Supplementary Material for Global-local Motion Transformer for Unsupervised Skeleton-based Action Learning

A Learned Positional Embedding

As illustrated in Fig. A, the positional embedding obtained through training effectively reflects the sequential information of the motion sequence. Fig. A presents similarity between learned positional embedding vectors $M_{t,j} \in \mathbb{R}^D$, where $t = 1, \cdots, T$ and $j = 1, \cdots, PK$. $M_{t,j}$ is a slice of the positional embedding tensor $M \in \mathbb{R}^{T \times PK \times D}$, where t and j denote the frame and joint index, respectively. Each tiled figure shows a cosine similarity between the positional embedding vector of the indicated joint and frame and that of all other indices. As indicated as red circles, closer frames tend to have similar positional embedding vectors. This implies that the positional embedding is effectively trained to reflect temporal order within the motion sequence.

B Effectiveness of Natural-speed Input

The performance of using the input sequence with a natural speed exceeds the performance of using the sampled 150 and 300 frames as an input, as presented in Table A. The results verify that the proposed scheme is effective in learning the natural dynamics between joints. Note that H-tarnsformer [1] leverages the input sequences as sampling to 150 frames. The result of sampling 300 frames is worse than that of sampling 150 frames, owing to the large distortion of the motion sequences.

C Learned Attention Maps

Fig B. and Fig. C are supplementary figures for Fig. 4 and Fig. 5 in the main manuscript, respectively. In Fig. B, we illustrates the average of learned temporal attention maps for every head in each GL-Transformer block. Likewise, Fig. C shows the average of learned spatial attention maps for every head in each GL-Transformer block. The attention maps are averaged over 300 motion sequences of the evaluation data.

Table A. Ablation study for verifying the effectiveness of natural-speed input sequences in the NTU-60 dataset with the linear evaluation protocol

Type	Accuracy(%)	
	xsub	xview
H-transformer [1]	69.3	72.8
Sampling (150 frames)	75.1	80.2
Sampling (300 frames)	73.2	79.0
Natural speed with attention mask	**76.3**	**83.8**

D Additional Study for Corrupted Input Sequence

Some of the previous works [3, 2, 1] in unsupervised skeleton-based action learning perform pretraining tasks with randomly corrupted input sequences. LongT GAN [3] and P&C [2] reconstruct the corrupted joints, and H-transformer [1] predicts the instantaneous velocity of the joints for the masked (i.e. corrupted) frames following the scheme in *masked token prediction* which is widely used in the transformer-based pretraining methods. We conduct additional experiments with randomly corrupted input sequences to see their effects on the proposed method. We randomly set the joint coordinates of the motion sequences to (0, 0, 0) and pretrain the proposed model with the corrupted sequences. Indeed, the accuracy gradually decreases while the proportion of the corrupted joint increases, as presented in Table B. Because our pretraining strategy is not to reconstruct the corrupted joints but to predict the displacement of all joints, the corrupted sequence rather causes the loss of information of the training data, resulting in lower performance.

Table B. Experimental results when the proposed model is pretrained on motion sequences with randomly corrupted joints. The experiments follow the linear evaluation protocol on NTU-60 dataset.

Corrupted joint proportion	Accuracy(%)	
	xsub	xview
0%	**76.3**	**83.8**
5%	75.7	82.5
10%	74.2	81.2
15%	73.7	79.9
20%	72.4	79.0

References

1. Cheng, Y.B., Chen, X., Chen, J., Wei, P., Zhang, D., Lin, L.: Hierarchical transformer: Unsupervised representation learning for skeleton-based human action recognition. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2021)
2. Su, K., Liu, X., Shlizerman, E.: Predict & cluster: Unsupervised skeleton based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9631–9640 (2020)
3. Zheng, N., Wen, J., Liu, R., Long, L., Dai, J., Gong, Z.: Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)

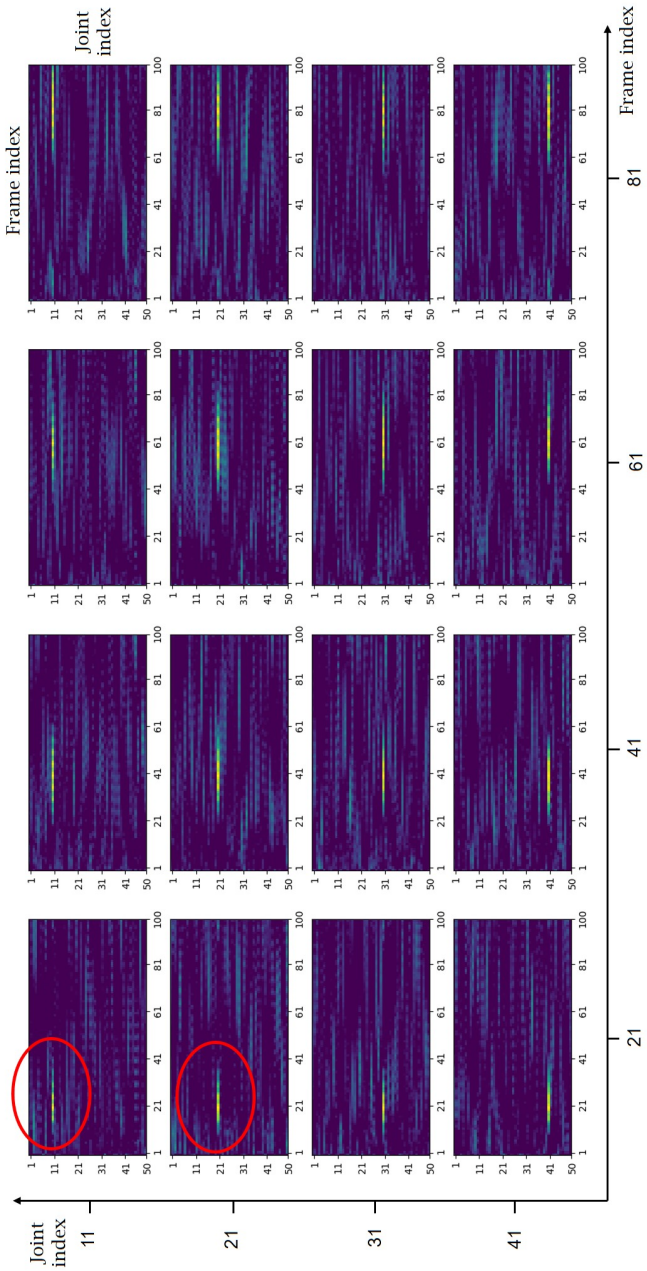

Fig. A. Similarity between learned positional embedding vectors. Each tiled figure shows the cosine similarity between the positional embedding vector of the indicated joint and frame and that of all other indices. Yellow color indicates a large cosine similarity

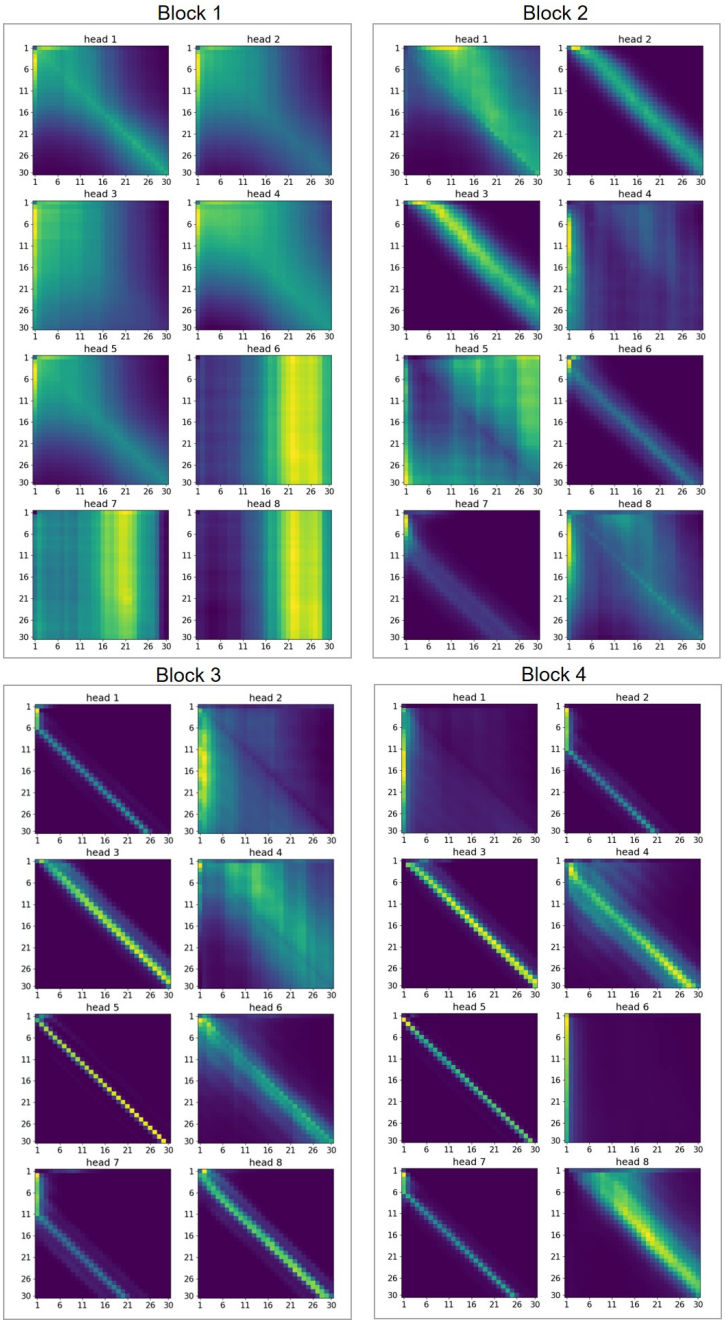

Fig. B. Learned temporal attention maps averaged over 300 evaluation sequences. Yellow color indicates a large value

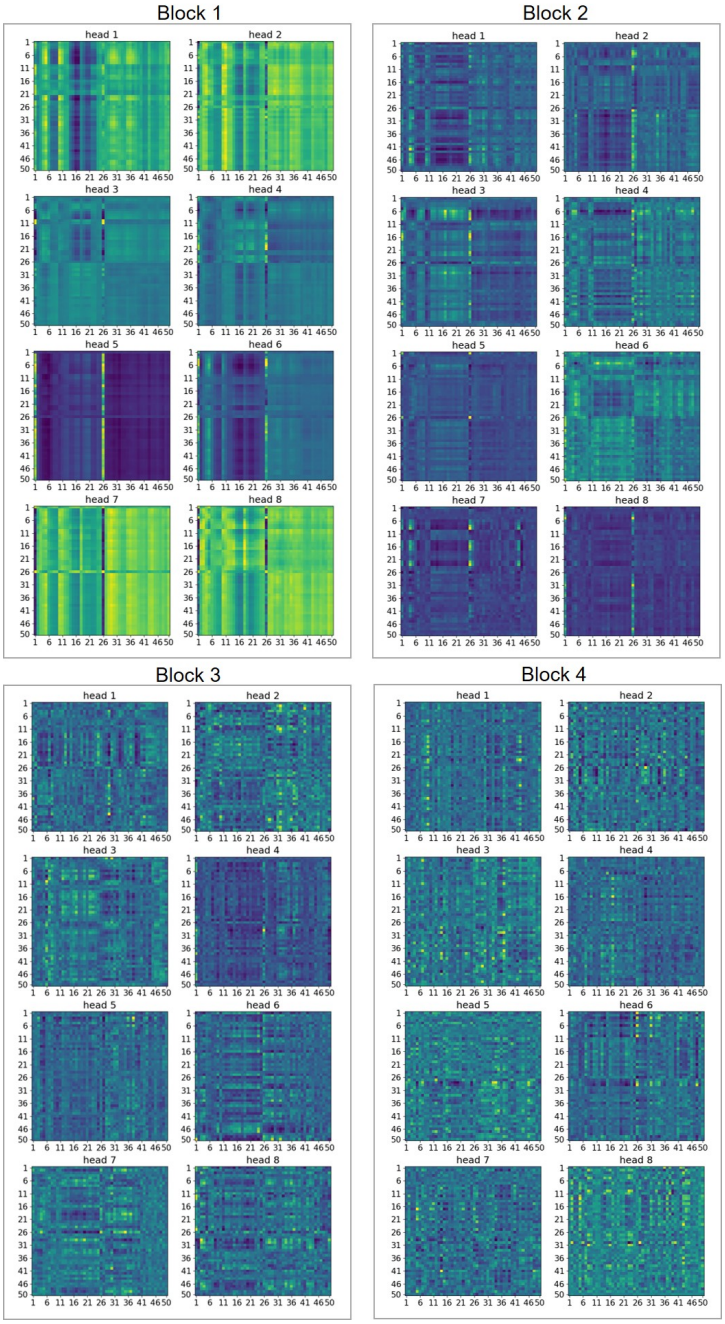

Fig. C. Learned spatial attention maps averaged over 300 evaluation sequences. Yellow color indicates a large value