highlight boxes).

6.4. Ablation Studies

Cross-attention enables learning human-object interdependencies. Tab. 2 shows that our human-object-contact cross-attention (Sec. 4.2) significantly improves performance by effectively sharing information between human, contact, and object sequence modalities. In Fig. 6, we see this encourages realistic contact between human and object.

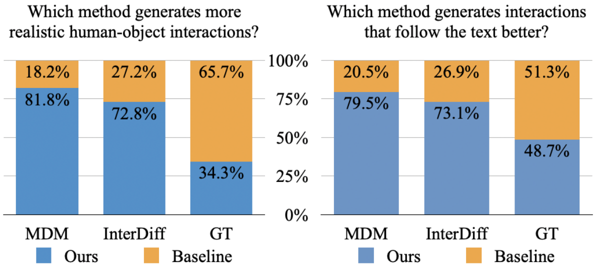


Figure 5. Perceptual User Study. Participants significantly favor our method over baselines, for overall realism and text coherence.

Contact prediction improves HOI generation performance. Predicting contact (Sec. 5) is crucial to generating more realistic human-object sequences, resulting in more realistic interactions between human and object (Fig. 6), and improved fidelity (Tab. 2). Notably, learning contact jointly with human and object motion improves overall quality, compared to a separately trained contact model used for inference guidance ("Separate contact pred.", Tab. 2).

Contact-based object transformation weighting improves generation performance. Weighting predicted object motion hypotheses with predicted contact (Sec. 4.3) improves HOI generation over naive object sequence prediction, both quantitatively in Tab. 2 ("No contact weighting") and visually as realistic human-object interactions in Fig. 6.

Contact-based guidance during inference helps produce physically plausible interactions. As shown in Fig. 6 and Tab. 2, using our guidance based on predicted contacts leads to a higher degree of fidelity and physical plausibility.

	BEHAVE				CHAIRS			
Approach	R-Prec. (top-3) ↑	FID ↓	Diversity →	MModality →	R-Prec. (top-3) ↑	FID ↓	Diversity →	MModality →
Real	0.81	0.17	6.80	6.24	0.87	0.02	9.91	6.12
No cross-attention	0.35	10.44	8.23	7.40	0.49	10.84	12.22	10.64
No contact prediction	0.41	9.64	10.10	**6.89**	0.41	8.53	11.56	9.15
Separate contact pred.	0.47	8.01	5.12	5.12	0.52	9.34	7.65	4.62
No contact weighting	0.55	8.54	6.52	5.29	0.64	7.55	8.56	5.45
No contact guidance	0.59	7.22	7.84	5.30	0.70	7.41	8.05	5.76
Ours	**0.62**	**6.31**	**6.63**	5.47	**0.74**	**6.74**	**8.91**	**5.94**

Table 2. Ablation on our design choices. Joint contact prediction with cross-attention encourages the generation of more natural HOIs, and our weighting scheme and inference-time contact guidance together enable the best generation performance.

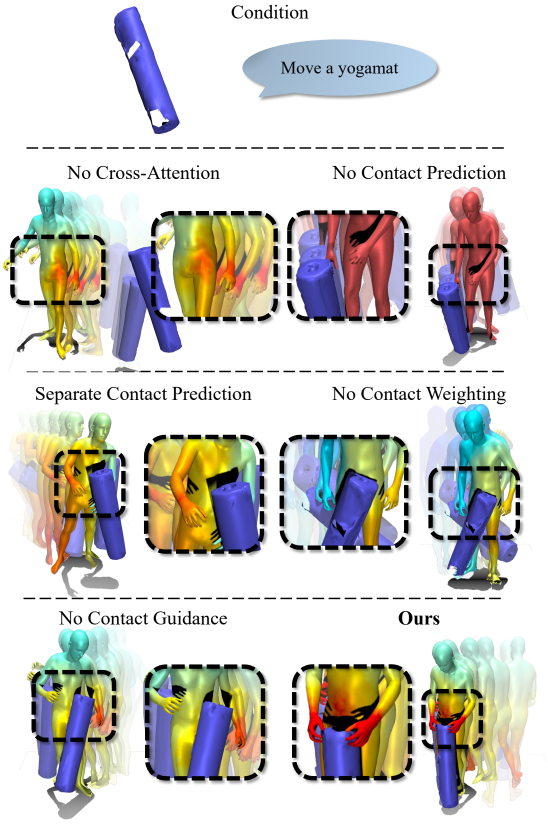


Figure 6. Visualization of ablations of our method design: Generation, weighting, and inference-time guidance work together to enable realistic interactions in our method, resolving artifacts such as object floating.

6.5. Applications

Human motion generation given object trajectory. Our approach can be directly applied to conditionally generate human sequences given object sequences as condition, as shown in Fig. 7. As our model learns a strong correspondence between object and human motion, facilitated by contact distance predictions, we are able to condition without any additional training.

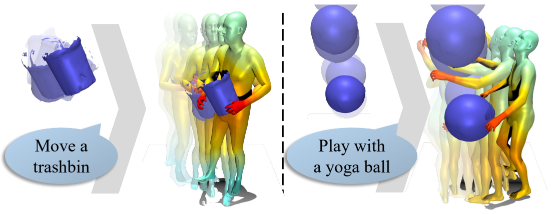


Figure 7. Given an object trajectory at inference time, our method can generate corresponding human motion without re-training.

Populating 3D scans. Fig. 8 shows that we can also apply our method to generate human-object interactions in static scene scans. Here, we use a scene from the ScanNet++ dataset [94], with their existing semantic object segmentation. This enables the potential to generate realistic human motion sequences only given a static scene environment.

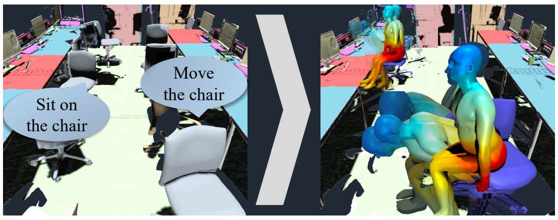


Figure 8. Application to static 3D scene scans. Our method can generate HOIs from segmented objects in such environments.

6.6. Limitations

While we have demonstrated the usefulness of joint contact prediction in 3D HOI generation, several limitations remain. For instance, our method focuses on realistic interactions with a single object. We show that this can be applied to objects in static 3D scans; however, we do not model multiple objects together, which could have the potential to model more complex long-term human behavior (e.g. cooking sequences). Additionally, our method requires expensive 3D HOI captures for training; a weakly supervised approach leveraging further supervision from 2D action data might be able to represent more diverse scenarios. Similarly, our method depends on manual text annotations; more specific prompts might lead to more control over generated HOIs.

7. Conclusion

We propose an approach to generating realistic, dynamic human-object interactions based on contact modeling. Our diffusion model effectively learns interdependencies between human, object, and contact through cross-attention along with our contact-based object transformation weighting. Our predicted contacts further facilitate refinement using custom diffusion guidance, generating diverse, realistic interactions based on text descriptions. Since our model learns a strong correlation between human and object sequences, we can use it to conditionally generate human motion from given object sequences. Extensive experimental evaluation confirms both fidelity and diversity of our generated sequences and shows improved performance compared to baselines.

Acknowledgements

This project is funded by the ERC Starting Grant SpatialSem (101076253), the Bavarian State Ministry of Science and the Arts coordinated by the Bavarian Research Institute for Digital Transformation (bidt), and the German Research Foundation (DFG) Grant "Learning How to Interact with Scenes through Part-Based Understanding".

References

[1] Hyemin Ahn, Esteve Valls Mascaro, and Dongheui Lee. Can we use diffusion probabilistic models for 3d motion prediction? In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*, pages 9837–9843. IEEE, 2023. 2

[2] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7143–7152. IEEE, 2019. 2

[3] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Trans. Graph.*, 42(4):44:1–44:20, 2023. 2

[4] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5222–5231. Computer Vision Foundation / IEEE, 2020. 2

[5] Samaneh Azadi, Akbar Shah, Thomas Hayes, Devi Parikh, and Sonal Gupta. Make-an-animation: Large-scale text-conditional 3d human motion generation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 14993–15002. IEEE, 2023. 2

[6] German Barquero, Sergio Escalera, and Cristina Palmero. Belfusion: Latent diffusion for behavior-driven human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2317–2327, 2023. 2

[7] Emad Barsoum, John R. Kender, and Zicheng Liu. HP-GAN: probabilistic 3d human motion prediction via GAN. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1418–1427. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[8] Gulcin Baykal, Halil Faruk Karagoz, Taha Binhuraib, and Gozde Unal. Protodiffusion: Classifier-free diffusion guidance with prototype learning. *CoRR*, abs/2307.01924, 2023. 2, 5

[9] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: dataset and method for tracking human object interactions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15914–15925. IEEE, 2022. 2, 5, 6, 17, 18

[10] Apratim Bhattacharyya, Bernt Schiele, and Mario Fritz. Accurate and diverse sampling of sequences based on a "best of many" sample objective. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8485–8493. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[11] Jona Braun, Sammy Joe Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. *CoRR*, abs/2309.07907, 2023. 3

[12] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, pages 387–404. Springer, 2020. 2

[13] Ling-Hao Chen, Jiawei Zhang, Yewen Li, Yiren Pang, Xiaobo Xia, and Tongliang Liu. Humanmac: Masked motion completion for human motion prediction. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 9510–9521. IEEE, 2023. 2

[14] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 18000–18010. IEEE, 2023. 1, 2

[15] Hsu-Kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*, pages 1423–1432. IEEE, 2019. 2

[16] Jeongjun Choi, Dongseok Shim, and H. Jin Kim. Diffupose: Monocular 3d human pose estimation via denoising diffusion probabilistic model. In *IROS*, pages 3773–3780, 2023. 2

[17] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 9760–9770. IEEE, 2023. 1, 2

[18] Christian Diller, Thomas A. Funkhouser, and Angela Dai. Forecasting characteristic 3d poses of human actions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15893–15902. IEEE, 2022. 2, 4

[19] Christian Diller, Thomas A. Funkhouser, and Angela Dai. Futurehuman3d: Forecasting complex long-term 3d human behavior from video observations. *CoRR*, abs/2211.14309, 2022. 2

[20] Danny Driess, Zhiao Huang, Yunzhu Li, Russ Tedrake, and Marc Toussaint. Learning multi-object dynamics with compositional neural radiance fields. In *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, pages 1755–1768. PMLR, 2022. 3

[21] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4346–4354. IEEE Computer Society, 2015. 2

[22] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven

full-body motion synthesis for human-object interactions. *Comput. Graph. Forum*, 42(2):1–12, 2023. 3

[23] Anand Gopalakrishnan, Ankur Arjun Mali, Dan Kifer, C. Lee Giles, and Alexander G. Ororbia II. A neural temporal model for human motion prediction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12116–12125. Computer Vision Foundation / IEEE, 2019. 2

[24] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. Contactopt: Optimizing contact to improve grasps. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 1471–1481. Computer Vision Foundation / IEEE, 2021. 3

[25] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 2021–2029. ACM, 2020. 6, 18

[26] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5142–5151. IEEE, 2022. 2, 6, 15, 17, 18

[27] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2282–2292. IEEE, 2019. 2

[28] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J. Black. Stochastic scene-aware motion prediction. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11354–11364. IEEE, 2021. 3

[29] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3d scenes by learning human-scene interaction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14708–14718. Computer Vision Foundation / IEEE, 2021. 2, 15

[30] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael J. Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *ACM SIGGRAPH 2023 Conference Proceedings, SIGGRAPH 2023, Los Angeles, CA, USA, August 6-10, 2023*, pages 63:1–63:9. ACM, 2023. 2

[31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2, 4

[32] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 16750–16761. IEEE, 2023. 2, 3

[33] Ashesh Jain, Amir R. Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5308–5317. IEEE Computer Society, 2016. 2

[34] Michael Janner, Yilun Du, Joshua B. Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pages 9902–9915. PMLR, 2022. 2

[35] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11087–11096. IEEE, 2021. 3

[36] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 9331–9342. IEEE, 2023. 2, 5, 17, 18

[37] James F. Mullen Jr., Divya Kothandaraman, Aniket Bera, and Dinesh Manocha. Placing human animations into 3d scenes by learning interaction- and geometry-driven keyframes. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pages 300–310. IEEE, 2023. 2, 3

[38] Sai Shashank Kalakonda, Shubh Maheshwari, and Ravi Kiran Sarvadevabhatla. Action-gpt: Leveraging large-scale language models for improved and generalized zero shot action generation. *CoRR*, abs/2211.15603, 2022. 2

[39] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. GMD: controllable human motion synthesis via guided diffusion models. *CoRR*, abs/2305.12577, 2023. 2, 5

[40] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. FLAME: free-form language-based motion synthesis & editing. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 8255–8263. AAAI Press, 2023. 2

[41] Paul G. Kry and Dinesh K. Pai. Interaction capture and synthesis. *ACM Trans. Graph.*, 25(3):872–880, 2006. 3

[42] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas J. Guibas. NIFTY: neural object interaction fields for guided human motion synthesis. *CoRR*, abs/2307.07511, 2023. 3, 4

[43] Jiye Lee and Hanbyul Joo. Locomotion-action-manipulation: Synthesizing human-scene interactions in complex 3d environments. In *IEEE/CVF International Conference on Com-*

puter Vision, ICCV 2023, Paris, France, October 1-6, 2023, pages 9629–9640. IEEE, 2023. 3

[44] Jiaman Li, Jiajun Wu, and C. Karen Liu. Object motion guided human motion synthesis. *ACM Trans. Graph.*, 42(6): 197:1–197:11, 2023. 3

[45] Quanzhou Li, Jingbo Wang, Chen Change Loy, and Bo Dai. Task-oriented human-object interactions generation with implicit neural representations. *CoRR*, abs/2303.13129, 2023. 3

[46] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, 2015. 4

[47] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9488–9496. IEEE, 2019. 2

[48] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIV*, pages 474–489. Springer, 2020. 2

[49] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Generating smooth pose sequences for diverse human motion prediction. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 13289–13298. IEEE, 2021. 2

[50] Wei Mao, Miaomiao Liu, Richard I. Hartley, and Mathieu Salzmann. Contact-aware human motion forecasting. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 2

[51] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Weakly-supervised action transition learning for stochastic human motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 8141–8150. IEEE, 2022. 3

[52] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4674–4683. IEEE Computer Society, 2017. 2

[53] Damian Mrowca, Chengxu Zhuang, Elias Wang, Nick Haber, Li Fei-Fei, Josh Tenenbaum, and Daniel L. K. Yamins. Flexible neural representation for physics prediction. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8813–8824, 2018. 3

[54] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pages 16784–16804. PMLR, 2022. 2

[55] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10975–10985. Computer Vision Foundation / IEEE, 2019. 5, 18

[56] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer VAE. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 10965–10975. IEEE, 2021. 3

[57] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: generating diverse human motions from textual descriptions. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXII*, pages 480–497. Springer, 2022. 2, 3

[58] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 77–85. IEEE Computer Society, 2017. 4, 18

[59] Sigal Raab, Inbal Leibovitch, Guy Tevet, Moab Arar, Amit H. Bermano, and Daniel Cohen-Or. Single motion diffusion. *CoRR*, abs/2302.05905, 2023. 1, 2, 4

[60] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 4

[61] Davis Rempe, Srinath Sridhar, He Wang, and Leonidas J. Guibas. Predicting the physical dynamics of unseen 3d objects. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 2823–2832. IEEE, 2020. 3

[62] Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 13756–13766. IEEE, 2023. 2

[63] Zhiyuan Ren, Zhihong Pan, Xin Zhou, and Le Kang. Diffusion motion: Generate text-guided 3d human motion by diffusion model. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE, 2023. 1, 2

[64] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 36(6):245:1–245:17, 2017. 6, 18

[65] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H. Bermano. Human motion diffusion as a generative prior. *CoRR*, abs/2303.01418, 2023. 2

[66] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 2256–2265. JMLR.org, 2015. 2

[67] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2

[68] Jiarui Sun and Girish Chowdhary. Towards globally consistent stochastic human motion prediction via motion diffusion. *CoRR*, abs/2305.12554, 2023. 2

[69] Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: generating 4d whole-body motion for hand-object grasping. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 13253–13263. IEEE, 2022. 3

[70] Yongyi Tang, Lin Ma, Wei Liu, and Wei-Shi Zheng. Long-term human motion prediction by modeling motion context and enhancing motion dynamics. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 935–941. ijcai.org, 2018. 2

[71] Purva Tendulkar, Dídac Surís, and Carl Vondrick. FLEX: full-body grasping without full-body grasps. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 21179–21189. IEEE, 2023. 3

[72] Guy Tevet, Brian Gordon, Amir Hertz, Amit H. Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to CLIP space. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXII*, pages 358–374. Springer, 2022. 3

[73] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 2, 4, 6, 7, 17, 18

[74] Sibo Tian, Minghui Zheng, and Xiao Liang. Transfusion: A practical and effective transformer-based diffusion model for 3d human motion prediction. *CoRR*, abs/2307.16106, 2023. 2

[75] Tze Ho Elden Tse, Zhongqun Zhang, Kwang In Kim, Ales Leonardis, Feng Zheng, and Hyung Jin Chang. S^{2}contact: Graph-based network for 3d hand-object contact estimation with semi-supervised learning. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part I*, pages 568–584. Springer, 2022. 3

[76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 4

[77] Weilin Wan, Lei Yang, Lingjie Liu, Zhuoying Zhang, Ruixing Jia, Yi-King Choi, Jia Pan, Christian Theobalt, Taku Komura, and Wenping Wang. Learn to predict how humans manipulate large-sized objects from interactive motions. *IEEE Robotics Autom. Lett.*, 7(2):4702–4709, 2022. 3

[78] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9401–9411. Computer Vision Foundation / IEEE, 2021. 3

[79] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12206–12215. Computer Vision Foundation / IEEE, 2021.

[80] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 20428–20437. IEEE, 2022. 2, 3

[81] Yin Wang, Zhiying Leng, Frederick W. B. Li, Shun-Cheng Wu, and Xiaohui Liang. Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 21978–21987. IEEE, 2023. 2, 4

[82] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. HUMANISE: language-conditioned human motion generation in 3d scenes. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 2, 3

[83] Dong Wei, Huaijiang Sun, Bin Li, Jianfeng Lu, Weiqing Li, Xiaoning Sun, and Shengxiang Hu. Human joint kinematics diffusion-refinement for stochastic motion prediction. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 6110–6118. AAAI Press, 2023. 2

[84] Dong Wei, Xiaoning Sun, Huaijiang Sun, Bin Li, Shengxiang Hu, Weiqing Li, and Jianfeng Lu. Understanding text-driven motion synthesis with keyframe collaboration via diffusion models. *CoRR*, abs/2305.13773, 2023. 2

[85] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. SAGA: stochastic whole-body grasping with contact. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October*

23-27, 2022, Proceedings, Part VI, pages 257–274. Springer, 2022. 3

[86] Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. *CoRR*, abs/2309.07918, 2023. 2, 3

[87] Zhaoming Xie, Jonathan Tseng, Sebastian Starke, Michiel van de Panne, and C. Karen Liu. Hierarchical planning and control for box loco-manipulation. *Proc. ACM Comput. Graph. Interact. Tech.*, 6(3):31:1–31:18, 2023. 2

[88] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14928–14940, 2023. 3, 4, 6, 7, 17

[89] Sirui Xu, Yu-Xiong Wang, and Liangyan Gui. Stochastic multi-person 3d motion forecasting. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 2, 3

[90] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. MT-VAE: learning motion transformations to generate multimodal human dynamics. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V*, pages 276–293. Springer, 2018. 2

[91] Siqi Yang, Zejun Yang, and Zhisheng Wang. Longdancediff: Long-term dance generation with conditional diffusion model. *CoRR*, abs/2308.11945, 2023. 2

[92] Zhao Yang, Bing Su, and Ji-Rong Wen. Synthesizing long-term human motions with diffusion models via coherent sampling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 3954–3964. ACM, 2023. 2

[93] Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 22479–22489. IEEE, 2023. 3

[94] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 8

[95] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IX*, pages 346–364. Springer, 2020. 2, 4

[96] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16010–16021, 2023. 2

[97] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2M-GPT: generating human motion from textual descriptions with discrete representations. *CoRR*, abs/2301.06052, 2023. 2

[98] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *CoRR*, abs/2208.15001, 2022. 2, 4

[99] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 364–373. IEEE, 2023. 2

[100] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. PLACE: proximity learning of articulation and contact in 3d environments. In *8th International Conference on 3D Vision, 3DV 2020, Virtual Event, Japan, November 25-28, 2020*, pages 642–651. IEEE, 2020. 2, 15

[101] Wanyue Zhang, Rishabh Dabral, Thomas Leimkühler, Vladislav Golyanik, Marc Habermann, and Christian Theobalt. ROAM: robust and object-aware motion generation using neural pose descriptors. *CoRR*, abs/2308.12969, 2023. 3, 4

[102] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. COUCH: towards controllable human-chair interactions. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part V*, pages 518–535. Springer, 2022. 3

[103] Yan Zhang and Siyu Tang. The wanderings of odysseus in 3d scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 20449–20459. IEEE, 2022. 2, 3

[104] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6194–6204, 2020. 2, 15

[105] Yan Zhang, Michael J. Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3372–3382. Computer Vision Foundation / IEEE, 2021. 3

[106] Zihan Zhang, Richard Liu, Kfir Aberman, and Rana Hanocka. Tedi: Temporally-entangled diffusion for long-term motion synthesis. *CoRR*, abs/2307.15042, 2023. 1, 2

[107] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VI*, pages 311–327. Springer, 2022. 2, 3

[108] Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, and Siyu Tang. Synthesizing diverse human motions in 3d indoor scenes. In *IEEE/CVF International Conference on*

Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023, pages 14692–14703. IEEE, 2023. 2, 3

[109] Mengyi Zhao, Mengyuan Liu, Bin Ren, Shuling Dai, and Nicu Sebe. Modiff: Action-conditioned 3d motion generation with denoising diffusion probabilistic models. *CoRR*, abs/2301.03949, 2023. 1, 2

[110] Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. CAMS: canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 585–594. IEEE, 2023. 3

[111] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C. Karen Liu, and Leonidas J. Guibas. GIMO: gaze-informed human motion prediction in context. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIII*, pages 676–694. Springer, 2022. 2, 3

[112] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. TOCH: spatio-temporal object-to-hand correspondence for motion refinement. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part III*, pages 1–19. Springer, 2022. 3

[113] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5745–5753. Computer Vision Foundation / IEEE, 2019. 4

[114] Guangxiang Zhu, Zhiao Huang, and Chongjie Zhang. Object-oriented dynamics predictor. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9826–9837, 2018. 3

Appendix

We show in this appendix additional qualitative (Sec. A) and quantitative (Sec. B) results, detail our baseline evaluation protocol (Sec. C), elaborate on the metrics used in the main paper (Sec. D), show the architecture used in our approach (Sec. E), and provide additional details regarding the data (Sec. F).

A. Additional Qualitative Results

A.1. Additional Interactions

We show additional generated 3D human-object interactions of our method in Fig. 10, with object geometry and text condition on the left, and our generated sequence on the right.

A.2. Same Prompt, Different Interactions

We evaluate the ability of our method to generate diverse interactions for a fixed text condition visually in Fig. 9, for text prompt "Move a stool" and "Sit on a stool". In the ground truth training data, move is only done with one or two hands, and feet; moving with the butt sometimes occurs for the text description "Sit on a stool".

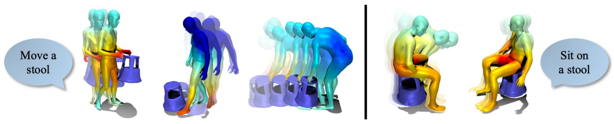


Figure 9. Our method is able to generate diverse human-object interactions for the same prompts.

B. Additional Quantitative Results

B.1. Evaluating Penetrations and Floating

Our method discourages penetration and floating implicitly, by enforcing correct contact distances as a soft constraint at train and test time. However, the exact fidelity and diversity of our results is hard to capture with any single metric. Thus, we evaluate multiple such metrics in the main paper (R-Precision, FID, Diversity, MultiModality), and conduct a perceptual user study to verify the metrics' expressiveness.

Here, we provide an additional evaluation based on intuitive physics-based metrics: Tab. 3 evaluates the mean ratio of frames with some penetration as well as the ratio of penetrating vertices overall, showing that penetrations typically happens with small body parts (e.g., hands, which also occurs in the ground-truth data). We also evaluate the ratio of frames and vertices with human and object not in contact, including floating and stationary objects, which is expected to be close to the dataset ratio.

Results show similar penetration and floating between our generations and ground-truth training data.

	BEHAVE				CHAIRS			
	Penetration		Non-Contact		Penetration		Non-Contact	
	Frames	Vertices	Frames	Vertices	Frames	Vertices	Frames	Vertices
Dataset	32.9%	4.1%	21.4%	86.2%	26.9%	1.1%	11.9%	70.4%
Ours	31.3%	3.0%	17.8%	93.3%	35.8%	4.2%	14.1%	74.3%

Table 3. Penetration and non-contact (including floating) ratios in terms of frames as well as overall vertices vs ground-truth data.

B.2. Evaluating Contact

Tab. 4 evaluates our contact predictions using precision/recall and distance metrics. We follow [29, 100, 104] to define contact if ≤5cm from object. We also report mean ℓ_1 error in contact distance predictions. All metrics are reported for body parts $\leq 1m$ of the object, to focus on contact scenarios. Better contact prediction corresponds with better HOI generations. Note that none of our baselines predict contact distances.

	BEHAVE			CHAIRS		
Approach	Precision ↑	Recall ↑	Distance ↓	Precision ↑	Recall ↑	Distance ↓
Separate contact pred.	23.4%	25.6%	0.53	58.6%	49.1%	0.24
No contact weighting	29.5%	33.5%	0.34	60.6%	63.4%	0.10
No contact guidance	46.3%	39.2%	0.31	64.2%	70.2%	0.12
Ours	**63.6%**	**59.5%**	**0.07**	**78.3%**	**84.5%**	**0.04**

Table 4. Evaluation of predicted contact distances, in terms on precision and recall ($\leq 5cm$ distance), as well as mean contact ℓ_1 error in meters.

B.3. Novelty of Generated Interactions

We perform an additional interaction novelty analysis to verify that our method does not simply retrieve memorized train sequences but is indeed able to generate novel human-object interactions. To do so, we generate $\approx$ 500 sequences from both datasets and retrieve the top-3 most similar train sequences, as measured by the l_2 distance in human body and object transformation parameter space.

Fig. 11 shows the top-3 closest train sequences, along with a histogram of l_2 distances computed on our test set of $\approx$ 500 generated sequences. In red, we mark the intra-trainset distance between samples in the train set. We observe that the distance between our generated sequences and the closest train sequence is mostly larger than the intra-train distance. Thus, our method is able to produce samples that are novel and not simply retrieved train sequences.

B.4. SMPL Bodies vs. HumanML3D Skeletons

We observe slight pose jitter and foot skating in our ground-truth training data (especially BEHAVE, captured with Kinect sensors). As a result, our model reflects some of these effects. Skeleton representations such as HumanML3D [26] could tackle these artifacts, but do not work with contact as effectively as SMPL bodies. Nevertheless, we train ours with HumanML3D parameters for comparison in Tab. 5 (fitting SMPL after for comparable evaluation) which leads to degraded performance due to less effective contact guidance.

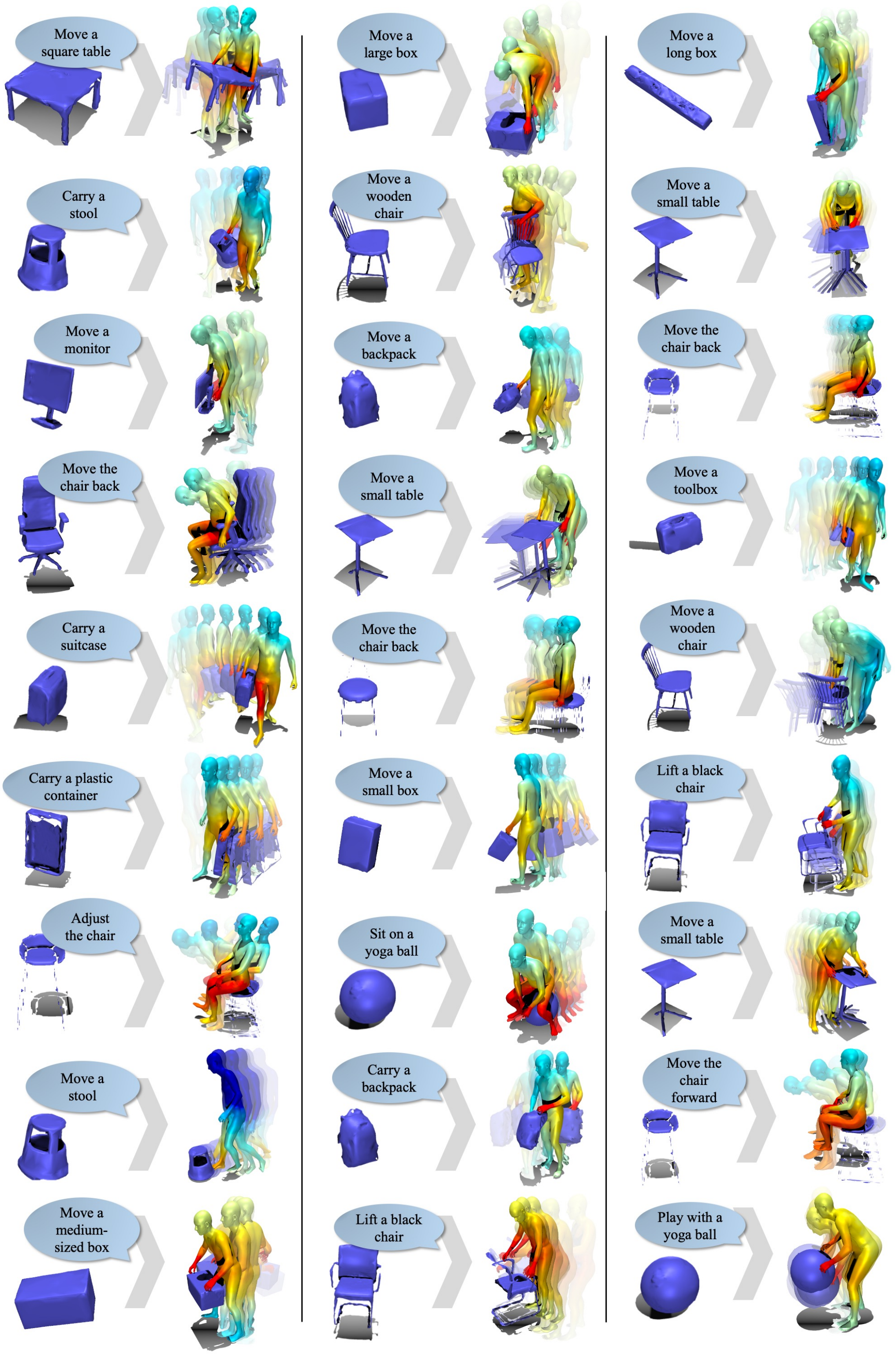


Figure 10. Additional qualitative evaluation. Our method produces diverse and realistic 3D human-object interaction sequences, given object geometry and short text description of the action. The sequences depict high-quality human-object interactions by modeling contact, mitigating floating and penetration artifacts.

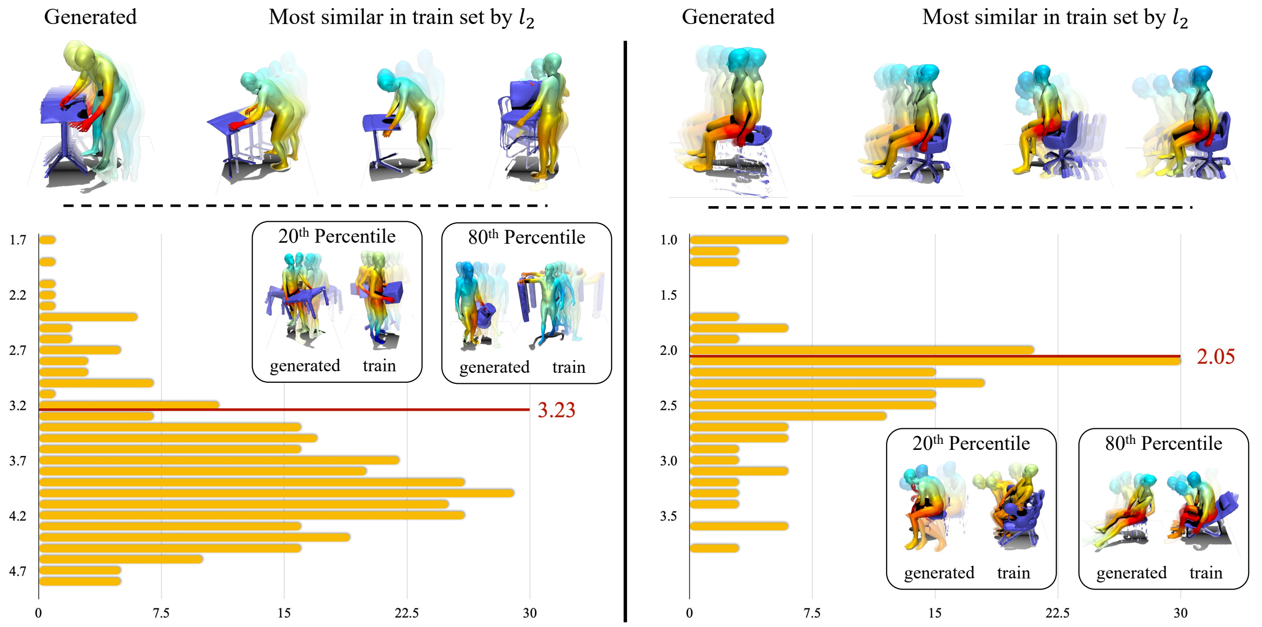


Figure 11. Human-Object Interaction Sequence Novelty Analysis. Performed on BEHAVE [9] (left) and CHAIRS [36] (right). We retrieve top-3 most similar sequences from the train set, and plot a histogram of distances to the closest train sample. While sequences at the 20th percentile still resemble the generated interactions, there is a large gap in the 80th percentile. We show the intra-trainset distance in red. Our approach generates novel shapes, not simply retrieving memorized train samples.

C. Baseline Evaluation Setup

There is no previous approach to modeling 3D human-object interactions from text and object geometry for direct comparison. Thus, we compare to the two closest methods, and compare to them in multiple settings, for a fair comparison.

The most related approach is InterDiff [88]. Their setting is to generate a short sequence of human-object interactions, from an observed such sequence as condition, with geometry but no text input. Their goal is to generate one, the most likely, sequence continuing the observation. We use their full approach, including the main diffusion training together with the post-processing refinement step. We compare in two different settings: First, in their native setup, running their method unchanged and modifying ours to take in geometry and past sequence observation instead of text (Motion-Cond. HOI in Tab. 1 main). Then, we modify their approach to take in geometry and text, replacing their past motion encoder with our CLIP-based text encoder (Text-Cond. HOI in Tab. 1 main). We observe that our method is able to outperform InterDiff in both scenarios, for both datasets.

We additionally compare to MDM [73], a recent diffusion-based state-of-the-art human motion generation approach. Their approach is based on a transformer encoder formulation, using each human body as a token in the attention. We run their method on SMPL parameters and first compare in their native setting, only predicting human motion. We compare to the human motion generated by our method which is trained to generate full human-object interactions (Text-Cond. Human Only in Tab. 1 main). We also compare to human motion sequences generated by InterDiff in this setting. We see that our method is able to outperform both baselines even in this setting, demonstrating the added benefit of learning interdependencies of human and object motion. For the comparison in our setting, we modify MDM by adding additional tokens for the objects to the attention formulation. Our approach performs more realistic and diverse sequences in both settings which better follow the text condition.

	BEHAVE				CHAIRS			
Representation	R-Prec. (top-3) ↑	FID ↓	Diversity →	MModality →	R-Prec. (top-3) ↑	FID ↓	Diversity →	MModality →
Ours (HumanML3D)	0.33	11.94	2.15	3.75	0.48	12.83	4.39	5.11
Ours	**0.62**	**6.31**	**6.63**	**5.47**	**0.74**	**6.45**	**8.91**	**5.94**

Table 5. Ours (using SMPL bodies) vs. using HumanML3D [26] skeletons and fitting SMPL bodies afterwards. While HumanML3D is designed to reduce jitter and foot skating, it leads to degraded performance in our scenario due to less effective contact guidance.

D. Fidelity and Diversity Metrics

We base our fidelity and diversity metrics R-Precision, FID score, Diversity, and MultiModality on practices established for human motion generation [25, 26, 73], with minor modifications: We use the same networks used by these previous approaches, and adapt the input dimensions to fit our feature lengths, $F = 79$ when evaluating human body motion only, and $F = 79 + 128 + 9 = 216$ (SMPL parameters, contact distances, object transformations) for full evaluation in the human-object interaction scenario.

E. Architecture Details

Fig. 12 shows our detailed network architecture, including encoder, bottleneck, and decoder formulations.

F. Data Details

F.1. Datasets

CHAIRS [36] captures 46 subjects as their SMPL-X [55] parameters using a mocap suit, in various settings interacting with a total of 81 different types of chairs and sofas, from office chairs over simple wooden chairs to more complex models like suspended seating structures. Each captured sequence consists of 6 actions and a given script; the exact separation into corresponding textual descriptions was manually annotated by the authors of this paper. In total, this yields $\approx$ 1300 sequences of human and object motion, together with a textual description. Every object geometry is provided as their canonical mesh; we additionally generate ground-truth contact and distance labels based on posed human and object meshes per-frame for each sequence. We use a random 80/10/10 split along object types, making sure that test objects are not seen during training.

BEHAVE [9] captures 8 participants as their SMPL-H [64] parameters captured in a multi-Kinect setup, along with the per-frame transformations and canonical geometries of 20 different object with a wide range, including yoga mats and tables. This yields $\approx$ 130 longer sequences. We use their original train/test split.

F.2. Object Geometry Representation

We represent object geometry as a point cloud, to be processed by a PointNet [58] encoder. For this, we sample $N = 256$ points uniformly at random on the surface of an object mesh. Each object category is sampled once as a pre-processing step and kept same for train and inference.

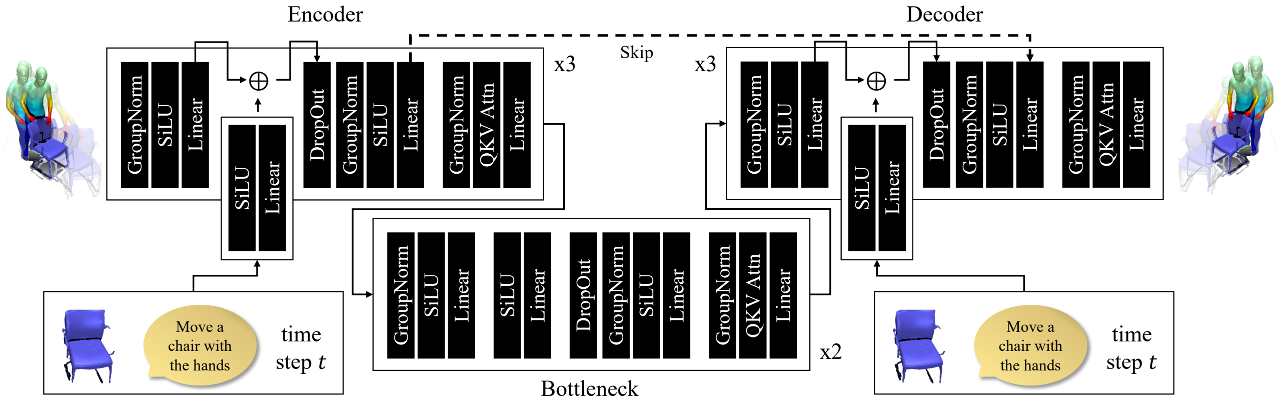


Figure 12. Network architecture specification.

List of Figures

List of Tables