Learning Constrained Dynamic Correlations in Spatiotemporal Graphs for Motion Prediction — Supplementary Material —

Jiajun Fu, Fuxing Yang, Yonghao Dang, Xiaoli Liu and Jianqin Yin*

***Abstract*—This supplementary material provides network details, an evaluation of matrix decomposition, extra experiments, more visualization results on three benchmark datasets, and a symbol table.**

I. Network Details

TABLE I
THE DETAILED ARCHITECTURE OF DSTD-GCN. J IS SET TO 22 ON THE HUMAN3.6M DATASET, 25 ON THE CMU MOCAP DATASET, AND 23 ON THE 3DPW DATASET. T IS SET TO 35. C_{in} IS SET TO 6, C_{out} IS SET TO 3, AND C_m IS SET TO 64. IN THE "PARAMETERS", THE VALUE IN THE BRACKET AFTER EACH PARAMETER INDICATES THE PARAMETER'S SIZE.

Layers	Operations	Parameters	Output Sizes
Encoding Basic Conv	Parallel DS-GCs	$\boldsymbol{A}^s(J\times J), \boldsymbol{W}(C_{in}\times C_m)$	$J\times T\times C_m$
	DT-GC	$\boldsymbol{A}^t(T\times T), \boldsymbol{W}(C_m\times C_m)$	$J\times T\times C_m$
Basic Block $\times 5$	Parallel DS-GC	$\boldsymbol{A}^s(J\times J), \boldsymbol{W}(C_m\times C_m)$	$J\times T\times C_m$
	DT-GC	$\boldsymbol{A}^t(T\times T), \boldsymbol{W}(C_m\times C_m)$	$J\times T\times C_m$
Decoding Basic Conv	Parallel DS-GCs	$\boldsymbol{A}^s(J\times J), \boldsymbol{W}(C_m\times C_m)$	$J\times T\times C_m$
	DT-GC	$\boldsymbol{A}^t(T\times T), \boldsymbol{W}(C_m\times C_{out})$	$J\times T\times C_{out}$

The details of our proposed DSTD-GCN are illustrated in Table I. Our model contains an encoding Basic Conv unit, five Basic blocks, and a decoding Basic Conv unit. The Basic block contains one Basic Conv unit. All Basic Conv units have the same structure. The input spatial information consists of the original three-dimensional coordinates and offsets positions from the last input pose.

II. More Experiment Results

A. Correlation Decomposition in SpatioTemporal Decompose Graph Convolution

In this section, we show that SpatioTemporal Decompose Graph Convolution is with full-rank decomposition on the three benchmark datasets. First, we derive the full-rank decomposition theorem for spatiotemporal-unshared graph convolutions. Then, we show that the Spatiotemporal Decompose Graph Convolution, as a specific spatiotemporal-unshared graph convolution, satisfies the theorem on the three benchmark datasets.

We check the rank property for the matrix decomposition. As the STD-GC belongs to spatiotemporal-unshared GC, we analyze the rank property based on the formulation of the spatiotemporal-unshared GC. We reformulate the spatiotemporal-unshared GCs with respect to the spatiotemporal adjacency matrix product. Recall that a human motion sequence can be represented as a graph $\mathcal{G}^{st}$. The vertex set $\mathcal{V}^{st}$ contains $J\times T$ nodes. Besides, the spatiotemporal correlation is encoded in an adjacency matrix $\boldsymbol{A}^{st}\in\mathbb{R}^{JT\times JT}$, where the index for joint p of frame m is $p*T+m$. We denote the spatial and temporal correlations in spatiotemporal-unshared GCs as $\boldsymbol{A}^s\in\mathbb{R}^{T\times J\times J}$ and $\boldsymbol{A}^t\in\mathbb{R}^{J\times T\times T}$, respectively. Then, as shown in Fig. 1, spatiotemporal-unshared GCs decompose spatiotemporal adjacency matrix into the following:

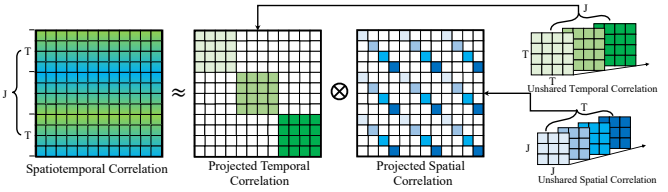

Fig. 1. Spatiotemporal correlation decomposition based on matrix product. We project the spatial and temporal correlations to $\mathbb{R}^{JT\times JT}$ to formulate a matrix product. The brightness of the color indicates the data source of the spatial and temporal correlations. $\otimes$ is matrix multiplication

$$\boldsymbol{A}^{st}\approx\boldsymbol{A}^{t(p)}\otimes\boldsymbol{A}^{s(p)}, \tag{1}$$

where $\boldsymbol{A}^{s(p)}$, $\boldsymbol{A}^{t(p)}$ is the projection of $\boldsymbol{A}^s$, $\boldsymbol{A}^t$ in $\mathbb{R}^{JT\times JT}$, and $\otimes$ is matrix multiplication. The projection[1] is defined as:

$$\begin{aligned} a^s_{mpq}&\rightarrow a^{s(p)}_{((p*T+m)(q*T+m)},\\ a^t_{pmn}&\rightarrow a^{t(p)}_{(p*T+m)(p*T+n)}. \end{aligned} \tag{2}$$

Based on the matrix product formulation, we find that the rank of spatiotemporal-unshared GCs is determined by the rank of the spatial and temporal adjacency matrices. Formally, we can derive that Eq. 1 is a full rank decomposition if and only if $\boldsymbol{A}^{s(p)}$ and $\boldsymbol{A}^{t(p)}$ are full rank matrices. This is equivalent to the situation that T vanilla spatial adjacency matrices in $\boldsymbol{A}^s$ and J vanilla temporal adjacency matrices in $\boldsymbol{A}^t$ are with full rank.

Theorem 1. *A Spatiotemporal-unshared GC decomposes the spatiotemporal correlation with full rank if and only if the T vanilla spatial adjacency matrices in $\boldsymbol{A}^s$ and J vanilla temporal adjacency matrices in $\boldsymbol{A}^t$ are all with full rank.*

With this theorem, we now evaluate the rank property of the STD-GC on the three benchmark datasets (Human3.6M,

* Jianqin Yin is the corresponding author.
All authors are with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China. (email: jqyin@bupt.edu.cn)
Digital Object Identifier 10.1109/TNNLS.2023.3277476

[1]Items with no projection are set as 0 in the full matrix.

TABLE II
COMPARISON BETWEEN ST-GC AND STD-GC ON HUMAN3.6M, CMU MOCAP, AND 3DPW DATASETS

GC	Human3.6M									CMU Mocap									3DPW								
	80	160	320	400	560	1000	Average	Params.	Full rank	80	160	320	400	560	1000	Average	Params.	Full rank	200	400	600	800	1000	Average	Params.	Full rank
ST-GC	11.87	25.45	51.62	62.80	80.83	113.31	57.64	4.23M	✓	8.70	16.16	31.34	39.16	53.91	85.30	39.09	5.44M	✓	26.78	53.26	76.08	93.96	106.13	71.25	6.01M	✓
STD-GC	11.79	25.60	52.17	63.35	81.23	113.78	57.99	0.39M	✓	8.14	15.45	30.71	38.51	53.75	86.10	38.78	0.45M	✓	25.58	51.76	75.00	94.07	108.16	70.92	0.49M	✓

CMU Mocap, and 3DPW). We use the experiment settings in the "GC Comparison" part (Sect. IV-C of the manuscript). We set the Basic Conv unit as ST-GC and STD-GC, respectively, and compare their performance. The results are shown as Table II. ST-GC and STD-GC achieve similar performance. Here, the spatiotemporal adjacency matrices in ST-GC are all with full rank. Besides, all the spatial and temporal adjacency matrices are with full rank. According to Theorem 1, STD-GC decomposes spatiotemporal correlation with full rank and doesn't suffer the low-rank constraint problem. In addition to the rank property, STD-GC achieves similar performance to ST-GC.

B. Results of the "Running" Scenario on CMU Mocap Dataset

We report the prediction error of the "running" scenario on CMU Mocap dataset. As shown in Table III, our model achieves the best or second-best performance.

TABLE III
COMPARISON OF "RUNNING" PREDICTIONS ON CMU MOCAP DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, WHILE THE SECOND BEST RESULTS ARE SHOWN IN UNDERLINE.

Action	Running					
Millisecond	80	160	320	400	560	1000
Residual sup. [1]	23.46	39.94	62.26	67.46	73.21	78.17
DMGNN [2]	17.55	30.09	45.74	49.34	53.65	81.60
FC-GCN [3]	15.03	25.21	39.05	42.36	43.28	53.56
Traj-CNN [4]	14.39	23.41	37.51	39.51	40.68	51.38
STS-GCN [5]	13.26	21.84	33.98	37.71	40.95	44.76
MSR-GCN [6]	13.17	**20.91**	29.87	33.35	38.22	43.57
Ours	**12.04**	20.97	**28.10**	**33.31**	**37.16**	**43.55**

C. Comparison of Different STS-GCN Implementations

We compare our implementation with the official one[2] under our settings on CMU Mocap dataset. The results are shown in Table IV. There are two differences between the STS-GCN [5] in this paper and their official implementation.

The first difference is the experiment setting. STS-GCN adopted a new error calculation in their paper [5], which is different from the mainstream (our setting). [1], [3], [4], [6]–[8]. We directly present the instant errors for a frame at a specific frame while STS-GCN calculates the average MPJPE across all frames before this frame. Take the error calculation on the CMU Mocap dataset as an example, we calculate the error of the 25-th frame as the error of the "1000ms", while the paper calculates the average error of the first 25 frames. With the average error calculation, the paper reported a substantially lower error than our settings' results.

[2]https://github.com/FraLuca/STSGCN

The second difference is the model architecture: the official implementation adopts an encoder-decoder architecture to directly generate the results from input representation. In contrast, our implementation utilizes the prediction framework, which gradually recovers output motion from a concatenation of input and padding output. As shown in Table IV, our implementation achieves better performance than the official one. This reflects two characteristics of the prediction framework: Firstly, the prediction framework can effectively model complex interactions between the input and output sequence. The encoder-decoder architecture separately models the input and output of the motion sequence, where only high-level information is exchanged. Since features from different levels abstract the sequence at multiple scales, cross-talk between input and output at multiple levels can complete the motion representation. In the prediction framework, features from the input and output implicitly interact at multiple levels for strong feature representation. Secondly, the complete sequence modeling can obtain global context information and generate results with more temporal consistency. Therefore, the short-term prediction errors are large for the official implementation. For clarity, we use our implementation in this paper.

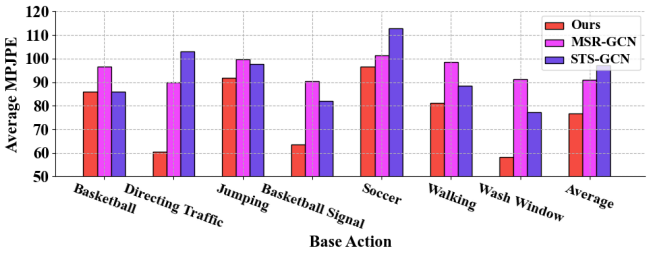

Fig. 2. Comparison of unseen action prediction.

D. Unseen Action Prediction

To further understand the effectiveness of our proposed model, we conduct experiments on unseen action data. On each of the seven actions in the CMU Mocap dataset, we train STS-GCN [5], MSR-GCN [6], and our model. Then, we test these models on all actions. Our method outperforms STS-GCN and MSR-GCN with lower MPJPE. Fig. 2 shows the comparison results. Our method outperforms two state-of-the-art models by a large margin. It illustrates that our DSTD-GCN can generate sample-specific spatiotemporal correlations that may be used to improve human motion prediction in response to the individualized motions associated with various actions. On the other hand, the model's prediction error is 35.93 when trained with full action data. Although our model outperforms the two baselines in unseen action data by a large margin, there is still a considerable gap in the performance under full action settings. We still have much room to adapt the model for unseen human motions.

TABLE IV
COMPARISONS OF PREDICTION RESULTS OF DIFFERENT STS-GCN IMPLEMENTATION ON CMU MOCAP DATASET

Action	Basketball						Basketball Signal						Directing Traffic						Jumping					
Millisecond	80	160	320	400	560	1000	80	160	320	400	560	1000	80	160	320	400	560	1000	80	160	320	400	560	1000
Ours	10.23	18.67	36.93	45.98	61.19	91.36	2.96	5.52	12.12	16.12	25.15	50.88	5.95	11.99	27.55	36.75	57.05	111.53	15.66	30.63	59.13	71.87	93.32	125.94
Official	21.16	28.79	48.55	57.51	72.42	97.77	10.91	14.69	26.71	34.16	49.97	86.89	12.49	18.15	36.37	47.12	69.77	121.31	34.85	47.12	69.77	90.81	112.12	136.24

Action	Soccer						Walking						Wash Window						Average					
Millisecond	80	160	320	400	560	1000	80	160	320	400	560	1000	80	160	320	400	560	1000	80	160	320	400	560	1000
Ours	11.30	20.45	39.04	48.88	69.12	102.54	6.87	11.29	18.13	21.06	26.12	37.86	5.44	10.84	23.90	30.72	44.00	71.42	8.33	15.62	30.97	38.77	53.70	84.50
Official	21.09	29.36	52.93	64.64	86.74	127.03	11.30	13.47	18.68	22.90	27.72	39.96	12.50	17.05	31.37	38.80	51.95	77.54	17.75	24.09	41.65	52.47	68.03	97.93

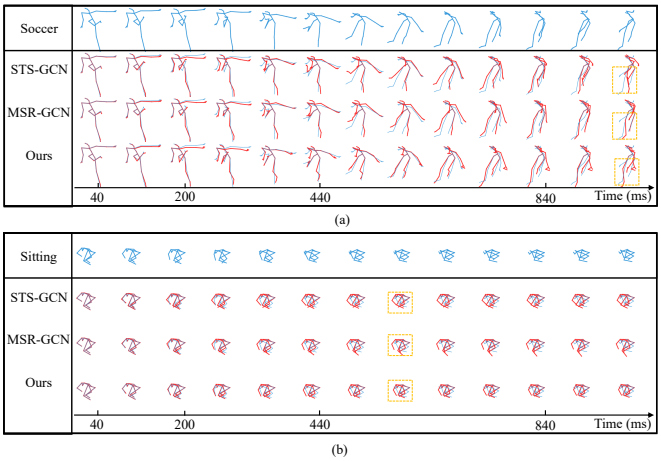

Fig. 3. Visualization of failure cases. The blue and red poses indicate ground truths and predictions, respectively. The orange boxes indicate inaccurate predictions.

III. VISUALIZATION RESULTS

A. Failure Cases

Although our method successfully models the constrained dynamic correlations in human motion, there are still some failure cases in extreme motions, and rare human poses. Specifically, we present some failure predictions of our method, STS-GCN, and MSR-GCN, in Fig. 3. On the one hand, the human motions in the "soccer" sequence are extreme, and all models fail to predict the accurate movement of the arms and feet. On the other hand, all models cannot accurately predict the body movement in the "sitting" sequence since the crouching pose is rare across the dataset. Although our model's predictions are closer to the ground truth, we have much for further improvements.

B. Visual Comparison of Predicted Sequences

In this section, we provide more visual comparisons of different models over three benchmark datasets. The visualizations are shown in Fig. 4, 5, and 6 (The figures are in the following two pages). We compare our model with STS-GCN [5], and MSR-GCN [6]. In each of these examples, our model's predictions are more similar to the ground truth.

IV. SYMBOL TABLE

We present a table for most of the symbols in the paper, as shown in Table V.

TABLE V
THE SYMBOLS AND THEIR MEANINGS IN THIS PAPER.

Symbol	Meaning
j, p, q, q_1, q_2	Joint indices.
t, m, n, n_1, n_2	Frame/Time indices.
i, i_1, i_2	Sample indices.
$\boldsymbol{A}^{st}$	A spatiotemporal adjacency matrix in $\mathbb{R}^{JT\times JT}$
$\boldsymbol{A}^{s}$	A spatial adjacency matrix from $\mathbb{R}^{T\times J\times J}$ or $\mathbb{R}^{J\times J}$
$\boldsymbol{A}^{t}$	A spatial adjacency matrix from $\mathbb{R}^{J\times T\times T}$ or $\mathbb{R}^{T\times T}$
$a^{st}_{(pm)(qn)}$ $a^{st(i)}_{(pm)(qn)}$	The correlation strength from joint p of frame m to joint q of frame n in $\boldsymbol{A}^{st}$ The superscript i indicates that the strength will change for different samples.
a^{s}_{npq} $a^{s(i)}_{npq}$	The correlation strength in frame n from joint p to joint q in $\boldsymbol{A}^{s}$ from $\mathbb{R}^{T\times J\times J}$. The superscript i indicates that the strength will change for different samples.
a^{s}_{pq}	The correlation strength from joint p to joint q in $\boldsymbol{A}^{s}$ from $\mathbb{R}^{J\times J}$.
a^{t}_{qmn} $a^{t(i)}_{qmn}$	The correlation strength of joint q from frame m to frame n in $\boldsymbol{A}^{t}$ from $\mathbb{R}^{J\times T\times T}$. The superscript i indicates that the strength will change for different samples.
a^{t}_{mn}	The correlation strength from frame m to frame n in $\boldsymbol{A}^{t}$ from $\mathbb{R}^{T\times T}$.
$\boldsymbol{X}, \boldsymbol{Y}$	Feature tensors from $\mathbb{R}^{J\times T\times C}$ for the entire motion sequence.
$\boldsymbol{x}_{qn}, \boldsymbol{y}_{qn}$	Feature vectors for joint q of frame n.
$\mathcal{M}^s, \mathcal{A}^s, \mathcal{U}^s$	Functions in the Dynamic Spatial Graph Convolution.
$\boldsymbol{M}^s, \boldsymbol{A}^s, \boldsymbol{F}^s$	Feature tensors in the Dynamic Spatial Graph Convolution.
$\mathcal{M}^t, \mathcal{A}^t, \mathcal{U}^t$	Functions in the Dynamic Temporal Graph Convolution.
$\boldsymbol{M}^t, \boldsymbol{A}^t, \boldsymbol{F}^t$	Feature tensors in the Dynamic Temporal Graph Convolution.
$\boldsymbol{W}, \boldsymbol{W}_1, \boldsymbol{W}_2$	Matrices for feature transformation.

REFERENCES

[1] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2891–2900.
[2] M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, and Q. Tian, "Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 214–223.
[3] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9489–9497.
[4] X. Liu, J. Yin, J. Liu, P. Ding, J. Liu, and H. Liu, "Trajectorycnn: A new spatio-temporal feature learning network for human motion prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 6, pp. 2133–2146, 2021.
[5] T. Sofianos, A. Sampieri, L. Franco, and F. Galasso, "Space-time-separable graph convolutional network for pose forecasting," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 11 209–11 218.
[6] L. Dang, Y. Nie, C. Long, Q. Zhang, and G. Li, "Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 11 467–11 476.
[7] Q. Cui, H. Sun, and F. Yang, "Learning dynamic relationships for 3d human motion prediction," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2020, pp. 6519–6527.
[8] Z. Liu, P. Su, S. Wu, X. Shen, H. Chen, Y. Hao, and M. Wang, "Motion prediction using trajectory cues," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 13 299–13 308.

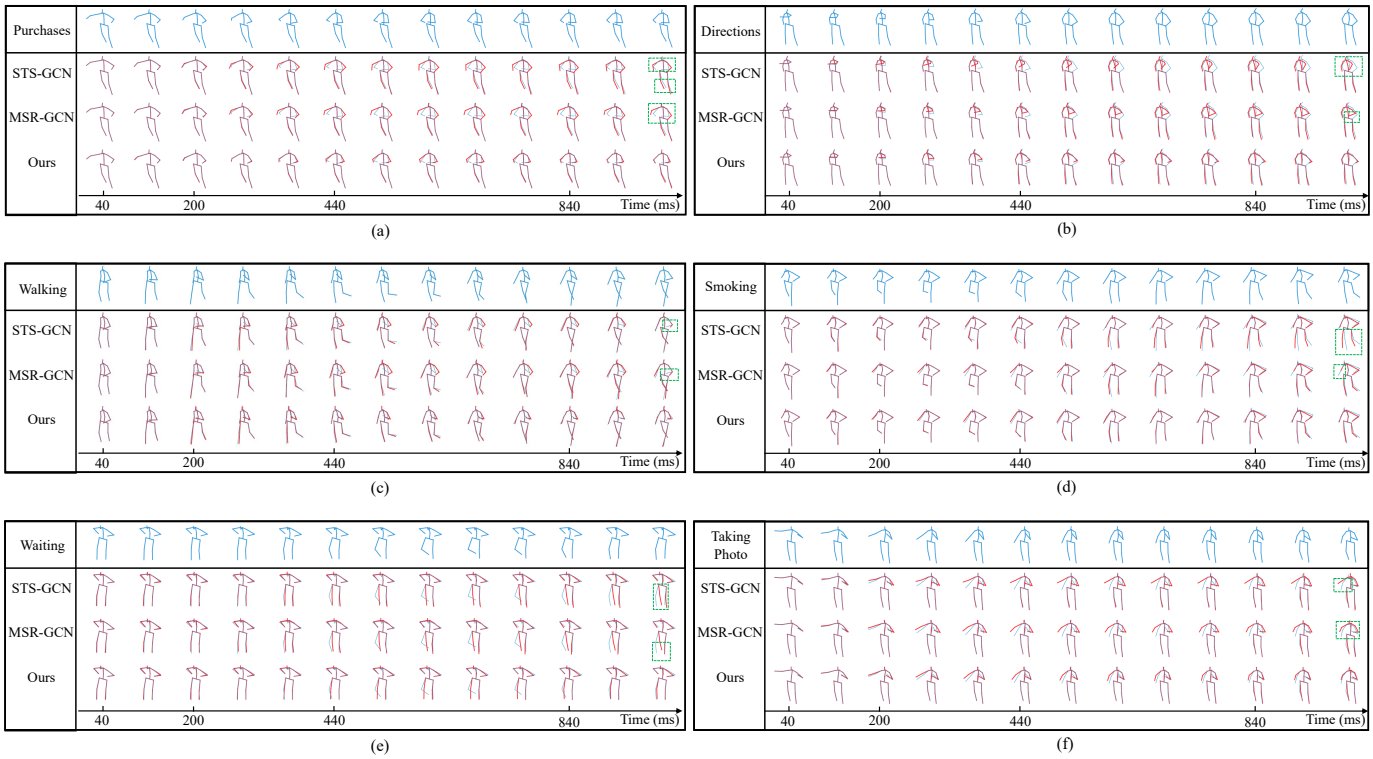

Fig. 4. Visualization of prediction examples on the Human3.6M dataset. The blue and red poses indicate ground truths and predictions, respectively. (a) Purchases. (b) Directions. (c) Walking. (d) Smoking. (e) Waiting. (f) Taking Photo. The green boxes indicate our improvements.

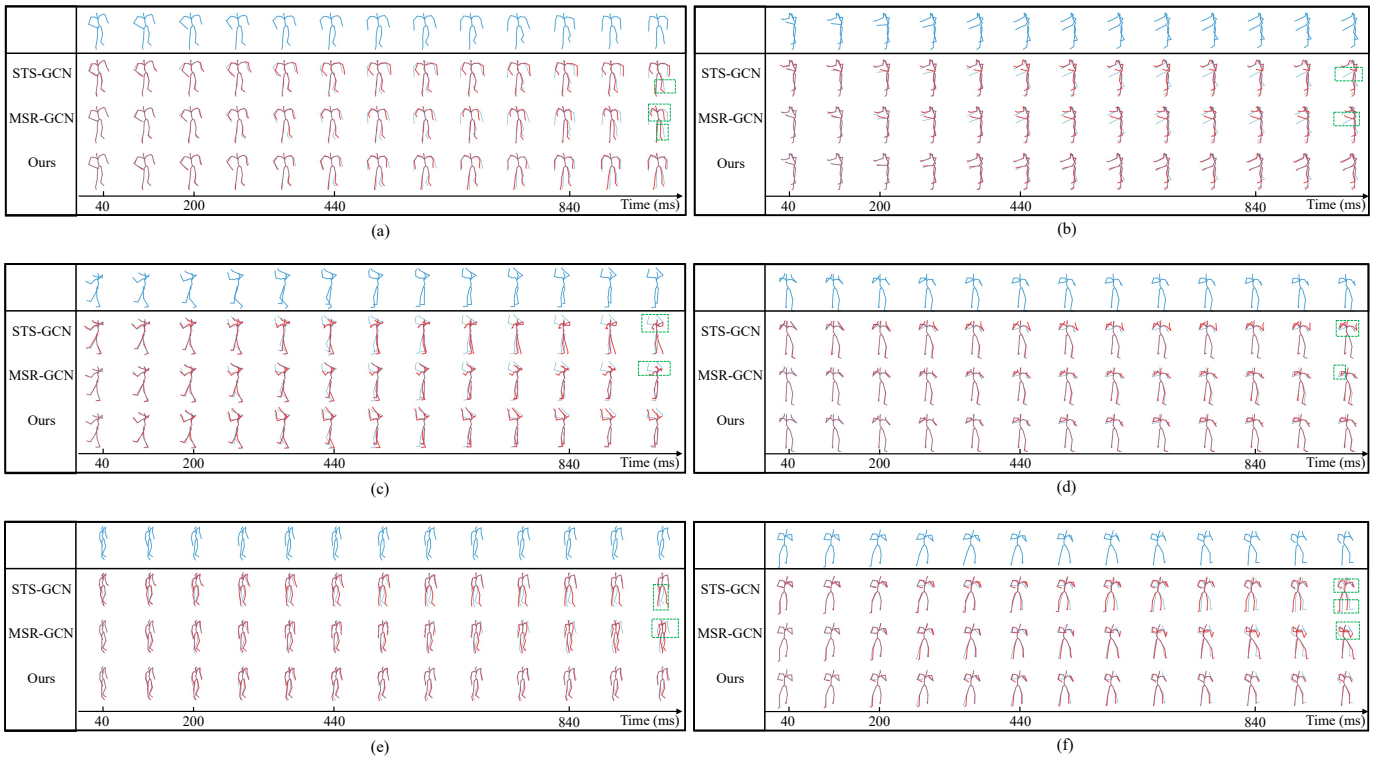

Fig. 5. Visualization of prediction examples on the 3DPW dataset. The blue and red poses indicate ground truths and predictions, respectively. The green boxes indicate our improvements.

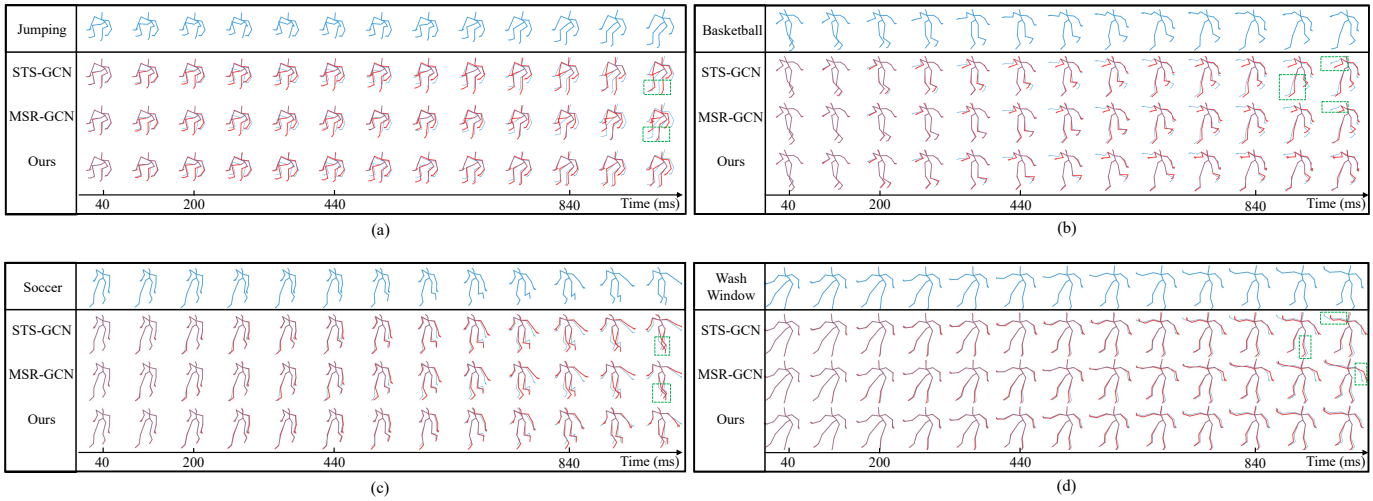

Fig. 6. Visualization of prediction examples on the CMU-Mocap dataset. The blue and red poses indicate ground truths and predictions, respectively. (a) Jumping. (b) Basketball. (c) Soccer. (d) Wash Window. The green boxes indicate our improvements.