Supplementary Material

Anonymous ICME submission

I. Analysis of the Absence of Skeletons

Table I presents the results under the condition of missing skeleton data, where only half of the joints are available for motion sequence prediction. This scenario is particularly challenging because it significantly limits the information available for modeling the spatial and temporal dependencies that are crucial for accurate human motion prediction. AvgDescend stands for Performance Degradation, indicating the extent to which each method's performance declines under these harsh conditions.

Previous methods have shown considerable difficulties in maintaining accuracy when confronted with such incomplete data, often resulting in substantial performance degradation. However, our proposed method exhibits a remarkably lower performance degradation compared to these previous approaches. Despite the severe reduction in available joint data, our model successfully minimizes the impact of missing skeletons, preserving a high level of predictive accuracy. This indicates that our approach is capable of effectively compensating for missing information through its robust feature extraction and trend learning mechanisms.

Specifically, the minimal decline in performance observed with our method highlights its superior robustness, as it can still capture essential spatial-temporal correlations even when the skeletal information is incomplete. This robustness is crucial for real-world applications where sensor failures, occlusions, or other issues might lead to incomplete motion data. Our method's ability to maintain stable and accurate predictions under such challenging conditions demonstrates its practicality and resilience in real-world scenarios.

TABLE I
Prediction with missing skeletal joints on CMU-Mocap.

milliseconds	80	160	320	400	AvgDescend ↓
Res-sup [1]	32.1	51.5	93.1	113.4	23.4%
STSGCN [2]	12.4	22.1	40.9	55.5	29.7%
MSRGCN [3]	10.9	20.2	38.6	53.4	33.0%
PGBIG [4]	9.9	18.3	35.6	50.2	25.8%
SPGSN [5]	10.6	17.7	34.5	47.4	24.2%
Ours	**8.2**	**14.8**	**30.1**	**40.3**	**13.8%**

II. Analysis of the Impact of the Skeleton Enhancement Coefficient

We conduct a detailed investigation into the impact of various enhancement coefficients on model training, as illustrated in Figure 1. The results clearly demonstrate that both excessively high and excessively low coefficients adversely affect model performance. When the enhancement coefficient is set too high, the addition of numerous virtual features tends to introduce noise and overfitting, thereby hindering the model's ability to generalize effectively. Conversely, a coefficient that is too low fails to provide sufficient data feature richness, limiting the diversity and depth of the information available for the model to learn. This imbalance highlights the importance of selecting an optimal enhancement coefficient to ensure robust and effective training.

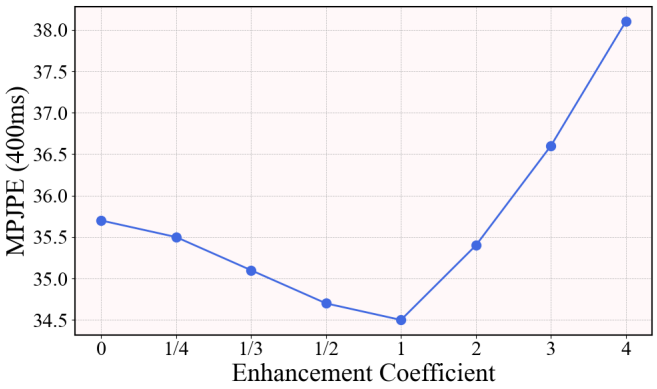

Fig. 1. Effect of different enhancement coefficients in the model training on CMU-Mocap.

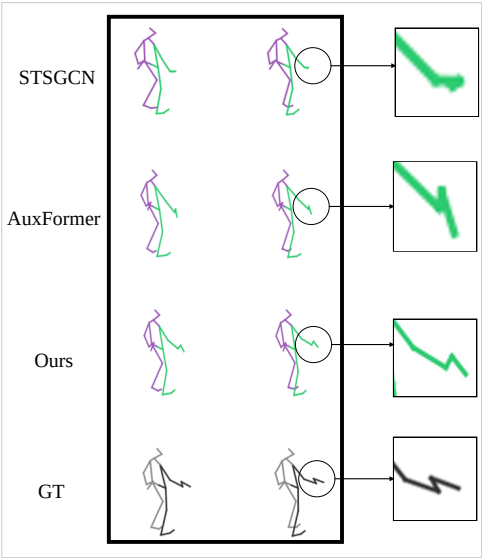

Fig. 2. The magnified view of the yellow box shows that the predicted result captures a more realistic bend of the arm and models the hand with greater precision. The visualization at the bottom represents the real situation.

III. More visualization results of Human3.6M

To begin, we present a more detailed feature analysis and compare the visualization results. As shown in Figure 2, our method predicts more accurate movements, with the characteristics of the limb closer to the truth of the ground. In contrast, other methods exhibit significant deviations in hand orientation and degree of arm bending.

The visualization of prediction results in this section demonstrates the proposed model's capabilities. The result is shown in the Figure 3 – 6 (next page). It can be observed that most actions exhibit significant dynamic variations rather than gradually converging toward an average posture. This indicates that the model is effective in capturing and representing the complex spatial-temporal dynamics of the actions.

References

[1] Julieta Martinez, Michael J Black, and Javier 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] 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*, 2021, pp. 11209–11218.

[3] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li, "Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11467–11476.

[4] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li, "Progressively generating better initial guesses towards next stages for high-quality human motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[5] Maosen Li, Siheng Chen, Zijing Zhang, Lingxi Xie, Qi Tian, and Ya Zhang, "Skeleton-parted graph scattering networks for 3d human motion prediction," in *European conference on computer vision*. Springer, 2022, pp. 18–36.

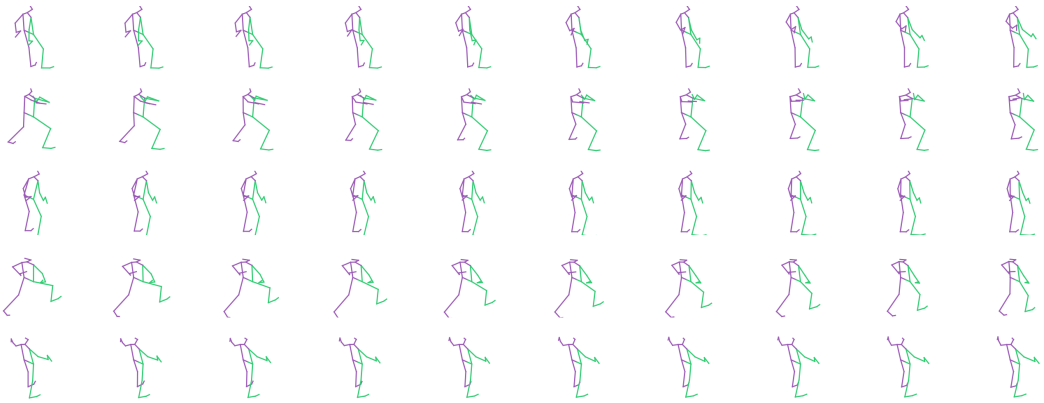

Fig. 3. The action of posing.

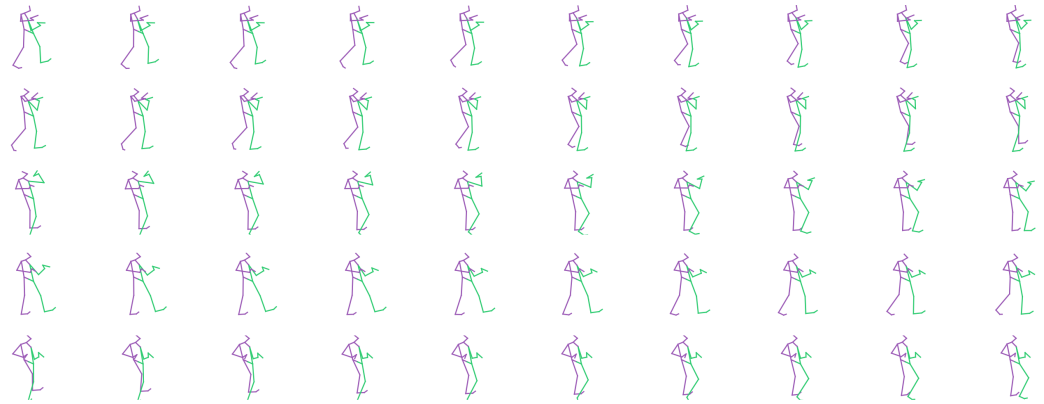

Fig. 4. The action of eating.

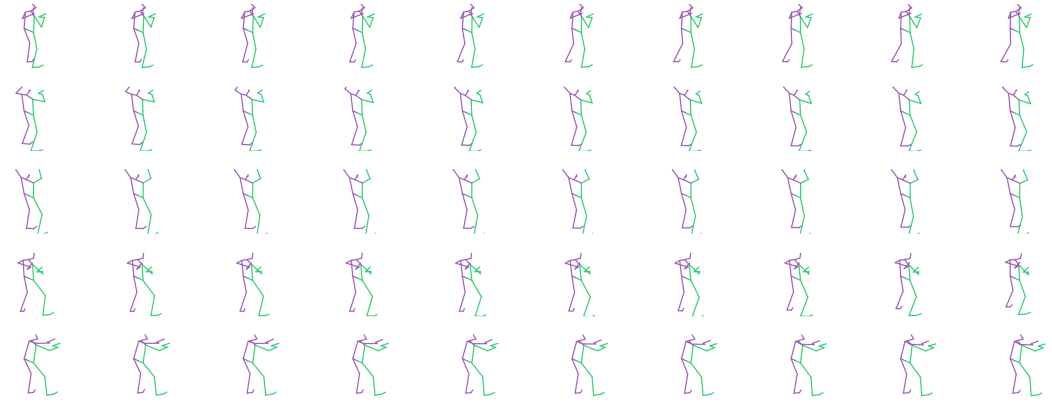

Fig. 5. The action of walkingdog.

Fig. 6. The action of walkingdog.