Supplemental Material for Forecasting Characteristic 3D Poses of Human Actions

In this supplemental, we show additional qualitative results (Sec. 1), additional quantitative analysis (Sec. 2), detail our network architecture specification (Sec. 3), provide additional details regarding the dataset (Sec. 4) as well as our training setup (Sec. 5), and discuss potential negative societal impacts of our method (Sec. 6).

1. Additional Qualitative Results.

We show additional qualitative results of our method in Fig. 2, which demonstrate the diversity of our characteristic pose predictions for a given input sequence. Our approach not only effectively models the multi-modal nature of characteristic poses, but also captures the final target action pose (highlighted pose prediction).

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

2. Additional Quantitative Results.

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

Autoregressive Joint Order. We determined the order of the joints for the autoregressive prediction empirically; most ambiguity occurred in active end-effectors (i.e. right and left hands), whereas the rest of the body tended to have lower variability. In Tab. 1, we compare our original approach of (right hand, left hand, rest) with two alternatives: (left hand, right hand, rest), and (full autoregressive from human kinematic chain following left/right hands). Our method is robust to these orderings (though diversity of the rest of the body except hands decreases with autoregression through the kinematic chain).

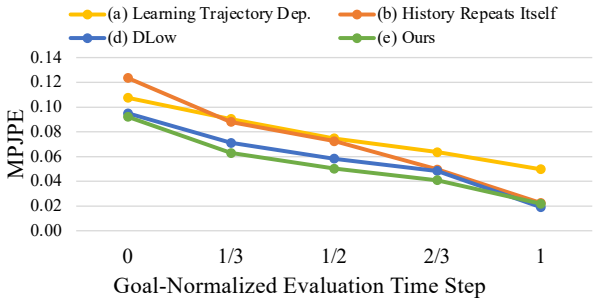

Figure 1. MPJPE comparison to baselines, evaluating with the input sequence at different points in time: from the start of the sequence (0) to N frames before the target characteristic pose (1).

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

Table 1. Ablation analysis on autoregressive order on GRAB data.

Grid Resolution and Offset Prediction. We show additional ablations on the effect of grid resolution and offset prediction in Tab 2 on GRAB data; A resolution of 16^3 performs better than 8^3 or 32^3. Our offset prediction helps mitigate grid artifacts even at 32^3.

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

Table 2. Ablation analysis on heatmap grid size and offset prediction on GRAB data.

Per-Bodypart MPJPE. In Tab. 6, we show our final pose prediction performance in MPJPE, broken down per bodypart, as compared to sequential baselines.

Characteristic Pose Forecasting with Ground Truth Action Labels. In Tab. 3, we additionally evaluate our approach using ground truth action labels as input to provide additional contextual information.

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

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

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

Table 3. Comparison of ours to an ablation with ground truth action labels as additional input.

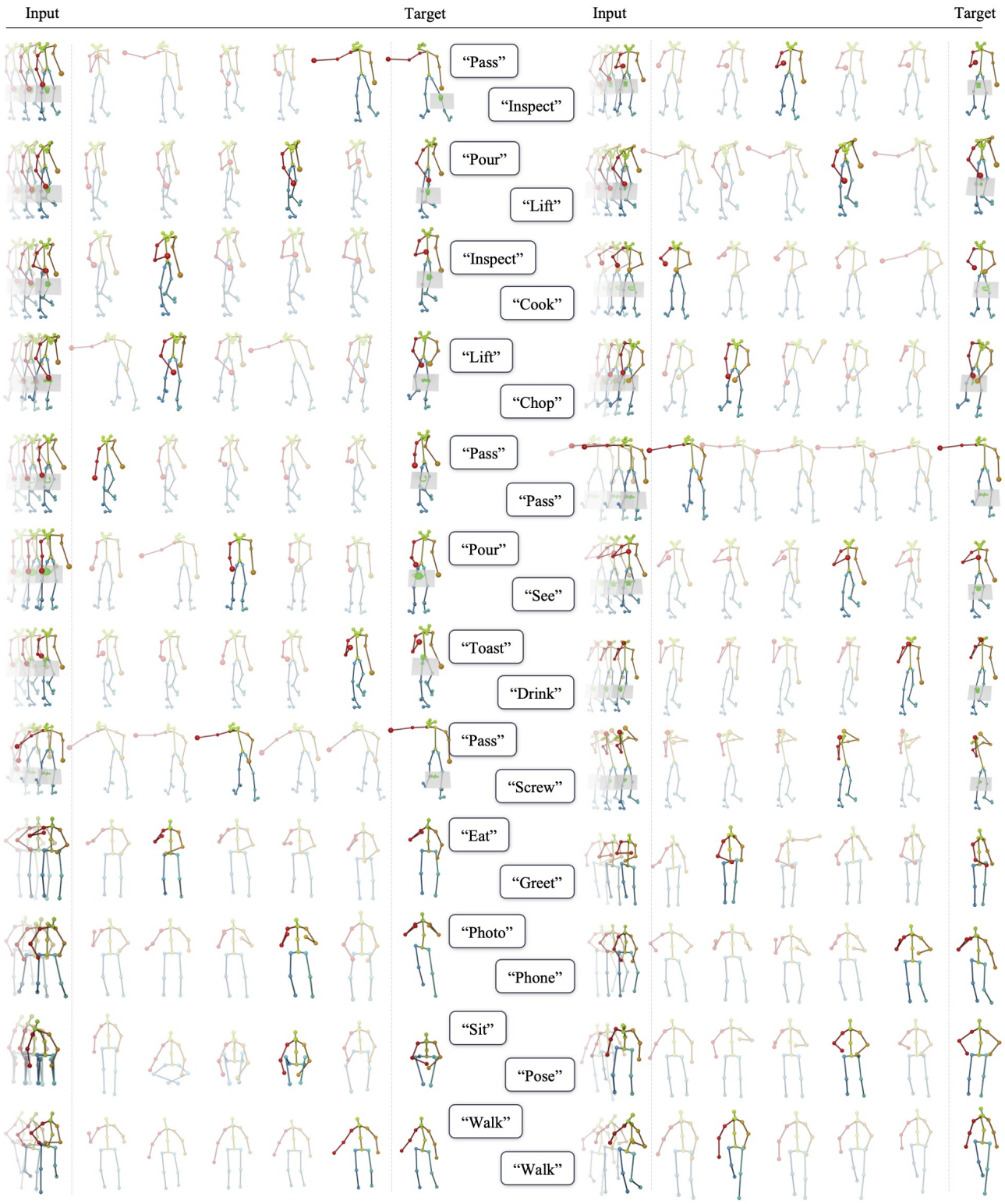

Figure 2. Additional qualitative results, showing the for each action sequence the inputs (left), our diverse set of predictions (middle) and the target action pose (right). Our final pose prediction is highlighted for each action sequence.

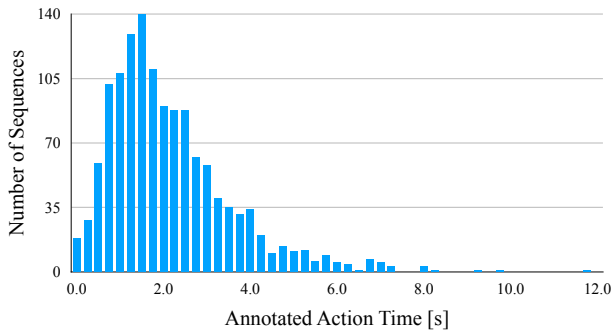

Figure 3. Times at which characteristic poses occur for GRAB.

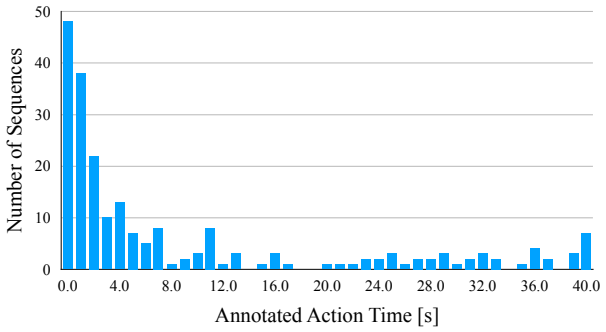

Figure 4. Times at which char. poses occur for Human3.6M.

3. Architecture Details

Fig. 6 details our network specification from input (left) to heatmap and offsets output (right). For each GRU layer, we provide the hidden dimension and number of layers in parentheses, for normalization layers the dimension to be normalized over, for dropout layers the dropout probability p, and for convolutions the number of input and output channels as well as kernel size (ks), stride (str), and padding (pad). We apply cross-entropy (CE) losses at a heatmap resolution of 8^3 and at the final resolution of 16^3; for the offsets prediction, we concatenate the offsets volume generated from the last input skeleton after 5 convolution blocks and supervise the final predictions with an ℓ_1 loss.

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

4. Dataset

GRAB Pose Layout. Since GRAB [8] not only provides a human skeleton representation but full body shape parameters, we preprocess all pose sequences by first extracting relevant joints for our approach. For this, we chose the 3d

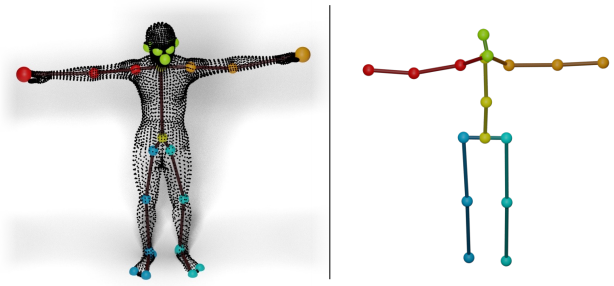

Figure 5. GRAB [8] body and our extracted skeleton joints overlaid (left); 17-joint skeleton based on Human3.6M [4] (right).

OpenPose [3] layout as it describes the prevalent body joints and is widely used for representing 3d poses. Note that we do not apply the OpenPose method on 2d data; we only use their joint definitions in 3d. We extract 25 body joints from the SMPL-X body given by the GRAB dataset [8] using the correspondences shown in Tab. 5. Additionally, we denote in Tab. 5 the correspondences of joints to body parts, for the body part analysis in Tab. 6. Fig. 5 (left) visualizes our joint selection, overlaying the body shape given in GRAB as a point cloud over the 25-joint skeleton.

Human3.6M Pose Layout. For all our experiments on Human3.6M [4], we use 17 pose joints, visualized in Fig. 5 (right). Tab. 4 describes the exact joints used as well as the correspondences of joints to body parts, as used in Tab. 6.

Visualization Details. While our approach is agnostic to context or action, we visualize the context provided by GRAB [2, 8] (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. 7, and note the strong time differences at which the characteristic poses occur. Furthermore, Fig. 3 and Fig. 4 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.

5. Additional Training Details

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

for the class imbalances, by the inverse of their log-scaled occurrence, and a weight of 0.1 for the no-probability bin.

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

Table 4. Joint Correspondences for Human3.6M

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

Table 5. Joint Correspondences for GRAB

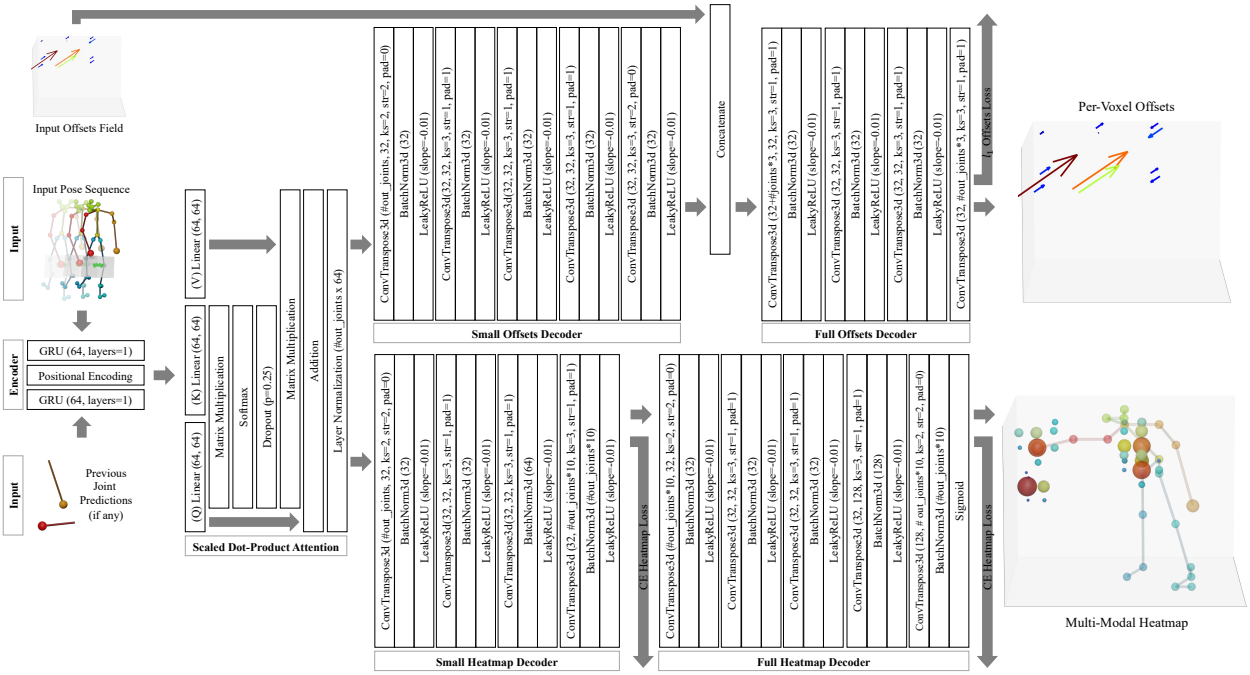

Figure 6. Our network architecture with details for encoder, scaled dot-product attention, as well as heatmap and offsets decoders.

	GRAB						H3.6M					
Method	R. Arm ↓	L. Arm ↓	R. Leg ↓	L. Leg ↓	Spine ↓	Head ↓	R. Arm ↓	L. Arm ↓	R. Leg ↓	L. Leg ↓	Spine ↓	Head ↓
L. T. D. [6]	0.165	0.115	0.058	0.057	0.028	0.085	0.225	0.225	0.135	0.146	0.108	0.123
H. R. I. [5]	0.160	0.113	0.056	0.055	0.026	0.079	0.199	0.191	**0.079**	0.088	0.040	0.089
DLow [9]	0.146	0.109	0.052	0.050	0.024	0.068	0.174	0.169	0.108	0.112	0.044	0.096
Ours	**0.105**	**0.084**	**0.045**	**0.045**	**0.020**	**0.057**	**0.147**	**0.122**	0.091	**0.085**	**0.033**	**0.066**

Table 6. Characteristic 3d pose prediction performance comparison to baselines, broken down by body part MPJPE.

State-of-the-art comparisons. We use the official code with default settings of the methods we compare to ([6], [5], and [9]). We train all methods from scratch on our characteristic 3d pose dataset, setting the number of input frames to 10 and the number of output frames to 100. From the predicted sequence, we evaluate the pose at a timestep predicted by the baselines themselves as characteristic pose and compare it to the target. This scenario is the closest to our approach, as predicting characteristic 3d poses involves which pose is the characteristic pose.

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

For DLow [9], we add one linear layer to the final feature output of each of the 100 steps, followed by a ReLU, reducing each step's output dimension to 10. Then, one additional linear layer summarizes the combined output of all steps ($100 * 10$) down to a vector of size 100.

In the case of History Repeats Itself [5], we add a classification head consisting of one linear layer, a 1d batch norm, a ReLU, and one additional linear layer to the output of their last Graph Convolution Block (GCN). While the first linear layer keeps the original dimensionality of 100, the second linear layer reduces the dimension from #graph_nodes $*$ 100 down to 100.

Finally, for Learning Trajectory Dependencies [6], we apply the same architecture and add a linear layer, a 1d batch norm, a ReLU, and a second linear layer after the final GCN. Here, we first reduce the per-node feature dimension from 256 to 100 and combine the features of all nodes with the second linear layer, going from #graph_nodes $*$ 100 down to 100.

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

To evaluate the diversity and quality of multi-modal outputs, 10 samples are taken from a probabilistic method for each input sequence, and we report diversity in terms of MPJPE between samples as well as the Inception Score, following [1].

6. Potential Negative Societal Impacts

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

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

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

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

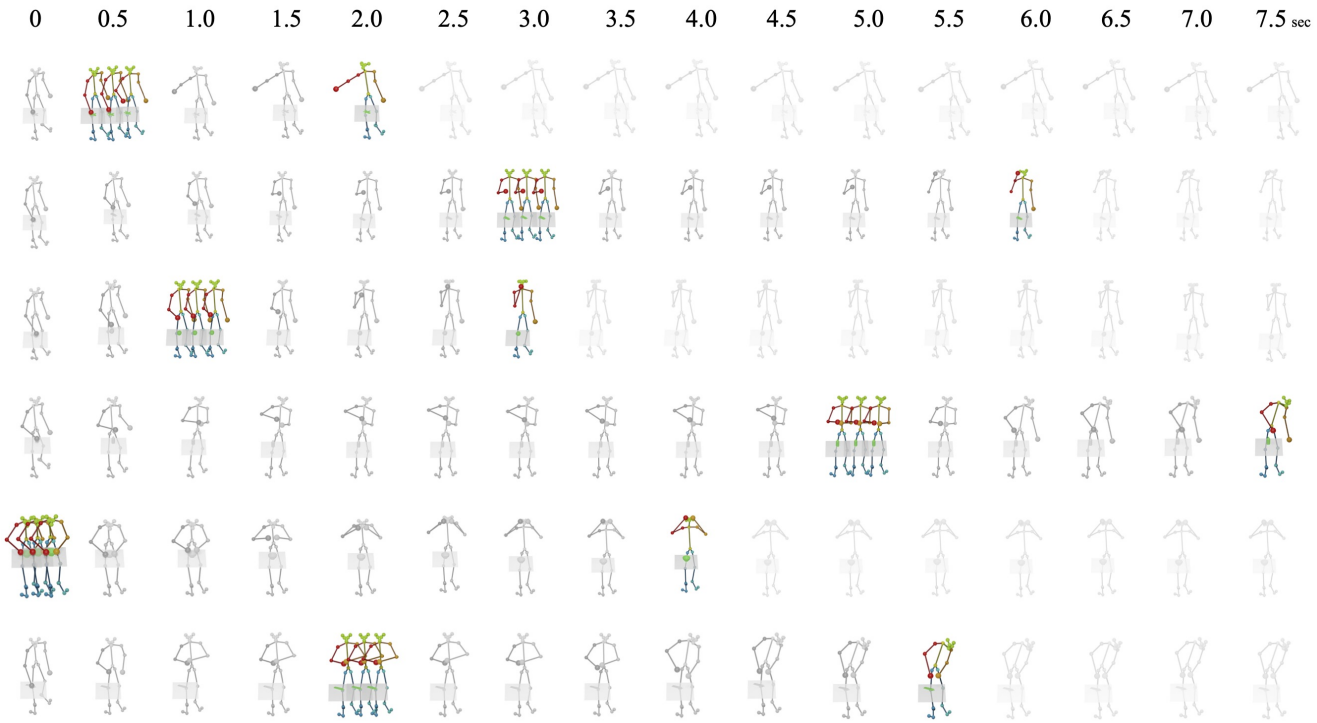

Figure 7. Sample input-target pairs (colored) for our characteristic 3d pose forecasting task, with temporal snapshots along the sequence (grayscale). Each snapshot is half a second apart. Depicted as input is the last frame of the respective input sequence.

References

[1] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5223–5232, 2020. 5

[2] 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. 3

[3] 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, 4

[4] 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. 3

[5] 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. 5

[6] 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. 5

[7] 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. 4

[8] 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. 3

[9] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*, pages 346–364. Springer, 2020. 5