CST: Character State Transformer for Object-Conditioned Human Motion Prediction Supplementary Materials

Kuan-Wei Tseng[1] Rei Kawakami[1] Satoshi Ikehata[1,2] Ikuro Sato[1,3]
[1]Institute of Science Tokyo [2]National Institute of Informatics [3]Denso IT Laboratory
Tokyo, Japan
https://kuan-wei-tseng.github.io/CST

1. More Implementation Details

1.1. CMU MoCap dataset

The original CMU Motion Capture Database [1] provides data in *tvd*, *c3d*, and *amc* formats. For development purposes, we used *bvh* files sourced from CGSpeed [2] to enhance accessibility. To ensure data quality, we eliminated duplicate and unreliable joints as recommended by the original database, narrowing down to 18 specific joints: *lowerback*, *chest*, *chest2*, *neck*, *head*, *headend*, *right* and *left shoulders*, *elbows*, *wrists*, *hips*, *knees*, and *ankles*. Additionally, the original dataset comprises over 4M poses. From this, we selectively extracted about poses from subjects numbered 7, 8, 9, 13, 14, 15, 16, 35, 40, 41, and 91. These selections specifically include motions such as walking, turning, and sitting, as detailed in the database's descriptions. We split the data into training and validation set at a 8:2 ratio. As a result, about 500k poses are used for pre-training.

1.2. Model Architectures

Human Joint Transformer. The dimension of the initial human joint embedding for each joint is 256. The main part of the HJT is implemented with the transformer encoder layer in PyTorch, with nhead = 8 (number of heads) and dim_feedforward = 256 (the dimension of the feedforward network model).

Spatiotemporal State Transformer. We utilize multihead attention in PyTorch with num_heads = 8 (number of heads) to construct self-attention and cross-attention layers for SST. Same as the transformer, we also adopt dropout with value 0.1. For the cross-attention layer, the query for the trajectory sequence and goal sequence are the goal sequence and trajectory sequence, respectively. The key and value are from the same source. The token dimension is 256, while the hidden layer dimension of the MLP is 512.

Table 1. Parameters of the Mask Recovery Head.

Layer Name	Layer Type	Input Features	Output Features
hjt_linear	Linear	$N_j \times 256$	$N_j \times 12$
linear1	Linear	$N_j \times 12$	512
linear2	Linear	512	512
linear3	Linear	512	$N_j \times 12$

Mask Recovery Head. The model consists of linear layers that include bias terms where N_j is the number of joint in each skeleton. The details of each layer are summarized in the Table 1 below:

2. More Experimental Results

The extended table for Table 3 in the main paper which compares different pre-training condition can be found in Table 3. Besides, we present more qualitative results on the SAMP dataset in Fig. 1. We will release the demo with interactive object will be released along with our code. Please refer to the screen shots in the following pages.

References

[1] Carnegie Mellon University - CMU Graphics Lab - motion capture library — mocap.cs.cmu.edu. http://mocap.cs.cmu.edu. [Accessed 13-03-2024]. 1

[2] cgspeed - Motion Capture — sites.google.com. https://sites.google.com/a/cgspeed.com/cgspeed/motion-capture. [Accessed 13-03-2024]. 1

[3] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J. Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11374–11384, October 2021. 2

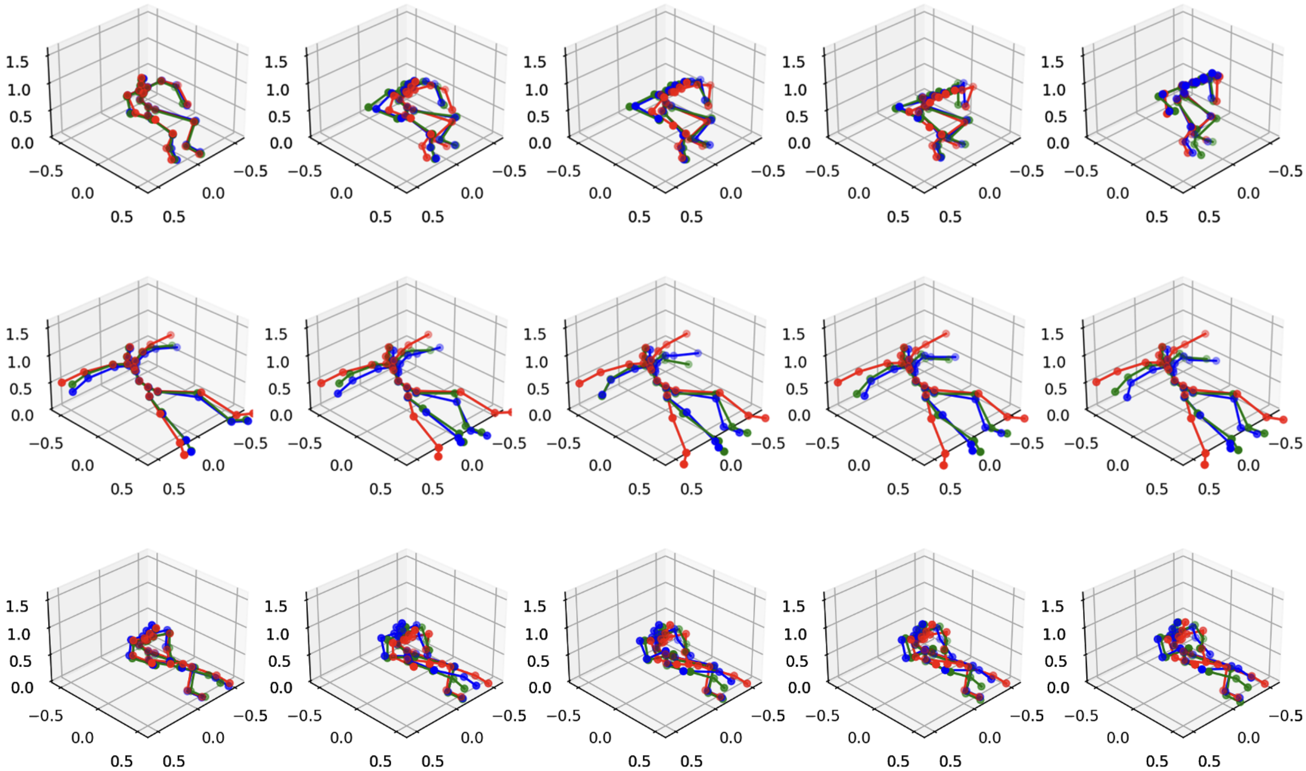

Figure 1. More Qualitative Results on the SAMP dataset [3]. In each row, we arrange from left to right in chronological order. We color the ground-truth, motion generated by SAMP and motion generated by our method in red, blue, and green, respectively.

Table 2. Comparison of different pre-training conditions.

Dataset	**# Joints**	**Masking Ratio**	**MPJPE ↓**			**MPJRE ↓**		
			1-step	**5-step**	**10-step**	**1-step**	**5-step**	**10-step**
from sratch	N/A	N/A	0.184	0.341	0.403	13.417	28.192	33.369
HumanAct12	24	10%	0.181	0.325	0.377	13.227	26.364	30.923
HumanAct12	24	20%	0.176	0.320	**0.370**	13.465	26.250	30.409
HumanAct12	24	30%	0.180	0.317	0.368	13.375	26.375	30.640
CMU MoCap	18	10%	0.175	0.323	0.387	12.793	25.500	30.000
CMU MoCap	18	20%	**0.169**	**0.318**	0.376	**12.411**	**25.119**	**29.593**
CMU MoCap	18	30%	0.175	0.321	0.382	12.634	25.431	30.517