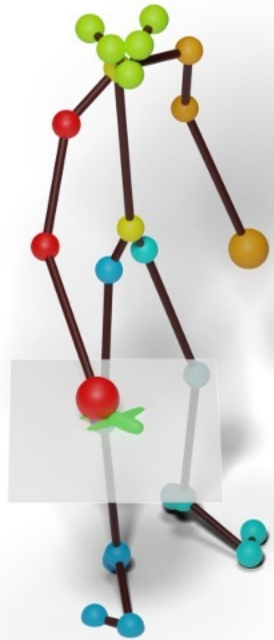


# 3D Human Behavior Generation through Action and Interaction Synthesis



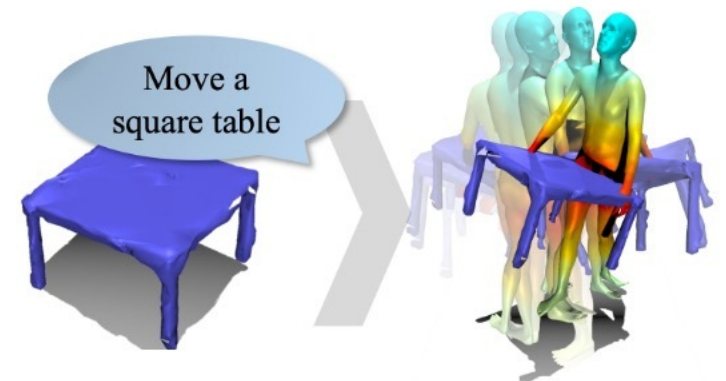
**Christian Diller**  
Supervisor: Prof. Angela Dai

**Tuesday, 10<sup>th</sup> December 2024**



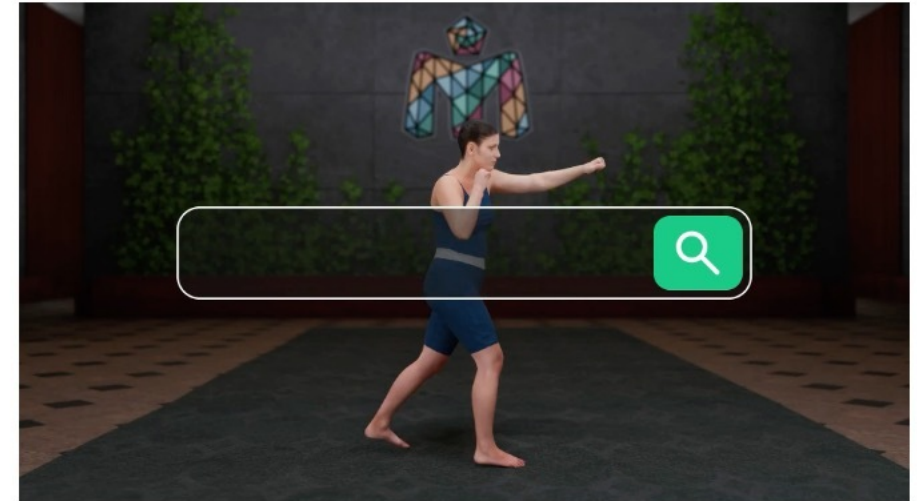
# Motivation: Understanding Human Behavior in 3D

- **Human behavior understanding is important for perception**
  - Higher-order understanding of human-machine interaction
  - Anticipatory action vs. perceptual reaction
- **Human environments are made by humans for humans**
- **Human motion generation in 3D**
  - Allows for more fine-grained actions, e.g., grasping objects
  - Enables direct interactions with an environment



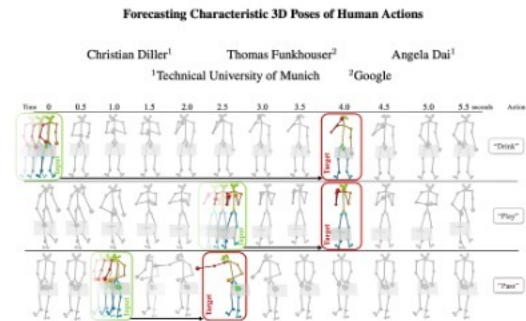
# Applications

- **Human-centered assistive systems**
  - Interaction between humans and robots in a shared physical space
  - Assistance robotics in medicine and care
- **Autonomous driving**
  - Forecasting interactions between cars & pedestrians
- **Content Creation**
  - Plausible human motion from sparse input (e.g., text)



# 3D Human Behavior Generation: Action & Interaction

## Efficient Action Representation



**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 drifting 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 conflates 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

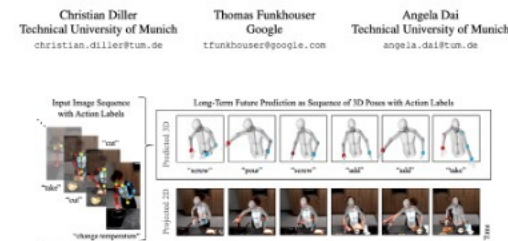
method, we construct a dataset of manually annotated characteristic 3d poses. Our experiments with this dataset suggest that our proposed probabilistic approach outperforms state-of-the-art methods by 26% on average.

### 1. Introduction

Future human pose forecasting is fundamental towards a comprehensive understanding of human behavior, and consequently towards achieving higher-level perceptions 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

## Complex Action Sequences

### FutureHuman3D: Forecasting Complex Long-Term 3D Human Behavior from Video Observations



**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 coarse 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 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 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 actions and 3D pose prediction: our joint approach outperforms each task treated individually, enables robust longer-term sequence prediction, and outperforms alternative approaches to forecast actions and characteristic 3D poses.

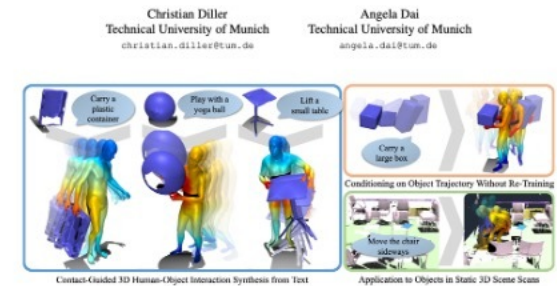
### 1. Introduction

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

## Human-Object Interactions

### CG-HOI: Contact-Guided 3D Human-Object Interaction Generation



**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 synthesis of realistic, coherent HOIs. Extensive evaluation shows that our joint contact-based human-object interaction approach generates realistic and physically plausible sequences, and we show two ap-

plications 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

## Forecasting Characteristic 3D Poses [1]

## FutureHuman3D [2]

## CG-HOI [3]

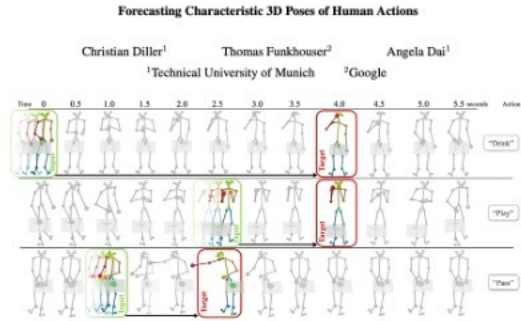
[1] Diller, Christian, Thomas Funkhouser, and Angela Dai. "Forecasting characteristic 3d poses of human actions." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[2] Diller, Christian, 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.

[3] Diller, Christian, and Angela Dai. "Cg-hoi: Contact-guided 3d human-object interaction generation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

# 3D Human Behavior Generation: Action & Interaction

## Efficient Action Representation



**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 conflates 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

## Complex Action Sequences

### FutureHuman3D: Forecasting Complex Long-Term 3D Human Behavior from Video Observations



**Figure 2.** We propose a novel generative approach to model long-term future human behavior by jointly forecasting a sequence of linear action labels and their associated realizations as 3D body poses. The novel application of our autoregressive method only requires weak supervision and past observations in the form of 2D RGB video data, together with a database of reconstructed 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 acquire in 3D (coverage, safety, expensive, etc.) but easy to acquire in 2D (simple RGB cameras). Thus, we design our method to only require 2D RGB data while being able to generate 3D human motion sequences. We use an autoregressive 3D pose task 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 that would be a practically intractable manner, jointly producing high-level coarse action labels together with their fine-level fine-grained realizations as characteristic 3D body poses. We illustrate that these low-action representations are capable to capture, and even predict, benefits both action and pose forecasting. Our experiments demonstrate the complementary nature of pose action and 3D pose prediction: our pose approach outperforms each task reward individually, enables robust long-term sequence prediction, and outperforms 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, virtual reality, and more. For instance, a monitoring system might issue early warnings of potentially dangerous behaviors, and a robotic assistant can take time-saving 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 critical steps 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 failures when performed in a location, they become dangerous when done near 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 human bodies.

To support these types of applications, we focus 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

## Human-Object Interactions

### CG-HOI: Contact-Guided 3D Human-Object Interaction Generation



**Figure 3.** We present an approach to generate realistic 3D human-object interactions (HOI) from a text description and given static object geometry to be interacted with (chairs). Our main insight is to explicitly model contact (visualized as red on the body mesh, blue 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 sequences (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 (HOI) from text. We model the motion of both human and object in an independent fashion, as semantically rich human motion words support an evaluation without user intervention. Our key insight is that explicitly modeling contact between the human body surface and object geometry can be used as strong priors 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, self-supervised through cross-attention. We show coverage this learned contact prior guidance during inference conditions of realistic, coherent HOIs. Extensive evaluation shows that our joint contact-guided human-object interaction approach generates realistic and physically plausible sequences, and we show two ap-

plifications highlighting the capabilities of our method. Our diffusion on a given object trajectory, we can generate the corresponding human motion without re-training, allowing strong 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 motion capture for animation, intuitive robotic systems, scene layout planning, or human behavior simulation. Crucially, human interacts in interdependency with the objects it being interacted with the object movement of a chair or table, for instance, encourages 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 wobbly chair, carrying a backpack). Existing works typically focus solely on generating dynamic human, and thereby disregarding their interdependency

## Forecasting Characteristic 3D Poses [1]

## FutureHuman3D [2]

## CG-HOI [3]

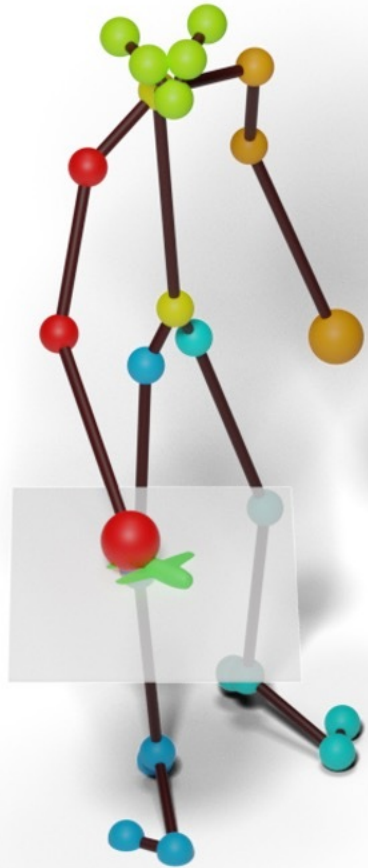
[1] Diller, Christian, Thomas Funkhouser, and Angela Dai. "Forecasting characteristic 3d poses of human actions." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[2] Diller, Christian, 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.

[3] Diller, Christian, and Angela Dai. "Cg-hoi: Contact-guided 3d human-object interaction generation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

# Forecasting Characteristic 3D Poses of Human Actions

How to efficiently represent 3D human motion sequences?



**Christian Diller**

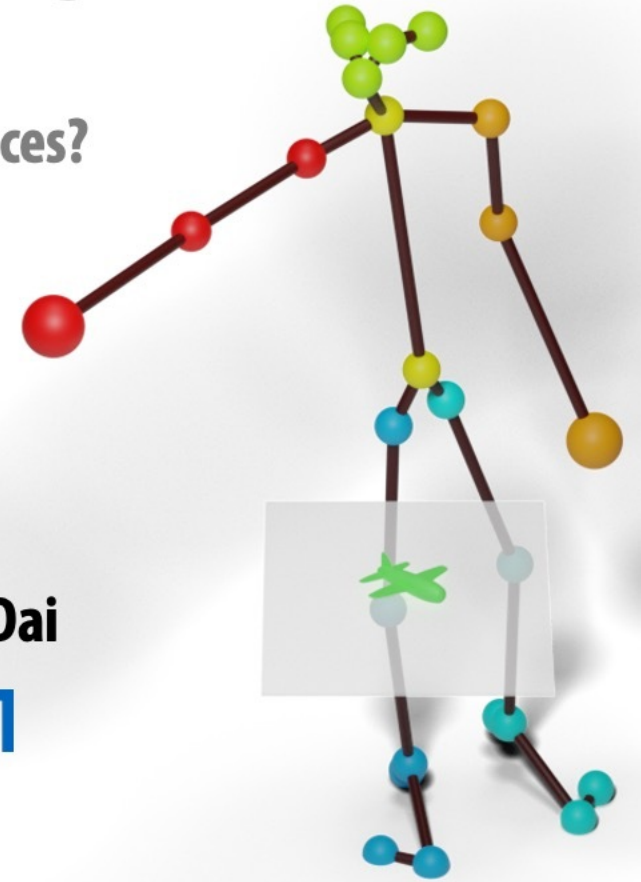
**TUM**

**Thomas Funkhouser**

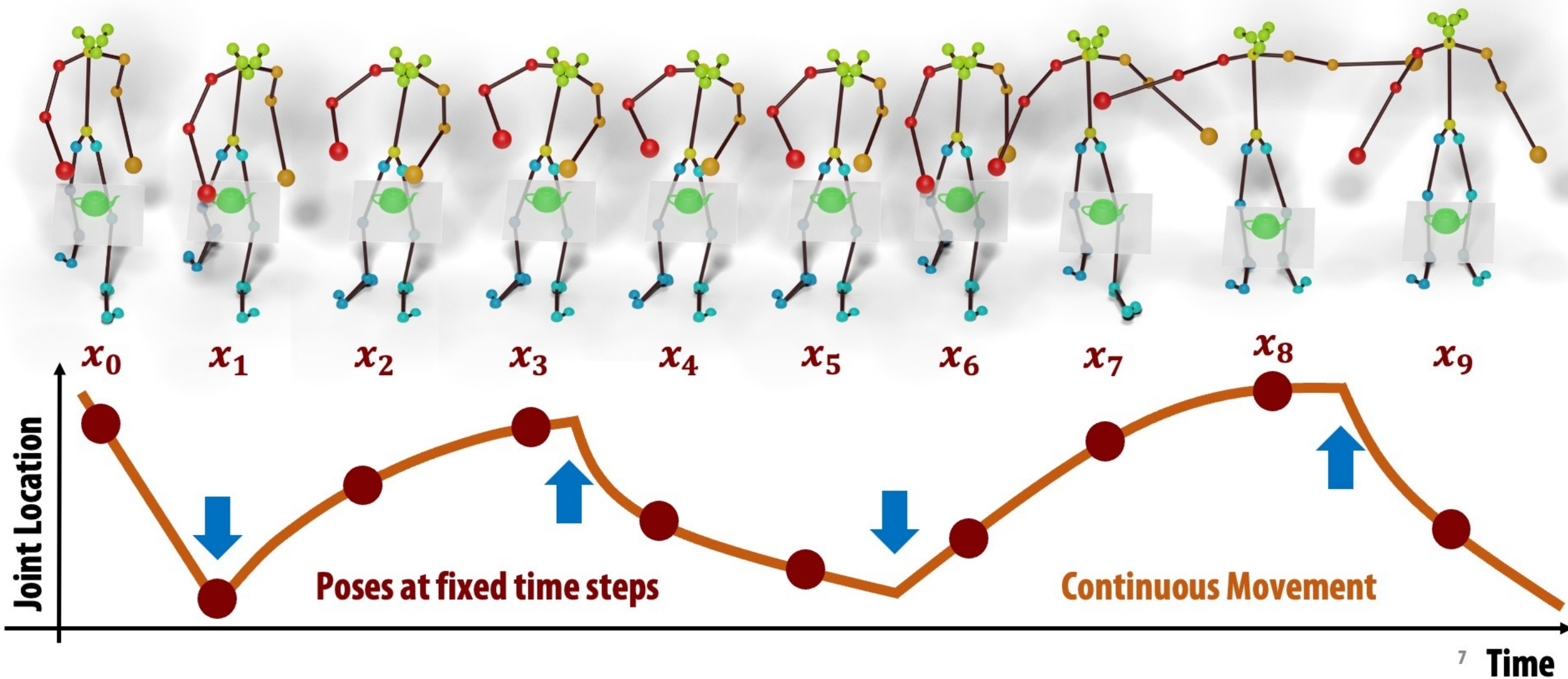


**Angela Dai**

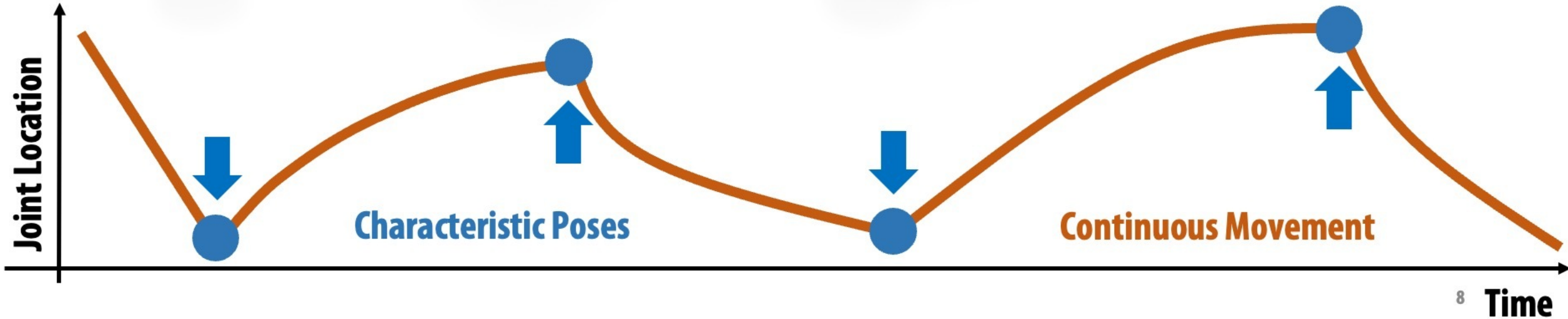
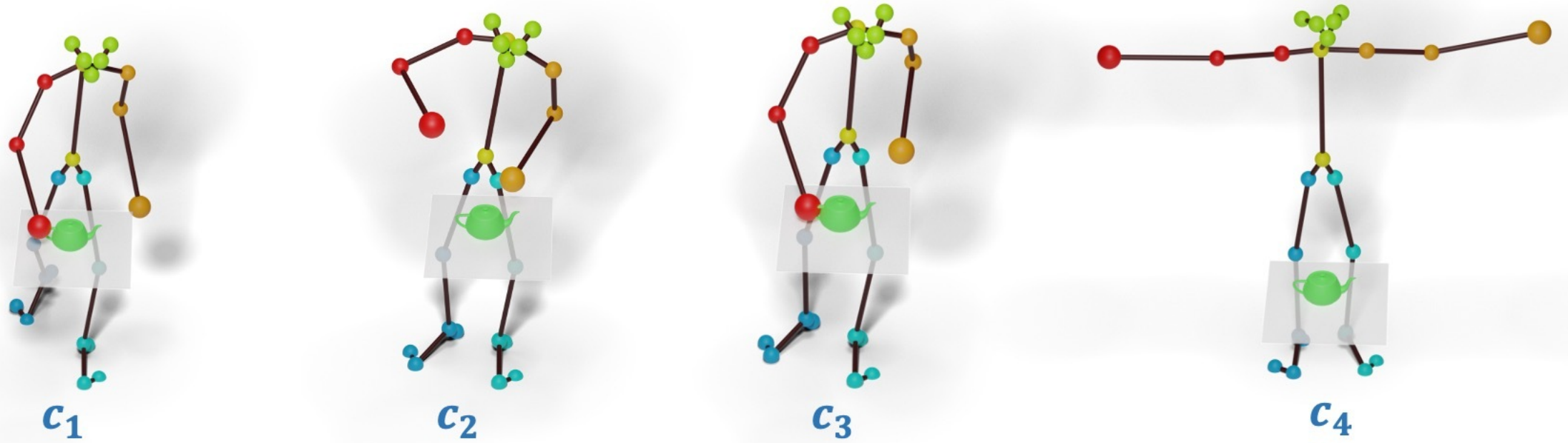
**TUM**



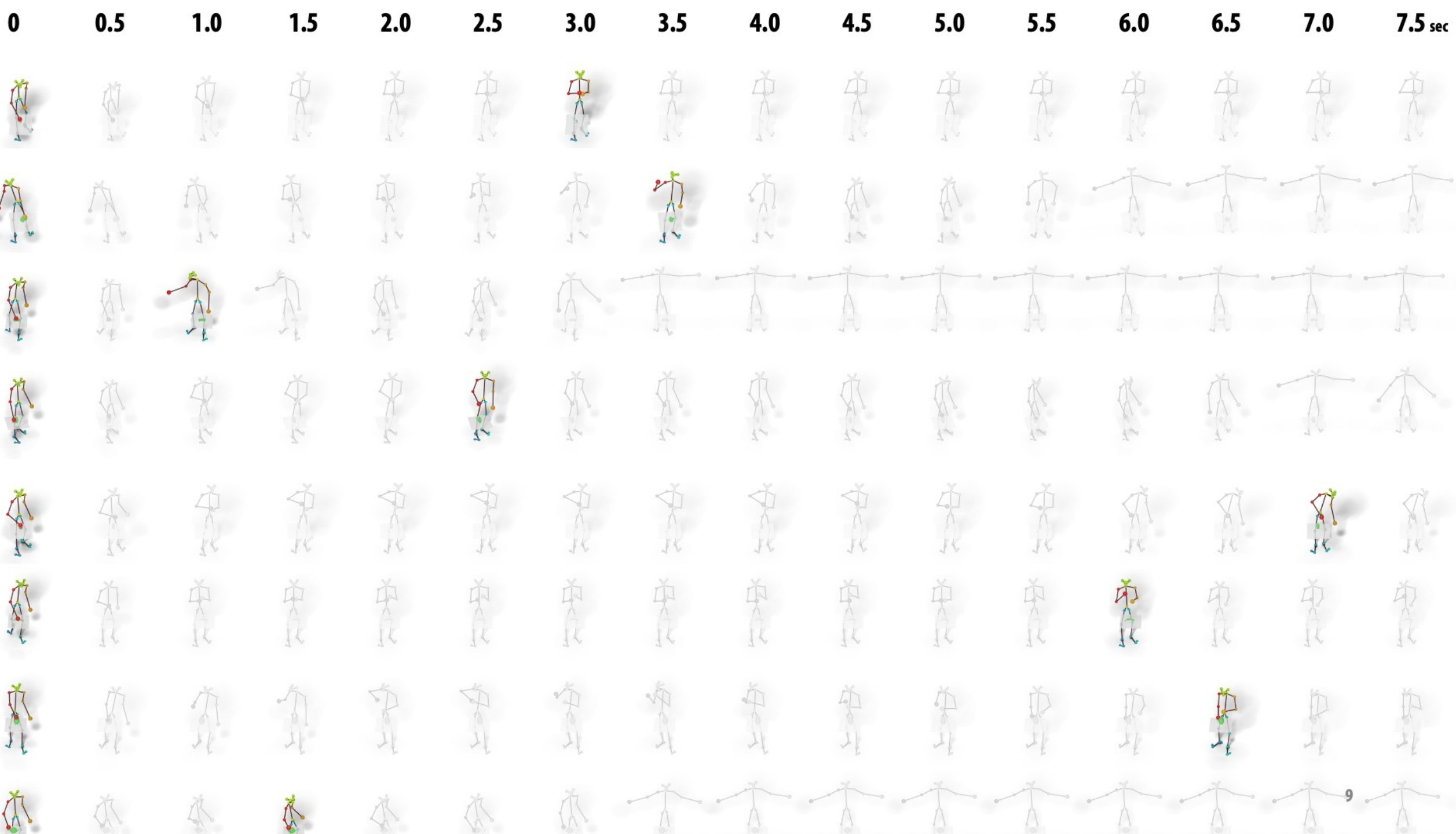
# Time-Based Future Human Motion Prediction



# Forecasting Characteristic 3D Human Poses







# Task: Characteristic 3D Poses for Action Goals



**Observation**

**Action Goal**

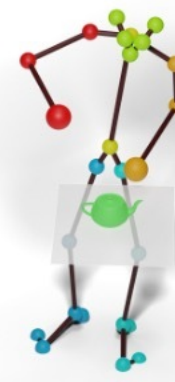


**Input**

**Target**

**Observation**

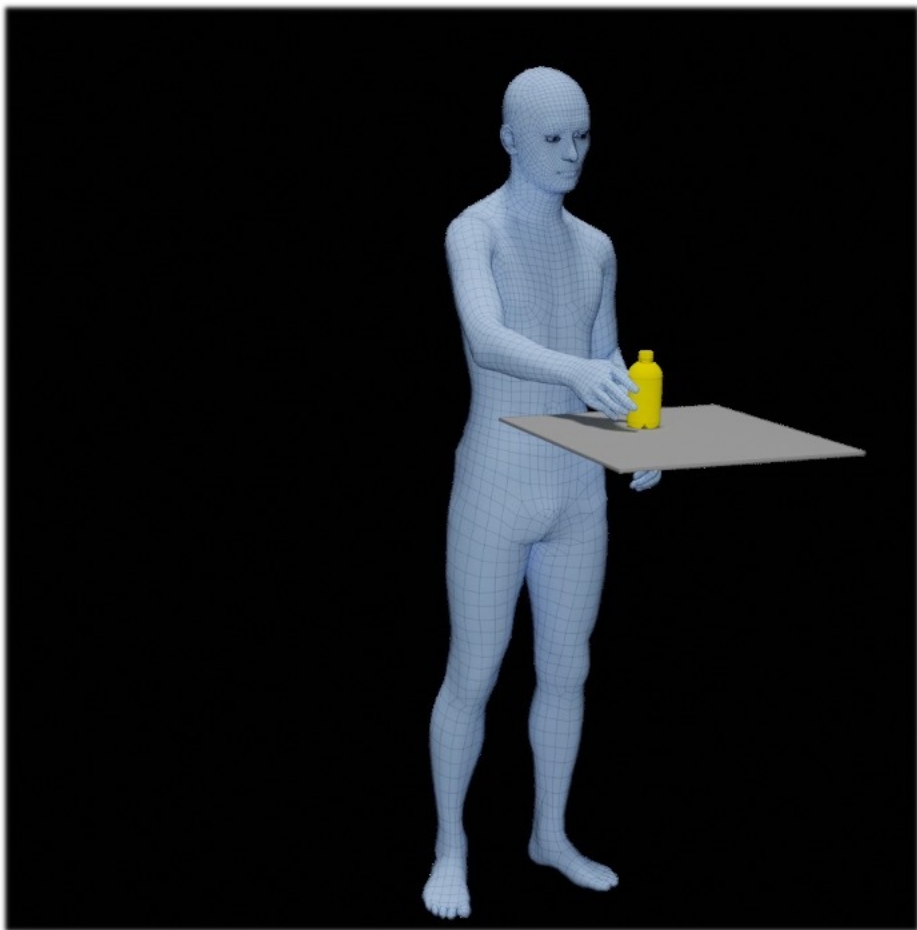
**Action Goal**



**Input**

**Target**

# Dataset: Characteristic 3D Poses on GRAB [1]



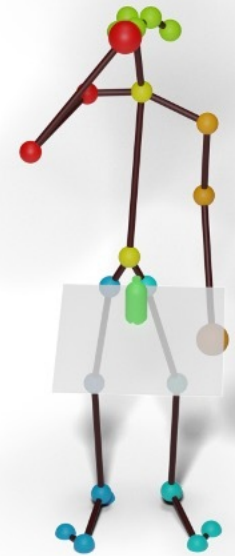
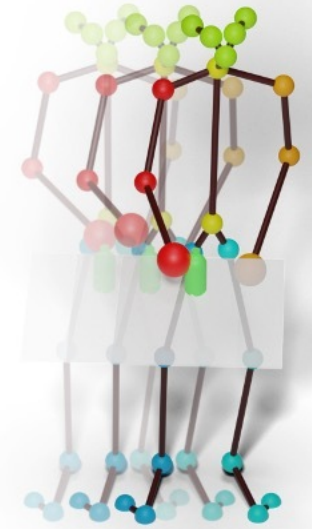
Input



"Drink"



Target

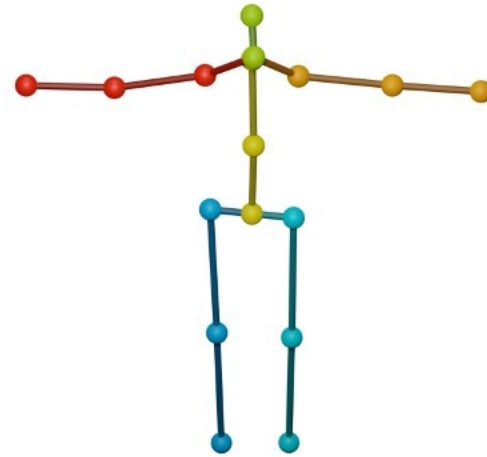


Original GRAB [1] Dataset

3D Skeleton Sequence

Pose Annotations

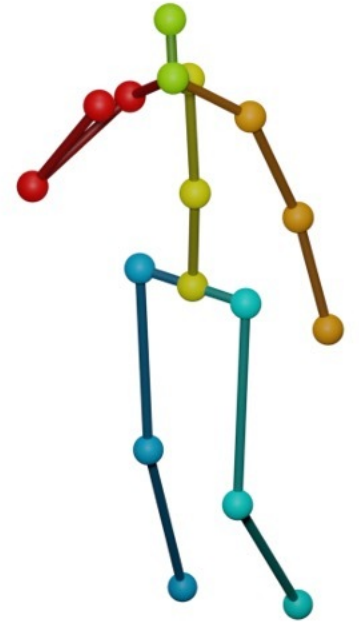
# Dataset: Characteristic 3D Poses on Human3.6m [1]



**"Phoning"**



**Characteristic  
Pose**

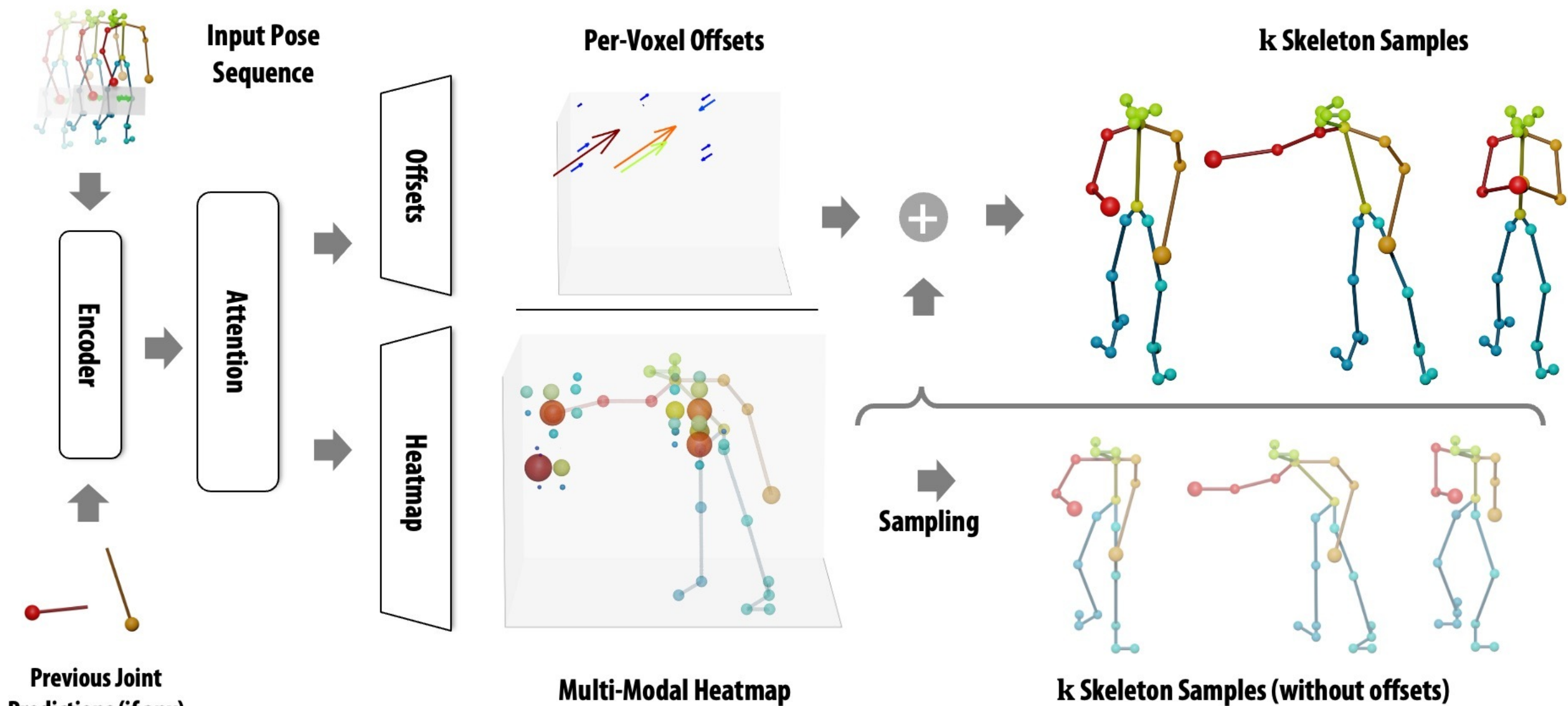


**Original Human3.6M [1] Dataset**

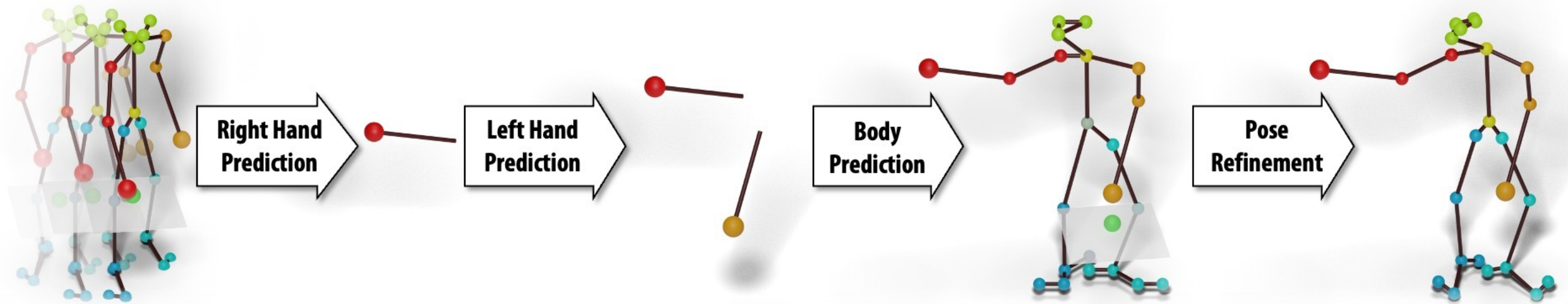
**3D Skeleton Sequence**

**Pose Annotations**

# Method: Architecture



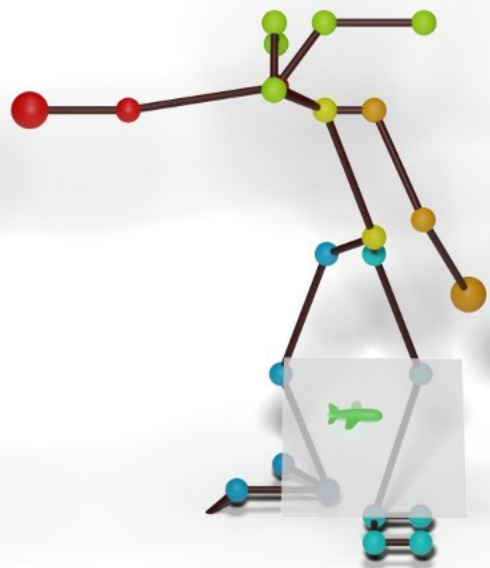
# Method: Autoregressive Prediction



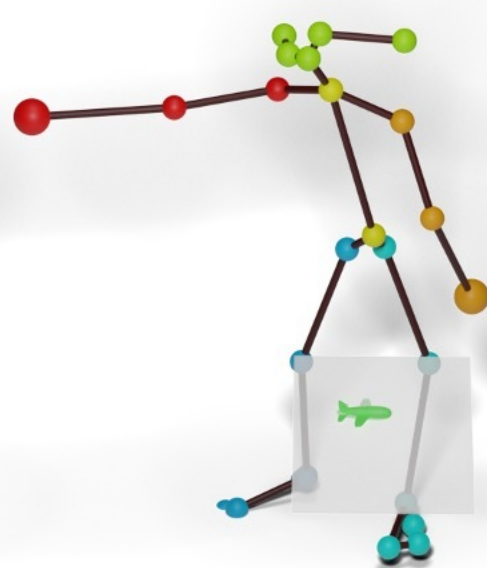
# Method: Pose Refinement

- End-Effector Locations
- Bone-Lengths, as observed in input
- Joint angles, as observed in input
- Heatmap joint probability

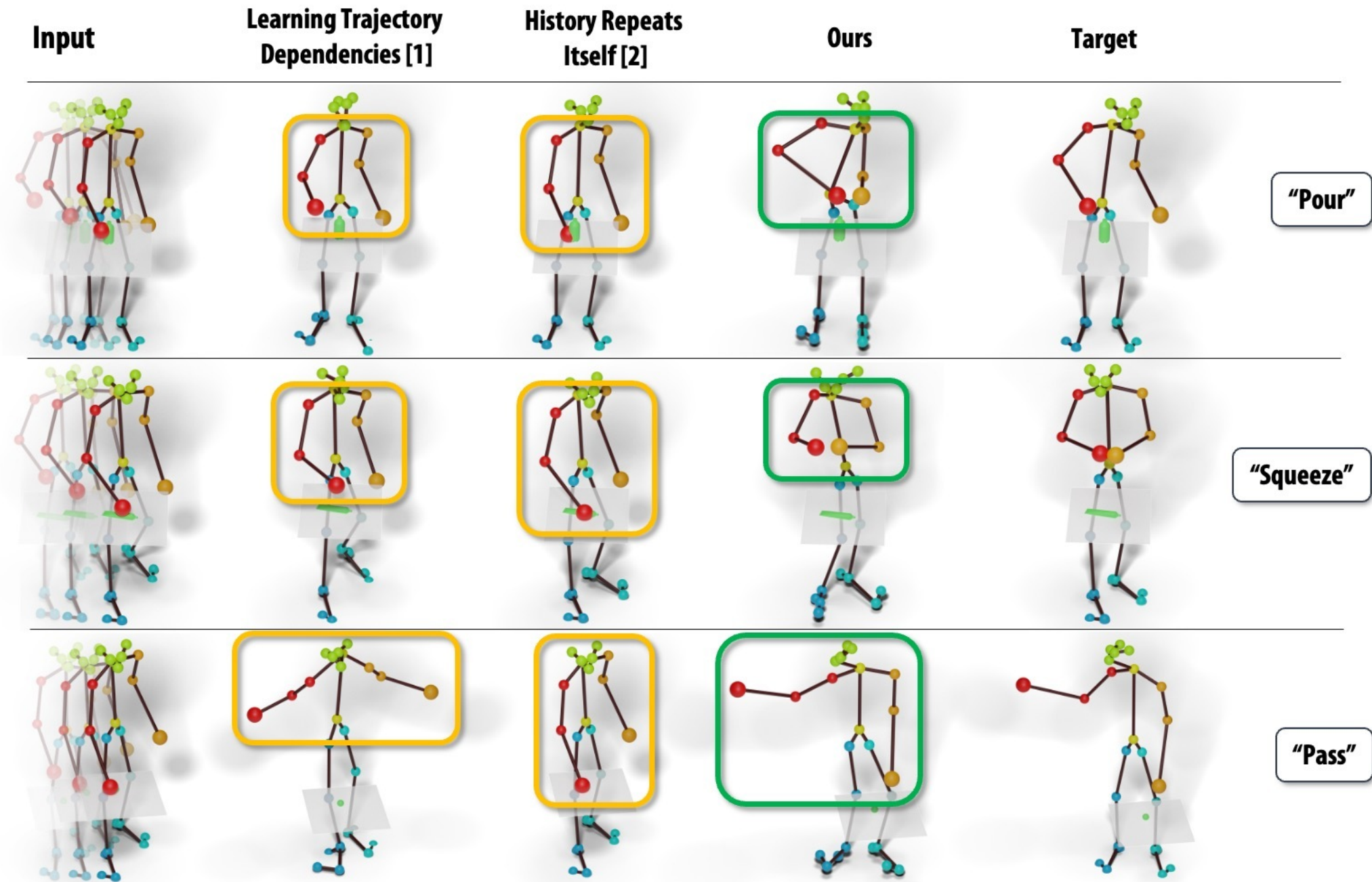
$$E_R(\mathbf{x}, \mathbf{e}, \mathbf{b}, \theta, H) =$$



**Initial Prediction**



**Refined**



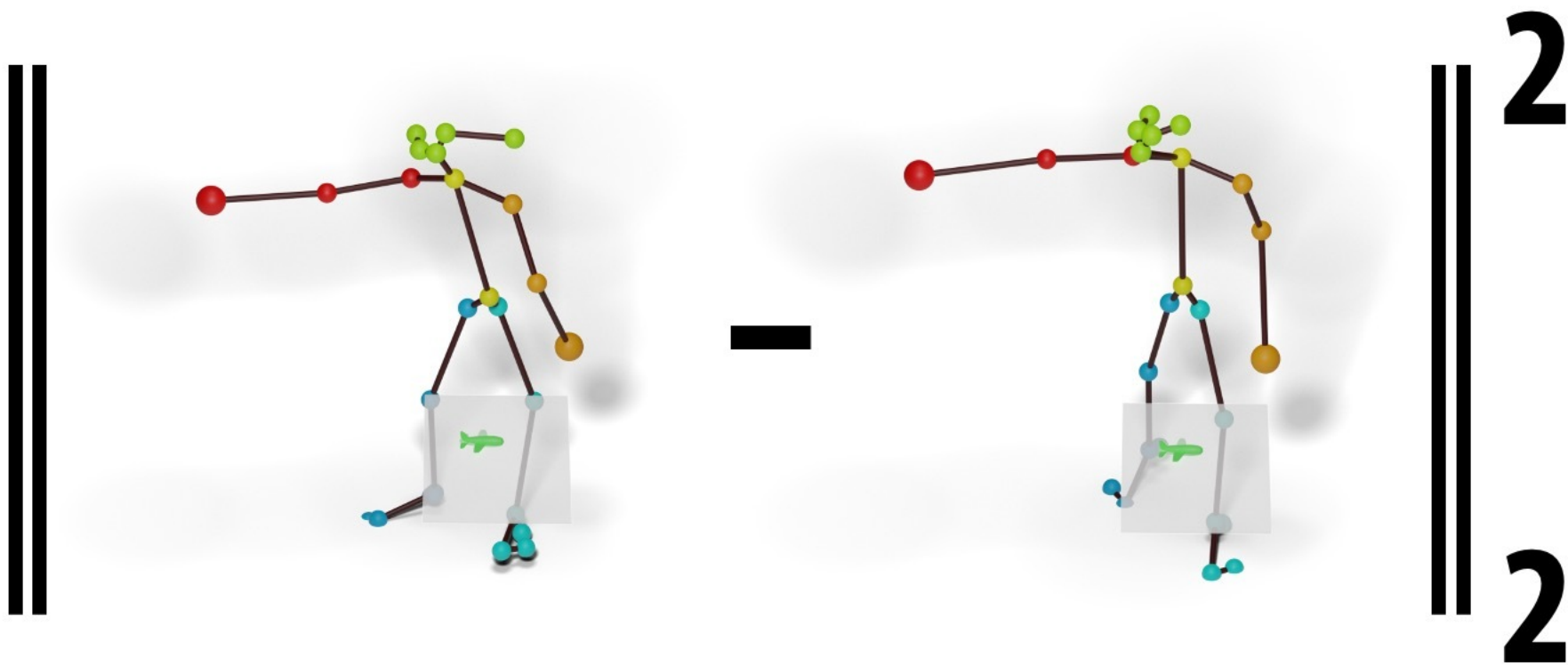
[1] Mao, Wei, et al. "Learning trajectory dependencies for human motion prediction." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[2] Mao, Wei, Miaomiao Liu, and Mathieu Salzmann. "History repeats itself: Human motion prediction via motion attention." European Conference on Computer Vision. Springer, Cham, 2020.



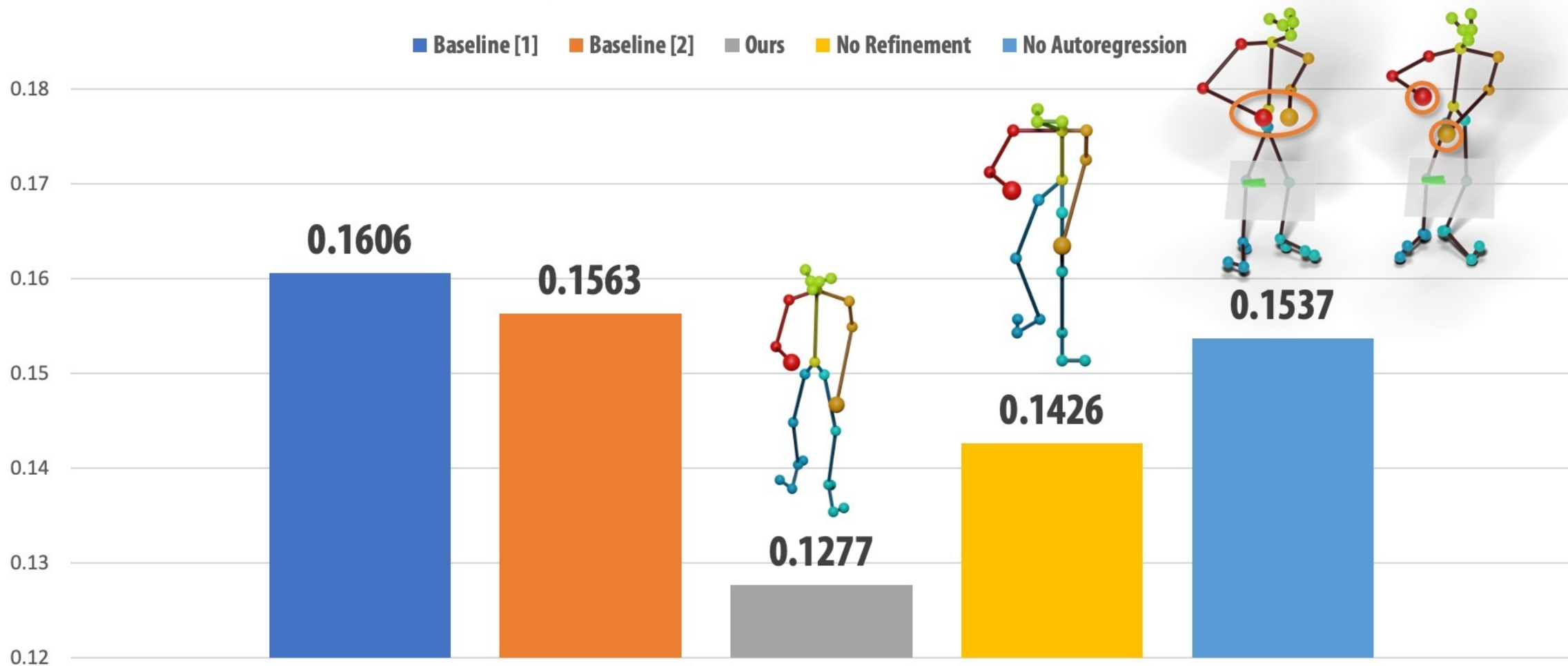
# Results: Mean Per-Joint Position Error

$$E_{\text{MPJPE}} = \frac{1}{25} \sum_{j=1}^{25} \|p'_j - p_j\|_2^2$$



# Results: Quantitative

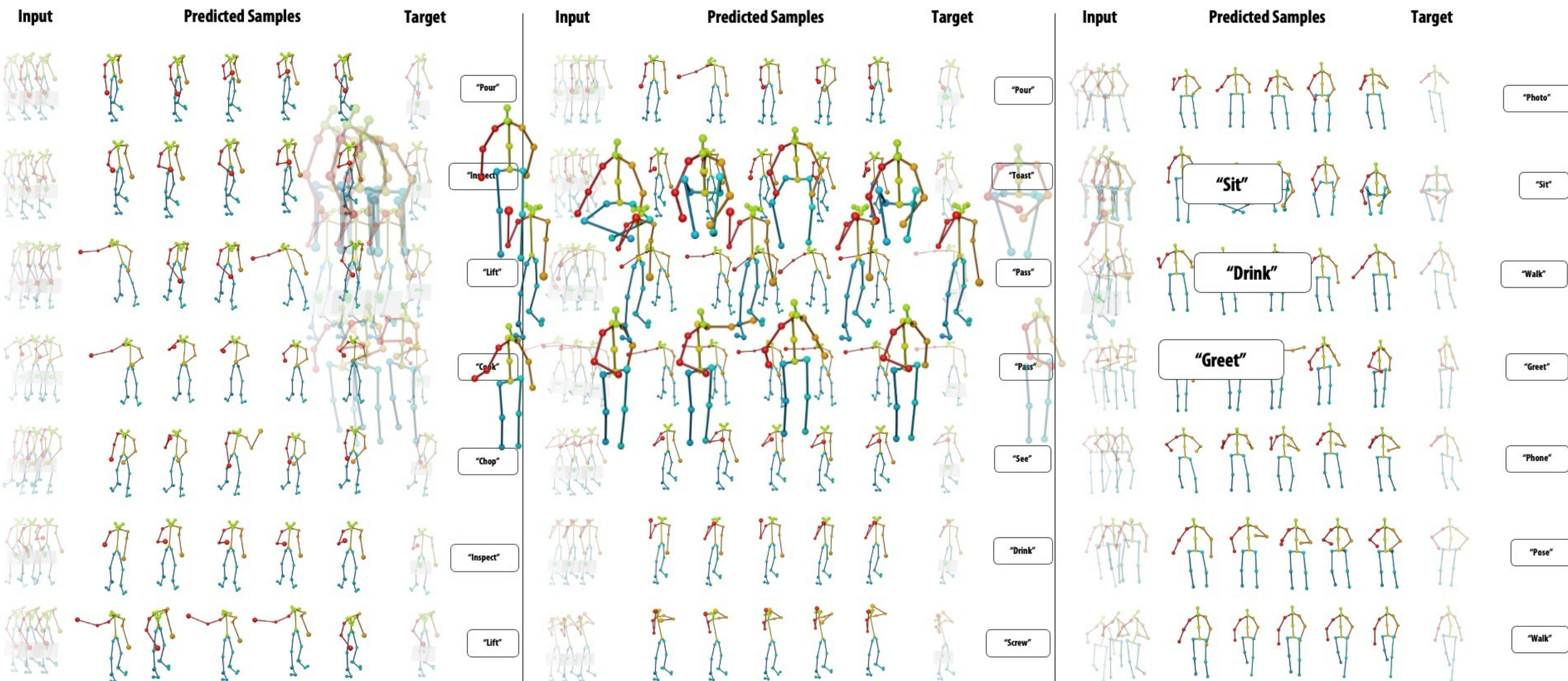
MPJPE = Mean Per-Joint Position Error



[1] Mao, Wei, et al. "Learning trajectory dependencies for human motion prediction." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[2] Mao, Wei, Miaomiao Liu, and Mathieu Salzmann. "History repeats itself: Human motion prediction via motion attention." European Conference on Computer Vision. Springer, Cham, 2020.

# Results: Qualitative – Multi-Modal Predictions



# 3D Human Behavior Generation: Action & Interaction

## Efficient Action Representation

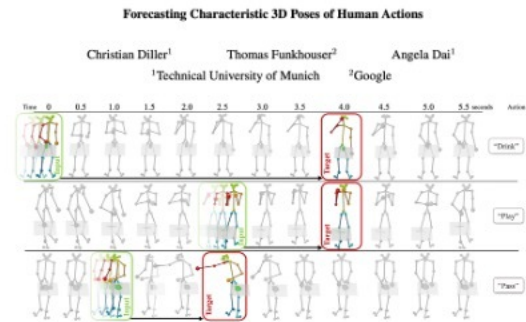


Figure 1. For a real world 3d skeleton sequence of a human performing an action, we propose to forecast the semantically meaningful characteristic 3d pose, representing the action goal for this sequence. As input, we take a short observation of a sequence of consecutive poses leading up to the target 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 conflates 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

## Complex Action Sequences



Figure 1. We propose a novel generative approach to model long term future human behavior by jointly forecasting a sequence of linear action labels and their associated realizations as 3D body poses. The novel application of our autoregressive method only requires weak supervision and past observations in the form of 2D RGB video data, together with a database of reconstructed 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 acquire as 3D coverage (e.g., sequences, sensors) has yet to acquire in 3D (single RGB cameras). Thus, we design our method to only require 2D RGB data while being able to generate 3D human motion sequences. We use a differentiable 3D pose-to-pose 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 that would be a semantically hierarchical manner, jointly producing high-level coarse action labels together with their fine-level fine-grained realizations as characteristic 3D body poses. We illustrate that these low-action representations are capable to capture, 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 long-term sequence prediction, and outperforms 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, virtual reality, and more. For instance, a monitoring system might issue early warnings of potentially dangerous behaviors, and a robotic assistant can take time-saving 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 critical steps 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 failures when performed in a location, they become dangerous when done later to an active table top 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 propose 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

## Human-Object Interactions

### CG-HOI: Contact-Guided 3D Human-Object Interaction Generation



Figure 1. We present an approach to generate realistic 3D human-object interactions (HOI) from a text description and given static object geometry to be interacted with (chairs). Our main insight is to explicitly model contact (visualized as red on the body mesh, blue 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 sequences (top right), and generate interactions in static scenes (bottom right).

### Abstract

We propose CG-HOI, the first method to address the task of generating dynamic 3D human-object interactions (HOI) from text. We model the motion of both human and object in an independent fashion, as semantically rich human motion words happen in evaluation without user intervention. Our key insight is that explicitly modeling contact between the human body surface and object geometry can be used as strong priory 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, self-conditioned through cross-attention. We show coverage this learned contact first guidance during inference consists of realistic, coherent HOIs. Extensive evaluation shows that our joint contact-guided human-object interaction approach generates realistic and physically plausible sequences, and we show two ap-

plifications highlighting the capabilities of our method. Our diffusion on a given object trajectory, we can generate the corresponding human motion without training, allowing strong strong human-object interdependency learning. Our approach is also flexible, and can be applied to static real-world 3D scenes again.

### 1. Introduction

Generating human motion sequences in 3D is important for many real-world applications, e.g. efficient motion capture for animation, synthetic robotic systems, scene layout planning, or human behavior simulation. Crucially, human interacts in interdependency with the objects it being interacted with the object movement of a chair or table, for instance, encourages 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 wobbly chair, carrying a backpack). Existing works typically focus solely on generating dynamic human, and thereby disregarding their interdependency

## Forecasting Characteristic 3D Poses [1]

## FutureHuman3D [2]

## CG-HOI [3]

[1] Diller, Christian, Thomas Funkhouser, and Angela Dai. "Forecasting characteristic 3d poses of human actions." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[2] Diller, Christian, 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.

[3] Diller, Christian, and Angela Dai. "Cg-hoi: Contact-guided 3d human-object interaction generation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

# 3D Human Behavior Generation: Action & Interaction

## Efficient Action Representation

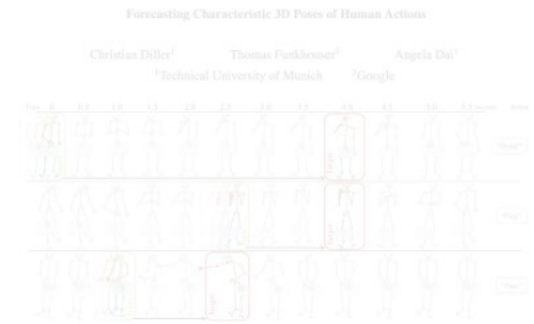


Figure 1. For a real-world 3d motion sequence of a human performing an action, we propose to forecast the semantically meaningful characteristic 3d pose representing the action goal for this sequence. To do so, we take a short observation of a sequence of consecutive poses leading up to the target characteristic pose. Thus, we propose to take a goal-oriented approach, predicting the key moments characterizing human behavior, instead of predicting continuous motion, which can occur in varying speeds with predictions more easily drifting for longer time (1-3) predictions. We develop an attention-driven prediction approach to capture the most likely states 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 more so differ, the focus is more intermediate and multi-component and universal aspects of human action. Instead, we define a semantically meaningful pose prediction and that decouples the predicted pose from time, taking inspiration from goal-oriented behavior. To predict characteristic poses, we propose a prediction 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 poses. To evaluate our

method, we construct a dataset of manually annotated characteristic 3d poses. Our experiments with this dataset suggest that our proposed prediction approach outperforms state-of-the-art methods by 20% 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 real-world interactions with humans, such as autonomous robots or vehicles. In fact, prediction is considered to play a foundational part in intelligence [1, 2, 3]. 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 actions towards the forecasted future. For example, a robotic surgical assistant should predict in advance where to place a tool to avoid the surgeon's next action, what sensor

## Complex Action Sequences

### FutureHuman3D: Forecasting Complex Long-Term 3D Human Behavior from Video Observations

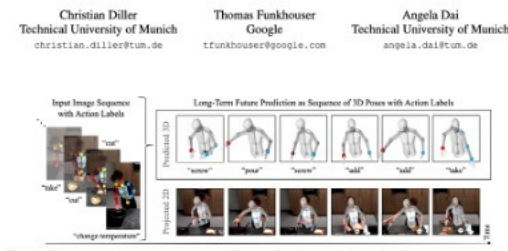


Figure 1. We propose a novel generative approach to model long-term future human behavior by jointly forecasting a sequence of coarse action labels and their concrete realizations as 3D body poses. For broad applicability, our autoregressive method only requires weak supervision and past observations 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 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 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 outperforms 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 behavior, 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

## Human-Object Interactions

### CG-HOI: Contact-Guided 3D Human-Object Interaction Generation



Figure 1. We present an approach to generate realistic 3D human-object interactions (HOI) from a text description and prior state, object geometry to be interacted with (e.g., chair). Our main insight is to explicitly model contact (in addition to collision) on the body mesh, close contact to self, in tandem with human and object sequences, in a joint diffusion process. In addition to synthesizing HOI from text, we can also synthesize human motions conditioned on given object sequences (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 (HOI) from text. We model the motion of both human and object in an interdependent fashion, as semantically rich human motion needs happens in isolation without any interaction. Our key insight is that explicitly modeling contact between the human body surface and object geometry can be used as strong priory 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 show that this learned contact first guidance during inference constrains of realistic, coherent HOI. Extensive evaluation shows that our joint contact-based human-object interaction approach generates realistic and physically plausible sequences, and we show two ap-

plications highlighting the capabilities of our method. Our diffusion on a given object trajectory, we can generate the corresponding human motion without training, allowing 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 motion planning, animation, robotic systems, scene layout planning, or human behavior simulation. Crucially, human interactions in interdependencies with the object(s) being interacted with the object structure of a chair or table, 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 human, and thereby disregarding their interdependencies

## Forecasting Characteristic 3D Poses [1]

## FutureHuman3D [2]

## CG-HOI [3]

[1] Diller, Christian, Thomas Funkhouser, and Angela Dai. "Forecasting characteristic 3d poses of human actions." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[2] Diller, Christian, 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.

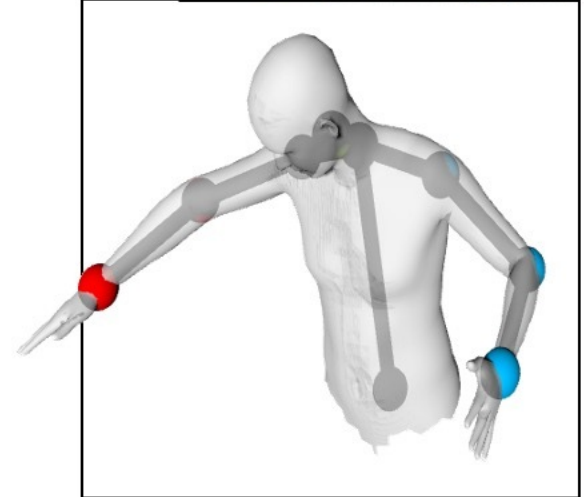
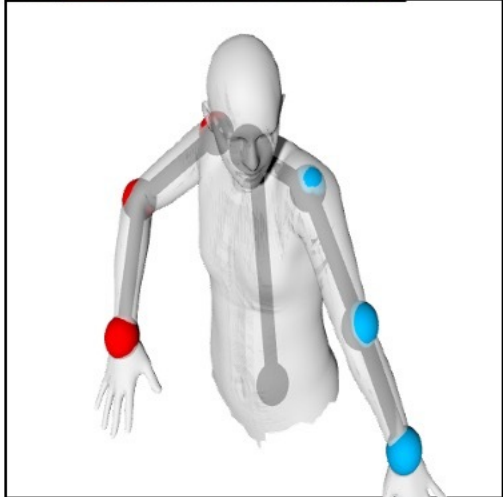
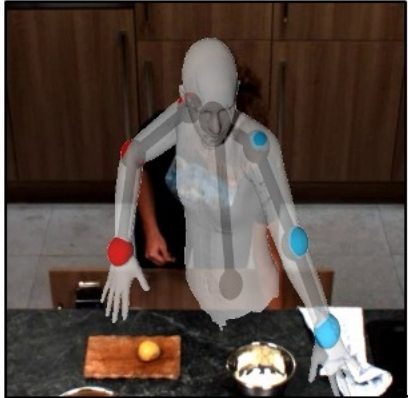
[3] Diller, Christian, and Angela Dai. "Cg-hoi: Contact-guided 3d human-object interaction generation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

# FutureHuman3D: Forecasting Complex Long-Term 3D Human Behavior from Video Observations

How can we learn complex long-term action sequences with limited 3D data?

“take”

“pour”



Christian Diller



Thomas Funkhouser

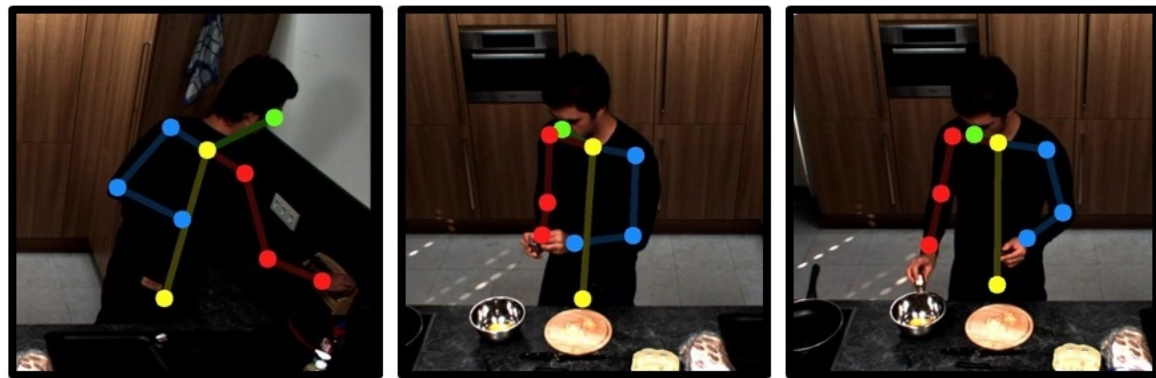


Angela Dai



# Related Work: Action Forecasting

... "take" "screw" "spice"

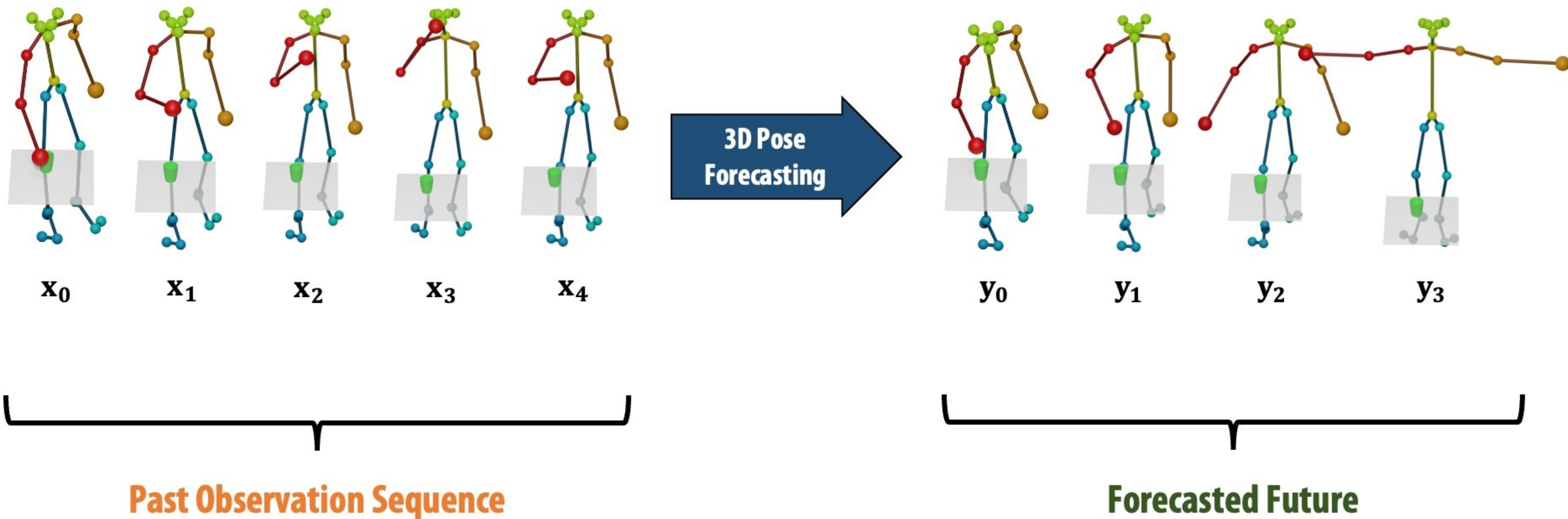


"screw" ...

Past Observation Sequence

Forecasted Future

# Related Work: 3D Pose Forecasting





# Task: Future Actions & 3D Poses from 2D

## 2D RGB Images + Action Labels



"take"



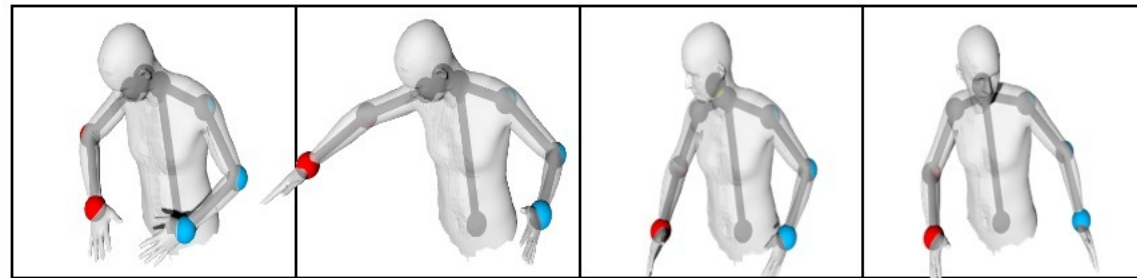
"cut"



"cut"

Joint 3D Pose &  
Action Forecasting

## 3D Pose Sequence + Action Labels



"screw"

"pour"

"screw"

"add"

Past Observation Sequence

Forecasted Future

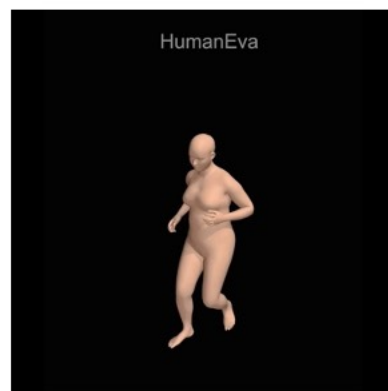
# Data: Uncorrelated 2D and 3D Human Poses

## 2D Action Sequences



- Take
- Wash
- Take
- Take
- Take
- Close
- Take
- Take
- Peel
- Throw in Garbage
- Cut
- Add
- Throw in Garbage

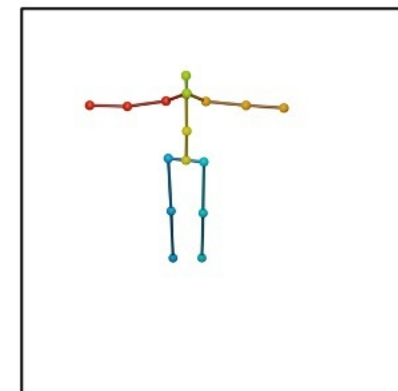
## 3D Pose Data



AMASS [1]



GRAB [2]



Human3.6m [3]



**No correspondence**

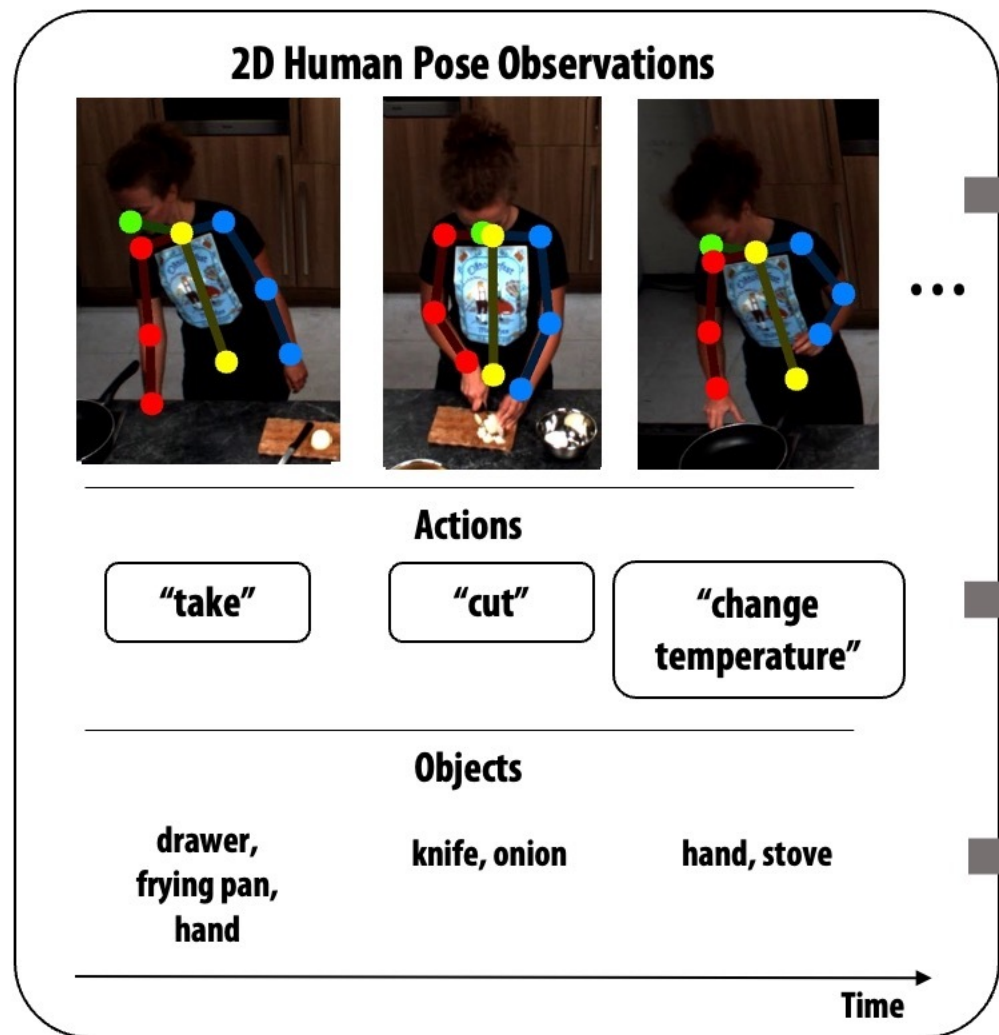
[1] Mahmood, Naureen, et al. "AMASS: Archive of motion capture as surface shapes." Proceedings of the IEEE/CVF international conference on computer vision. 2019.

[3] Ionescu, Catalin, et al. "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments." IEEE transactions on pattern analysis and machine intelligence 36.7 (2013): 1325-1339.

[2] Taheri, Omid, et al. "GRAB: A dataset of whole-body human grasping of objects." European conference on computer vision. Springer, Cham, 2020.

# Method: Architecture

## Input Sequence



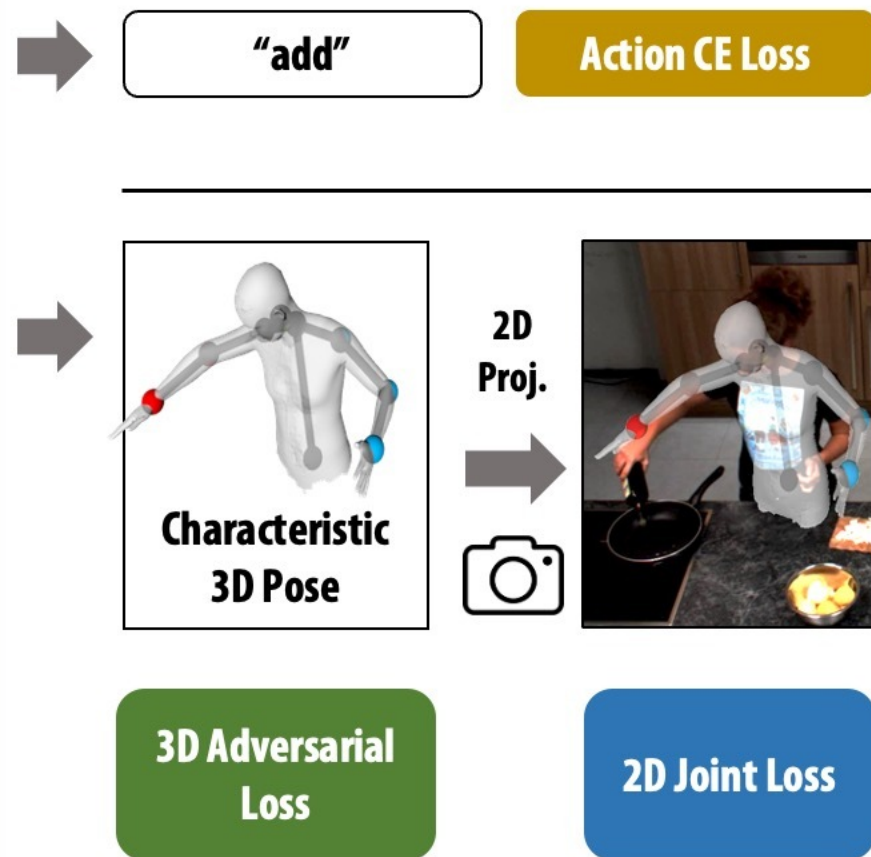
Pose History Encoder

Action Encoder

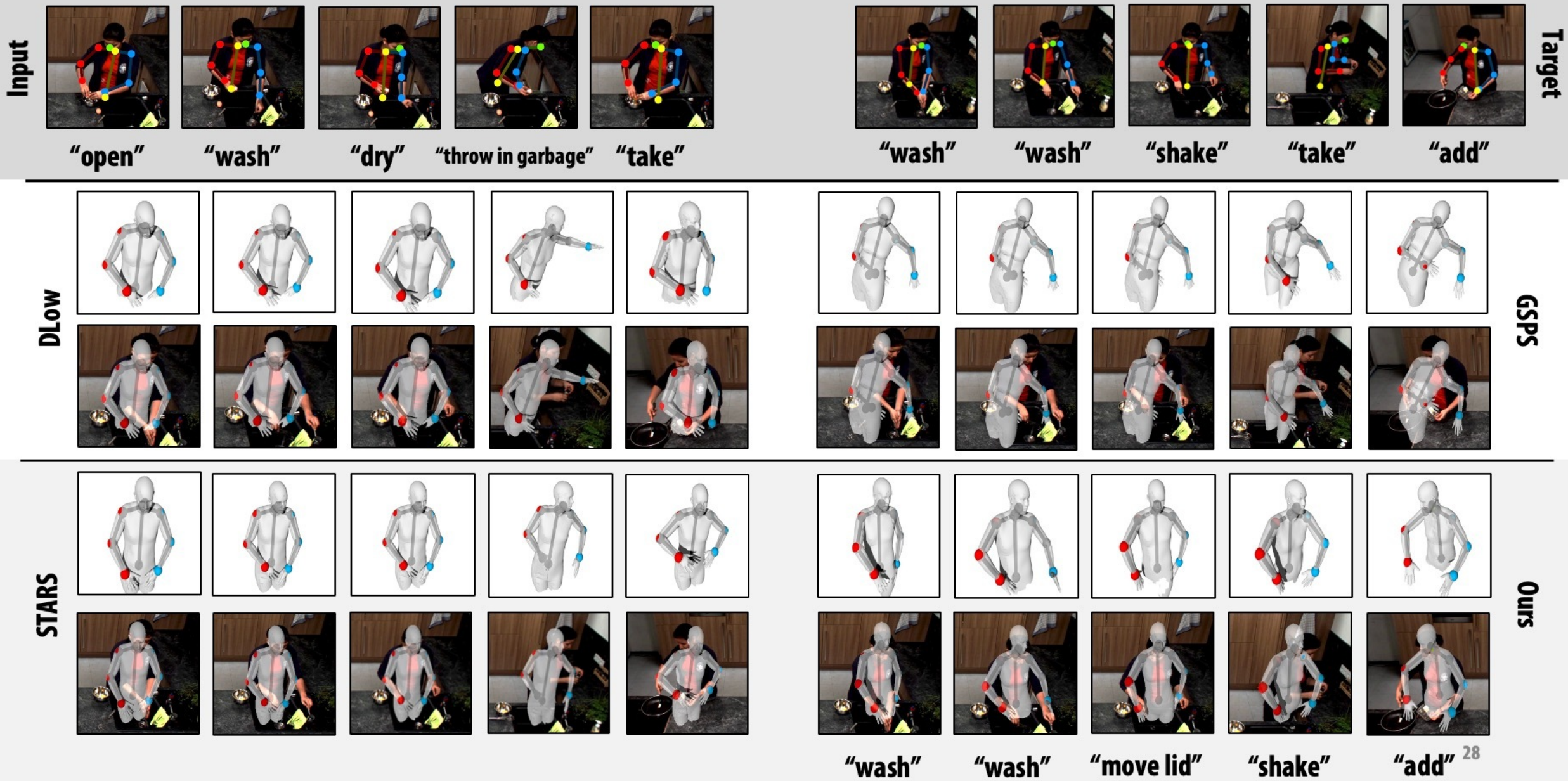
Object Encoder

MLP Decoder

## Forecasted 3D Pose + Action

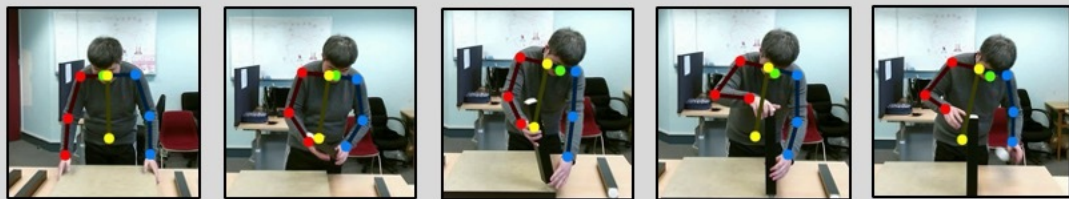


# Results: Qualitative 3D Pose & Action – Cooking



# Results: Qualitative 3D Pose & Action – Furniture Assembly

Input



“rotate”

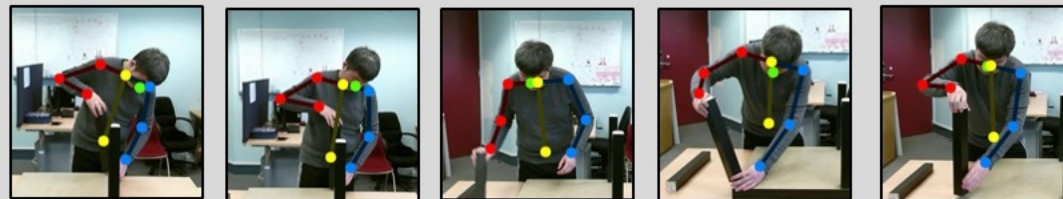
“pick up”

“align”

“spin”

“pick up”

Target



“align”

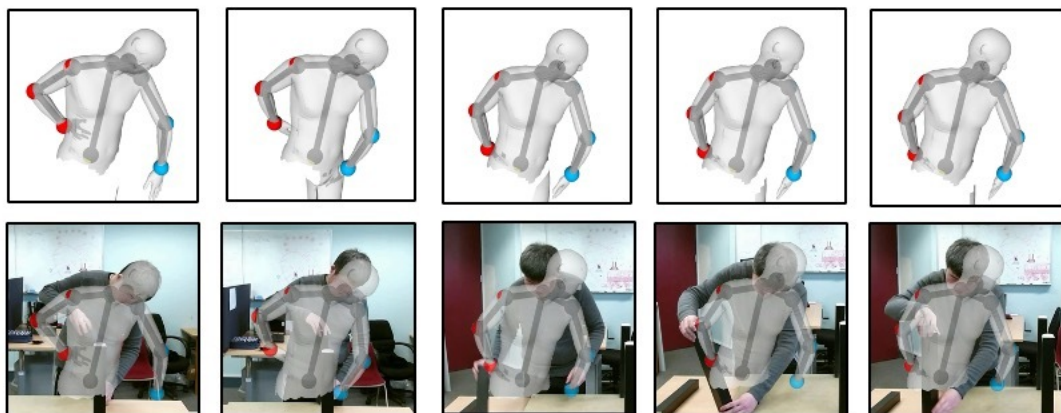
“spin”

“pick up”

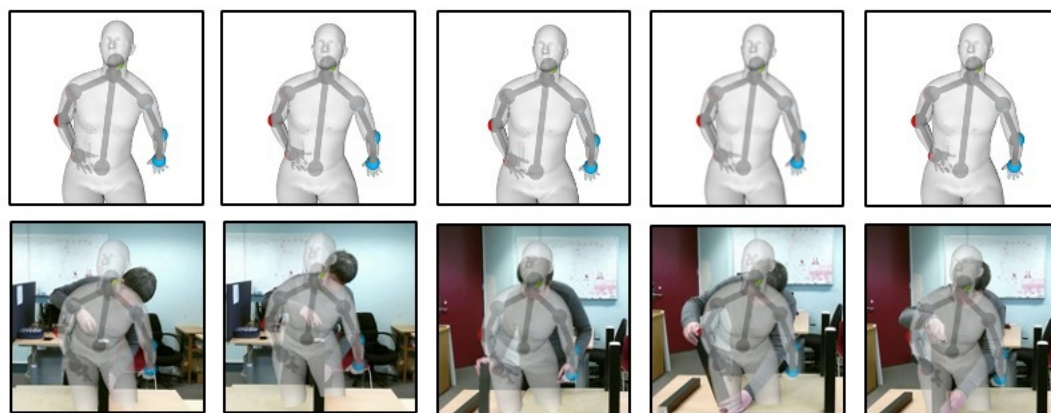
“align”

“spin”

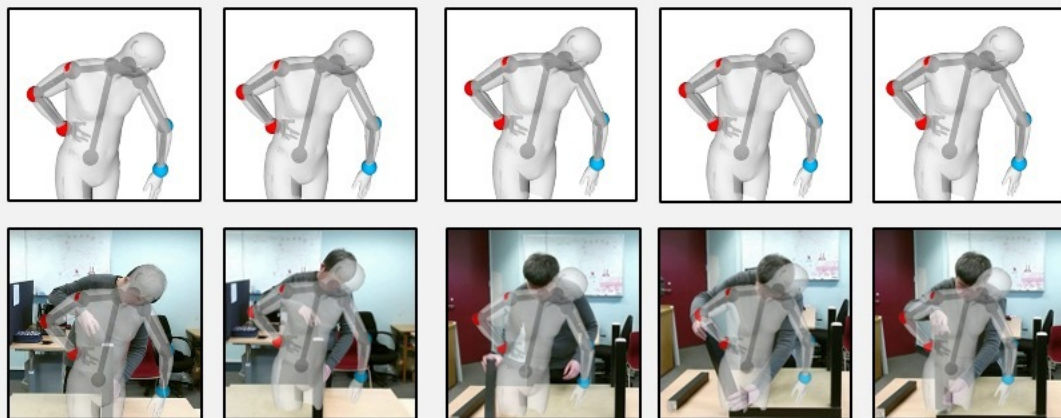
DLOW



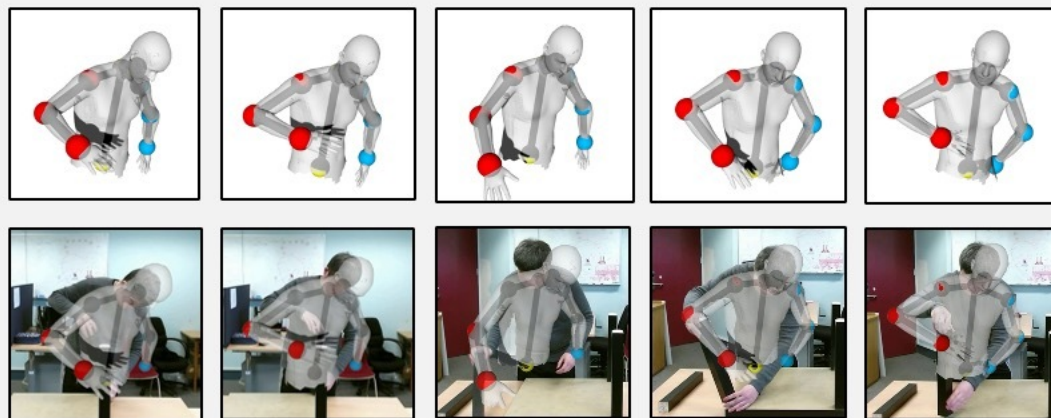
GSPS



STARS



Ours



“align”

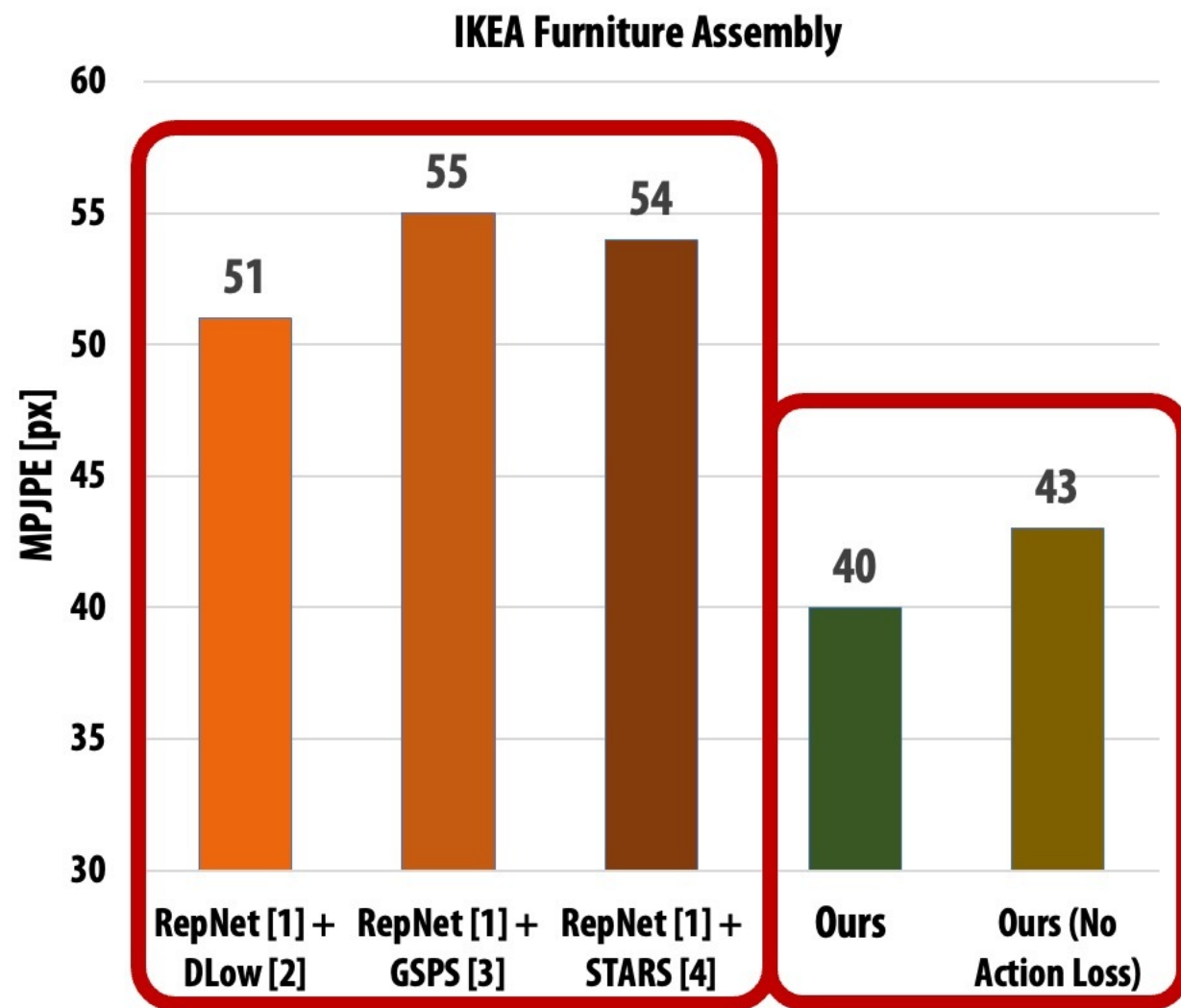
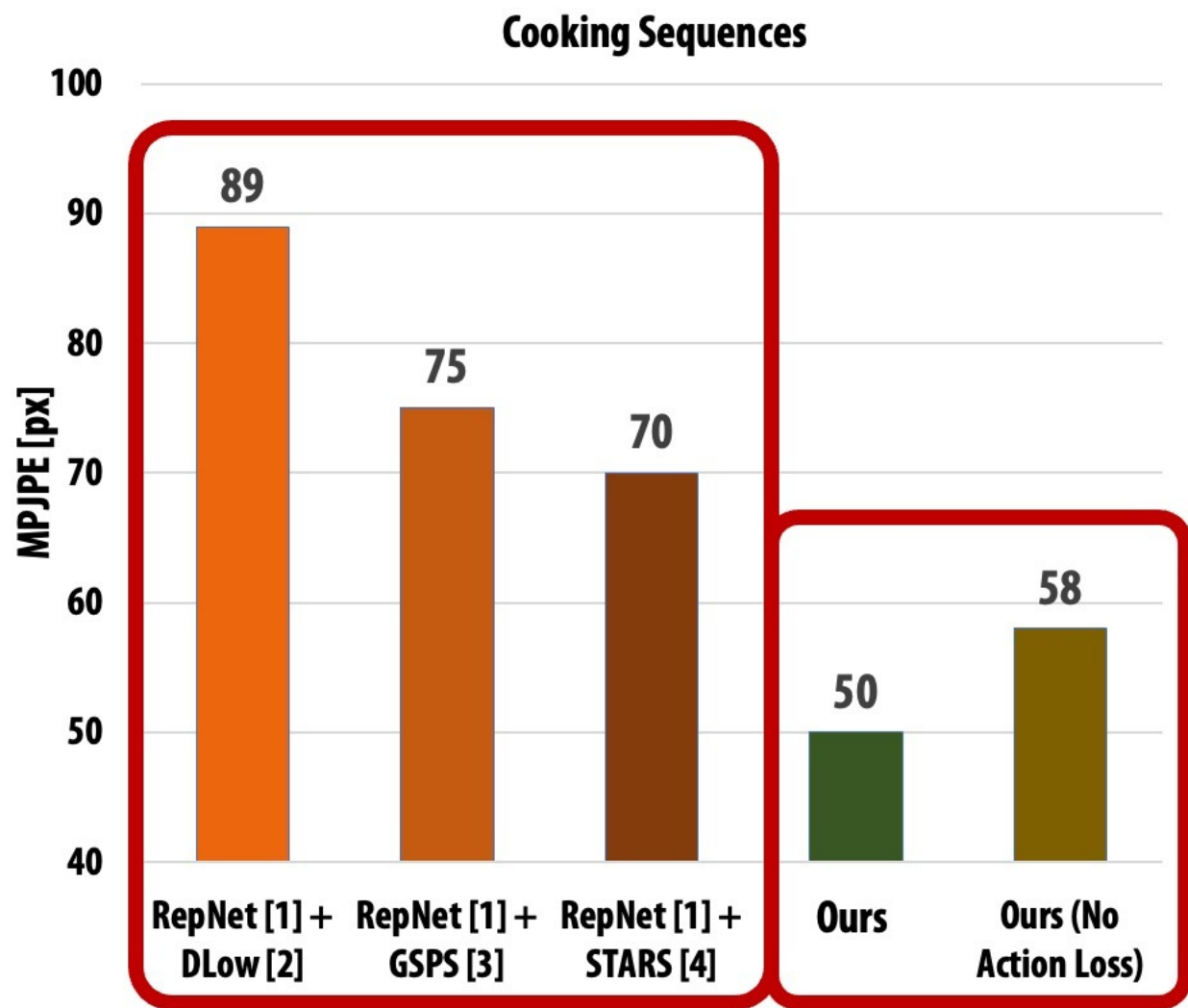
“spin”

“align”

“spin”

“spin”<sup>29</sup>

# Results: 3D Pose Forecasting – 2D Joint Error



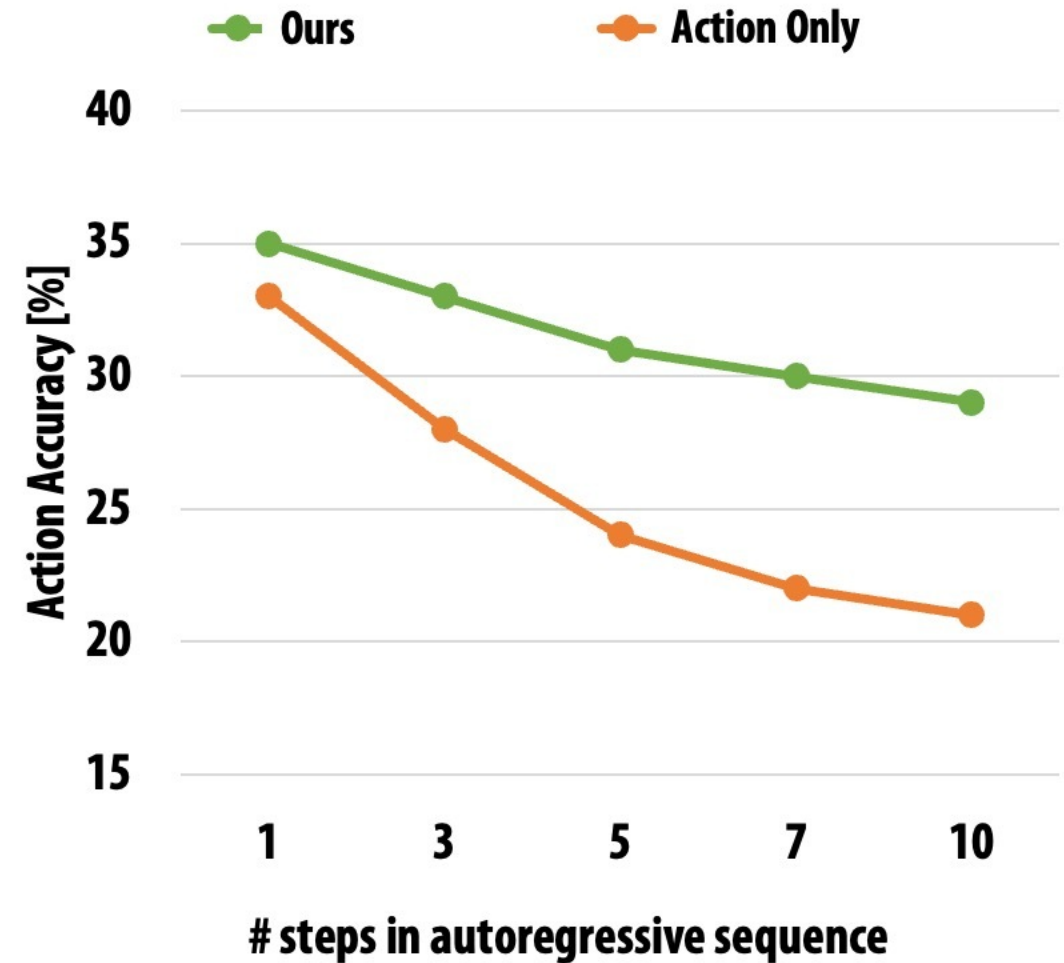
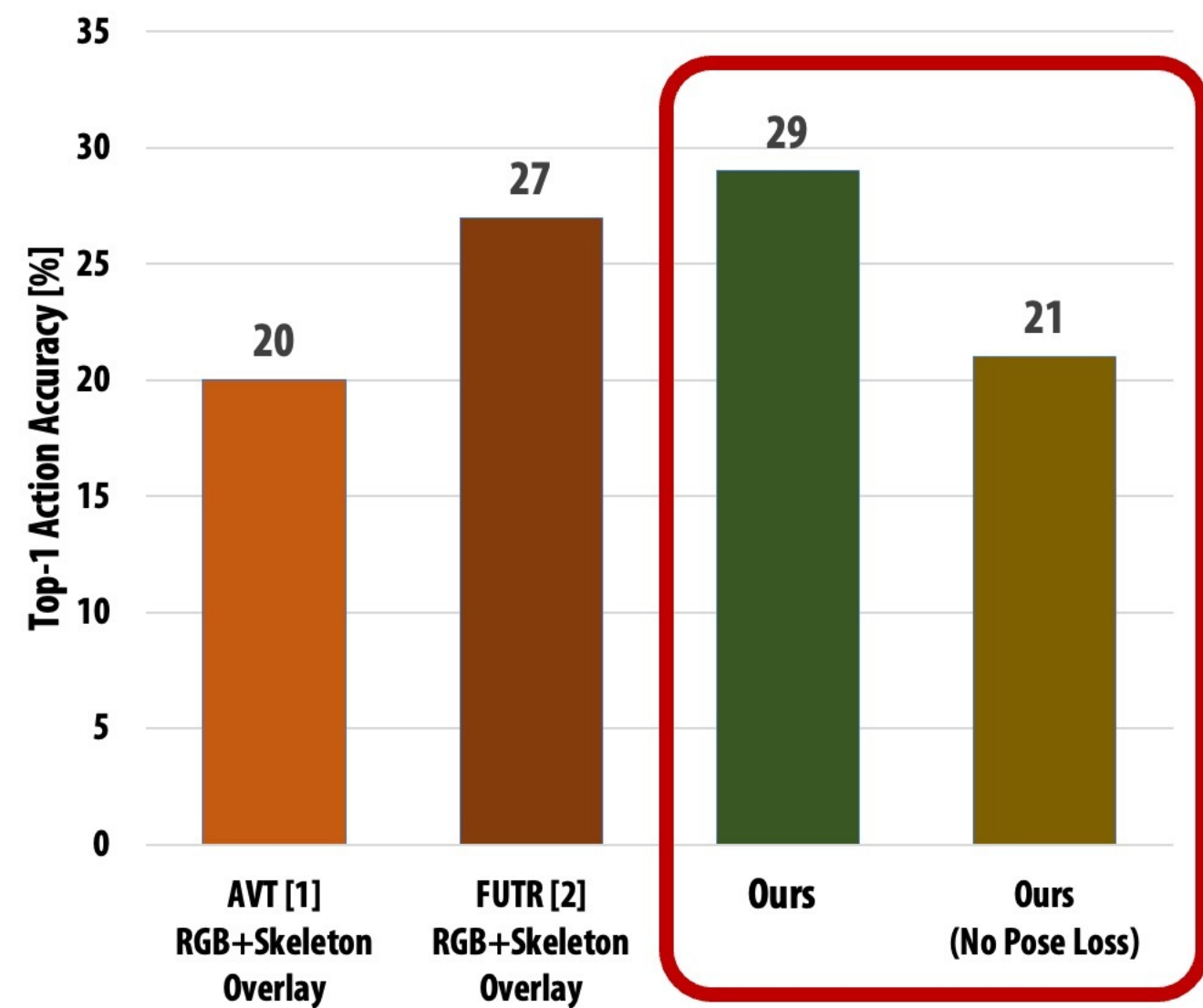
[1] Wandt, Bastian, and Bodo Rosenhahn. "Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[2] Yuan, Ye, and Kris Kitani. "Dlow: Diversifying latent flows for diverse human motion prediction." ECCV 2020.

[3] Mao, Wei, Miaomiao Liu, and Mathieu Salzmann. "Generating Smooth Pose Sequences for Diverse Human Motion Prediction." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[4] Xu, Sirui, Yu-Xiong Wang, and Liang-Yan Gui. "Diverse Human Motion Prediction Guided by Multi-level Spatial-Temporal Anchors." ECCV 2022.

# Results: Action Forecasting – Action Accuracy



[1] Girdhar, Rohit, and Kristen Grauman. "Anticipative video transformer." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[2] Gong, Dayoung, et al. "Future transformer for long-term action anticipation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2022.

# 3D Human Behavior Generation: Action & Interaction

## Efficient Action Representation

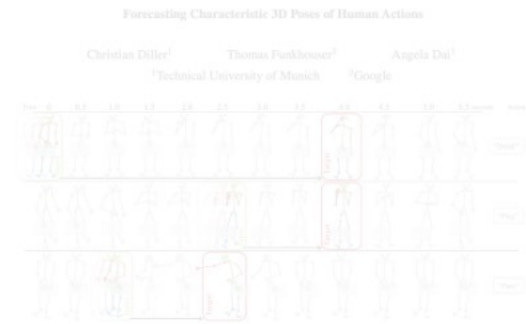


Figure 1. For a real-world 3d motion sequence of a human performing an action, we propose to forecast the semantically meaningful characteristic 3d poses representing the critical part of this sequence. In input, we take a short observation of a sequence of consecutive poses leading up to the target characteristic pose. Thus, we propose to take a goal-oriented approach, predicting the key moments characterizing human behavior, instead of predicting continuous motion, which can occur in varying speeds with predictions more easily drifting for longer time (1-3) predictions. We develop an attention-driven probabilistic approach to capture the most likely states 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 more so define the poses by dense intermediate and multi-component and uninformative aspects of human motion. Instead, we define a semantically meaningful pose prediction and that decouples the predicted pose from time, taking inspiration from goal-oriented 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 poses. 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 significantly outperforms state-of-the-art methods by 20% 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 real-world interactions with humans, such as autonomous robots or vehicles. In fact, prediction is considered to play a foundational part in intelligence [1, 2, 3]. 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 actions towards the forecasted future. For example, a robotic surgical assistant should predict in advance where to place a tool to avoid the surgeon’s next action, what sensor

## Complex Action Sequences

### FutureHuman3D: Forecasting Complex Long-Term 3D Human Behavior from Video Observations

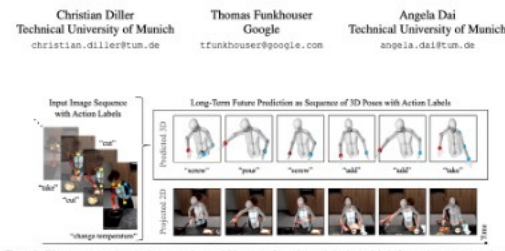


Figure 1. We propose a novel generative approach to model long-term future human behavior by jointly forecasting a sequence of coarse action labels and their concrete realizations as 3D body poses. For broad applicability, our autoregressive method only requires weak supervision and past observations 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 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 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 outperforms 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 behavior, 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

## Human-Object Interactions

### CG-HOI: Contact-Guided 3D Human-Object Interaction Generation



Figure 1. We present an approach to generate realistic 3D human-object interactions (HOI) from a text description and prior state, object geometry to be interacted with (e.g., chair). Our main insight is to explicitly model contact (modeled as forces on the body mesh, close contact is soft, in contact with human and object surfaces, in a pose diffusion process). In addition to synthesizing HOI 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 (HOI) from text. We model the motion of both human and object in an independent fashion, as semantically rich human motion words happen in isolation without any interaction. Our key insight is that explicitly modeling contact between the human body surface and object geometry can be used as strong priors 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, semi-correlated through cross-attention. We show that this learned contact first guidance during inference constrains of realistic, coherent HOI. Extensive evaluation shows that our joint contact-based human-object interaction approach generates realistic and physically plausible sequences, and we show two ap-

plications highlighting the capabilities of our method. Our diffusion on a given object trajectory, we can generate the corresponding human motion without training, allowing 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 motion planning, animation, robotic systems, scene layout planning, or human behavior simulation. Crucially, human interactions in interdependencies with the object(s) being interacted with the object (e.g., a chair or table, for instance, constrains the possible human motions with the object (e.g., sitting, lying), 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 human, and thereby disregarding their interdepend-

## Forecasting Characteristic 3D Poses [1]

## FutureHuman3D [2]

## CG-HOI [3]

[1] Diller, Christian, Thomas Funkhouser, and Angela Dai. "Forecasting characteristic 3d poses of human actions." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[2] Diller, Christian, 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.

[3] Diller, Christian, and Angela Dai. "Cg-hoi: Contact-guided 3d human-object interaction generation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.



# 3D Human Behavior Generation: Action & Interaction

## Efficient Action Representation

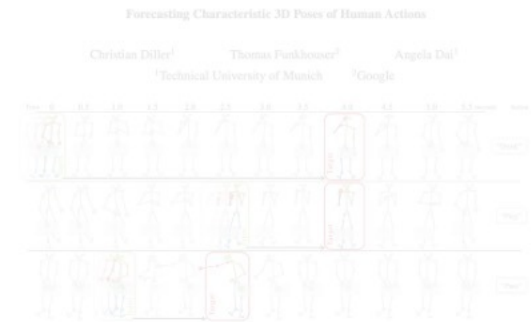


Figure 1. For a real-world 3d motion sequence of a human performing an action, we propose to forecast the semantically meaningful characteristic 3d pose representing the action goal for this sequence. To do so, we take a short observation of a sequence of consecutive poses leading up to the target characteristic pose. Thus, we propose to take a goal-oriented approach, predicting the key moments characterizing human behavior, instead of predicting continuous motion, which can occur in varying speeds with predictions more easily drifting for longer time (> 1s) predictions. We develop an attention-driven prediction approach to capture the most likely states 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 fully 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 focuses lower poses at fixed time intervals. Although more so define the poses by dense intermediate and multi-component and sequential aspects of human action. Instead, we define a semantically meaningful pose prediction task that decouples the predicted pose from time, taking inspiration from goal-oriented behavior. To predict characteristic poses, we propose a prediction 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 poses. To evaluate our

method, we construct a dataset of manually annotated characteristic 3d poses. Our experiments with this dataset suggest that our proposed prediction approach outperforms state-of-the-art methods by 20% in coverage.

### 1. Introduction

Future human pose forecasting is fundamental towards a comprehensive understanding of human behavior, and consequently towards addressing higher-level perception in real-world interactions with humans, such as autonomous robots or vehicles. In fact, prediction is considered to play a foundational part in intelligence [1, 2, 3]. In particular, predicting the 3d pose of a human in the future lays a basis for both semantic and semantic understanding of human behavior, and for an agent to take fine-grained anticipatory actions 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

## Complex Action Sequences



Figure 1. We propose a novel generative approach to model long-term future human behavior by jointly forecasting a sequence of future action labels and their associated conditions in 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 reconstructed 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 (coverage, costs, equipment, safety) but easy to acquire in 2D (simple RGB cameras). Thus, we design our method to only require 2D RGB data while being able to generate 3D human motion sequences. We use a differentiable 3D 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 make this in a semantically hierarchical manner, jointly predicting high-level coarse action labels together with their fine-level fine-grained realizations as characteristic 3D body poses. We illustrate that these low-action representations are easy to learn, and pose prediction benefits both action and pose forecasting. Our experiments demonstrate the complementary nature of pose action and 3D pose prediction: our pose approach outperforms each task trained individually, enables longer-range sequence prediction, and outperforms alternative approaches to forecast actions and characteristic 3D poses.

### 1. Introduction

Predicting future human behavior is fundamental to real-world intelligence, with many applications in content creation, robotics, virtual reality, and more. For instance, a monitoring system might learn early warnings of potentially dangerous behaviors, and a robotic assistant can use this coming 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 critical steps to learn 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 to a high level is not sufficient, it must also reason about where the action will occur. Actions such as “grab a tool” are likely failures when performed in a location 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 forecasted humans.

To support these types of applications, we now 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

## Human-Object Interactions

### 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

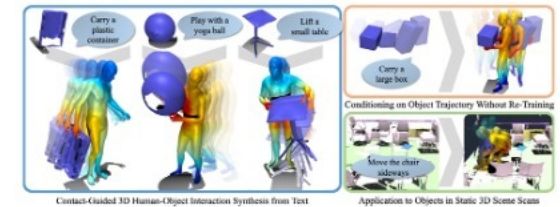


Figure 1. We present an approach to generate realistic 3D human-object interactions (HOIs), from a text description and given static object geometry to be interacted with (left). Our main insight is to explicitly model contact (visualized as colors on the body mesh, closer contact in red), tandem with 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 synthesis of realistic, coherent HOIs. Extensive evaluation shows that our joint contact-based human-object interaction approach generates realistic and physically plausible sequences, and we show two ap-

lications highlighting the capabilities of our method. Conditioned on a given object trajectory, we can generate the corresponding 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 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

## Forecasting Characteristic 3D Poses [1]

## FutureHuman3D [2]

## CG-HOI [3]

[1] Diller, Christian, Thomas Funkhouser, and Angela Dai. "Forecasting characteristic 3d poses of human actions." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[2] Diller, Christian, 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.

[3] Diller, Christian, and Angela Dai. "Cg-hoi: Contact-guided 3d human-object interaction generation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

# CG-HOI: Contact-Guided 3D Human-Object Interaction Generation

How to model realistic  
human-object interactions?



Move the  
chair

**Christian Diller**



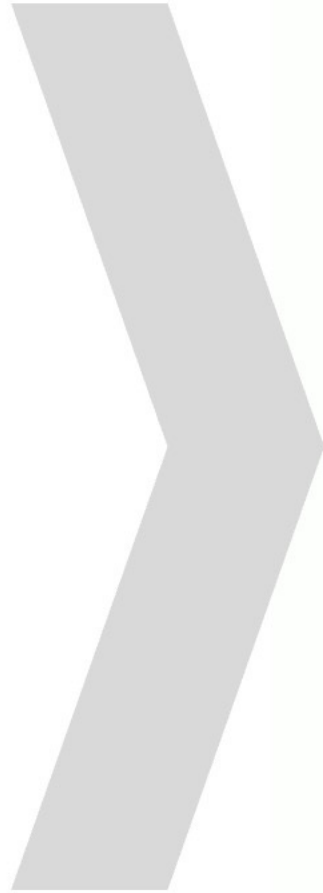
Carry a  
suitcase

**Angela Dai**

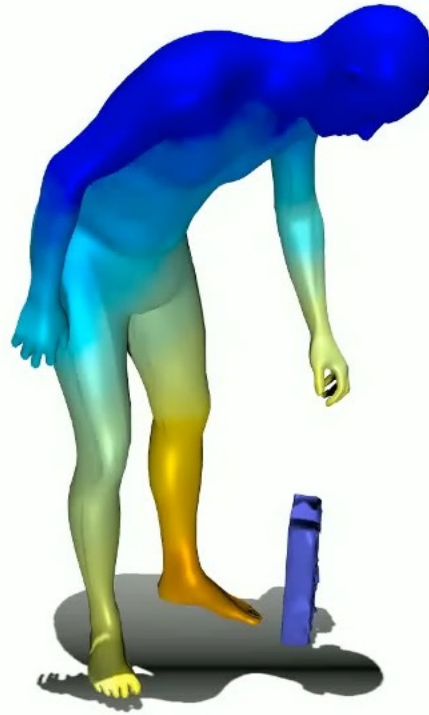


# Task: Joint Human-Object Motion Generation

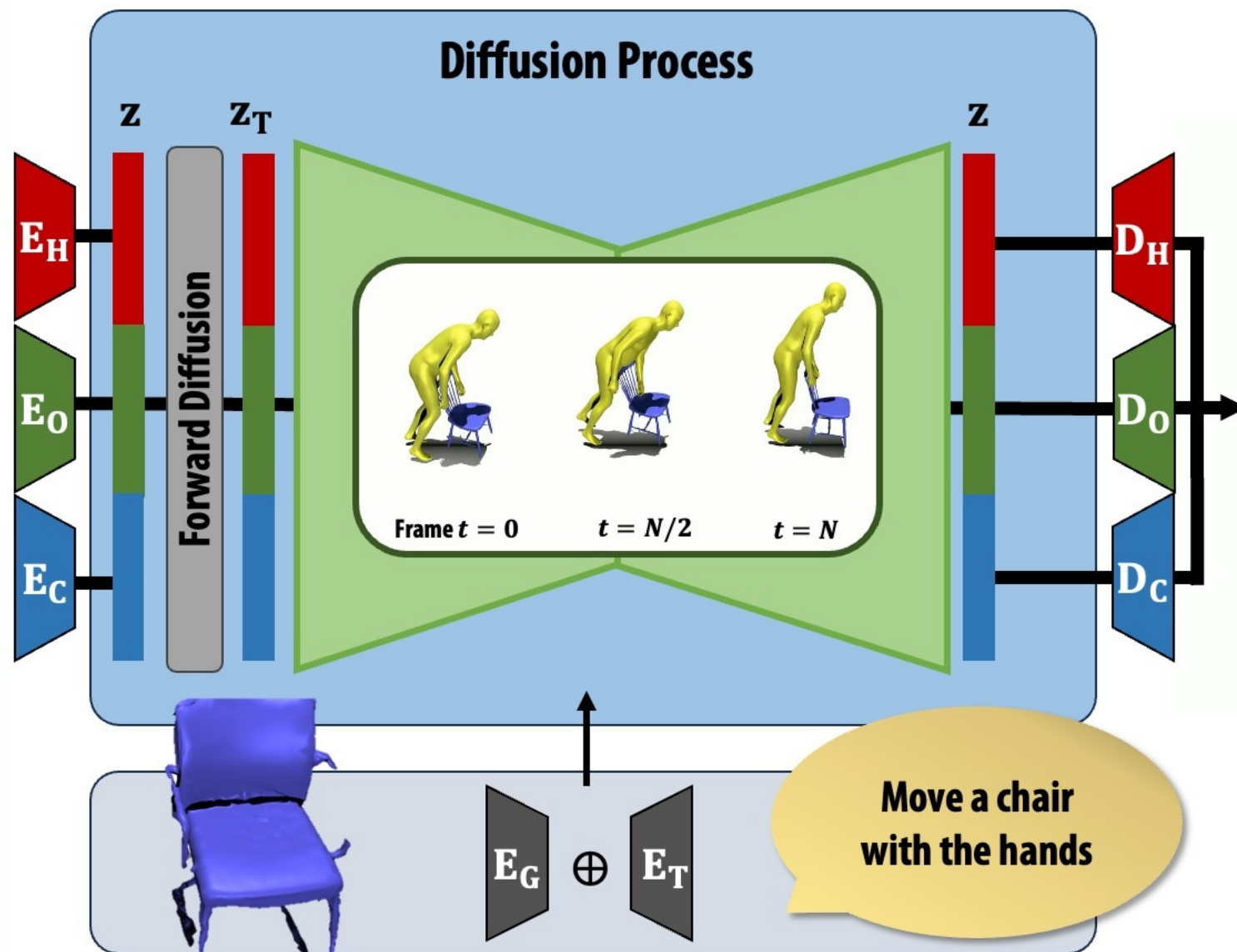
Move a chair with  
the hand



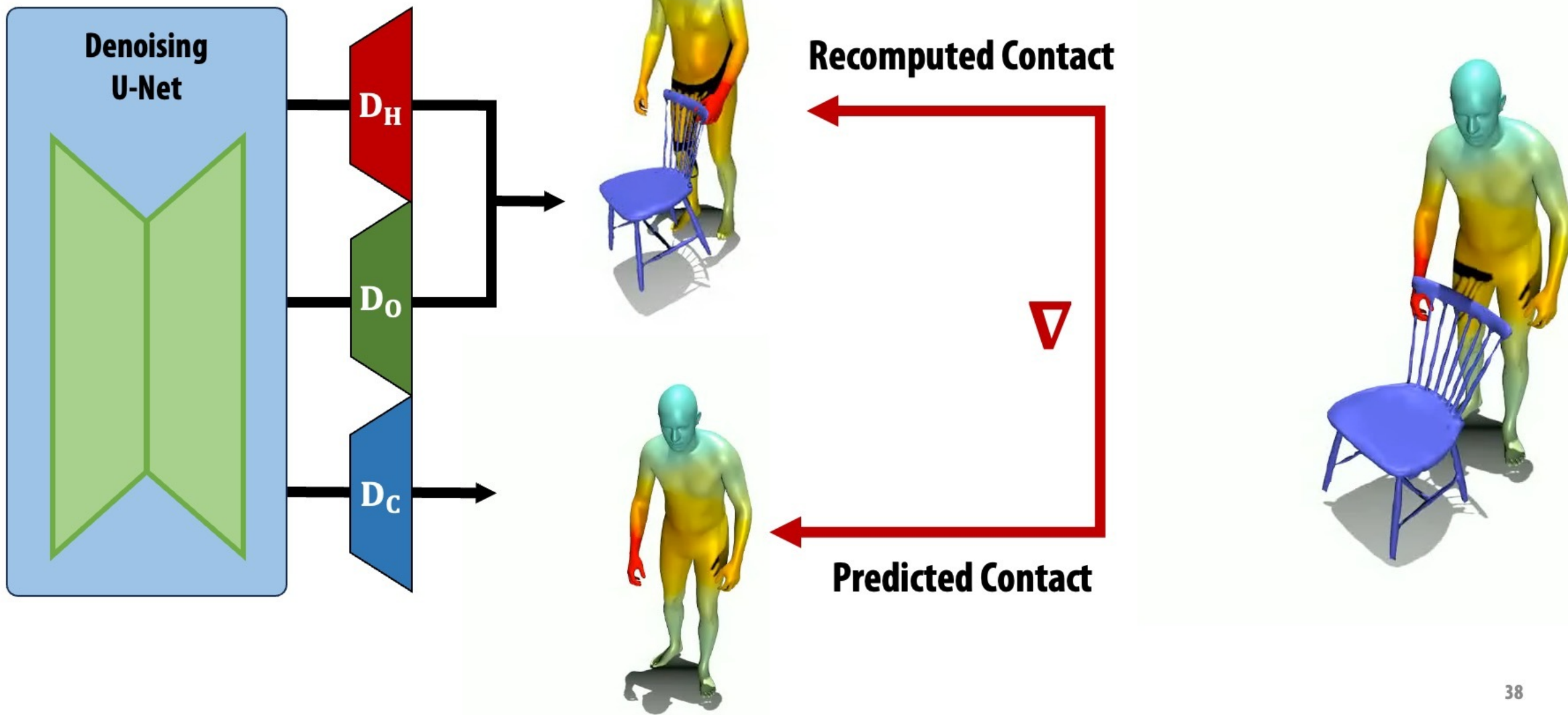
# Approach: Contact Modeling



# Method: Joint Training



# Method: Inference



# Results: Qualitative



Carry a suitcase



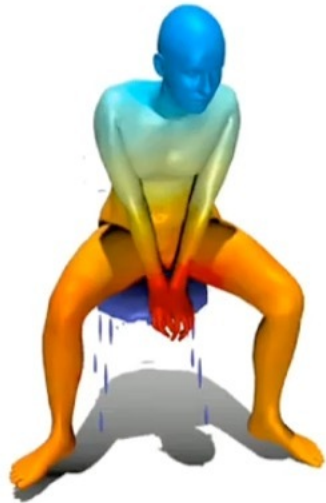
Condition

Move the chair backwards



Condition







# Results: Qualitative Comparison to Baseline MDM [1]



Move a small  
table

**Condition**



**MDM [1]**



**Ours**



Carry a  
backpack in  
the hand



# Results: Qualitative Comparison to Baseline InterDiff [1]



Move a chair  
with the hand

**Condition**



**InterDiff [1]**



**Ours**



Carry a trash  
bin



# Results: Ablation Study

Move a  
yogamat



**No Contact Modeling**

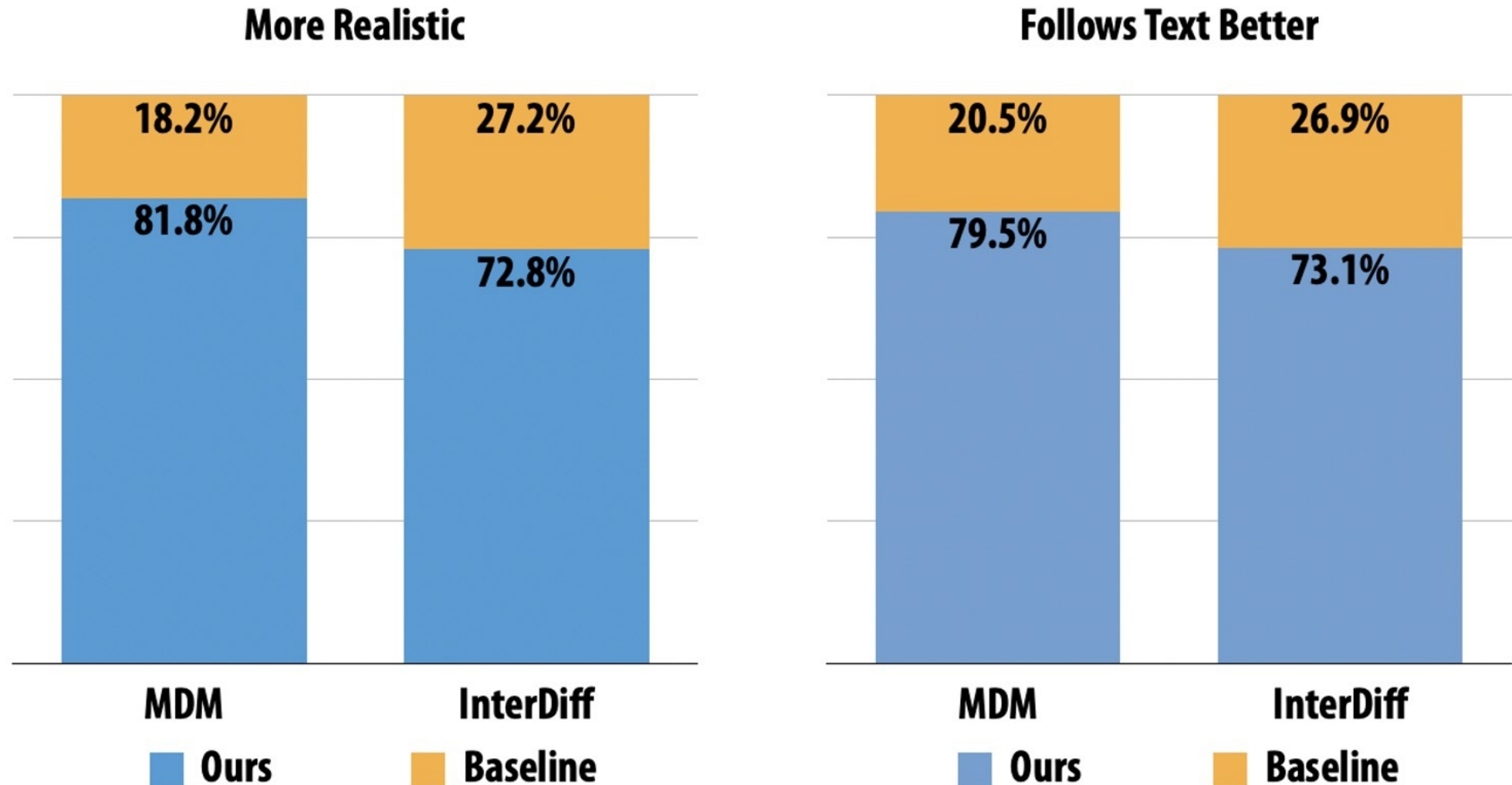


**No Contact Guidance**



**Ours (Full)**

# Results: Quantitative – User Study

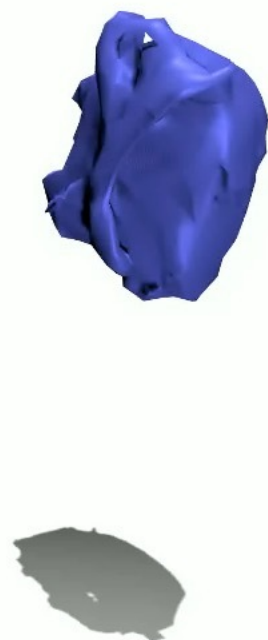


# Results: Quantitative

		BEHAVE				CHAIRS			
Task	Approach	R-Prec. (top-3) $\uparrow$	FID $\downarrow$	Diversity $\rightarrow$	MModality $\rightarrow$	R-Prec. (top-3) $\uparrow$	FID $\downarrow$	Diversity $\rightarrow$	MModality $\rightarrow$
	Real (human)	0.73	0.09	4.23	4.55	0.83	0.01	7.34	3.00
Text-Cond. Human Only	MDM [71]	0.52	4.54	5.44	5.12	0.72	5.99	6.83	3.45
	InterDiff [84]	0.49	5.36	3.98	3.98	0.63	6.76	5.24	2.44
	<b>Ours</b>	<b>0.60</b>	<b>4.26</b>	<b>4.92</b>	<b>4.10</b>	<b>0.78</b>	<b>5.24</b>	<b>7.90</b>	<b>3.22</b>
	Real	0.81	0.17	6.80	6.24	0.87	0.02	9.91	6.12
Motion- Cond. HOI	InterDiff [84]	0.68	3.86	5.62	5.90	0.67	4.83	7.49	4.87
	<b>Ours</b>	<b>0.71</b>	<b>3.52</b>	<b>6.89</b>	<b>6.43</b>	<b>0.79</b>	<b>4.01</b>	<b>8.42</b>	<b>6.29</b>
Text- Cond. HOI	MDM [71]	0.49	9.21	6.51	8.19	0.53	9.23	6.23	7.44
	InterDiff [84]	0.53	8.70	3.85	4.23	0.69	7.53	5.23	4.63
	<b>Ours</b>	<b>0.62</b>	<b>6.31</b>	<b>6.63</b>	<b>5.47</b>	<b>0.74</b>	<b>6.45</b>	<b>8.91</b>	<b>5.94</b>

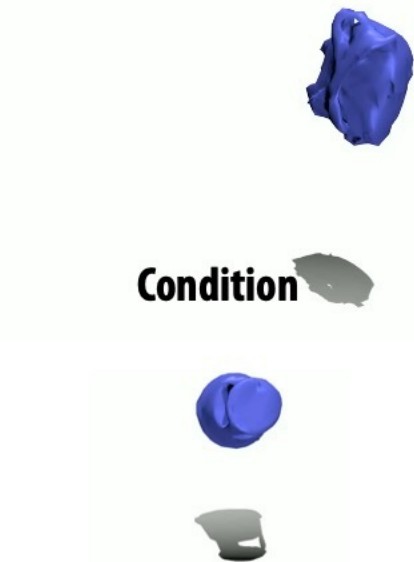
# Application: Object Trajectory Guidance

Generation



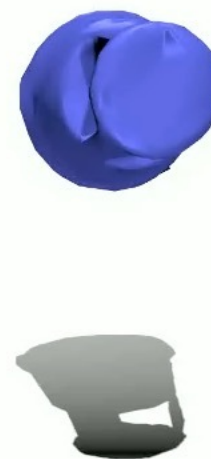
Carry a backpack on the back

Condition



Move a stool

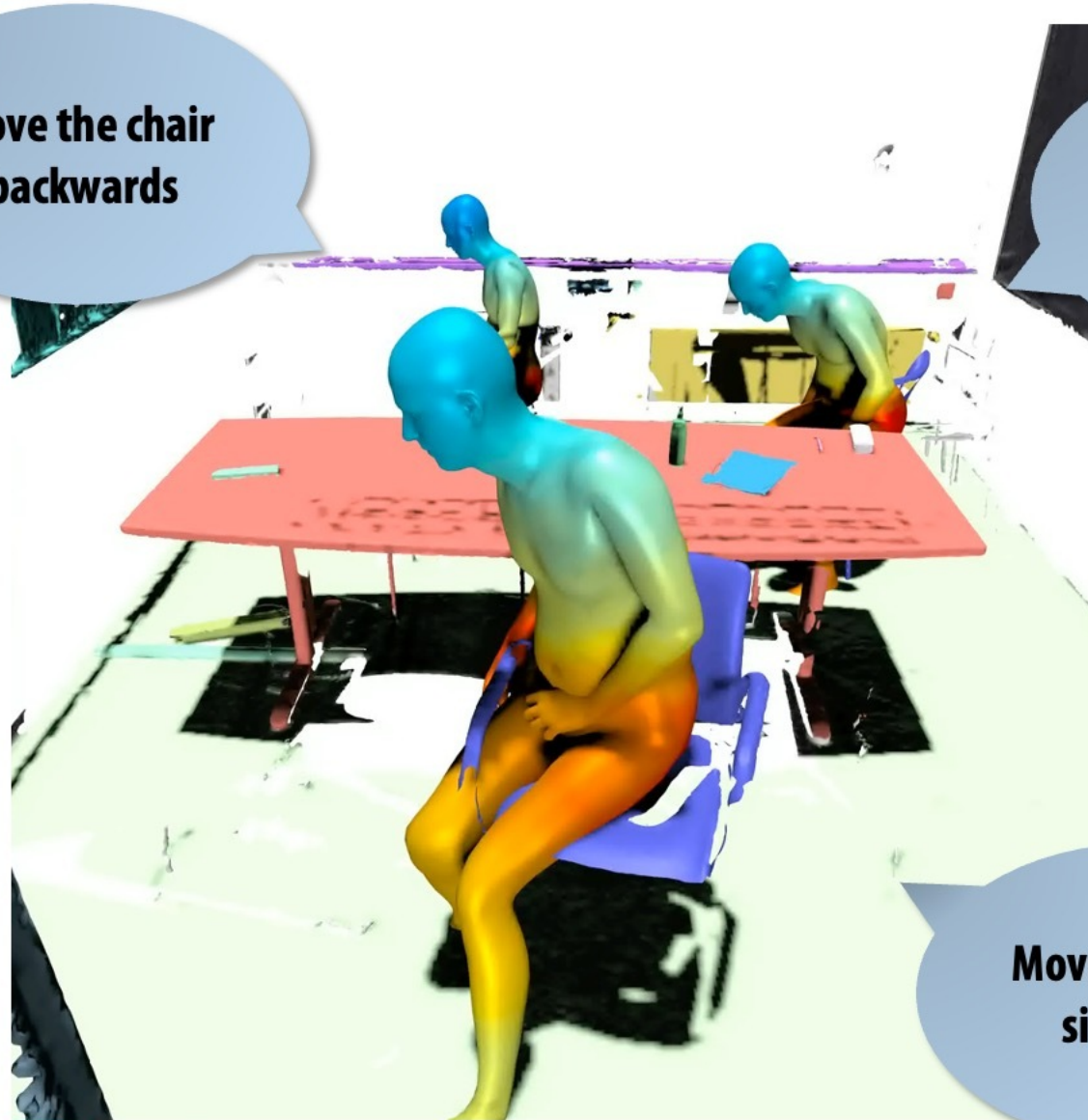
Generation



# Application: 3D Static Scene Population

**Move the chair  
backwards**

**Adjust the chair**



**Move the chair  
sideways**

# 3D Human Behavior Generation: Action & Interaction

## Efficient Action Representation

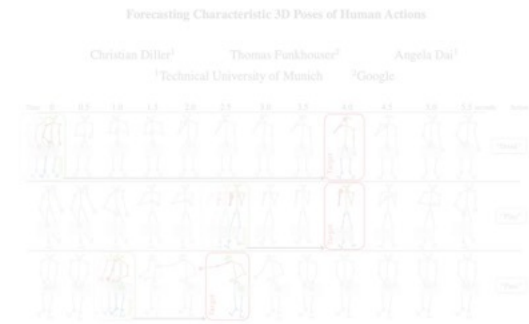


Figure 1. For a real-world 3D motion sequence of a human performing an action, we propose to forecast the semantically meaningful characteristic 3D poses representing the action goal for this sequence. To begin, we take a short observation of a sequence of consecutive poses leading up to the target characteristic pose. Thus, we propose to take a goal-oriented approach, predicting the key elements characterizing human behavior, instead of predicting continuous motion, which can occur in varying speeds with predictions more easily diverging for longer time (> 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 users do desire this from a more formalized and well-organized 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 poses. To evaluate our

method, we construct a dataset of manually annotated characteristic 3D poses. Our experiments with this dataset indicate that our proposed probabilistic approach significantly outperforms the current methods by 20% 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 real-world interactions with humans, such as autonomous vehicles or vehicles. In fact, prediction is considered to play a foundational part in intelligence [1], [2], [3]. 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 to place a tool to assist the surgeon's next action, what would

## Complex Action Sequences

### FutureHuman3D: Forecasting Complex Long-Term 3D Human Behavior from Video Observations



Figure 1. We propose a novel generative approach to model long-term future human behavior by jointly forecasting a sequence of future action labels and their corresponding motions in 3D body poses. For broad applications, our autoregressive method only requires weak supervision and past observations in the form of 2D RGB video data, together with a database of reconstructed 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 more downstream applications. The required ground truth data is best to require as 3D coverage (i.e., repetitive actions) but easy to acquire in 2D (single RGB camera). Thus, 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 sub-actions. We tackle this by a semantically hierarchical manner, jointly predicting high-level coarse action labels together with their fine-level fine-grained realizations as characteristic 3D body 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 trained individually, enables robust longer-term sequence prediction, and supersedes alternative approaches to forecast actions and characteristic 3D poses.

### 1. Introduction

Predicting future human behavior is fundamental to real-world intelligence, with many applications in contact analysis, robotics, virtual reality, and more. For instance, a monitoring system might issue early warnings of potentially dangerous behavior, and a robotic assistant can use fine-grained 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 designed 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 an active table top 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

## Human-Object Interactions

### 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

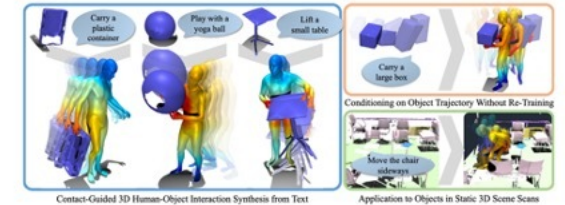


Figure 1. We present an approach to generate realistic 3D human-object interactions (HOIs), from a text description and given static object geometry to be interacted with (left). Our main insight is to explicitly model contact (visualized as codes on the body mesh, closer contact in red), in tandem with 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 interactive 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 synthesis of realistic, coherent HOIs. Extensive evaluation shows that our joint contact-based human-object interaction approach generates realistic and physically plausible sequences, and we show two ap-

lications 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

## Forecasting Characteristic 3D Poses [1]

## FutureHuman3D [2]

## CG-HOI [3]

[1] Diller, Christian, Thomas Funkhouser, and Angela Dai. "Forecasting characteristic 3d poses of human actions." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[2] Diller, Christian, 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.

[3] Diller, Christian, and Angela Dai. "Cg-hoi: Contact-guided 3d human-object interaction generation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.



# 3D Human Behavior Generation: Action & Interaction

## Efficient Action Representation

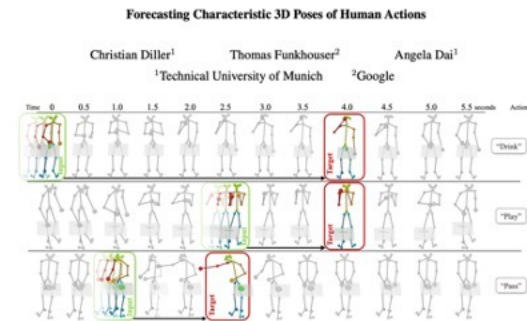


Figure 1. For a real-world 3d skeleton sequence of a human performing an action, we propose to forecast the semantically meaningful characteristic 3d pose, representing the action goal for this sequence. As input, we take a short observation of a sequence of consecutive poses leading up to the target 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 conflates 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 perceptions 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

## Complex Action Sequences

### FutureHuman3D: Forecasting Complex Long-Term 3D Human Behavior from Video Observations

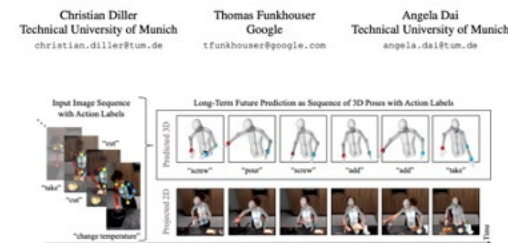


Figure 1. We propose a novel generative approach to model long-term future human behavior by jointly forecasting a sequence of coarse action labels and their concrete realizations as 3D body poses. For broad applicability, our autoregressive method only requires weak supervision and past observations 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 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 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 outperforms 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

## Human-Object Interactions

### CG-HOI: Contact-Guided 3D Human-Object Interaction Generation

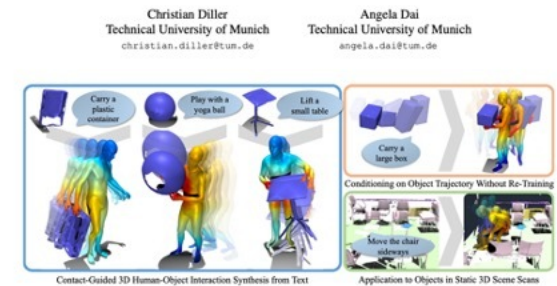


Figure 1. We present an approach to generate realistic 3D human-object interactions (HOIs), from a text description and given static object geometry to be interacted with (left). Our main insight is to explicitly model contact (visualized as codes on the body mesh, closer contact in red), in tandem with 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 synthesis of realistic, coherent HOIs. Extensive evaluation shows that our joint contact-based human-object interaction approach generates realistic and physically plausible sequences, and we show two ap-

lications 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

## Forecasting Characteristic 3D Poses [1]

## FutureHuman3D [2]

## CG-HOI [3]

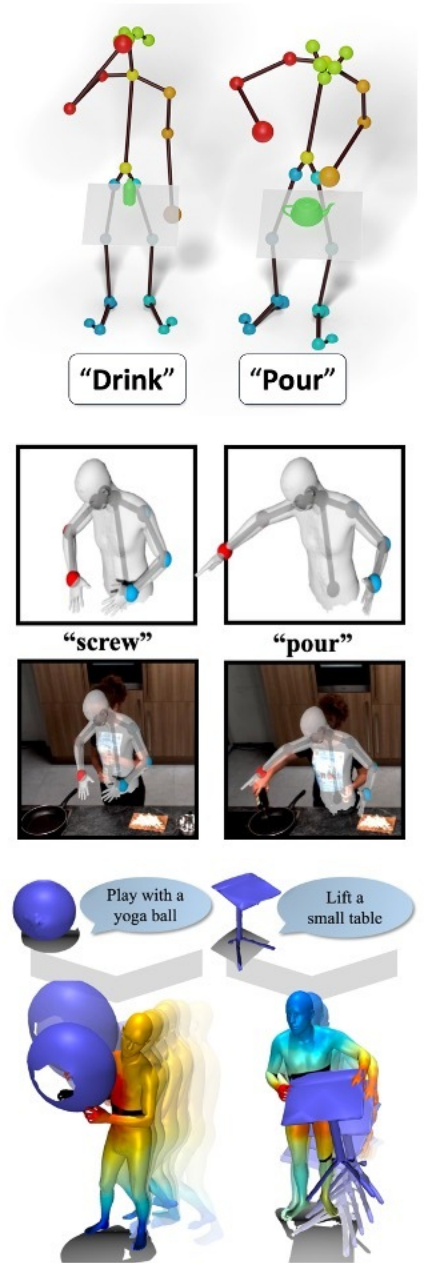
[1] Diller, Christian, Thomas Funkhouser, and Angela Dai. "Forecasting characteristic 3d poses of human actions." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[2] Diller, Christian, 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.

[3] Diller, Christian, and Angela Dai. "Cg-hoi: Contact-guided 3d human-object interaction generation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

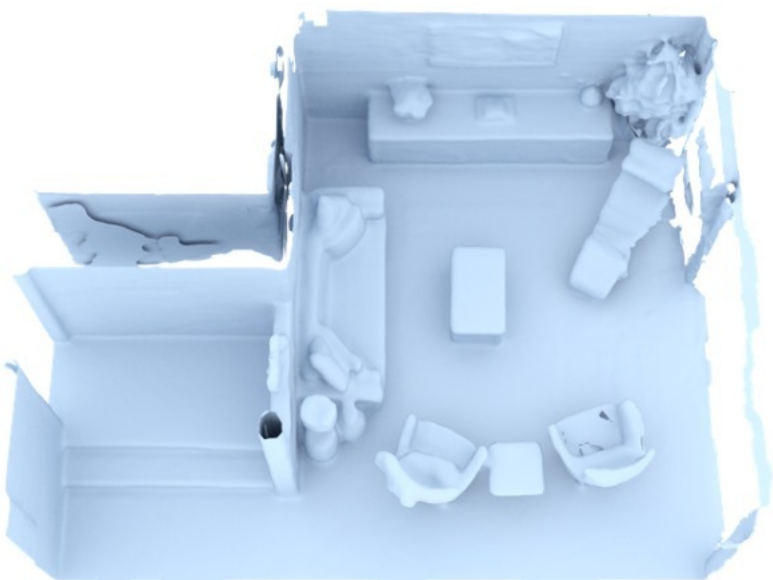
# Summary and Conclusion

- **Predicting future characteristic 3D poses of human actions**
  - Probabilistic approach for capturing the most likely future 3D action poses
- **Forecasting complex long-term 3D human behavior from 2D**
  - Joint action and 3D pose forecasting of composite long-term behavior
- **Contact-Guided 3D Human-Object Interactions**
  - Realistic human-object interaction generation from text and geometry



# Outlook: 3D Scene Understanding

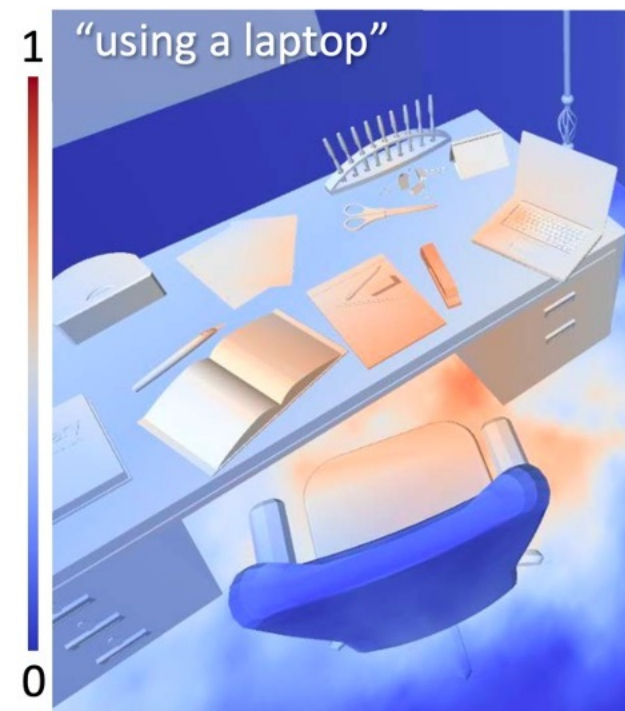
Reconstruction [1]



Semantic Instance Segmentation [2]



Affordance Prediction [3]

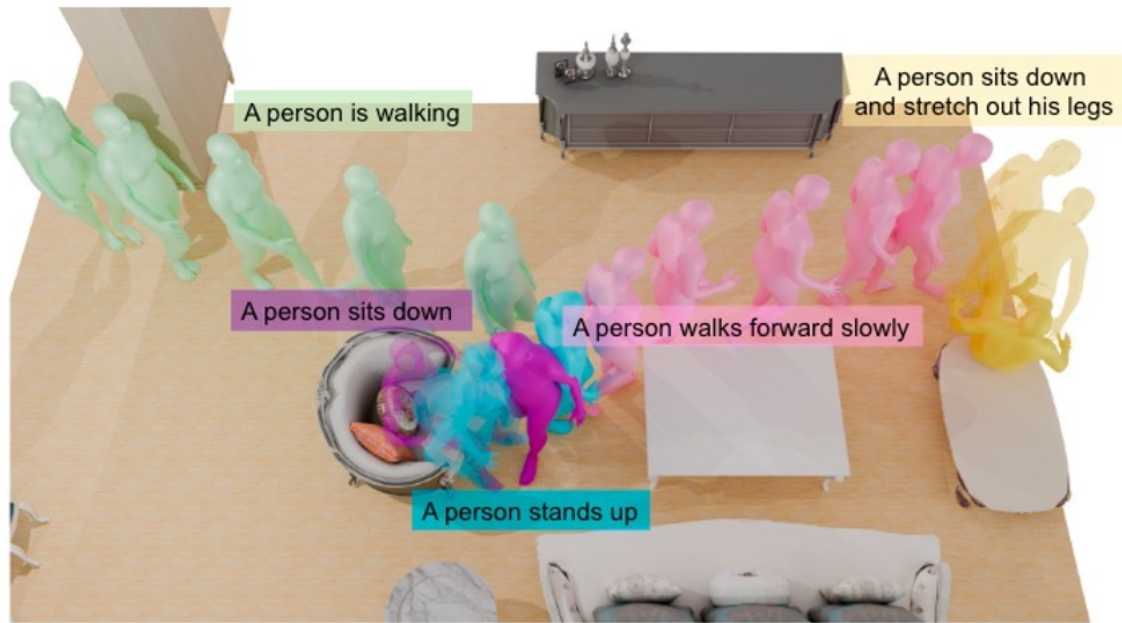


[1] Dai, Angela, Christian Diller and Matthias Nießner. "SG-NN: Sparse Generative Neural Networks for Self-Supervised Scene Completion of RGB-D Scans." 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020): 846-855.

[2] Hou, Ji, Angela Dai and Matthias Nießner. "3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans." 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019): 4416-4425.

[3] Savva, Manolis, Angel X. Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. "SceneGrok: Inferring action maps in 3D environments." ACM transactions on graphics (TOG) 33, no. 6 (2014): 1-10.

# Outlook: Dynamic Human Interactions in 3D Scenes



**Text-based motion and interaction [1]**

*The person walks forward from the curtain to pick up his guitar.*



*The person cartwheels towards the campfire from the table.*



**Zero-shot path-finding with large language models [2]**

[1] Yi, Hongwei, et al. "Generating human interaction motions in scenes with text control." European Conference on Computer Vision. Springer, Cham, 2024.

[2] Qu, Haoxuan, Ziyang Guo, and Jun Liu. "GPT-Connect: Interaction between Text-Driven Human Motion Generator and 3D Scenes in a Training-free Manner." arXiv preprint arXiv:2403.14947 (2024).

# Thank You!



**Prof. Angela Dai**



**Prof. Michael Black**

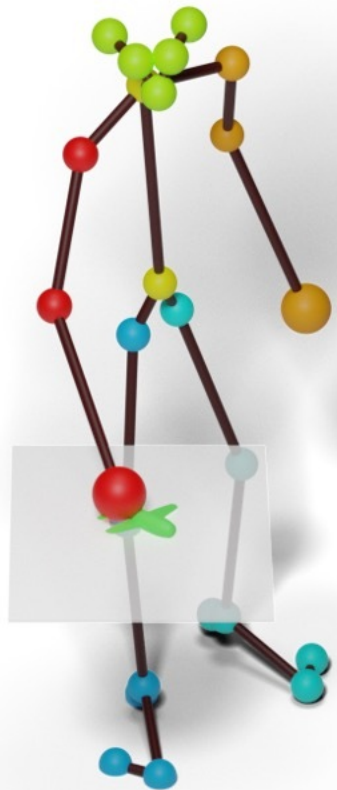


**Prof. Stefan Leutenegger**



**Prof. Thomas Funkhouser**

# 3D Human Behavior Generation through Action and Interaction Synthesis



**Christian Diller**  
Supervisor: Prof. Angela Dai

**Tuesday, 10<sup>th</sup> December 2024**

