# Spatio-Temporal Gating-Adjacency GCN for Human Motion Prediction -Supplementary Material

In this supplementary material, we will show more details about the experiment setting and demonstrate more experiment results. First, we will show the details of the used state-of-the-art dataset and implementation in Sec. 1. Then, we will analyze the fusion block of our method in Sec. 2. Finally, the more qualitative evaluation of our methods will be shown in Sec. 3, 4, 5.

## 1. More Details about Datasets and Implementation

### 1.1. Datasets

**Human 3.6M** Human 3.6M is the most used benchmark dataset in the field of motion prediction. Human 3.6M has 3.6 million 3D poses, consisting of 15 motion categories from 7 subjects(each character has 32 joint points), such as walking, eating, direction, discussion, etc. We downsample the frame rate to 25Hz. Following the previous works [1], we use subject 1,6,7,8,9 for training, subject 11 for validation, and the subject 5 for testing.

**AMASS** The Archive of Motion Capture as Surface Shapes(AMASS) dataset is a recently published human motion dataset, which gathers 18 existing mocap datasets, such as CMU, KIT, and BMLrub. AMASS uses SMPL to represent a human by a shape vector and joint rotation angles. Following [1], we apply forward kinematics to skeletons to obtain poses in 3D, and we discard the hand joints and the 4 static joints, leading to an 18-joint human pose. We downsample the frame rate to 25Hz as for Human 3.6M. Then, we select 8 datasets from AMASS for training, 4 datasets for validation, and 1 dataset(BMLrub) for testing.

**3DPW** The 3D Pose in the Wild dataset consists of both indoor and outdoor actions, which contains 51,000 frames captured at 30Hz. We only use 3DPW to test the generalization of the models trained on AMASS.

### 1.2. Implementation Details

The whole encoder consists of 6 GAGCN layers, each of which has residual connections and batch normalization layer. The input channels of them are 3, 32, 64, 64, 32, 3, respectively. All the gating networks are 3-layers fully connected network with 256, 64, $n$ or $m$ hidden units. The decoder consist of 4 TCN layers and each layer's kernel size is 3. We use Nvidia 3090 to train our network for a total 200 epochs. We use Adam Optimizer with an initial learning rate of 0.01 which decays by 10% every 20 epochs.The batch size is 128. According to the experimental settings of previous works, we take the past 10 frames as input to predict the future 25 frames.

## 2. Effect of Fusion Block

To demonstrate the effect of fusion block, we set up two contrast experiments against our method(shown in Table 1). The results show that using the Kronecker product to fuse spatio-temporal features is a better choice than summation and concatenation.

| | Human 3.6M-average | | | | | |
|---|---|---|---|---|---|---|
| Fusion | 80 | 160 | 320 | 400 | 560 | 1000 |
| $A_s \oplus A_t$ | 11.8 | 18.8 | 34.7 | 42.4 | 55.1 | 75.9 |
| $A_s \& A_t$ | 11.0 | 18.3 | 33.8 | 40.9 | 52.7 | 74.4 |
| $A_s \otimes A_t$ | **10.1** | **16.9** | **32.5** | **38.5** | **50.0** | **72.9** |

Table 1. Ablation study for the effect of fusion method. "$\oplus$" means to add the spatial and temporal features directly. "&" denotes that we concatenate the spatial and temporal features. "$\otimes$" indicates the Kronecker product.

## 3. Visualization of Predicted Sequence on AMASS

We visualize the predicted sequence on AMASS and compared them with Ground Truth in Fig. 1. Our predictions are almost identical to Ground Truth over the entire time horizons.

## 4. Visualization of Temporal Blending Coefficients

We randomly select 16 sequences from a single motion type to compute the average temporal blending coefficients. Then we do the same operation on several motion types and visualize them(seeing Fig. 2). We can see that there is a clear difference in the blending coefficients distribution for

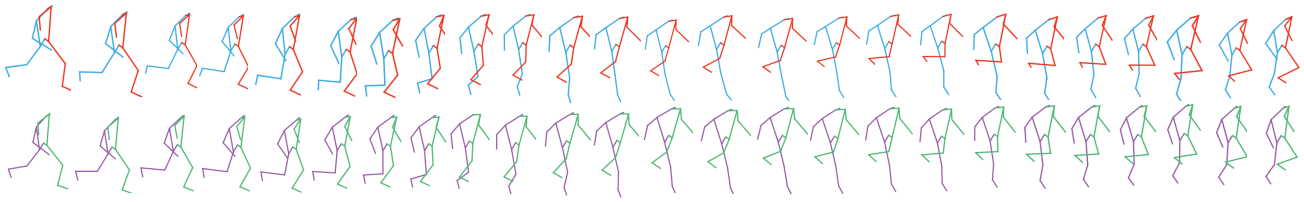


Figure 1. Visualization of predicted sequences against Ground Truth sequences on AMASS for all time horizons. The green and purple lines indicate prediction and the red and blue lines indicate the corresponding Ground Truth.

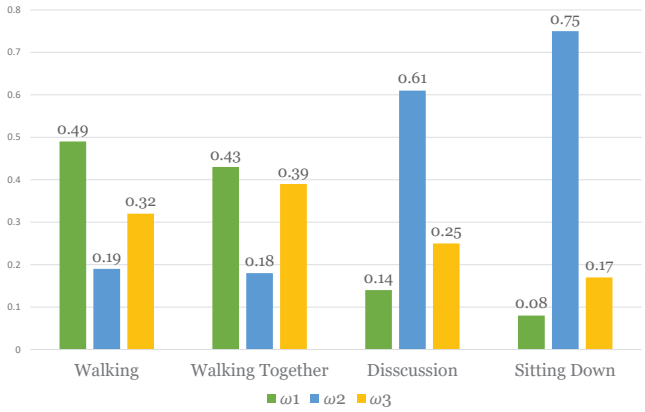
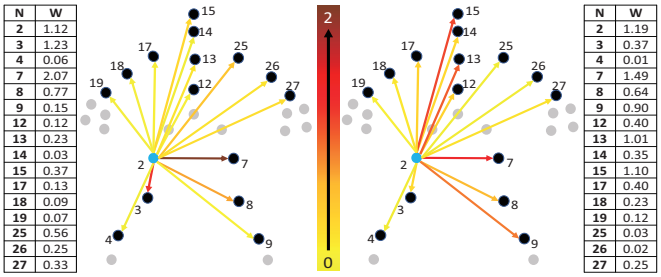



Figure 2. Visualization of average temporal blending coefficients for "Walking", "Walking Together", "Discussions", "Sitting Down". $\omega_1, \omega_2, \omega_3$ denote the 3 blending coefficients, respectively. Different action types(like "Walking", "Discussions", and "Sitting Down") have different coefficients distribution while the coefficients of similar motions are similarly distributed(like "Walking" and "Walking Together".

different motion types. Periodic motion types(like "Walking", "Walking Together") have high values of $\omega 1$ and $\omega 3$ while non-periodic motions(like "Discussions", "Sit-ting Down") have high values of $\omega 2$.

## 5. Visualization of Spatial and Temporal Adjacency Matrix

**Spatial Adjacency Matrix** We visualize the spatial adjacency of the "right knee" in "Walking"(left in Fig. 3) and "Posing"(right in Fig. 3). Given the different historical sequences(such as "Walking" and "Posing"), the spatial adjacency matrices are adaptive, i.e. different weights distributions. The weights between node 2 and 3 are relatively high because they have skeletal connections, and he weights between node 2 and 7 are relatively high because they are symmetric in the skeleton. For "Walking" in the left, the weights between the left hand(25,26,27) and right knee(2) is higher than the weight between the right hand(17,18,19) and right



| N | W |
|---|---|
| 2 | 1.12 |
| 3 | 1.23 |
| 4 | 0.06 |
| 7 | 2.07 |
| 8 | 0.77 |
| 9 | 0.15 |
| 12 | 0.12 |
| 13 | 0.23 |
| 14 | 0.03 |
| 15 | 0.37 |
| 17 | 0.13 |
| 18 | 0.09 |
| 19 | 0.07 |
| 25 | 0.56 |
| 26 | 0.25 |
| 27 | 0.33 |



| N | W |
|---|---|
| 2 | 1.19 |
| 3 | 0.37 |
| 4 | 0.01 |
| 7 | 1.49 |
| 8 | 0.64 |
| 9 | 0.90 |
| 12 | 0.40 |
| 13 | 1.01 |
| 14 | 0.35 |
| 15 | 1.10 |
| 17 | 0.40 |
| 18 | 0.23 |
| 19 | 0.12 |
| 25 | 0.03 |
| 26 | 0.02 |
| 27 | 0.25 |

Figure 3. Visualization of spatial adjacency of the chosen reference joint in "Walking"(left) and "Posing"(right). The dots represent the nodes of the skeleton in Human 3.6M, the lines indicate the weights of the two joints, i.e. their dependencies. The blue dots represent our chosen reference joint, i.e. the "right knee". The black dots represent selected visual joints. The gray dots represent other nodes. We use the gradient of color to represent the weight value, while the first column in the lists on both sides represents the number of the selected visual joints, and the second column represents the weight of their connection to the reference joint.

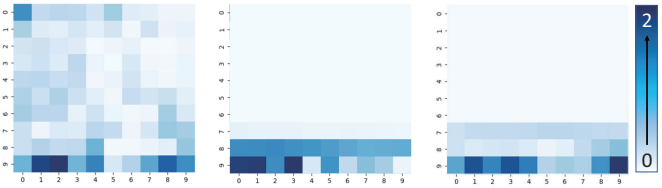



Figure 4. Visualization of Temporal Adjacency Matrix.We choose joint node "right knee" in "Walking" action type to demonstrate the temporal adjacency matrix. From left to right are the temporal adjacency matrices of the first, third and sixth GAGCN layers respectively.

knee, because the left hand and right knee move at the same time during walking. For "Posing", the movement amplitude of the head(13,14,15) and the right hand(17,18,19) are larger(seeing Figure 3 in the submission), which can better express the characteristics of "Posing". Thus the weights between the head and right knee is higher, and the weights between the right hand and right knee is higher than the

weights between the left hand and right knee. These results demonstrate that the spatial adjacency matrix is not only adaptive, but also conforms to the natural law of human motion.

**Temporal Adjacency Matrix** The temporal adjacency matrix has been visualized in Fig. 4. The heat map represents the weight between the input 10 frames of joint node "right knee" in "Walking" action type. From left to right are the temporal adjacency matrices of the first, third, and sixth GAGCN layers respectively. In the first layer, the weights are throughout the graph. As the number of layers increases, the high weight values gradually converge towards the last few frames(the last few rows in the figure). Indicating that the prediction effect highly relies on the last few frames of the historical sequence, which is also in line with our intuition of motion prediction.

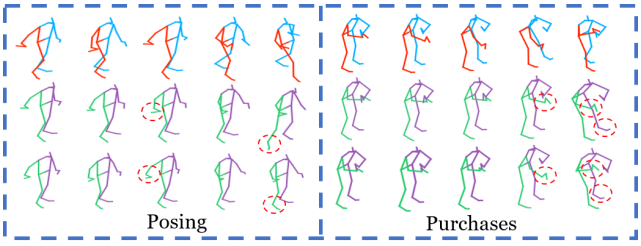



Figure 5. Prediction visualizations of the two most challenging motion types to predict in Human 3.6M: "Posing" and "Purchase". From top to bottom are Ground Truth, STSGCN's, and our results, respectively. In the long term prediction (the last three frames), our prediction results are closer to GT, especially in the places circled in the red dotted line.

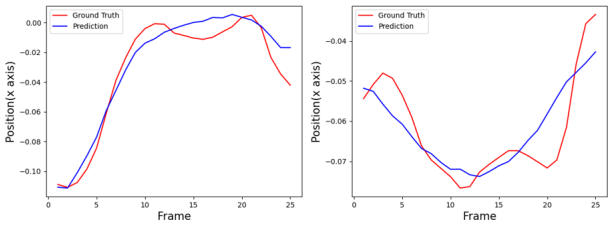



Figure 6. The predicted(blue) and Ground Truth(red) joint position curves of left hip(left) and left wrist(right) on "Walking" motion. From the figure, we can see that the smoothness of our prediction results is the same as that of Ground Truth.

## 6. Qualitative Comparison

Because the results of our work and STSGCN's are clearly superior to other work used for comparison, here we only show the **qualitative comparison** with STSGCN [2]. The visualizations are shown in Fig. 5. As for the **smoothness** of the predicted motion, we use the trajectory curve of the joint nodes to illustrate it, shown in Fig 6.

## References

[1] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*, pages 474–489. Springer, 2020. 1

[2] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-time-separable graph convolutional network for pose forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11209–11218, 2021. 3