Supplementary: Joint-Relation Transformer for Multi-Person Motion Prediction

Anonymous ICCV submission

Paper ID 7193

1. 3DPW dataset

Drift problem. When processing and visualizing the original data of 3DPW-SoMoF, discovered a significant issue with unnatural drift caused by camera movement, see the **ATTACHMENT**. Even if the person in the scene is not moving at all, the absolute position of the individual in the world coordinate system exhibits strange and erratic drift. This phenomenon is unrelated to human movement, making it challenging to accurately forecast future motions by studying past human behavior.

Previous work usually uses 3DPW for single-person motion estimation tasks. And the drifting problem is not so significant in this task, since the person's pose is generally centered to the central joint (*e.g.*, the hip joint) when processing the data, that is, the human trajectory is not considered and they only take care of the change in pose when predicting the future motion. However, in multi-person motion prediction tasks, the absolute position is quite important since the relative distance between joints derived from their absolute positions contains crucial information for interaction modeling.

We thus obtain a clean version of 3DPW-SoMoF by estimating the camera movement and subtracting it. The comparison between the original data and the corrected data can be found in the attachment.

Metric. The metric we mainly use for comparison on 3DPW(both 3DPW-SoMoF and 3DPW-SoMoF/RC) in the paper is VIM, which is first proposed in the paper[1]. They claim that the VIM is just the simple MPJPE metric except that the invisible joints (if exist) are not penalized and are discarded by considering the truth. However, their released code shows that it wrongly calculates the mean 3J-dimensional distance between the ground truth and predicted joint positions after flattening the joint and coordinate dimensions instead of the standard MPJPE calculation. Although this metric can measure the difference between the predicted result and the ground truth, it lacks a precise physical interpretation and appears meaningless. However, in order to compare with the results of previous work such as TRiPOD[1], we still choose to use this metric.

Table 1. Experimental results in MPJPE on the 3DPW-SoMoF/RC test sets. The best results are highlighted in bold. Our method achieves the best performance on the 3DPW-SoMoF/RC test sets.

Methods	3DPW-SoMoF/RC					
	AVG	100	240	500	640	900
Zero Velocity	16.72	5.65	10.27	18.02	21.58	28.07
LTD [5] $'^{2019}$	15.21	4.78	8.81	16.62	19.92	25.92
DViTA [2]$'^{2021}$	14.27	3.65	7.70	15.22	18.89	25.89
MRT [4]$'^{2021}$	12.66	4.85	8.59	14.05	16.19	19.61
SoMoFormer [3]$'^{2022}$	10.55	2.60	5.94	11.77	14.29	18.15
Ours	**9.53**	**2.18**	**5.05**	**10.50**	**12.98**	**16.93**

MPJPE on 3DPW-SoMoF/RC. We also provide the MPJPE result on the 3DPW-SoMoF/RC dataset to have a more meaningful comparison, as shown in Tab. 1. Our model outperforms all the previous works at all timesteps.

Performance Analysis. Our model outperforms most state-of-the-art methods on all datasets except SoMoFormer[3] on 3DPW-SoMoF. Since our model directly takes the absolute positions of joints as input and all information updates and fusions are based on this initial input information, the quality of the input data will greatly affect the performance of our model, *i.e.*, the drift problem of the 3DPW-SoMoF dataset brings great trouble to our method. While SoMoFormer avoids this problem by modeling the overall trajectory information as an additional positional embedding, and the joints input only contains the pose relative to the central joints. We can see that i) the human body drift in the 3DPW-SoMoF dataset is unnatural and abnormal, which makes the comparison on the 3DPW-SoMoF dataset not very meaningful; ii) although troubled with the drifting problem, our model still outperforms most previous methods and obtained competitive results compared to SoMoFormer on 3DPW-SoMoF, which demonstrates the strong ability of our model in predicting future motions; iii) our model outperforms all existing models on 3DPW-SoMoF/RC, CMU-Mocap, and MuPoTS-3D, which illustrates that when the input data quality is guaranteed, our model exhibits the best prediction performance.

2. Baseline Methods

Zero Velocity. As a widely used baseline method, zero velocity predicts future velocity as zero, providing predicted results for a stationary state. Technically, we use the last observed frame as the prediction for the future and calculate the corresponding metrics.

LTD [5]. LTD adopts the DCT transformation on the input sequence and design graph convolutional layers to fuse these coefficients. Before the inverse DCT transformation to the output sequence, they use the residual connection to the input DCT coefficients, expecting the model to learn the residual movement. In the 3DPW-SoMoF/RC dataset, LTD has 4 GCN layers to avoid the over-smoothing problem with a hidden size of 256. We repeat the last observed frame for 14 times such that the input and output have the same sequence length of 30. We train the model for 128 epochs with an initial learning rate of 1×10^{-3}. We adopt the step schedule which decays the learning rate by 0.8 every 10 epochs.

DViTA [2]. They decouple the trajectory from the pose and model the single-person motion using the LSTM for trajectory prediction and VAE for pose prediction. We follow the model setting and train the model using our training schedule as LTD.

MRT [4]. MRT uses the Transformer-based encoder and decoder to fuse the DCT coefficients. Note that the input sequence and output sequence have different lengths and we use the full-connected layers (MLP) to fill this gap. In the 3DPW-SoMoF/RC dataset, DViTA has an embedding dimension of 8, a hidden size of 64, and a latent dimension of 32. We use the same training schedule in our experiments without pre-train.

SoMoFormer [3]. They also uses the Transformer-based encoder and decoder to fuse the DCT coefficients. However, SoMoFormer firstly pads the input sequence using the last observed frame such that the input and the output have the same length. For SoMoFormer in the 3DPW-SoMoF/RC dataset, we pre-train the model on AMASS dataset and finetune the model following our training schedule.

3. More Ablation Studies

Relation Function. We verify the effect of the proposed relation function by comparing it with two kinds of variants: i). the relation function only contains the first-order term, *i.e.*, $f_{\text{RF}}(\mathbb{F}_R) = \mathbb{F}_R \mathbf{W}_l$; ii). the relation function only contains the second-order term, *i.e.*, $f_{\text{RF}}(\mathbb{F}_R) = \sum(\mathbb{F}_R \mathbf{W}_q^1 \odot \mathbb{F}_R \mathbf{W}_q^2)$. The result in MPJPE on CMU-Mocap is shown in Tab. 2. We can see that together with both the first-order term and the second-order term, the model achieves the best performance.

Table 2. Ablation of relation function in MPJPE on CMU-Mocap test sets.

First-order	Second-order	1s	2s	3s
✓		8.9	14.7	19.4
	✓	8.5	14.2	18.8
✓	✓	**8.3**	**13.9**	**18.5**

Table 3. Ablation study of training loss. 'Baseline' means the model trained only with the prediction loss, '+ Recon. Loss' represents the model trained with additional reconstruction loss, '+ DS Loss' is the model trained with additional deep supervision loss. Our model is trained with both reconstruction loss and deep supervision, which is denoted as '+ DS & Recon. Loss'.

Methods	3DPW-SoMoF/RC					
	AVG	100	240	500	640	900
Baseline	40.8	9.7	22.1	45.1	54.8	72.1
+ Recon. Loss	40.4	9.6	22.0	44.7	54.8	70.9
+ DS Loss	40.1	9.6	22.0	44.6	54.2	70.0
+ DS & Recon. Loss	**39.5**	**9.5**	**21.7**	**44.1**	**53.4**	**68.8**

Train Loss. When calculating the training loss, we introduce deep supervision loss to prevent the overfitting problem caused by the deep network, additionally, we also introduce the reconstruction loss to enhance the model's understanding of historical information and utilize it to improve performance. To validate these two kinds of losses, we compare our model with three models trained with different losses: i) baseline model that trains only with the prediction losses, *i.e.*, the total loss is calculated as $\mathcal{L} = \mathcal{L}_J(\widehat{\mathbf{Y}}_{\text{NJ}}) + \mathcal{L}_R(\widehat{\mathbb{R}}_{\mathbf{Y}})$; ii) the model trained with an additional reconstruction loss, *i.e.*, the total loss can be formulated as $\mathcal{L} = \mathcal{L}_J(\widehat{\mathbf{X}}_{\text{NJ}}, \widehat{\mathbf{Y}}_{\text{NJ}}) + \mathcal{L}_R(\widehat{\mathbb{R}}_{\mathbf{X}}, \widehat{\mathbb{R}}_{\mathbf{Y}})$; ii) the model training with the additional deep supervision, *i.e.*, $\mathcal{L} = \mathcal{L}_J(\widehat{\mathbf{Y}}_{\text{NJ}}) + \mathcal{L}_R(\widehat{\mathbb{R}}_{\mathbf{Y}}) + \mathcal{L}_{\text{DS}}$. Tab. 3 shows the result, it can conclude that both the deep supervision loss and the reconstruction loss contribute to the prediction results, while with both the two losses, the model achieves the best performance.

Pre-train. Due to the small sample size of the original 3DPW-SoMoF dataset, it is insufficient for training the Transformer, following the previous works[4, 3], we conduct pre-train on the AMASS dataset instead of directly training on 3DPW-SOMoF. Tab. 4 shows the result with/without pre-train. The pre-train operation on a large data set improves the accuracy of single-person action prediction and provides a good foundation for expanding to multi-person scenarios.

Structural Hyperparameter. To explore the most suitable structural hyperparameters, we conduct heavy experiments to study the model performance under different structural hyperparameters including the choice of layer number L, the heads number D_H, and the hidden size D. Re-

Table 4. Ablation study of pre-train.

Methods	3DPW-SoMoF/RC					
	AVG	100	240	500	640	900
w/o. Pre-train	45.7	12.0	26.8	51.8	61.9	76.2
w/ Pre-train	**39.5**	**9.5**	**21.7**	**44.1**	**53.4**	**68.8**

Table 5. Ablation of structural hyperparameter in VIM on 3DPW-SoMoF/RC test sets. L represents the number of joint-relation fusion layers, D_H is the number of heads and D is the hidden size of joint and relation feature.

Hyperparam			3DPW-SoMoF/RC						Param $\times 10^6$
L	D_H	D	AVG	100	240	500	600	900	
2	8	128	40.7	9.8	22.5	45.6	55.1	70.3	2.6
6	8	128	**39.5**	9.6	22.0	44.5	53.8	**67.9**	4.6
4	1	128	40.2	9.8	22.4	45.3	54.6	69.0	3.6
4	2	128	40.0	9.7	22.0	44.5	54.3	69.6	3.6
4	4	128	40.2	9.8	22.3	44.6	54.3	69.9	3.6
4	8	64	40.3	9.8	22.1	44.6	54.6	70.5	1.9
4	8	256	39.8	**9.4**	21.9	44.3	53.8	69.7	9.8
4	8	128	**39.5**	9.5	**21.7**	**44.1**	**53.4**	68.8	3.6

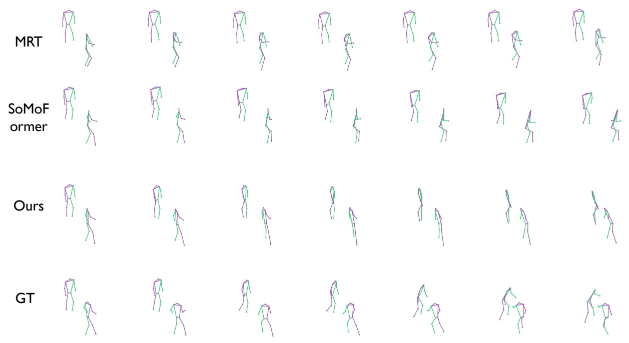


Figure 1. Another prediction sample.

sults are shown in Tab. 5. Fewer model layers or attention heads, and smaller hidden size will all lead to a decline in model performance. Although with 6 layers or 256 hidden size, the model achieves competitive performance, it also brings a heavy calculation and time consumption. Taking performance and time consumption into consideration, we selected the chosen setting with $L = 4, D_H = 8, D = 128$.

4. Visualization

Visualization of prediction result. We provide another prediction result where two persons are playing basketball (hard prediction case), see Fig. 1. All models fail to predict the future sequence while our method provides a relatively more precise and reasonable prediction.

Visualization of attention. Here we provide more visualization of the attention matrix. We first present the visualization results of the first layer of the attention matrix under 4 different motion samples, as Fig. 2 shows. Different motions have different intra-person and inter-person interaction patterns. We also present the attention visualization of different layers of the dancing sequence, as Fig. 3 shows.

References

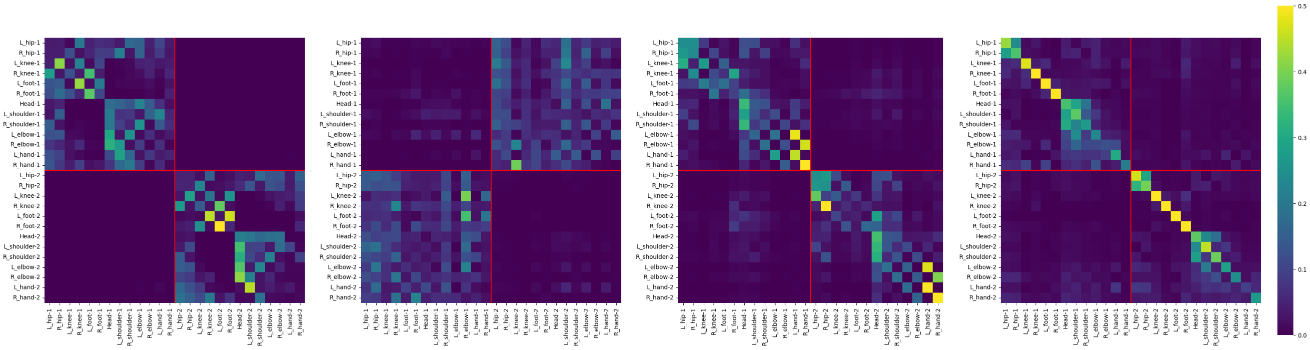
[1] Vida Adeli, Mahsa Ehsanpour, Ian Reid, Juan Carlos Niebles, Silvio Savarese, Ehsan Adeli, and Hamid Rezatofighi. Tripod: Human trajectory and pose dynamics forecasting in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13390–13400, 2021. 1

[2] Behnam Parsaeifard, Saeed Saadatnejad, Yuejiang Liu, Taylor Mordan, and Alexandre Alahi. Learning decoupled representations for human pose forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2294–2303, October 2021. 1, 2

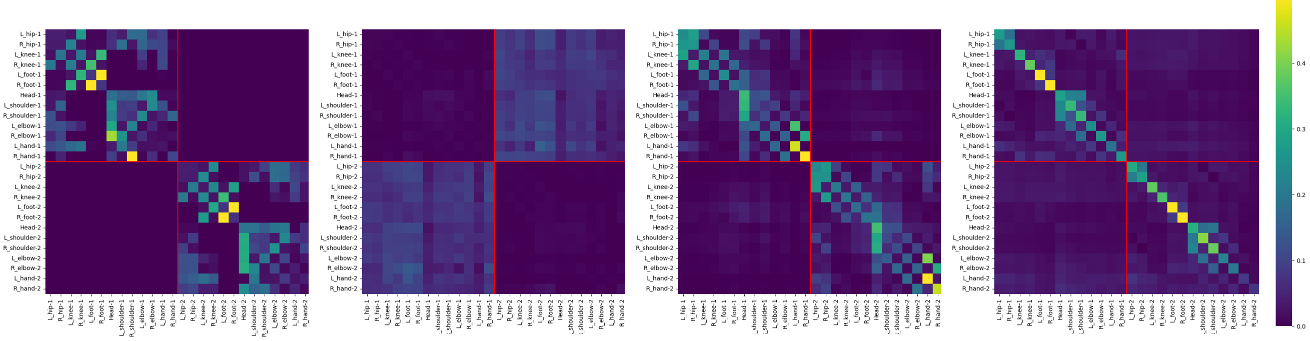
[3] Edward Vendrow, Satyajit Kumar, Ehsan Adeli, and Hamid Rezatofighi. Somoformer: Multi-person pose forecasting with transformers. *arXiv preprint arXiv:2208.14023*, 2022. 1, 2

[4] Jiashun Wang, Huazhe Xu, Medhini Narasimhan, and Xiaolong Wang. Multi-person 3d motion prediction with multi-range transformers. *Advances in Neural Information Processing Systems*, 34:6036–6049, 2021. 1, 2

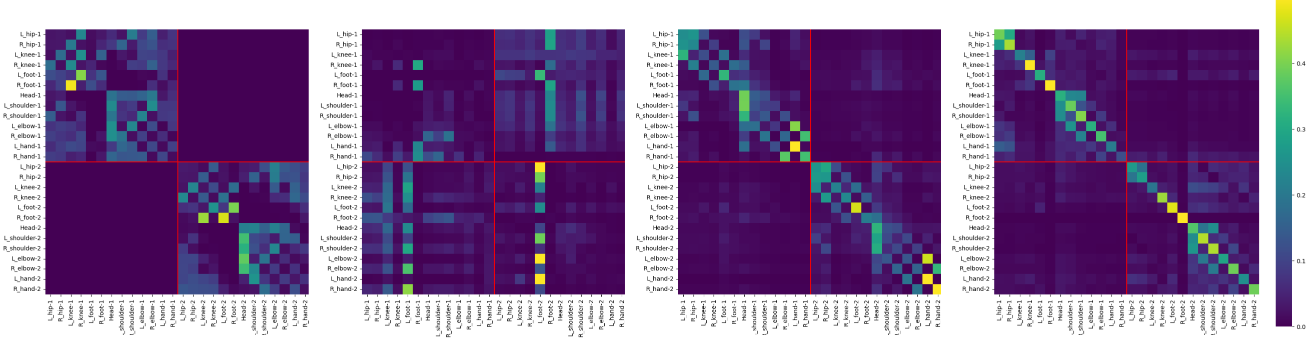
[5] Mao Wei, Liu Miaomiao, Salzemann Mathieu, and Li Hongdong. Learning trajectory dependencies for human motion prediction. In *ICCV*, 2019. 1, 2



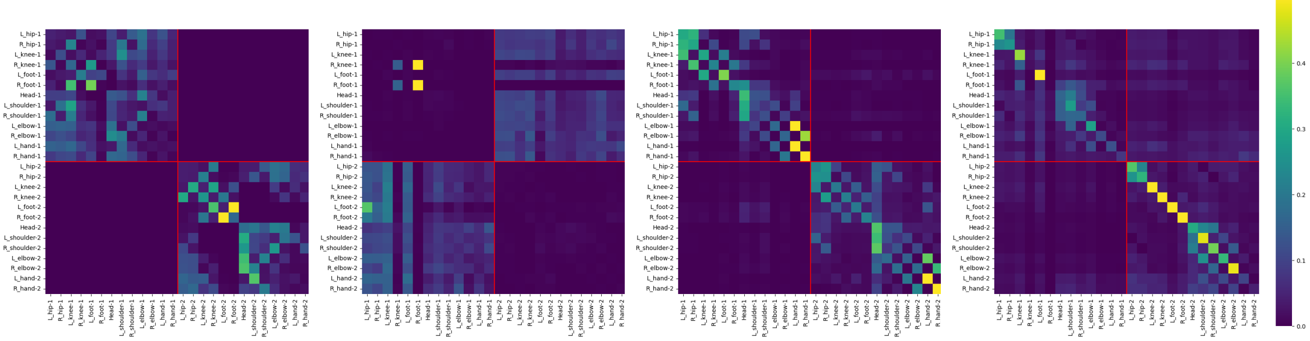
(a) dancing



(b) arguing

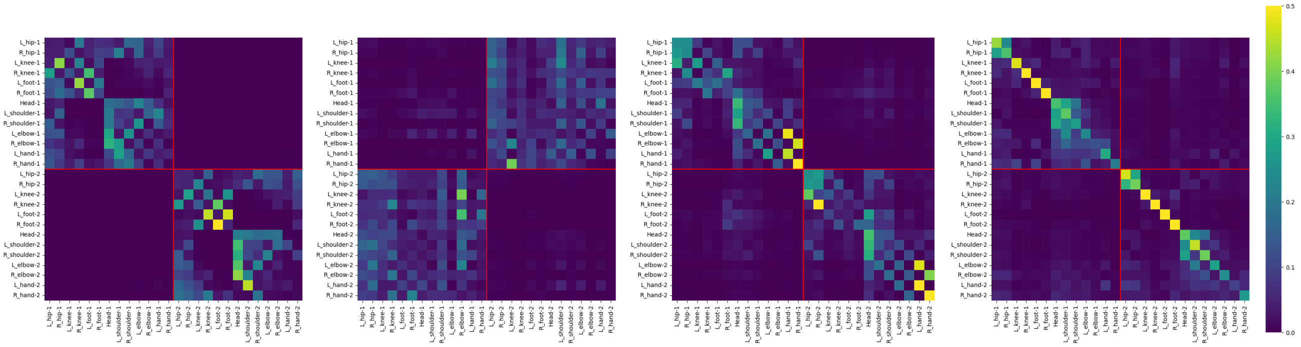


(c) basketball

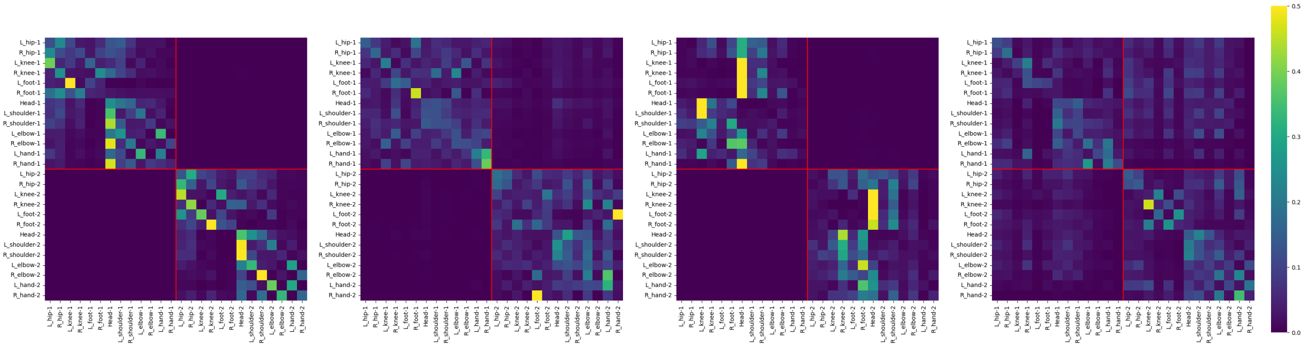


(d) capoeira

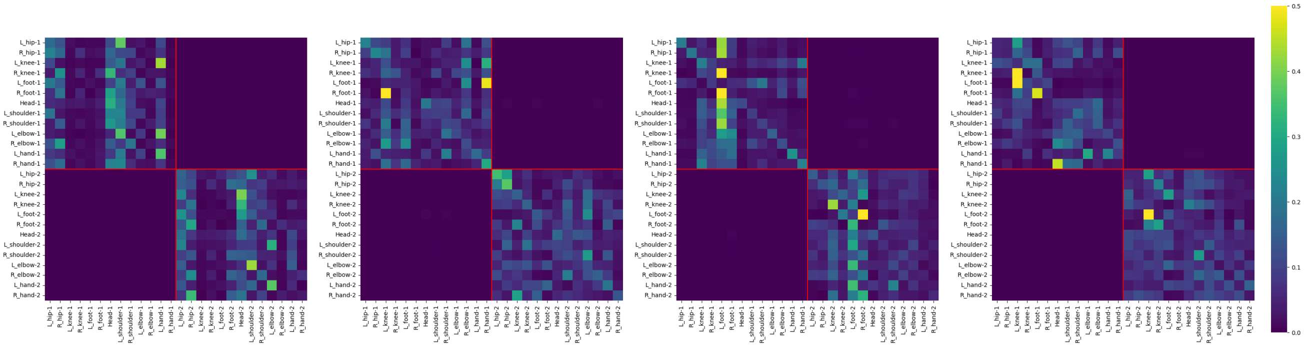
Figure 2. Attention visualization of different sequences.



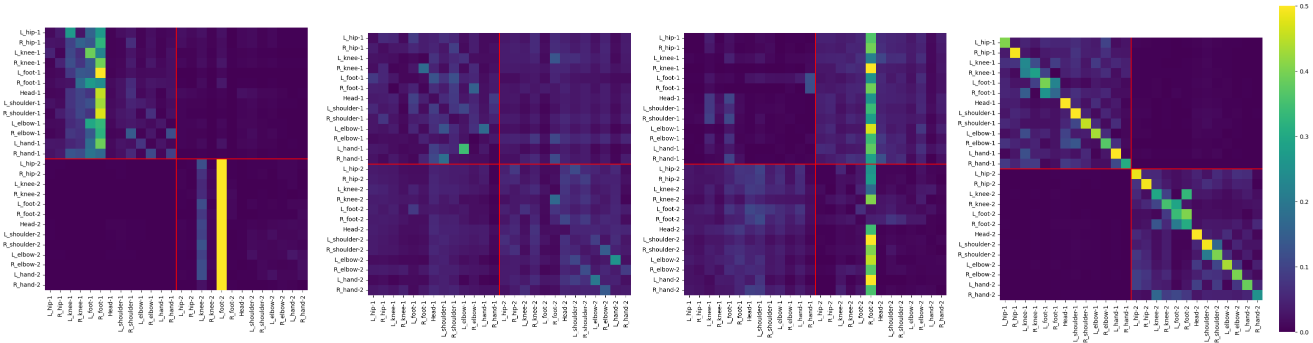
(a) Layer 1



(b) Layer 2



(c) Layer 3



(d) Layer 4

Figure 3. Attention visualization of different layers