

Forecasting Characteristic 3D Poses of Human Actions

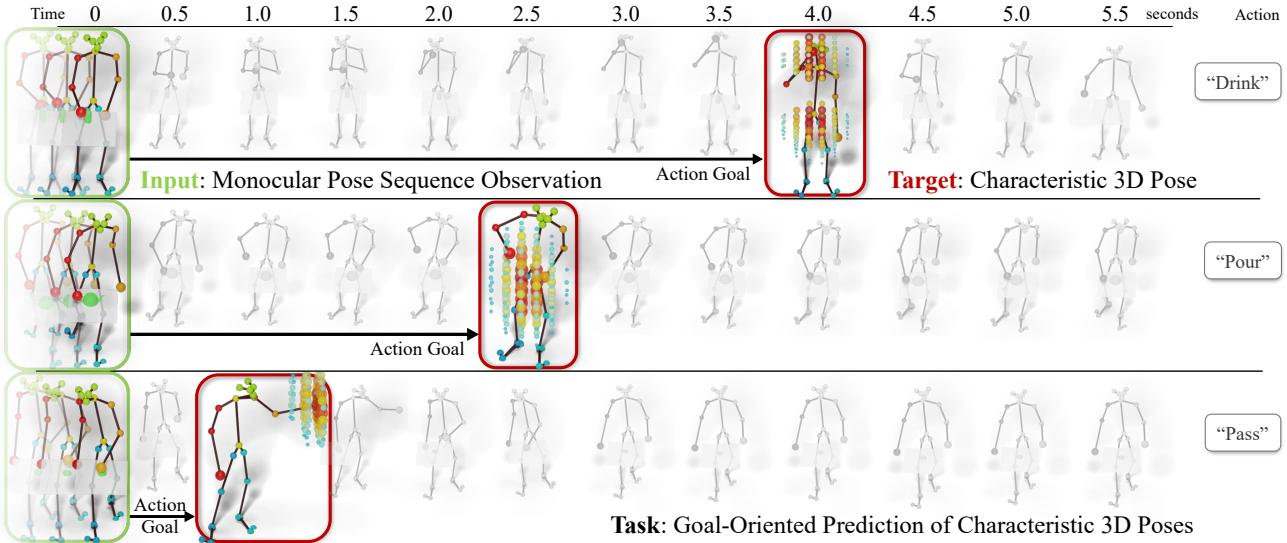
Christian Diller¹Thomas Funkhouser²Angela Dai¹¹Technical University of Munich²Google

Figure 1: From a monocular pose sequence observation, we propose to forecast a future *characteristic 3D pose* defining a semantically meaningful key moment in a future action of the observed person. Instead of predicting continuous motion, which can occur at varying speeds with predictions more easily diverging for longer-term (>1 s) predictions, we propose to take a *goal-oriented* approach, predicting the key moments characterizing future behavior. To capture these characteristic poses, which often occur at longer-term intervals in the future, we develop a 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 monocular video observation of a person, to predict a future 3D pose of that person in a likely action-defining, characteristic pose – for instance, from observing a person reaching for a banana, predict the pose of the person eating the banana. Prior work on human motion prediction estimates future poses at fixed time intervals. Although easy to define, this frame-by-frame formulation confounds temporal and intentional aspects of human action. Instead, we define a semantically meaningful pose prediction task that decouples the predicted pose from time, taking inspiration from goal-directed behavior. To predict characteristic poses, we propose a probabilistic approach that first models the possible multi-modality in the distribution of likely characteristic poses. It then samples future pose hypotheses from the predicted distribution in an autoregressive fashion to model dependencies between joints and finally optimizes

the resulting pose with bone length and angle constraints. 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 22% 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, 15, 11]. 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 potentially take anticipatory action towards the forecasted future.

Recently, we have seen notable progress in the task of future 3D human motion prediction – from an initial observation of a person, forecasting the 3D behavior of that person up to ≈ 1 second in the future [12, 17, 25, 24, 23]. Various methods have been developed, leveraging RNNs [12, 17, 25, 14], graph convolutional neural networks [24, 21], and attention [29, 23]. However, these approaches all take a temporal approach towards forecasting future 3D human poses, and predict poses at fixed time intervals to imitate the fixed frame rate of camera capture. This makes it difficult to predict longer-term (several seconds) behavior, which requires predicting both the time-based speed of movement as well as the higher-level goal of the future action.

Thus, we propose to decouple the temporal and intentional behavior, and introduce a new task of forecasting *characteristic 3D poses* of a person’s future action: from a monocular pose sequence observation of a human, the goal is to predict the future pose of the person in its characteristic, action-defining moment. Fig. 2 visualizes the difference between this new task and the traditional, time-based approach: our task is to predict characteristic poses at action-defining moments (blue dots) rather than at fixed time-intervals (red dots). As shown in Fig. 1, the characteristic 3D poses are more semantically meaningful and rarely occur at exactly the same times in the future.

We believe that predicting these characteristic 3D poses takes an important step towards forecasting human action, by understanding the objectives underlying a future action or movement separately from the speed at which they occur.

Since future characteristic 3D poses often occur at longer-term intervals (> 1 s) in the future, there may be multiple likely modes of the characteristic poses, and we must capture this multi-modality in our forecasting. Rather than deterministic forecasting, as is the prevalent approach towards 3D human pose forecasting [24, 23, 21], we develop an attention-driven prediction of probability heatmaps representing the likelihood of each human pose joint in its future location. This enables generation of multiple, diverse hypotheses for the future pose. To generate a coherent pose

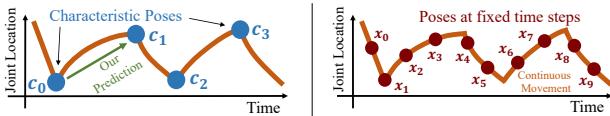


Figure 2: These plots depict the salient difference between our new task (left) and the traditional one (right). The red curve depicts the motion of one joint (e.g., hand position as a person drinks from a glass). It represents a typical piecewise continuous motion, which has discrete action-defining characteristic poses at cusps of the motion curves (e.g., grasping the glass on the table, putting it to ones mouth, etc.) separating smooth trajectories connecting them (e.g., raising or lowering the glass). Our task is to predict future characteristic poses (blue dots on left) rather than in-between poses at regular time intervals (red points on right).

prediction across all pose joints’ potentially multi-modal futures, we make autoregressive predictions for the end effectors of the actions (e.g., predicting the right hand, then the left hand conditioned on the predicted right hand location) – this enables a tractable modeling of the joint distribution of the human pose joints.

To demonstrate our proposed approach, we introduce a new benchmark on Characteristic 3D Pose prediction. We annotate characteristic keyframes in action sequences from the GRAB [28] and Human3.6M [16] datasets. Experiments with this benchmark show that our probabilistic approach outperforms alternative methods that take a deterministic approach to future 3D pose prediction by 22% on average.

In summary, we present the following contributions:

- We propose the task of forecasting *Characteristic 3D Poses*: predicting likely action-defining future moments from a sequence observation of a person, towards goal-oriented understanding of pose forecasting.
- We introduce an attention-driven, probabilistic approach to tackle this problem and model the most likely modes for characteristic poses, and show that it outperforms state-of-the-art, deterministic methods.
- We autoregressively model the multi-modal distribution of future pose joint locations, casting pose prediction as a product of conditional distributions of end effector locations (e.g., hands), and the rest of the body.
- We introduce a dataset and benchmark on our Characteristic 3D Pose prediction, comprising 1334 annotated characteristic pose frames from the GRAB [28] and 210 from the Human3.6M [16] datasets.¹

2. Related Work

Sequential Human Motion Forecasting. Many works have focused on the task of human motion forecasting, cast as a sequential prediction problem to predict a sequence of human poses according to the fixed frame rate capture of a camera. For this sequential task, recurrent neural networks have been widely employed for human motion forecasting [12, 17, 25, 1, 10, 31, 13]. Such approaches have achieved impressive success in shorter-term prediction (up to 1 second, occasionally several seconds for longer term predictions), but the RNN summarization of history into a fixed-size representation struggles to maintain the long-term dependencies needed for forecasting further into the future.

To address some of the drawbacks of RNNs, non-recurrent models have also been adopted, encoding temporal history with convolutional or fully connected networks [5, 20, 24], or attention [29, 23]. Li et al. [32] proposed an

¹Code and dataset to be released.

auto-conditioned approach enabling synthesizing pose sequences up to 300 seconds of periodic-like motions (walking, dancing). However, these works all focus on frame-by-frame synthesis, with benchmark evaluation of up to 1000 milliseconds. Instead of a frame-by-frame synthesis, we propose a goal-directed task to capture perception of longer-term human action, which not only lends itself towards forecasting more semantically meaningful key moments, but enables a more predictable evaluation: as seen in Fig. 1, there can be significant ambiguity in the number of pose frames to predict towards a key or goal pose, making frame-based evaluation difficult in longer-term forecasting.

Multi-Modal Trajectory Prediction. While 3D human motion forecasting has typically been addressed in a deterministic fashion, various works in 2D trajectory prediction of cars or pedestrians have developed methods to capture multiple modes in possible future predictions. Social LSTM [2] estimates a bivariate Gaussian distribution for the 2D location of a person at each time step in a predicted trajectory. Conditional variational autoencoders have also been proposed to learn sampling models for generating multiple hypotheses for person and car movements [19, 6]. Liang et al. [22] presented a probabilistic model for multi-future person trajectory prediction in 2D, first predicting an initial heatmap probability distribution over a coarse 2D grid, and then predicting a refined location. To generate our multi-modal distribution predictions for a characteristic 3D pose, we also adopt a heatmap representation for probability distribution of a pose joint, and couple it with an autoregressive formulation for generating a self-consistent pose prediction composed of many pose joints.

Goal-oriented Forecasting. While a time-based, frame-by-frame prediction is the predominant approach towards future forecasting tasks, several works have proposed to tackle goal-oriented forecasting. Recently, Jayaraman et al. [18] proposed to predict “predictable” future video frames in a time-agnostic fashion, and represent the predictions as subgoals for a robotic tasks. Pertsch et al. [27] predict future keyframes representing a future video sequence of events. Cao et al. [7] plan human trajectories from an image and 2D pose history, first predicting 2D goal locations for a person to walk to in order to synthesize the path. Inspired by such goal-based abstractions, we aim to represent 3D human actions as its key, characteristic poses.

3. Method Overview

Given a sequence of 3D pose observations of a person, our aim is to predict the future characteristic 3D pose of that person, characterizing the intent of the person’s future action. We take $N = 25$ joint locations for each pose of the input sequence in the widely used OpenPose [8] layout

for samples from GRAB and $N = 32$ in the native layout for Human3.6M. We then predict the same number of joint locations for the forecasted pose, by predicting a joint distribution of probability heatmaps for each joint, and sampling pose hypotheses from the joint distribution. By representing probability heatmaps for the joint predictions, we can capture multiple different modes in likely characteristic poses, enabling more accurate future pose prediction.

From the input sequence of pose joints, we develop a neural network architecture to predict a probability heatmap over a volumetric 3D grid for each joint, corresponding to likely future positions of that joint. We model these predictions conditionally in an autoregressive fashion in order to tractably model the joint distribution over all pose joint locations. This enables a consistent pose prediction over the set of pose joints, as a set of joints may have likely modes that are unlikely to be seen all together (e.g., maintaining a stationary hip while stepping forward with the right foot, or stepping forward with the left foot may both be valid poses, but sampling independently might lead to a pose prediction with both right and left foot forward but no hip movement, which is not a physically stable pose for a human). To sequentialize the pose joint prediction autoregressively, we first predict probability heatmaps for the end effectors in our dataset – right hand first, then left hand conditioned on the right hand prediction, followed by the rest of the body joints.

4. Capturing Multi-Modality with Heatmap Predictions

We aim to learn to predict likely future locations for an output pose joint j , characterized by a probability heatmap H_j over a volumetric grid of possible pose joint locations. We take as input 10 frames of pose observations characterized by the 3D locations of each pose’s N joints (as well as condition on any already predicted output pose joints), and construct an attention-driven neural network to learn the different dependencies between human skeleton joints to inform the final heatmap prediction.

Fig. 3 shows an overview of our network architecture. We represent the body joints of the input pose sequence as a concatenation of their 3d locations over time. As a first step, they are then processed with an MLP which learns a higher-dimensional body joint representation. Any prior joints (in the autoregressive scenario) are represented with their 3d location and similarly processed with the MLP. We then compute an attention map representing dependencies of the intended joint prediction with the input set of pose joints. This way, the network learns not only how different joints in the skeleton affect each other directly (in situations where one joint is directly connected to another in the kinematic chain, eg. right elbow affects right hand placement) but also learns to exploit more subtle dependencies like how

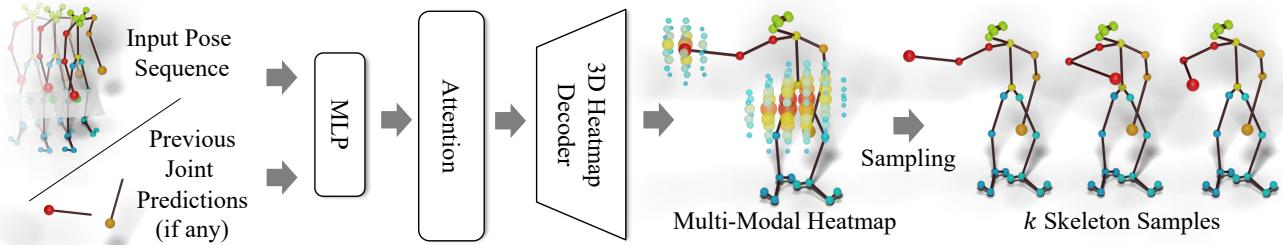


Figure 3: Overview of our approach for characteristic 3D pose prediction. From an input observed pose sequence, as well as any prior joint predictions, we leverage attention to learn inter-joint dependencies, and decode a 3D volumetric heatmap representing the probability distribution for the next joint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

the placement of feet will influence the final location of the hands. Following the formalism of Scaled Dot-Product Attention [30], popularized in natural language processing, our attention maps are computed from a query \mathbf{Q} and a set of key-value pairs \mathbf{K} and \mathbf{V} . During training, the MLP learns representations for \mathbf{Q} , \mathbf{K} , and \mathbf{V} which are shared between all joints. This allows us to project all joints into the same embedding space where we can then compare the joint of interest (represented by \mathbf{Q}) with all other joints (\mathbf{K}) to inform which parts of \mathbf{V} (the learned latent representation for all joints which will be passed to the decoder) are relevant for this joint of interest.

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D}}\right)\mathbf{V} = \mathbf{AV}, \quad (1)$$

Intuitively, the similarity between key and query defines which parts of a learned pose skeleton representation are important for the desired prediction. Formally, this is defined in Equation 1: The value representation \mathbf{V} is weighed per-element by the result of the dot-product between \mathbf{Q} and \mathbf{K} (scaled by the dimension of the embedding vector D and a softmax operation). In our case, the attention map \mathbf{A} has a dimensionality of $n_j \times N$ with n_j indicating the number of joints to be predicted. Any prior joint predictions for autoregressive prediction are considered as an additional node to our attention map, giving the attention map dimension $n_j \times (N + n_p)$ for n_p prior joints.

Based on the attention scoring, we then use a series of five 3D convolutions to decode an output probability heatmap H_j over a volumetric grid (in our experiments, we use a grid of size 16^3 , spanning $2m^3$, centered on the hip joint). For a detailed specification of our network architecture, we refer to the appendix.

Loss A value in the grid of H_j at location $H_j(x, y, z)$ corresponds to a probability of the joint j being at location (x, y, z) in the future characteristic pose. We predict $H_j(x, y, z)$ as a classification problem by discretizing the output values into $n_{discr} = 10$ bins in the $[0, 1]$ space.

We then use a cross entropy loss with the discretized target heatmap to train our heatmap predictions. In our experiments, we found that this classification formulation for H_j produced better results than an ℓ_2 or ℓ_1 regression loss, as it mitigated tending towards the average or median.

Note that for real-world data captured of human movement, we do not have a full ground truth probability distribution for the future characteristic pose, but rather a set of paired observations of input pose to the target pose. To generate target heatmap data from a single future observation in the training data, we apply a Gaussian kernel (size 5, $\sigma = 3$) over the target joint location. We then aim to learn multimodality by generalizing across train set observations, and show that our formulation can effectively model multiple modes across single future observations in Section 7.

4.1. Training Details

We train our models on a single NVIDIA GeForce RTX 2080Ti. We use an ADAM optimizer with a linear warmup schedule for 4000 steps; learning rate is then kept at 0.002. We use a batch size of 250, as a larger batch size helps with training our attention mechanism. Our model trains for up to 8 hours until convergence. During training, pose joint predictions conditioned on prior joint predictions are trained using the ground truth locations of the prior joints.

5. Autoregressive Joint Prediction for Pose Forecasting

We formulate our joint prediction for characteristic pose forecasting in an autoregressive fashion, in order to effectively model the interdependencies between pose joints; since we predict heatmaps for each pose joint location, and the joint locations are not independent of each other, we cannot simply independently sample from each joint probability heatmap. Fig. 6 visualizes an example of pose inconsistencies caused by independent joint sampling.

Thus, we model the joint distribution of pose joints autoregressively, as visualized in Fig. 4: we first predict end

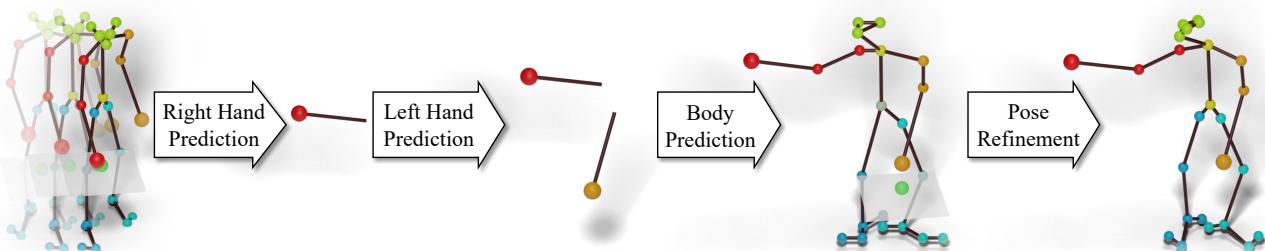


Figure 4: To model joint dependencies within the human skeleton, we sample joints in an autoregressive manner by first predicting the end-effectors (right and left hand), then the rest of the body; pose refinement then improves skeleton consistency.

effector joints, followed by other body joints. For our experiments, we find that the right and left hands tend to have a large variability, so we first predict the right hand, then the left hand conditioned on the right hand location, followed by the rest of the body joints. Empirically, we found that the order of those body joints does not make a difference as the hands tend to define the rest of the body pose. To sample from a joint heatmap prediction, we first sample local maxima in the heatmap, followed by random sampling of the heatmap as a probability distribution.

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

$$\begin{aligned} E_R(\mathbf{x}, \mathbf{e}, \mathbf{b}, \theta, H) = & \\ w_e \|\mathbf{x}_e - \mathbf{e}\|_2 + w_b \|\text{bonelengths}(x) - \mathbf{b}\|_1 & \\ + w_a \|\text{angles}(x) - \theta\|_1 + w_h \sum_j (1 - H_j) & \end{aligned} \quad (2)$$

where \mathbf{x} the raw predicted pose skeleton as a vector of N 3D joint locations; \mathbf{b} and θ the bone lengths and joint angles, respectively, of the initially observed pose skeleton; H_j the heatmap probability for each joint; \mathbf{e} the sampled end effector locations; and w_e, w_b, w_a, w_h weighting parameters (in all our experiments, we use $w_e = 1.0, w_b = 1.0, w_a = 3.0, w_h = 0.1$). We then optimize for \mathbf{x} under this objective to obtain our final pose prediction.

6. Characteristic 3D Pose Dataset

To train and evaluate the task of characteristic 3D pose forecasting, we introduce a dataset of annotated characteristic poses, built on the GRAB [28] and Human3.6M [16] datasets.

- **Human3.6M** is a commonly used dataset for human pose forecasting, comprising 210 actions performed by

11 professional actors in 17 scenarios for a total of 3.6 million frames. Joint locations in 3D are obtained via a high-speed motion capture system, and we directly use their 32-joint human skeleton in our method.

- **GRAB** is a recent dataset with over 1 million frames in 1334 sequences of 10 different actors performing a total of 29 actions with various objects. Each actor starts in a T-Pose, moves towards a table with an object, performs an action with the object, and then steps back to the T-Pose. The human motions are captured using modern motion capture techniques, with an accuracy in the range of a few millimeters. GRAB provides SMPL-X [26] parameters from which we extract 25 body joints in OpenPose [8] format.

We then annotate the timesteps of the captured sequences corresponding to characteristic poses: for GRAB, we annotate the initial observation as the 10-frame sequence leading up to the first physical contact with the object, and the future characteristic pose as a pose that defines the action taken by the person with the object; for Human3.6M, the initial 10 frames of every sequence are used as input observation and a pose representative of the action as characteristic pose. Several example input sequence-characteristic

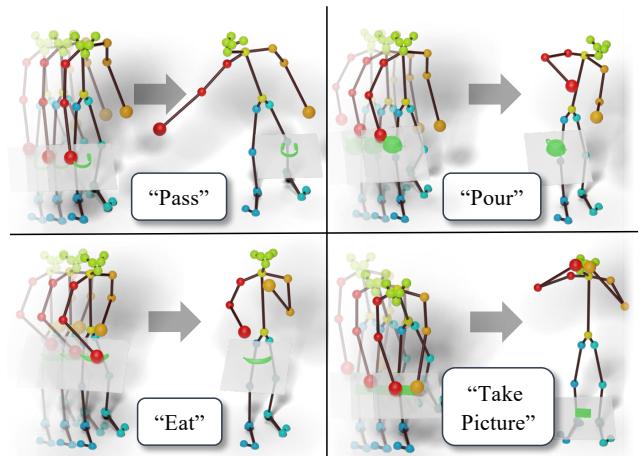


Figure 5: Several example input observations and target characteristic 3D poses from our annotated dataset.

pose pairs are visualized in Fig. 5. Annotations were performed by the authors, within a time span of three days. We define a characteristic pose as the point in time when the action is most articulated, i.e. right before the actor starts returning back to another pose. In the examples shown in Fig. 5, this means the frame where the hand is furthest from the person when passing, closest to the head when eating or taking a picture, and most tilted when pouring. For sequences containing multiple occurrences of the same action, like lifting, we chose the repetition with most articulation, e.g. when the object is lifted highest. In the case of Human3.6M, where there are sometimes multiple possible options for characteristic poses, we pick the first one that is representative of the action, e.g., the first sitting pose.

Evaluation We use a train/val/test split by actor in each dataset. For GRAB we have 8 train actors, 1 val actor, and 1 test actor, resulting in 992 train, 197 val, and 145 test sequences. For Human3.6M, we follow the split of [23]: 5 train, 1 val, and 1 test actor, giving 150 train, 30 val, and 30 test sequences. To evaluate our task of characteristic 3D pose prediction, we aim to consider the multi-modal nature of the task. Since we do not have ground truth probability distributions available, and only a single observed characteristic pose for each input pose observation, we aim to evaluate the quality of the most similar sampled pose from the distribution, following evaluation for future trajectory prediction of pedestrians and cars [9, 22]. Following [9], at test time, for a single target future observation from an initial pose, we evaluate $k = 6$ sampled poses from our predicted distribution and take the minimum error pose, using the mean per-joint position error (MPJPE) proposed in [16]:

$$E_{\text{MPJPE}} = \frac{1}{N} \sum_{j=1}^N \|p'_j - p_j\|_2^2 \quad (3)$$

where p'_j is the location of joint j in the predicted pose with N joints, and p_j the location of j in the target pose.

As MPJPE is an average over all samples, it can be difficult to distinguish if an evaluated method tends towards low-error predictions with several high outliers, or is typically producing the error reported by the MPJPE. Thus, we also evaluate the percentage of samples below the two thresholds of 0.15m and 0.25m to better capture the distribution of MPJPEs over samples.

Additionally, to evaluate the predicted probability distribution independent of the joint sampling, we evaluate the negative log-likelihood (NLL) of the target joints being drawn from the predicted distribution. This evaluates the predicted heatmaps independent of random sampling.

7. Results

We evaluate our method against alternative approaches for the task of characteristic 3D pose prediction, using our

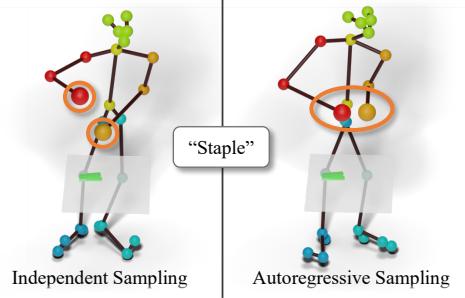


Figure 6: Effect of independent (left) vs. autoregressive sampling (right, ours). Independent sampling tends to result in potentially valid but globally inconsistent joints in a pose (e.g., hands should be close together for stapling), whereas our autoregressive sampling produces a more coherent pose.

annotated dataset built from the real-world GRAB [28] and Human3.6M [16] datasets. Note that ground truth probability distributions for characteristic poses are unavailable (only one observation per input pose), so we aim to capture the quality of a small set of samples (6), and the negative log-likelihood (NLL) that the target joints were drawn from the predicted distribution.

Comparison to time-based state-of-the-art forecasting.

In Tab. 1, we compare to state-of-the-art approaches for frame-based future human motion prediction, Learning Trajectory Dependencies [24] and History Repeats Itself [23], which use a graph neural network and an attention-based model, respectively, to predict human poses from input pose observations. We train these state-of-the-art sequential approaches on our datasets, given an input sequence of 10 frames, to predict a t -second pose sequence. We use $t = 12$ for GRAB (360 frames at 30fps) and $t = 40$ for Human3.6M (400 frames at 10fps) to ensure that the characteristic pose falls within each target sequence. From their predicted sequences, we then report the MPJPE of the pose with the lowest error w.r.t to the ground truth characteristic pose. In comparison, our approach to capture the likely mode of the characteristic pose enables sampling of more accurate future pose predictions.

Tab. 2 evaluates per-bodypart performance rather than the average over all $N = 25$ joints. Note that while state-of-the-art approaches perform well in bodyparts with little to no motion (e.g., legs and hip), our approach notably improves in regions with larger motion such as the arms. We refer to the appendix for a more fine-grained analysis as well as joint-bodypart correspondences.

Example qualitative results are shown in Fig. 8: [23] tends to predict very small joint movements; [24] predicts larger but often inconsistent or incomplete movement for arms (e.g., passing in both directions simultaneously in row 3 or right hand only half way in row 4). In contrast,

		GRAB			Human3.6m		
	Method	MPJPE ↓	0.15m ↑	0.25m ↑	MPJPE ↓	0.15m ↑	0.25m ↑
Statistical	Average Train Pose (entire dataset)	0.3513	0.45	3.37	0.4947	0.00	0.00
	Average Train Pose (mean per-action average)	0.3403	0.07	5.64	0.3869	0.00	7.14
	Zero Velocity (entire dataset)	0.1854	40.69	75.17	0.9981	0.00	0.00
	Zero Velocity (mean per-action average)	0.1695	42.56	93.28	1.0174	0.00	0.00
Algorithmic	Learning Trajectory Dependencies [24]	0.1606	51.72	87.59	0.2798	3.33	53.33
	History Repeats Itself [23]	0.1563	50.34	77.24	0.2630	36.67	50.00
	Ours (Deterministic)	0.1632	51.72	80.69	0.2877	16.66	53.33
	Ours (before Refinement)	0.1426	64.83	98.62	0.2116	43.33	73.33
	Ours	0.1277	78.62	98.62	0.1936	50.00	73.33

Table 1: Characteristic 3D pose prediction performance, compared to statistical and algorithmic baselines, in terms of MPJPE [meters] and the percentage of samples with $\text{MPJPE} < 0.15m$ and $< 0.25m$.

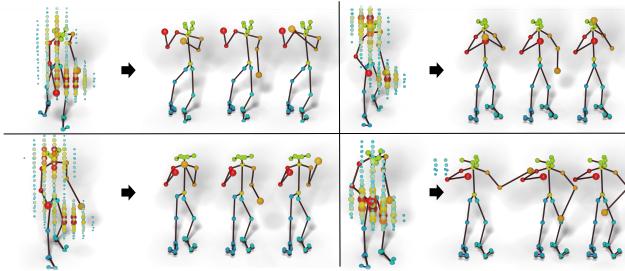


Figure 7: Examples of sampled poses from our heatmap predictions (for each ex, left: last frame in input sequence superimposed with predicted heatmap for left hand, right: output pose samples).

our probabilistic approach predicts the characteristic modes more faithfully.

Comparison to statistical baselines. We also compare with two statistical baselines: the average target train pose over the entire dataset and per-action, and a zero-velocity baseline (i.e., the error of simply using the last input pose as prediction), which was shown by Martinez et al. [25] to be competitive with and sometimes outperform state of the art. Our approach outperforms these statistical baselines, indicating learning of strong characteristic pose patterns.

Does a probabilistic prediction help? In addition to comparing to state-of-the-art alternative approaches which make deterministic predictions, we compare in Tab. 1 with our model backbone with a deterministic output head (an MLP) replacing the volumetric heatmap decoder which regresses offset positions for each pose joint relative to the

input positions. Removing our heatmap predictions similarly fails to effectively capture the characteristic modes; our probabilistic, heatmap-based predictions notably improve performance.

Does autoregressive pose joint sampling help? We analyze the effect of our autoregressive pose joint sampling, quantitatively in Tab. 3, and qualitatively in Fig. 6. We compare against a version of our model trained to predict each pose joint heatmap independently, and sample pose joints independently, which often results in valid individual pose joint predictions that are globally inconsistent with the other pose joints. In contrast, our autoregressive pose joint sampling helps to generate a likely, consistent pose.

How diverse are the sampled poses? We show qualitative examples of our multi-modal predictions in Fig. 7, outlining the diversity of both heatmap predictions and sampled skeletons. We also evaluate our prediction diversity as MPJPE between our sampled outputs as 0.2911m on average for the entire pose and as 0.4145m for the end-effectors.

What is the effect of the number of pose samples? If we take more pose samples from our predicted joint distribution (from 6 to 32), we can, as expected, better predict the potential target characteristic pose.

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

Method	Right Arm ↓	Left Arm ↓	Right Leg ↓	Left Leg ↓	Hip ↓	Head ↓
Learning Trajectory Dependencies [24]	0.2539	0.2501	0.1318	0.1068	0.0972	0.1690
History Repeats Itself [23]	0.3241	0.2228	0.1036	0.0759	0.0940	0.1921
Ours (Deterministic)	0.3056	0.2067	0.1160	0.0881	0.1193	0.2066
Ours (before Refinement)	0.2205	0.1519	0.1262	0.1243	0.1188	0.1399
Ours	0.2147	0.1458	0.1099	0.1093	0.0882	0.1218

Table 2: Characteristic 3D pose prediction performance on GRAB, broken down by body part MPJPE [meters].

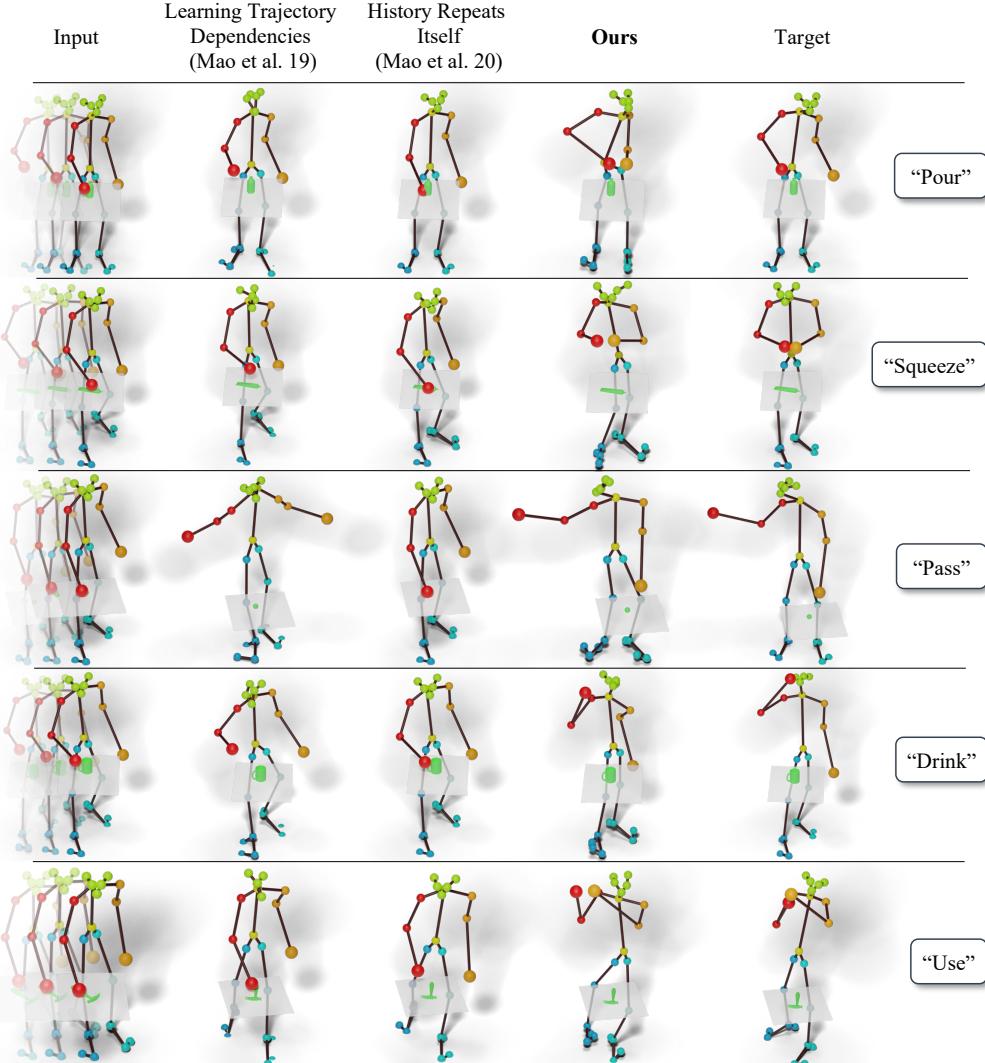


Figure 8: Qualitative results on characteristic 3D pose prediction. In comparison to state-of-the-art human pose forecasting approaches [24, 23], our method more effectively predicts likely intended action poses.

8. Conclusion

In this paper, we introduced a new method for predicting characteristic 3D poses – poses defining key future moments from single pose observations, towards goal-oriented prediction for human motion forecasting. We introduce a probabilistic approach to capturing the most likely modes in these characteristic poses, coupled with an autoregressive

formulation for pose joint prediction to sample consistent 3D poses from a predicted joint distribution. We trained and evaluated our approach on a new annotated dataset for characteristic 3D pose prediction, notably outperforming alternative approaches. We believe that this opens up many possibilities towards goal-oriented 3D human pose forecasting and understanding anticipation of human movements.

Autoregressive	Heatmap Loss	Samples (k)	MPJPE [meters] \downarrow	% Samples $< 0.15\text{m}$ \downarrow	% Samples $< 0.25\text{m}$ \downarrow	NLL \downarrow
✓	ℓ_1	$k = 6$	0.1802	31.72	87.59	6.622
✓	ℓ_2	$k = 6$	0.2033	30.34	76.55	6.706
✗	Cross Entropy	$k = 6$	0.1383	72.41	98.62	5.200
✓	Cross Entropy	$k = 6$	0.1277	78.62	98.62	5.125
✓	Cross Entropy	$k = 32$	0.0810	99.31	100.00	5.125

Table 3: Ablations study for varying heatmap losses, with and without autoregressive pose sampling, and varying number of samples taken for the evaluation.

9. Acknowledgements

We would like to thank the support of the Zentrum Digitalisierung.Bayern (Z.D.B).

References

- [1] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7143–7152. IEEE, 2019. [2](#)
- [2] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Fei-Fei Li, and Silvio Savarese. Social LSTM: human trajectory prediction in crowded spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 961–971. IEEE Computer Society, 2016. [3](#)
- [3] Moshe Bar. The proactive brain: memory for predictions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1235–1243, 2009. [1](#)
- [4] Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. ContactDB: Analyzing and predicting grasp contact via thermal imaging. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. [15](#)
- [5] Judith Bütepage, Michael J. Black, Danica Kragic, and Hedvig Kjellström. Deep representation learning for human motion prediction and classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1591–1599. IEEE Computer Society, 2017. [2](#)
- [6] Judith Bütepage, Hedvig Kjellström, and Danica Kragic. Anticipating many futures: Online human motion prediction and generation for human-robot interaction. In *2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018*, pages 1–9. IEEE, 2018. [3](#)
- [7] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 387–404. Springer, 2020. [3](#)
- [8] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. [3, 5, 14](#)
- [9] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3d tracking and forecasting with rich maps. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8748–8757. Computer Vision Foundation / IEEE, 2019. [6](#)
- [10] Hsu-Kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose fore-casting. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*, pages 1423–1432. IEEE, 2019. [2](#)
- [11] Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204, 2013. [1](#)
- [12] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4346–4354. IEEE Computer Society, 2015. [2](#)
- [13] Anand Gopalakrishnan, Ankur Mali, Dan Kifer, C. Lee Giles, and Alexander G. Ororbia II. A neural temporal model for human motion prediction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12116–12125. Computer Vision Foundation / IEEE, 2019. [2](#)
- [14] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José M. F. Moura. Adversarial geometry-aware human motion prediction. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, volume 11208 of *Lecture Notes in Computer Science*, pages 823–842. Springer, 2018. [2](#)
- [15] Jakob Hohwy. *The predictive mind*. Oxford University Press, 2013. [1](#)
- [16] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2014. [2, 5, 6, 14](#)
- [17] Ashesh Jain, Amir Roshan Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5308–5317. IEEE Computer Society, 2016. [2](#)
- [18] Dinesh Jayaraman, Frederik Ebert, Alexei A. Efros, and Sergey Levine. Time-agnostic prediction: Predicting predictable video frames. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [3](#)
- [19] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B. Choy, Philip H. S. Torr, and Manmohan Chandraker. DESIRE: distant future prediction in dynamic scenes with interacting agents. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2165–2174. IEEE Computer Society, 2017. [3](#)
- [20] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5226–5234. IEEE Computer Society, 2018. [2](#)
- [21] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural

- networks for 3d skeleton based human motion prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 211–220. IEEE, 2020. 2
- [22] Junwei Liang, Lu Jiang, Kevin P. Murphy, Ting Yu, and Alexander G. Hauptmann. The garden of forking paths: Towards multi-future trajectory prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10505–10515. IEEE, 2020. 3, 6
- [23] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIV*, volume 12359 of *Lecture Notes in Computer Science*, pages 474–489. Springer, 2020. 2, 6, 7, 8, 13, 15, 17, 18
- [24] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9488–9496. IEEE, 2019. 2, 6, 7, 8, 13, 15, 17, 18
- [25] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4674–4683. IEEE Computer Society, 2017. 2, 7
- [26] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10975–10985. Computer Vision Foundation / IEEE, 2019. 5, 14
- [27] Karl Pertsch, Oleh Rybkin, Jingyun Yang, Shenghao Zhou, Konstantinos Derpanis, Kostas Daniilidis, Joseph Lim, and Andrew Jaegle. Keyframing the future: Keyframe discovery for visual prediction and planning. In *Learning for Dynamics and Control*, pages 969–979. PMLR, 2020. 3
- [28] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, volume 12349 of *Lecture Notes in Computer Science*, pages 581–600. Springer, 2020. 2, 5, 6, 14, 15
- [29] Yongyi Tang, Lin Ma, Wei Liu, and Wei-Shi Zheng. Long-term human motion prediction by modeling motion context and enhancing motion dynamics. In Jérôme Lang, editor, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 935–941. ijcai.org, 2018. 2
- [30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 4
- [31] Borui Wang, Ehsan Adeli, Hsu-Kuang Chiu, De-An Huang, and Juan Carlos Niebles. Imitation learning for human pose prediction. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7123–7132. IEEE, 2019. 2
- [32] Yi Zhou, Zimo Li, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 2

Appendix

In this appendix, we show additional qualitative and quantitative results (Sec. A), detail our network architecture specification (Sec. B), and describe additional details regarding the dataset (Sec. C) as well as our training setup (Sec. D).

A. Additional Evaluation

Ablation Comparison before Refinement. Tab. 4 shows our ablations study before applying the pose refinement step; our design choices hold under no refinement as well.

Ablations on Human3.6M. The results of our full ablations study on Human3.6M are provided in Tab. 5, for predictions both after and before the pose refinement step.

Additional Qualitative Results. We show additional qualitative results of our method on GRAB in Fig. 9. Visual results on Human3.6M alongside predictions of baseline methods are shown in Fig. 10. In Fig. 11, we qualitatively compare our method after refinement with its predictions before refinement as well as with a deterministic output head.

Characteristic Pose Forecasting with Ground Truth Action Labels. In Tab. 7, we additionally evaluate our approach using ground truth action labels as input to provide additional contextual information. The ground truth action label is processed as an additional attention node. This action information moderately improves the characteristic pose prediction, as the action helps to indicate a narrower set of possible characteristic poses. In our original action-agnostic scenario, our approach nonetheless maintains plausible characteristic pose predictions across all actions.

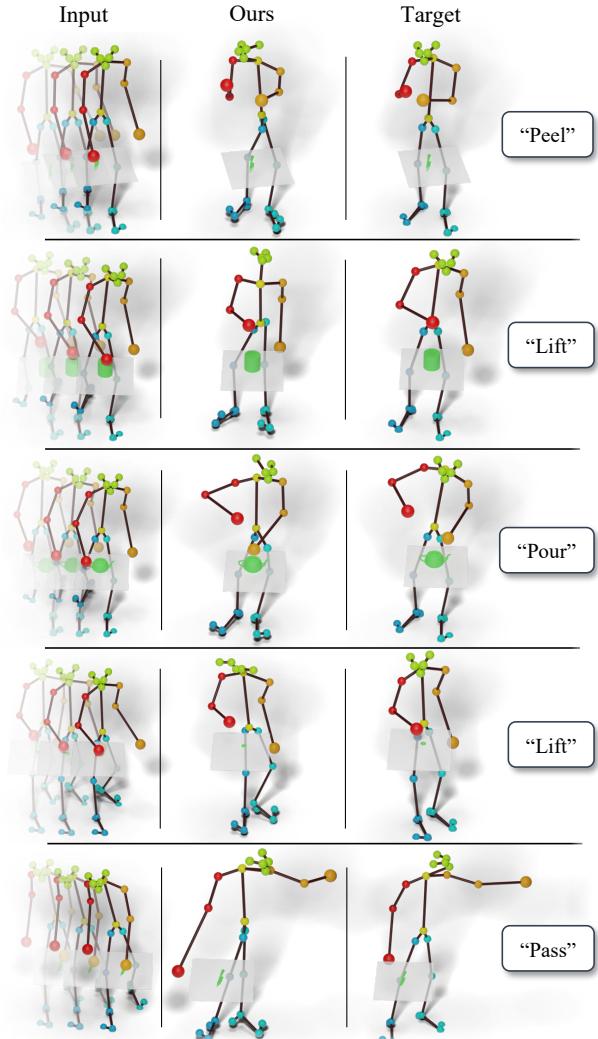


Figure 9: Additional Qualitative Results on GRAB.

Autoregressive	Heatmap Loss	Samples (k)	MPJPE [meters] \downarrow	% Samples $< 0.15\text{m}$ \uparrow	% Samples $< 0.25\text{m}$ \uparrow	NLL \downarrow
✓	ℓ_1	$k = 6$	0.2024	20.69	76.55	6.622
✓	ℓ_2	$k = 6$	0.2127	21.38	72.41	6.706
✗	Cross Entropy	$k = 6$	0.1537	55.17	97.93	5.200
✓	Cross Entropy	$k = 6$	0.1426	64.83	98.62	5.125
✓	Cross Entropy	$k = 32$	0.1002	99.31	100.00	5.125

Table 4: Characteristic 3D pose prediction performance ablations study on GRAB, without pose refinement.

Autoregressive	Heatmap Loss	Samples (k)	After Refinement			Before Refinement			NLL \downarrow
			MPJPE \downarrow	0.15m \uparrow	0.25m \uparrow	MPJPE \downarrow	0.15m \uparrow	0.25m \uparrow	
✓	ℓ_1	$k = 6$	0.2459	3.33	63.33	0.2724	3.33	53.33	7.841
✓	ℓ_2	$k = 6$	0.2734	3.33	60.00	0.2891	0.00	36.67	7.671
✗	Cross Entropy	$k = 6$	0.2032	46.67	86.67	0.2181	30.00	83.33	7.253
✓	Cross Entropy	$k = 6$	0.1936	50.00	73.33	0.2116	43.33	73.33	6.770
✓	Cross Entropy	$k = 32$	0.0975	86.67	100.00	0.1223	86.67	96.67	6.770

Table 5: Ablations study on Human3.6M for varying heatmap losses, with and without autoregressive pose sampling, and varying number of samples taken for the evaluation, in terms of MPJPE [meters] and the percentage of samples with MPJPE $< 0.15m$ and $< 0.25m$.

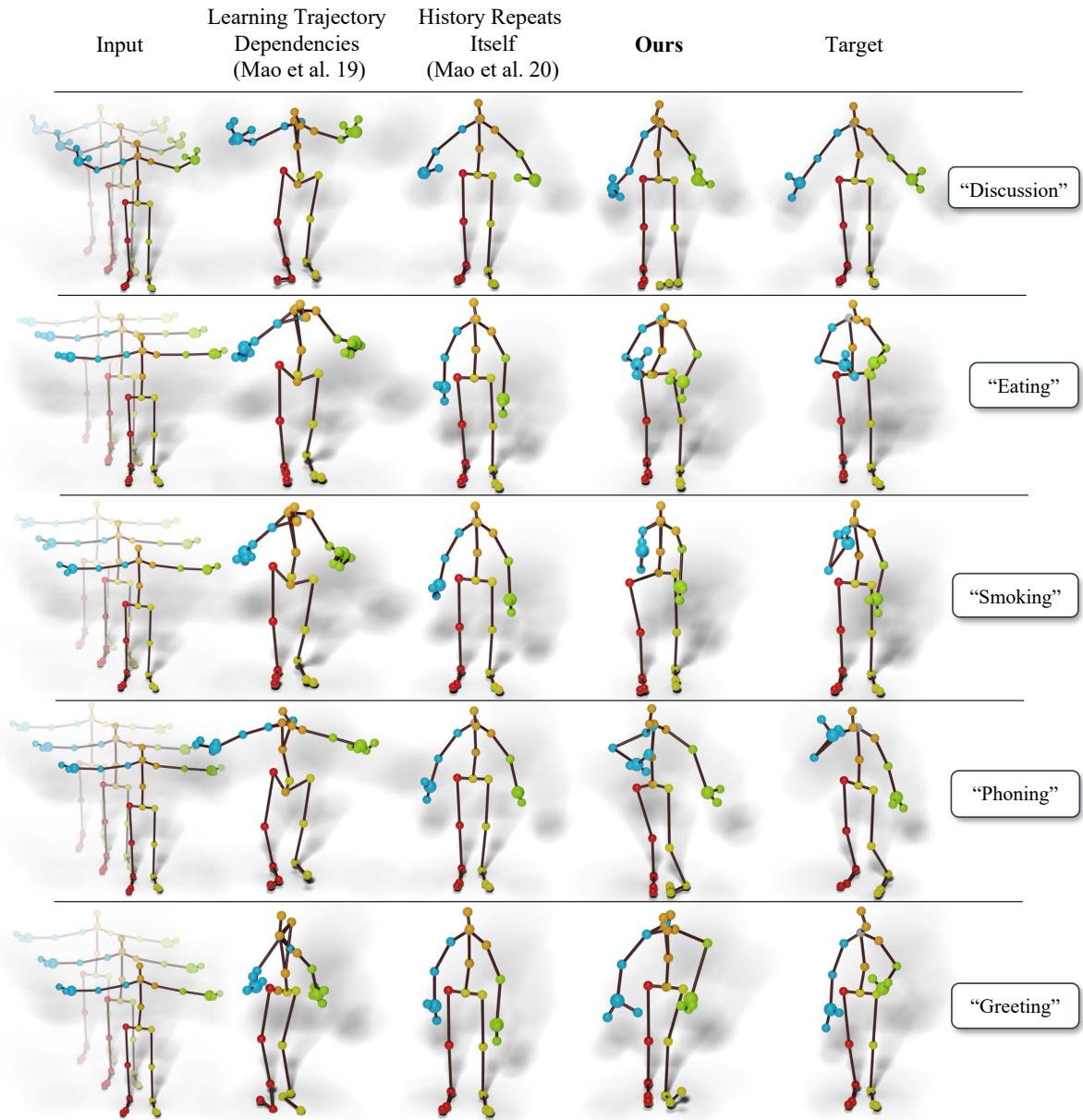


Figure 10: Qualitative results on characteristic 3D pose prediction on the Human3.6M dataset.

Method	Right Arm ↓	Left Arm ↓	Right Leg ↓	Left Leg ↓	Hip ↓	Head ↓
Learning Trajectory Dependencies [24]	0.3792	0.3887	0.1884	0.1729	0.2351	0.2189
History Repeats Itself [23]	0.3474	0.3031	0.2068	0.1748	0.2389	0.2453
Ours (Deterministic)	0.3969	0.3408	0.2165	0.1916	0.2480	0.2524
Ours (before Refinement)	0.3013	0.2483	0.1469	0.1458	0.1763	0.1782
Ours	0.2708	0.2275	0.1337	0.1383	0.1639	0.1699

Table 6: Characteristic 3D pose prediction performance on Human3.6m, broken down by body part MPJPE [meters].

B. Architecture Details

Fig. 12 details our network specification from input (top) to heatmap output and pose sampling (bottom). For each linear layer, we provide the number of input and output channels in parentheses, for normalization layers the dimension to be normalized over, for dropout layers the dropout probability p , and for convolutions the number of input and output channels as well as kernel size (ks), stride (str), and padding (pad).

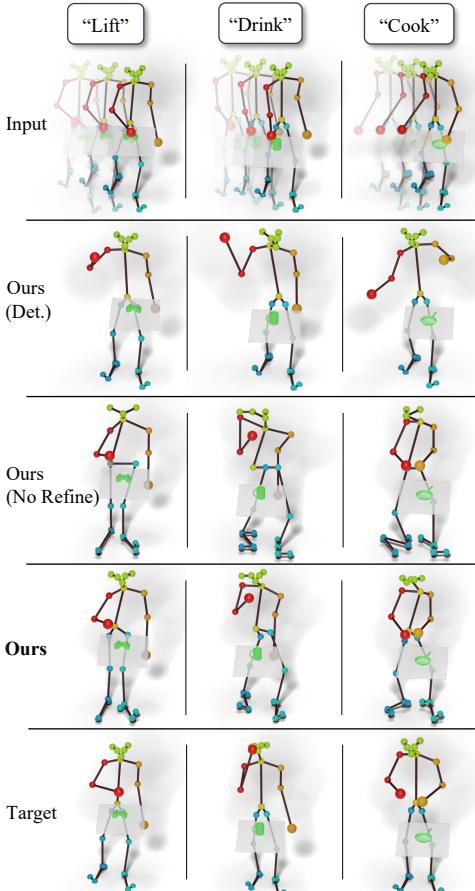


Figure 11: Visualization of our characteristic 3D pose predictions in comparison to our approach without pose refinement, and a deterministic prediction using our backbone.

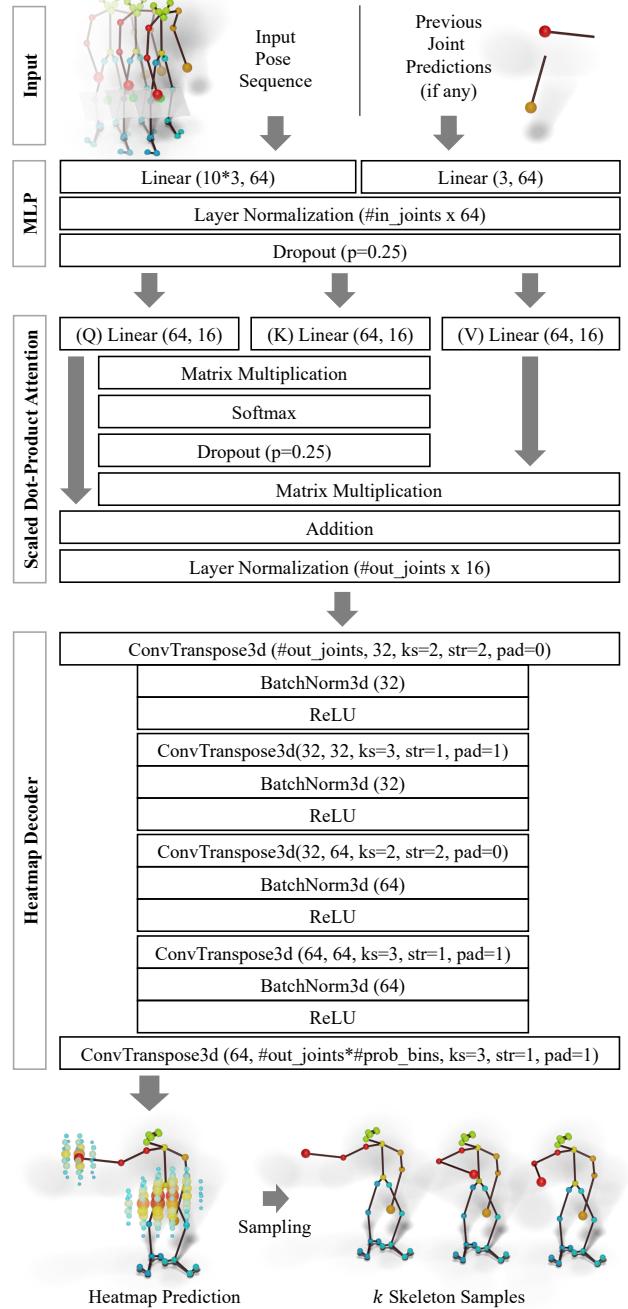


Figure 12: Our network architecture with details for MLP, scaled dot-product attention, and heatmap decoder.

		GRAB				Human3.6M			
Refined	GT Action Label	MPJPE ↓	0.15m ↑	0.25m ↑	NLL ↓	MPJPE ↓	0.15m ↑	0.25m ↑	NLL ↓
✗	✗	0.1426	64.83	98.62	5.125	0.2116	43.33	73.33	6.770
✗	✓	0.1296	82.76	98.62	5.006	0.2093	30.00	83.33	6.591
✓	✗	0.1277	78.62	98.62	5.125	0.1936	50.00	73.33	6.770
✓	✓	0.1144	87.59	99.31	5.006	0.1896	40.00	83.33	6.591

Table 7: Characteristic 3D pose prediction performance of our proposed method, compared to an ablation with ground truth action labels as input.

We take as input 25 joints in the case of GRAB and 32 joints for Human3.6M (#in_joints). The number of output joints (#out_joints) depends on whether the right or left hand is being predicted (#out_joints=1 for GRAB, #out_joints=5 for Human3.6M) or the rest of the body (#out_joints=23 for GRAB, #out_joints=22 for Human3.6M). In all our experiments, the number of probability bins (#prob_bins) is 10.

C. Dataset

Human3.6M 3D Pose Layout. For all our experiments on Human3.6M [16], we directly use their native joint layout with all 32 body joints as visualized in Fig. 13 (right). We denote in Tab. 9 the correspondences of joints to body parts, as used in Tab. 6.

GRAB Pose Layout. Since GRAB [28] not only provides a human skeleton representation but full body shape parameters, we preprocess all pose sequences by first extracting relevant joints for our approach. For this, we chose the OpenPose [8] layout as it describes the prevalent body joints and is widely used for representing 3D poses. We extract the 25 OpenPose body joints from the SMPL-X skeleton given by the GRAB dataset [28] using the correspondences shown in Tab. 8. Additionally, we denote in Tab. 8 the correspondences of joints to body parts, for the body part analysis in Tab. 2 of the main paper. Fig. 13 (left) visualizes our joint selection, overlaying the body shape given in GRAB as a point cloud over the 25-joint skeleton.

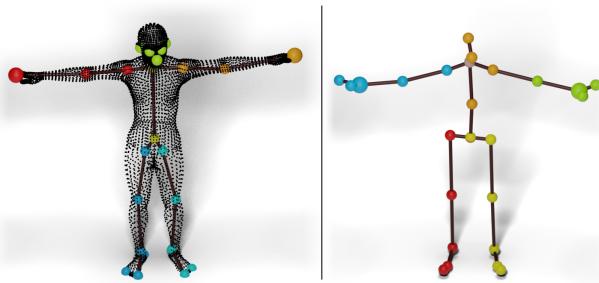


Figure 13: GRAB [28] body and our extracted skeleton joints overlaid (left); The native Human3.6M [16] skeleton (right).

		Ours (OpenPose [8])		Base (SMPL-X [26])	
		Idx	Label	Label	Idx
L. Arm	2	Right Shoulder	Right Shoulder	17	
	3	Right Elbow	Right Elbow	19	
	4	Right Finger	Right Index 3	42	
	5	Left Shoulder	Left Shoulder	16	
	6	Left Elbow	Left Elbow	18	
	7	Left Finger	Left Index 3	27	
	9	Right Hip	Right Hip	2	
Right Leg	10	Right Knee	Right Knee	5	
	11	Right Ankle	Right Ankle	8	
	22	Right Big Toe	Right Big Toe	63	
	23	Right Small Toe	Right Small Toe	64	
	24	Right Heel	Right Heel	65	
	12	Left Hip	Left Hip	1	
	13	Left Knee	Left Knee	4	
Left Leg	14	Left Ankle	Left Ankle	7	
	19	Left Big Toe	Left Big Toe	60	
	20	Left Small Toe	Left Small Toe	61	
	21	Left Heel	Left Heel	62	
	0	Nose	Nose	55	
	1	Neck	Neck	12	
	15	Right Eye	Right Eye	24	
Head	16	Left Eye	Left Eye	23	
	17	Right Ear	Right Ear	58	
	18	Left Ear	Left Ear	59	
	8	Mid-Hip	Pelvis	0	

Table 8: Joint Correspondences for GRAB

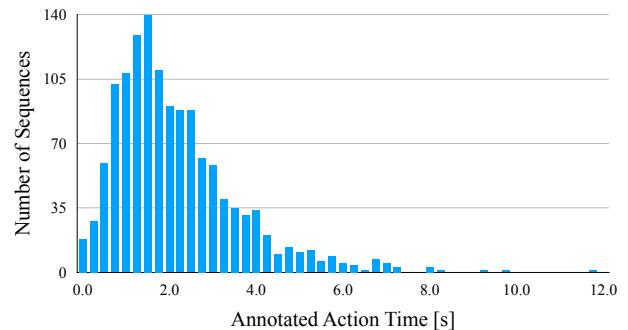


Figure 14: Times at which characteristic poses occur for GRAB.

B. Part	Label	Idx	Idx	Label	B. Part
R. Arm	R. Shoulder	25	17	L. Shoulder	L. Hand
	R. Elbow	26	18	L. Elbow	
	R. Hand	27	19	L. Hand	
	R. Hand	28	20	L. Hand	
	R. Thumb	29	21	L. Thumb	
	R. Finger	30	22	L. Finger	L. Hand
	R. Finger	31	23	L. Finger	
	R. Hip	1	6	L. Hip	L. Leg
	R. Knee	2	7	L. Knee	
	R. Heel	3	8	L. Heel	
R. Leg	R. Foot	4	9	L. Foot	
	R. Toe	5	10	L. Toe	
	Nose	14	12	Spine	
	Head	15	0	Hip	
	Neck	13	11	Hip	
Head	Neck	16			
	Neck	24			

Table 9: Joint Correspondences for Human3.6M

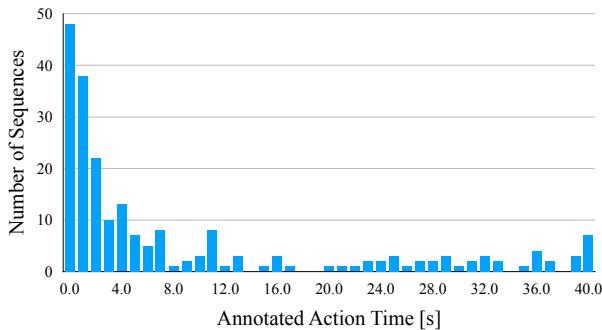


Figure 15: Times at which characteristic poses occur for Human3.6M.

Context and Action Labels. While our approach is agnostic to context or action, we visualize the context provided by GRAB [28, 4] (of the table and object) and action label provided by both GRAB and Human3.6M to help contextualize the pose visualizations. The context and action labels are not taken into account by the network or the evaluation, meaning that our approach infers plausible human action poses while being agnostic towards action and context.

Additional Characteristic 3D Pose Details. We show additional characteristic 3D poses in their original sequences in Fig. 16, and note the strong time differences at which the characteristic poses occur. Furthermore, Fig. 14 and Fig. 15 show the times during the sequences at which the characteristic 3D poses are annotated for GRAB and Human3.6M; these characteristic poses are distributed across a wide range (0-12 seconds and 0-40 seconds, respectively) of time.

D. Additional Training Details

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

State-of-the-art comparisons. We use the official code with default settings of the state-of-the-art methods we compare to ([24] and [23]). We train both from scratch on our characteristic pose dataset, setting the number of input frames to 10 and the number of output frames to 360 for GRAB and 400 for Human3.6M. From the predicted sequence, the pose with the smallest error wrt. the ground-truth characteristic pose is then evaluated.

Non-Maximum Suppression. Before sampling, we run a non-maximum suppression on the predicted heatmaps. Since we predict discretized heatmaps, we first convolve the heatmap with a box filter (kernel size 3). We then suppress non-maximum values in each neighborhood of 3^3 voxels, leading to a more diverse sampling.

E. Limitations

We believe that our characteristic pose forecasting takes important steps towards goal-oriented 3D understanding of future human behavior, but there remain many avenues for further development. For instance, the 3D scene context may provide additional strong cues for future pose prediction. Additionally, our probabilistic heatmaps are represented as dense volumetric grids, where a more efficient, sparse representation would help to scale to finer-grained resolutions. Another interesting future direction might be taking a self-supervised approach to predicting important characteristic moments in the future whereas our method relies on supervision on manually annotated poses. Finally, predicting several characteristic poses in sequence would help to enable modeling more complex, longer-term behaviors.

F. Video

We additionally provide a short explanatory video as part of the appendix at <https://youtu.be/vSxJg9z7cAM>.



Figure 16: Sample input-target pairs (colored) for our characteristic 3D pose forecasting task, with temporal snapshots along the sequence (grayscale). Each snapshot is half a second apart. Depicted as input is the last frame of the respective input sequence.

	Method	Nose ↓	Neck ↓	R-Shoulder ↓	R-Elbow ↓	R-Finger ↓	L-Shoulder ↓	
Statistical	Avg. Train Pose (entire dataset)	0.3571	0.3413	0.3666	0.4237	0.5339	0.3162	
	Avg. Train Pose (mean per-action avg.)	0.3417	0.3442	0.3547	0.3716	0.3801	0.3303	
	Zero Vel. (entire dataset)	0.2238	0.1633	0.1953	0.2704	0.5231	0.1691	
	Zero Vel. (mean per-action avg.)	0.1900	0.1554	0.1833	0.2339	0.4669	0.1535	
Algorithmic	Learning Trajectory Dependencies [24]	0.1818	0.1287	0.1481	0.2041	0.4095	0.1470	
	History Repeats Itself [23]	0.2013	0.1368	0.1776	0.2672	0.5274	0.1331	
	Ours (Deterministic)	0.1896	0.1674	0.1901	0.2647	0.4442	0.1501	
	Ours (before Refinement)	0.1482	0.1255	0.1315	0.1837	0.3463	0.1170	
	Ours	0.1199	0.0990	0.1202	0.1785	0.3455	0.1085	
	Method	L-Elbow ↓	L-Finger ↓	Mid-Hip ↓	R-Hip ↓	R-Knee ↓	R-Ankle ↓	
Statistical	Avg. Train Pose (entire dataset)	0.3571	0.4666	0.3140	0.3059	0.3108	0.3161	
	Avg. Train Pose (mean per-action avg.)	0.3527	0.3811	0.3162	0.3053	0.3125	0.3188	
	Zero Vel. (entire dataset)	0.2163	0.3930	0.1321	0.1303	0.1206	0.1494	
	Zero Vel. (mean per-action avg.)	0.2014	0.4165	0.1180	0.1165	0.0972	0.1223	
Algorithmic	Learning Trajectory Dependencies [24]	0.1911	0.4122	0.0972	0.0952	0.0935	0.1452	
	History Repeats Itself [23]	0.1723	0.3631	0.0940	0.0943	0.0932	0.1000	
	Ours (Deterministic)	0.1664	0.2814	0.1052	0.1059	0.0958	0.0814	
	Ours (before Refinement)	0.1322	0.2065	0.1188	0.1260	0.1088	0.1120	
	Ours	0.1219	0.2069	0.0882	0.1006	0.1050	0.0978	
	Method	L-Hip ↓	L-Knee ↓	L-Ankle ↓	R-Eye ↓	L-Eye ↓	R-Ear ↓	L-Ear
Statistical	Avg. Train Pose (entire dataset)	0.3141	0.3196	0.3445	0.3591	0.3549	0.3554	0.3454
	Avg. Train Pose (mean per-action avg.)	0.3199	0.3282	0.3575	0.3444	0.3412	0.3495	0.3418
	Zero Vel. (entire dataset)	0.1281	0.0980	0.0935	0.2241	0.2221	0.2048	0.1999
	Zero Vel. (mean per-action avg.)	0.1171	0.0912	0.0998	0.1953	0.1948	0.1891	0.1867
Algorithmic	Learning Trajectory Dependencies [24]	0.1046	0.0914	0.1075	0.1812	0.1671	0.1592	0.1557
	History Repeats Itself [23]	0.0901	0.0795	0.0671	0.2021	0.1995	0.1817	0.1758
	Ours (Deterministic)	0.0958	0.0797	0.0557	0.1920	0.1885	0.1868	0.1783
	Ours (before Refinement)	0.1228	0.1236	0.1030	0.1280	0.1397	0.1374	0.1461
	Ours	0.0930	0.1059	0.0909	0.1173	0.1219	0.1182	0.1320
	Method ↓	L-BigToe ↓	L-SmallToe ↓	L-Heel ↓	R-BigToe ↓	R-SmallToe ↓	R-Heel ↓	
Statistical	Avg. Train Pose (entire dataset)	0.3352	0.3514	0.3403	0.3163	0.3229	0.3153	
	Avg. Train Pose (mean per-action avg.)	0.3461	0.3649	0.3556	0.3131	0.3162	0.3191	
	Zero Vel. (entire dataset)	0.0955	0.0938	0.1009	0.1648	0.1644	0.1582	
	Zero Vel. (mean per-action avg.)	0.1028	0.1016	0.1081	0.1331	0.1317	0.1309	
Algorithmic	Learning Trajectory Dependencies [24]	0.1062	0.1059	0.1253	0.1491	0.1544	0.1532	
	History Repeats Itself [23]	0.0733	0.0717	0.0736	0.1140	0.1125	0.1076	
	Ours (Deterministic)	0.0554	0.0553	0.0616	0.0864	0.0893	0.0848	
	Ours (before Refinement)	0.1374	0.1384	0.1205	0.1294	0.1435	0.1377	
	Ours	0.1171	0.1398	0.1092	0.1050	0.1337	0.1169	

Table 10: Characteristic 3D pose prediction performance on GRAB, broken down by individual joint MPJPE [meters].

	Method	Hip	R-Hip	R-Knee	R-Heel	R-Foot	R-Toe	L-Hip	L-Knee
Statistical	Avg. Train Pose (entire dataset)	0.4575	0.4721	0.4696	0.4820	0.4838	0.4843	0.4445	0.4606
	Avg. Train Pose (mean per-action avg.)	0.3629	0.3771	0.4118	0.4268	0.4208	0.4210	0.3498	0.3256
	Zero Vel. (entire dataset)	0.8940	0.8003	1.0308	1.3872	1.4305	1.4090	0.9927	1.1959
	Zero Vel. (mean per-action avg.)	0.9242	0.8276	1.0372	1.3824	1.4356	1.4163	1.0252	1.2081
Algorithmic	Learning Trajectory Dependencies [24]	0.2388	0.2198	0.1606	0.1718	0.2048	0.1851	0.2187	0.1609
	History Repeats Itself [23]	0.2389	0.2472	0.1835	0.1999	0.2009	0.2026	0.2392	0.1631
	Ours (Deterministic)	0.2480	0.2600	0.1883	0.2085	0.2119	0.2140	0.2457	0.1813
	Ours (before Refinement)	0.1706	0.1907	0.1384	0.1394	0.1393	0.1268	0.1783	0.1291
	Ours	0.1626	0.1801	0.1116	0.1387	0.1202	0.1179	0.1741	0.1108
	Method	L-Heel	L-Foot	L-Toe	Hip	Spine	Neck	Nose	Head
Statistical	Avg. Train Pose (entire dataset)	0.4649	0.4641	0.4656	0.4575	0.4526	0.4752	0.4732	0.4827
	Avg. Train Pose (mean per-action avg.)	0.3142	0.3044	0.3091	0.3628	0.3561	0.3535	0.3658	0.3670
	Zero Vel. (entire dataset)	1.5175	1.5644	1.5611	0.8939	0.8310	0.7760	0.8091	0.8121
	Zero Vel. (mean per-action avg.)	1.5141	1.5627	1.5618	0.9241	0.8674	0.8151	0.8507	0.8540
Algorithmic	Learning Trajectory Dependencies [24]	0.1552	0.1463	0.1833	0.2313	0.2127	0.2123	0.2168	0.2209
	History Repeats Itself [23]	0.1534	0.1567	0.1616	0.2389	0.2450	0.2473	0.2466	0.2439
	Ours (Deterministic)	0.1749	0.1779	0.1784	0.2479	0.2514	0.2499	0.2510	0.2538
	Ours (before Refinement)	0.1423	0.1567	0.1228	0.1819	0.1902	0.1865	0.1838	0.1727
	Ours	0.1432	0.1396	0.1238	0.1652	0.1805	0.1659	0.1722	0.1676
	Method	Neck	L-Shoulder	L-Elbow	L-Hand	L-Hand	L-Thumb	L-Finger	L-Finger
Statistical	Avg. Train Pose (ent. dataset)	0.4752	0.4569	0.4498	0.4884	0.4884	0.5019	0.5212	0.5212
	Avg. Train Pose (mpa.)	0.3535	0.3402	0.3602	0.4073	0.4073	0.4127	0.4389	0.4389
	Zero Vel. (ent. dataset)	0.7760	0.8867	1.0354	1.0587	1.0587	1.0216	1.0986	1.0986
	Zero Vel. (mpa.)	0.8151	0.9271	1.0797	1.0308	1.0308	0.9930	1.0429	1.0429
Algorithmic	Learning Trajectory Dep. [24]	0.2087	0.2083	0.2768	0.4285	0.3911	0.4488	0.4623	0.5048
	History Repeats Itself [23]	0.2473	0.2411	0.2503	0.3058	0.3058	0.3337	0.3423	0.3423
	Ours (Deterministic)	0.2499	0.2507	0.2809	0.3481	0.3468	0.3629	0.3983	0.3983
	Ours (before Refinement)	0.1875	0.1726	0.2091	0.2806	0.2555	0.2506	0.2810	0.2886
	Ours	0.1706	0.1615	0.1881	0.2267	0.2263	0.2487	0.2661	0.2749
	Method	Neck	R-Shoulder	R-Elbow	R-Hand	R-Hand	R-Thumb	R-Finger	R-Finger
Statistical	Avg. Train Pose (ent. dataset)	0.4752	0.4794	0.4995	0.5737	0.5737	0.5650	0.6346	0.6346
	Avg. Train Pose (mpa.)	0.3535	0.3597	0.3460	0.4463	0.4463	0.4460	0.4979	0.4979
	Zero Vel. (ent. dataset)	0.7760	0.6701	0.6323	0.7644	0.7644	0.7491	0.8209	0.8209
	Zero Vel. (mpa.)	0.8151	0.7038	0.6539	0.8158	0.8158	0.8007	0.8915	0.8915
Algorithmic	Learning Trajectory Dep. [24]	0.2313	0.1922	0.2447	0.3987	0.4043	0.3981	0.5067	0.5098
	History Repeats Itself [23]	0.2473	0.2404	0.2501	0.3473	0.3473	0.3516	0.4477	0.4477
	Ours (Deterministic)	0.2499	0.2538	0.2946	0.4077	0.4070	0.4157	0.4995	0.5000
	Ours (before Refinement)	0.1885	0.1946	0.1979	0.3340	0.3341	0.3046	0.3540	0.3898
	Ours	0.1634	0.1738	0.1933	0.2685	0.2660	0.2752	0.3428	0.3760

Table 11: Characteristic 3D pose prediction performance on Human3.6M, broken down by individual joint MPJPE [meters].