Intent-Aware Co-Speech Gesture Generation via Semantic-Emotional Disentanglement and Alignment: Supplementary Material

CONTENTS

1. EXPERIMENTS

A. Implementation Details

Motion Prior.

We model full-body gestures in a separated quantized latent space to better capture heterogeneous motion patterns across different body parts. Following prior work [65], separating body and hand representations improves motion diversity. In addition, we further decouple the face and lower body from the upper body due to their distinct correlations with speech signals. Encoding the entire body using a single VQ-VAE may bias the model toward frequently occurring motions regardless of their relevance to audio, causing less frequent but semantically important gestures (e.g., elbow movements) to be overlooked.

Specifically, we employ four independent VQ-VAE models to encode motion sequences of the face, upper body, hands, and lower body, respectively. The separated quantized latent space is defined as

$$\mathcal{Q} = \left\{ \mathbf{q}_f, \mathbf{q}_u, \mathbf{q}_h, \mathbf{q}_l \right\} \tag{S1}$$

where the face features $\mathbf{g}_f \in \mathbb{R}^{T\times 106}$, upper-body features $\mathbf{g}_u \in \mathbb{R}^{T\times 78}$, hand features $\mathbf{g}_h \in \mathbb{R}^{T\times 180}$, and lower-body features $\mathbf{g}_l \in \mathbb{R}^{T\times 64\times 4}$ are quantized by their corresponding codebooks.

Each VQ-VAE is optimized by minimizing the quantization error

$$\mathbf{g}_i = \arg\min_{\mathbf{q}\in\mathcal{Q}_i} \|\mathbf{z}_i - \mathbf{q}\|_2^2, \tag{S2}$$

where $\mathbf{z}_i$ denotes the encoded latent representation of motion component i with a temporal window size of $w = 1$. The overall VQ-VAE objective is defined as

$$\mathcal{L}_{\text{VQ-VAE}} = \mathcal{L}_{\text{rec}}(\mathbf{g}, \hat{\mathbf{g}}) + \mathcal{L}_{\text{vel}}\left(\mathbf{g}', \hat{\mathbf{g}}'\right) + \mathcal{L}_{\text{acc}}\left(\mathbf{g}'', \hat{\mathbf{g}}''\right) + \|\,\text{sg}[\mathbf{z}] - \mathbf{q}\|_2^2 + \|\mathbf{z} - \text{sg}[\mathbf{q}]\|_2^2, \tag{S3}$$

where $\mathcal{L}_{\text{rec}}$, $\mathcal{L}_{\text{vel}}$, and $\mathcal{L}_{\text{acc}}$ denote reconstruction, velocity, and acceleration losses, respectively. The reconstruction loss combines geodesic distance and L1 loss. The stop-gradient operator $\text{sg}[\cdot]$ is applied to stabilize training, and the commitment loss weight is set to 1.0 in all experiments.

B. Evaluation Metrics

To assess the realism and fidelity of generated videos, we adopt the Fréchet Video Distance (FVD), an extension of the widely used Fréchet Inception Distance (FID) to the video domain. FVD quantifies the similarity between real and synthesized video distributions in a learned feature space. Specifically, video features are extracted using a pretrained Inflated 3D ConvNet (I3D) model, and their distributions are approximated as multivariate Gaussians. The FVD is then

defined as the Fréchet distance between the Gaussian distributions of real and generated video features:

$$\text{FVD} = \|\mu_r - \mu_g\|^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{\frac{1}{2}}\right) \tag{S4}$$

where (μ_r, Σ_r) and (μ_g, Σ_g) represent the mean and covariance of the real and generated video feature distributions, respectively. A lower FVD indicates that the generated videos are closer to real ones in terms of temporal dynamics and visual quality.

To further evaluate the realism of generated gestures, we employ the Fréchet Gesture Distance (FGD) [1], which measures the distributional similarity between real and synthesized body gestures. FGD adopts the same Fréchet distance formulation as FVD, but operates in a gesture feature space:

$$\text{FGD}(\mathbf{g}, \hat{\mathbf{g}}) = \left\|\mu_r - \mu_g\right\|^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{1/2}\right) \tag{S5}$$

where (μ_r, Σ_r) denote the mean and covariance of the latent feature distribution z_r for real gestures $\mathbf{g}$, while (μ_g, Σ_g) correspond to those of the synthesized gestures $\hat{\mathbf{g}}$.

Subsequently, Diversity[2] is quantified by computing the average L1 distance across multiple body gesture clips. Higher Diversity signifies greater variance within the gesture clips. We compute the average L1 distance across various N motion clips using the following equation:

$$\text{Diversity } = \frac{1}{2N(N-1)} \sum_{t=1}^{N} \sum_{j=1}^{N} \left\| p_t^i - \hat{p}_t^j \right\|_1 \tag{S6}$$

where p_t denotes the positions of joints in frame t. We assess diversity across the entire test dataset. Moreover, when calculating joint positions, translation is zeroed, indicating that L1 Diversity is exclusively concentrated on local motion dynamics.

The synchronization between the speech and motion is conducted using Beat Alignment Score (BAS) [3]. BC indicates a more precise synchronization between the rhythm of gestures and the audio's beat. We define the onset of speech as the audio's beat and identify the local minima of the upper body joints' velocity (excluding fingers) as the motion's beat. The synchronization between audio and gesture is determined using the following equation:

$$\text{BeatAlignScore} = \frac{1}{m} \sum_{i=1}^{m} \exp\left(-\frac{\min_{t_j^y \in B^y} \|t_i^x - t_j^y\|^2}{2\sigma^2}\right) \tag{S7}$$

where $B^x = \{t_i^x\}$ is the kinematic beats, $B^y = \{t_j^y\}$ is the audio beats and σ is a parameter to normalize sequences with different FPS. We set $\sigma = 3$ in all our experiments as the FPS of all our experiment sequences is 60.

Gesture-Emotion Accuracy (GA) measures the emotional expressiveness of synthesized gestures using a post-hoc emotion classifier trained on ground-truth motion data. Specifically, we train a multi-class emotion classifier on real SMPL-X motion sequences from the training split, using the same emotion annotations provided by the BEAT dataset. The classifier takes motion sequences as input and predicts one of the predefined emotion categories. During evaluation, the trained classifier is applied to synthesized gesture sequences generated by different methods, and GA is reported as the Top-1 classification accuracy at the sequence level. Importantly, the emotion classifier is trained independently of all gesture generation models and is kept fixed during evaluation, ensuring a fair and unbiased comparison across methods. Higher GA indicates better preservation of emotional cues in the generated gestures.

Semantic-Relevant Gesture Recall (SRGR) evaluates how well a model recalls semantically meaningful gestures by combining geometric correctness with human-annotated semantic relevance. Following Liu et al.[**?**] and the BEAT benchmark, SRGR weights the Probability of Correct Keypoints (PCK) between generated and ground-truth gestures using ground-truth semantic scores.

Formally, given ground-truth joint positions p_j^t and generated joint positions $\hat{p}_j^t$ at frame t and joint j, a joint is considered correctly recalled if $\left\| p_j^t - \hat{p}_j^t \right\|_2 < \delta$. SRGR is computed as:

$$\text{SRGR} = \lambda \cdot \frac{1}{TJ} \sum_{t=1}^{T} \sum_{j=1}^{J} \mathbf{1}\left(\left\| p_j^t - \hat{p}_j^t \right\|_2 < \delta\right), \tag{S8}$$

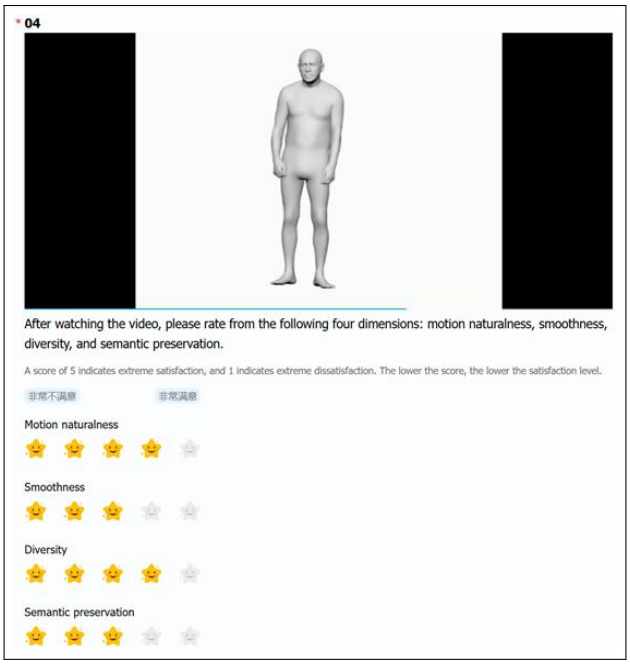

Fig. S1. Screenshot of user study website.

where $\mathbf{1}(\cdot)$ denotes the indicator function, T and J are the numbers of frames and joints, δ is the PCK threshold, and $\lambda \in [0, 1]$ is the ground-truth semantic relevance score for the corresponding clip.

The semantic relevance scores are provided by the BEAT dataset and are collected from 118 AMT annotators, covering four gesture types (beat, iconic, deictic, and metaphoric). By emphasizing accurate gesture recall in semantically important segments, SRGR better aligns with human perception of meaningful and diverse co-speech gestures.

C. User Study Details

In practice, objective measures may not always align with subjective human perception, particularly in novel contexts of collaborative voice and video generation. To gain deeper insights into the visual performance of our methods, we evaluated the visual performance of our generated videos through a user study.

To evaluate the overall quality of the body movements generated, we conducted a user study using 16 randomly sampled videos from the BEAT2 test set, each video is approximately one minute long. According to the standards established by the International Telecommunication Union (ITU) [4]. We invited 30 participants to rate the videos based on four dimensions: Natural, Diversity, Smootha, and Semantic preservation. Each criterion was rated on a scale from 1 to 5, with 5 representing the highest quality.

We created a Tencent questionnaire, as shown in the Figure S1, which includes test videos and four rating dimensions. We recorded the scores from all participants, cleaned the non-compliant data, and then calculated the average score for each dimension. Prior to participation, we provided training to the participants to ensure that their ratings were reasonable.

D. Efficiency Analysis

To assess the efficiency of our framework, we conducted a detailed comparison of inference speed across different methods. Leveraging the lightweight temporal modeling of Mamba and the discrete representation learning of VQ-VAE, our model reduces computational overhead while maintaining high-quality generation. We measured inference time on an NVIDIA GeForce RTX 3090 GPU for generating a 40-second video at 30 FPS, averaging over three runs, with results summarized in Table S1. As shown, EMAGE and MambaTalk require 171.43 s and 69.04 s, respectively, to generate the same sequence. Our method achieves an inference time of 41.15 s, significantly faster than EMAGE and comparable to MambaTalk, while still slightly slower

than GestureLSM (31.75 s). These results indicate that our framework is suitable for real-time or interactive applications, providing low-latency gesture generation with competitive efficiency.

Table S1. Time consumption comparison of inference (1 NVIDIA GeForce RTX 3090 GPU).

Methods	Inference(video of $\sim$ 40 sec)
EMAGE	$\sim$ 171.43 sec
MambaTalk	$\sim$ 69.04 sec
GestureLSM	$\sim$ 41.15 sec
Ours	$\sim$ **34.27** sec

2. LIMITATIONS

While the proposed framework demonstrates strong performance in intent-aware co-speech gesture generation, several limitations remain. First, our method relies on semantic and emotional annotations provided by the BEAT and BEAT2 datasets, where labels are assigned at the clip level. Although widely adopted, such annotations may not fully capture fine-grained, time-varying semantic or emotional cues within long speech segments, which may limit more precise intent modulation. Second, the notion of "intent" in this work is operationalized through disentangled semantic and emotional representations that primarily reflect observable motion-related cues. Higher-level pragmatic intentions, such as conversational goals or interactive context, are not explicitly modeled and remain an open challenge for future research.

REFERENCES

1. Y. Yoon, B. Cha, J.-H. Lee, *et al.*, "Speech gesture generation from the trimodal context of text, audio, and speaker identity," ACM Transactions on Graph. (TOG) **39**, 1–16 (2020).
2. X. Liu, Q. Wu, H. Zhou, *et al.*, "Learning hierarchical cross-modal association for co-speech gesture generation," in *Proceedings of the IEEE/CVF CVPR*, (2022), pp. 10462–10472.
3. R. Li, S. Yang, D. A. Ross, and A. Kanazawa, "Ai choreographer: Music conditioned 3d dance generation with aist++," in *Proceedings of the IEEE/CVF CVPR*, (2021), pp. 13401–13412.
4. R. BT, "Methodology for the subjective assessment of the quality of television pictures," Int. Telecommun. Union **4**, 19 (2002).