HOP: Heterogeneous Topology-based Multimodal Entanglement for Co-Speech Gesture Generation

Supplementary Material

1. More Details of Audio-Action Cross-modality Adaptation

The process of cross-modal adaptation between audio data and action data is illustrated in Fig. 2. Both modalities are transformed into a spatio-temporal graph structure and subsequently processed by the graph encoder [5]. The resulting fused features encapsulate action-related characteristics derived from gestures and rhythmic attributes extracted from the audio signals. Notably, the action features are segmented longitudinally into four parts, indicating that temporal action features with a time step of 4 have been effectively extracted from the action data.

In Fig. 1, we further analyze the role of the adaptive neighborhood matrix within the spatio-temporal graph encoder. To enhance the clarity of feature representation in the adaptive neighborhood matrix, we utilize the TED-Expressive datasets [4]. This dataset is particularly suitable as it includes a larger number of joints, each represented as nodes containing 3D joint features. The visualization reveals that certain columns in the matrix exhibit a higher density of high-value points, indicating that some nodes exert a stronger influence on other nodes, whereas others exhibit weaker interactions. For instance, column a displays a significantly higher concentration of high-value points compared to column b. This observation suggests that the joint action at node a likely represents a latent structural feature inherent in the action data.

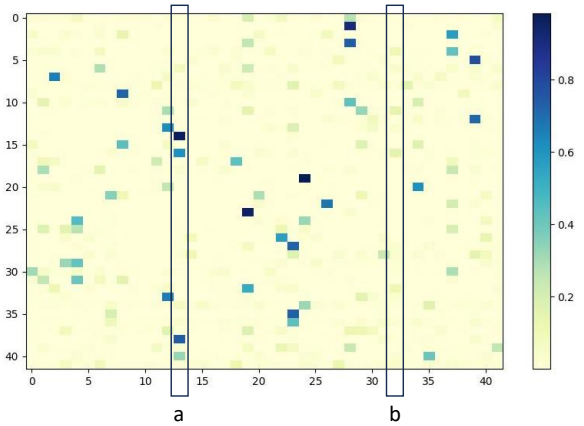

Figure 1. **The visualization of adaptive adjacency matrix.**

2. Comparison with Baselines

The visualized results of gestures generated by our method, alongside several baseline methods, are presented in Fig. 3. These visualizations illustrate the diversity of gestures produced on the TED dataset, with instances of overly sparse gesture actions highlighted in red. While the gestures generated by our approach are comparable to those of the method proposed in [4, 7, 8], in terms of BC and diversity metrics, the diversity visualization reveals that our method produces gestures with a more pronounced sense of rhythmic movement. Furthermore, compared to the method proposed in [1, 2, 6], the gestures generated by our approach exhibit greater vividness and convey richer semantic information.

3. More Details of HOP Model

Reprogramming Module. Mel-spectral features were extracted from the raw audio and transformed into a format compatible with the input space of the large language model (LLM) through a reprogramming methodology. The core of the reprogramming module [3] is the cross-attention mechanism, which facilitates the alignment and integration of audio features into the LLM framework. The detailed operations within the cross-attention mechanism, along with the corresponding output feature dimensions, are outlined in Table 1.

Operations	Feature Map Shapes
Input Mel-Spectral	$256 \times 34 \times 128$
Input word embeddings	1500×768
Query Linear(128,1024)	$256 \times 34 \times 1024$
Key Linear(768,1024)	1500×1024
Value Linear(768,1024)	1500×1024
Out Linear(1024,768)	$256 \times 34 \times 768$

Table 1. **Detailed feature Shapes in reprogramming module.**

Reprogramming Process. The reprogramming module [3] was first applied to the gesture generation task. To provide a more detailed explanation of the workflow and underlying mechanism of the reprogramming module, the corresponding algorithm is presented in Algorithm 1.

Spatial-Temporal Graph Encoder. The original audio data was processed using a sliding window approach, with a window size of 3400 and a step size of 2191. The audio data was then converted into a graph structure by increasing its dimensionality. Simultaneously, the action data was represented as a graph, with the number of joints serving as the nodes, each containing the 3D features of the respective joints. These graph representations of audio and action data were processed using Graph WaveNet, enabling the model

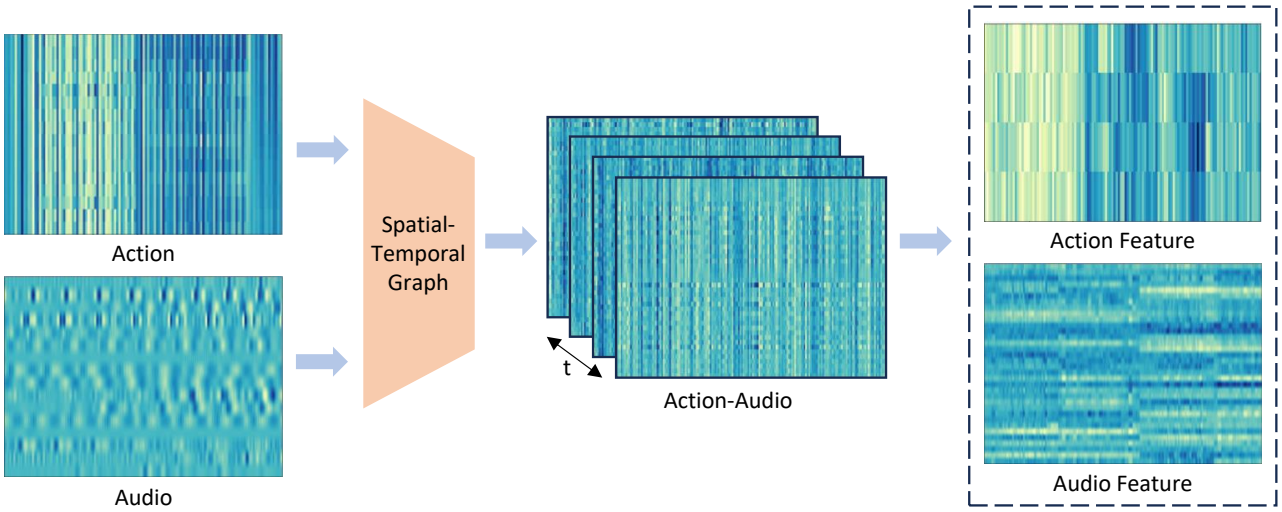

Figure 2. **Demonstration of the Audio-Action cross-modality adaptation process**. Audio and text data are cross-modally adapted using a spatial-temporal graph encoder [5], enabling the fusion of cross-modal features that incorporate action features and rhythmic characteristics from the audio.

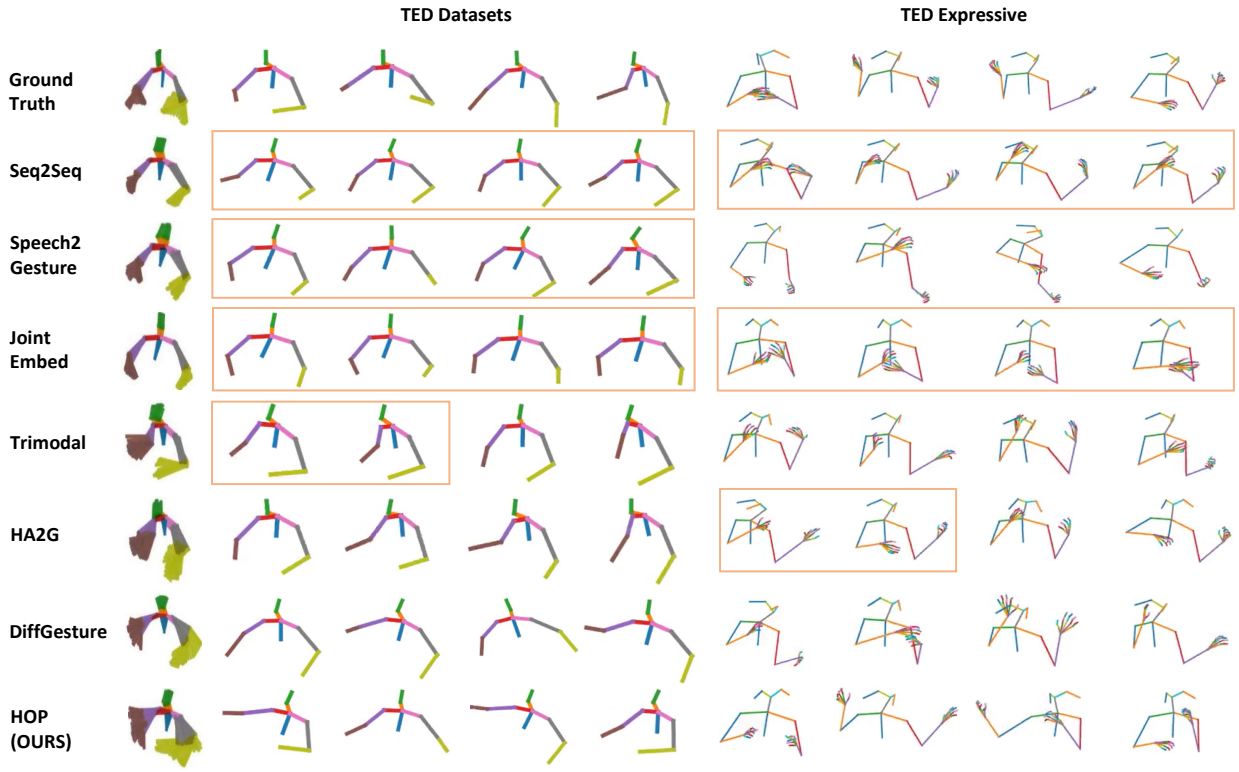

Figure 3. **Visualization of generated gestures.** We compared the gesture visualization results on these two datasets with those generated by BASELINE [1, 2, 4, 6–8]. Additionally, we present the visualization results of gesture diversity on the TED dataset, which clearly demonstrate that our approach significantly outperforms other methods in gesture diversity. Results with an insufficient number of gestures are highlighted with a red box.

to learn the rhythmic features from the audio and the action features from the gestures. The detailed operations and corresponding output feature dimensions are presented in Table 2.

Operations	Feature Map Shapes
Input Audio	1×36267
Input Action	$256 \times 16 \times 27$
Audio Matrix Converter	$256 \times 16 \times 9 \times 170$
Action Matrix Converter	$256 \times 16 \times 9 \times 3$
Graph Wavenet	$256 \times 173 \times 9 \times 4$

Table 2. **Detailed feature shapes in the spatial-temporal graph encoder.**

Algorithm 1: Reprogramming Layer Forward Pass

Input: $target\_embedding \in \mathbb{R}^{B \times L \times d}$, $source\_embedding \in \mathbb{R}^{S \times d\_llm}$, $value\_embedding \in \mathbb{R}^{S \times d\_llm}$

Output: Reprogrammed output embedding $output \in \mathbb{R}^{B \times L \times d\_llm}$

Initialization:

Initialize W_q, W_k, W_v, W_{out} for query, key, value, and output projections

Reshape $target\_embedding$:

$Q \leftarrow W_q \cdot target\_embedding \rightarrow \mathbb{R}^{B \times L \times H \times -1}$

Reshape $source\_embedding$:

$K \leftarrow W_k \cdot source\_embedding \rightarrow \mathbb{R}^{S \times H \times -1}$

Reshape $value\_embedding$:

$V \leftarrow W_v \cdot value\_embedding \rightarrow \mathbb{R}^{S \times H \times -1}$

Reprogramming:

Compute scaling factor: $scale \leftarrow 1/\sqrt{E}$

Compute attention scores: $scores \leftarrow Q \cdot K^\top$

Apply Softmax and Dropout: $A \leftarrow \text{Softmax}(scores \cdot scale)$

Compute reprogrammed embedding: $reprogrammed\_embedding \leftarrow A \cdot V$

Reshape $reprogrammed\_embedding$ to $(B, L, -1)$

Apply ReLU activation: $output \leftarrow \text{ReLU}(reprogrammed\_embedding)$

Final projection: $output \leftarrow W_{out} \cdot output$

return $output$

References

[1] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE, 2019. 1, 2

[2] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019. 1, 2

[3] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*, 2023. 1

[4] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10462–10472, 2022. 1, 2

[5] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. In *The 28th International Joint Conference on Artificial Intelligence (IJCAI)*, 2019. 1, 2

[6] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4303–4309. IEEE, 2019. 1, 2

[7] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, 39 (6):1–16, 2020. 1

[8] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. Taming diffusion models for audio-driven co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10544–10553, 2023. 1, 2