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

## A. Dataset

We evaluate our framework on the BEAT [1] dataset and its high-fidelity extension, BEAT2 [2]. BEAT contains multimodal recordings from 30 speakers with eight categorical emotion annotations, while BEAT2 provides temporally aligned SMPL-X motion parameters. Emotion labels from BEAT are associated with the corresponding motion sequences in BEAT2 at the clip level. Restricting the data to English-speaking participants results in 1,762 sequences from 25 speakers (12 female and 13 male), totaling approximately 35 hours of paired speech–motion data with high-fidelity finger motion. To ensure a fair comparison with prior arts such as EMAGE [2] and MambaTalk [3], we strictly follow their data partition protocol, splitting the dataset into training, validation, and testing sets with a ratio of 85%, 7.5%, and 7.5%, respectively. The training set is utilized to optimize the motion generation components, including the VQ-VAE motion prior and the intent-conditioned diffusion model

To further assess robustness and generalization in unconstrained scenarios, we additionally evaluate our method on the TED-Emotional dataset [4], which consists of 78,734 in-the-wild TED talk clips. The dataset is split into 66,679 training, 6,258 validation, and 5,797 test samples.

For training the semantic–emotional disentanglement module, we construct a specialized subset of approximately 6 hours. This subset consists of speech samples in which multiple speakers utter identical textual content under different emotion categories, and is used exclusively for cross-reconstruction training.

## B. 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 [5], 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

$$\mathcal{Q} = \{\mathbf{q}_f, \mathbf{q}_u, \mathbf{q}_h, \mathbf{q}_l\} \tag{1}$$

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. In our implementation, each codebook consists of $K = 512$ entries with a latent dimension of $d_z = 512$.

Each VQ-VAE is optimized by minimizing the quantization error

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

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

$$\begin{aligned}\mathcal{L}_{\mathrm{VQ-VAE}} = \mathcal{L}_{\mathrm{rec}}(\mathbf{g}, \hat{\mathbf{g}}) + \mathcal{L}_{\mathrm{vel}}(\mathbf{g}', \hat{\mathbf{g}}') + \mathcal{L}_{\mathrm{acc}}(\mathbf{g}'', \hat{\mathbf{g}}'') \\ + \|\,\mathrm{sg}[\mathbf{z}] - \mathbf{q}\|_2^2 + \|\mathbf{z} - \mathrm{sg}[\mathbf{q}]\|_2^2.\end{aligned} \tag{3}$$

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 $\mathrm{sg}[\cdot]$ is applied to stabilize training, and the commitment loss weight is set to 1.0 in all experiments.

**Audio Pre-processing.** All speech signals are resampled to 16 kHz and normalized by removing the DC component. The continuous audio streams are segmented into non-overlapping 10-second clips, following common practice in speech-driven gesture generation. Emotion annotations are assumed to be consistent within each clip. We discard residual fragments shorter than 10 seconds, as well as clips containing insufficient speech content (fewer than 300 audio frames after feature extraction). We extract 128-dimensional Mel-filterbank (FBANK) features using the Kaldi toolkit, with a Hanning window and a frame shift of 10 ms. To enable batch-wise training, all FBANK sequences are temporally aligned to a fixed length of $T_a = 1024$ frames. Sequences shorter than $T_a$ are zero-padded, while longer sequences are truncated. The length $T_a$ is chosen for computational convenience as a power of two.

**Training Configurations.** All experiments are implemented in PyTorch and conducted on a single NVIDIA A100 GPU using the Adam optimizer. Training proceeds in two stages. In Stage 1 (Disentanglement), the semantic and emotional encoders are jointly trained using the cross-reconstruction objective for 25 epochs with a batch size of 1, which is sufficient for the pairwise recombination strategy and reduces GPU memory consumption. The initial learning rate is set to $1 \times 10^{-5}$ and decayed by a factor of 0.85 starting from epoch 5 using a step scheduler. In Stage 2 (Generation), the pre-trained encoders are frozen, and the intent-conditioned latent diffusion model is trained for 12,000 epochs with a batch size of 32 and a fixed learning rate of $1 \times 10^{-4}$.

**Denoiser.** The denoising network shares the same Transformer architecture as the prior encoder, with a hidden dimension of 1024 for all layers. We employ a diffusion process with

1000 steps during training and 50 steps during inference. The noise variance schedule is defined by linearly spaced betas in the range [0.00085,0.012].

### C. 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:

$$\text{FVD} = \|\mu_r - \mu_g\|^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{\frac{1}{2}}\right) \tag{4}$$

where $(\mu_r, \Sigma_r)$ and $(\mu_g, \Sigma_g)$ 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) [6], 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:

$$\text{FGD}(\mathbf{g}, \hat{\mathbf{g}}) = \|\mu_r - \mu_g\|^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{1/2}\right) \tag{5}$$

where $(\mu_r, \Sigma_r)$ denote the mean and covariance of the latent feature distribution $z_r$ for real gestures $\mathbf{g}$, while $(\mu_g, \Sigma_g)$ correspond to those of the synthesized gestures $\hat{\mathbf{g}}$.

Subsequently, Diversity [7] 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:

$$\text{Diversity } = \frac{1}{2N(N-1)} \sum_{t=1}^{N} \sum_{j=1}^{N} \left\| p_t^i - \hat{p}_t^j \right\|_1 \tag{6}$$

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) [8]. 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:

$$\text{BeatAlignScore} = \frac{1}{m} \sum_{i=1}^{m} \exp\left(-\frac{\min_{t_j^y \in B^y} \|t_i^x - t_j^y\|^2}{2\sigma^2}\right) \tag{7}$$

where $B^x = \{t_i^x\}$ is the kinematic beats, $B^y = \{t_j^y\}$ is the audio beats and $\sigma$ 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:

$$\text{SRGR} = \lambda \cdot \frac{1}{TJ} \sum_{t=1}^{T} \sum_{j=1}^{J} \mathbf{1}\left(\left\|p_j^t - \hat{p}_j^t\right\|_2 < \delta\right), \tag{8}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function, $T$ and $J$ are the numbers of frames and joints, $\delta$ 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.

### D. Baselines

We compare the proposed IACG framework with six state-of-the-art co-speech gesture generation methods that differ in architectural design and control capability. Specifically, AMUSE [9] is an emotion-controllable approach that incorporates affective cues to modulate gesture generation; EMAGE [2] adopts expressive masked audio–gesture modeling to learn holistic spatio-temporal representations; GestureLSM [10] leverages latent shortcut connections to preserve fine-grained motion details and improve temporal coherence; MambaTalk [3] employs selective state space models to efficiently capture long-term motion dynamics for scalable high-fidelity

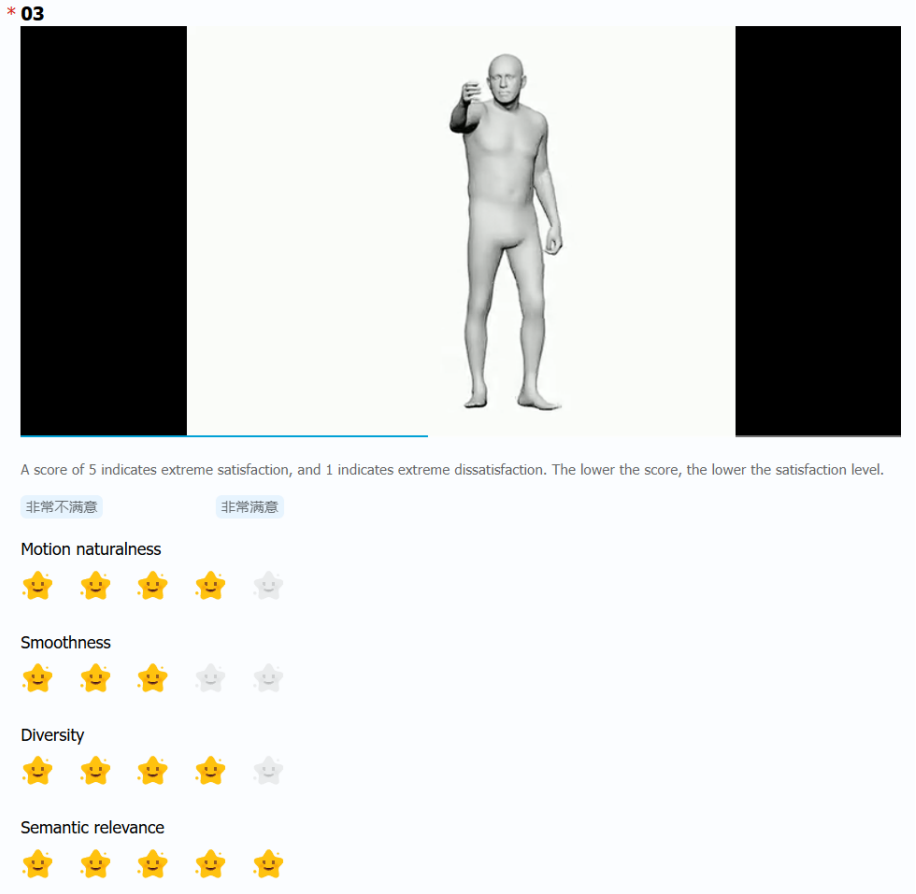


Fig. 1. Screenshot of user study website.

synthesis; ProbTalk [11] formulates co-speech motion generation as a probabilistic process with hierarchical discrete representations to produce diverse and coordinated full-body motions; and DiffSHEG [12] introduces a diffusion-based framework with unidirectional diffusion to achieve temporally consistent speech-driven gestures. Since AMUSE is the only baseline that supports explicit emotion editing, it is exclusively used for emotion-related evaluations, while all methods are compared on general motion quality, diversity, and semantic alignment metrics. Several recent approaches [13], [14] are excluded due to code unavailability or incompatibility with the BEAT dataset.

### *E. 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) [15]. 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 1, 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.

### *F. Qualitative Analysis of Ablation Study*

To intuitively demonstrate the contribution of each module in our IACG framework, we present a frame-by-frame visual comparison in Figure 2. The figure displays motion sequences generated by five ablated variants alongside the Ground Truth (leftmost) and our full model (rightmost).

**Impact of Dual-Stream Disentanglement (Cols. 2 & 3).** The second and third columns highlight the necessity of explicit semantic and emotional modeling.

- **w/o Semantic Stream:** As shown in Col. 2, removing the semantic encoder results in *generic and repetitive gestures*. The character performs random hand movements

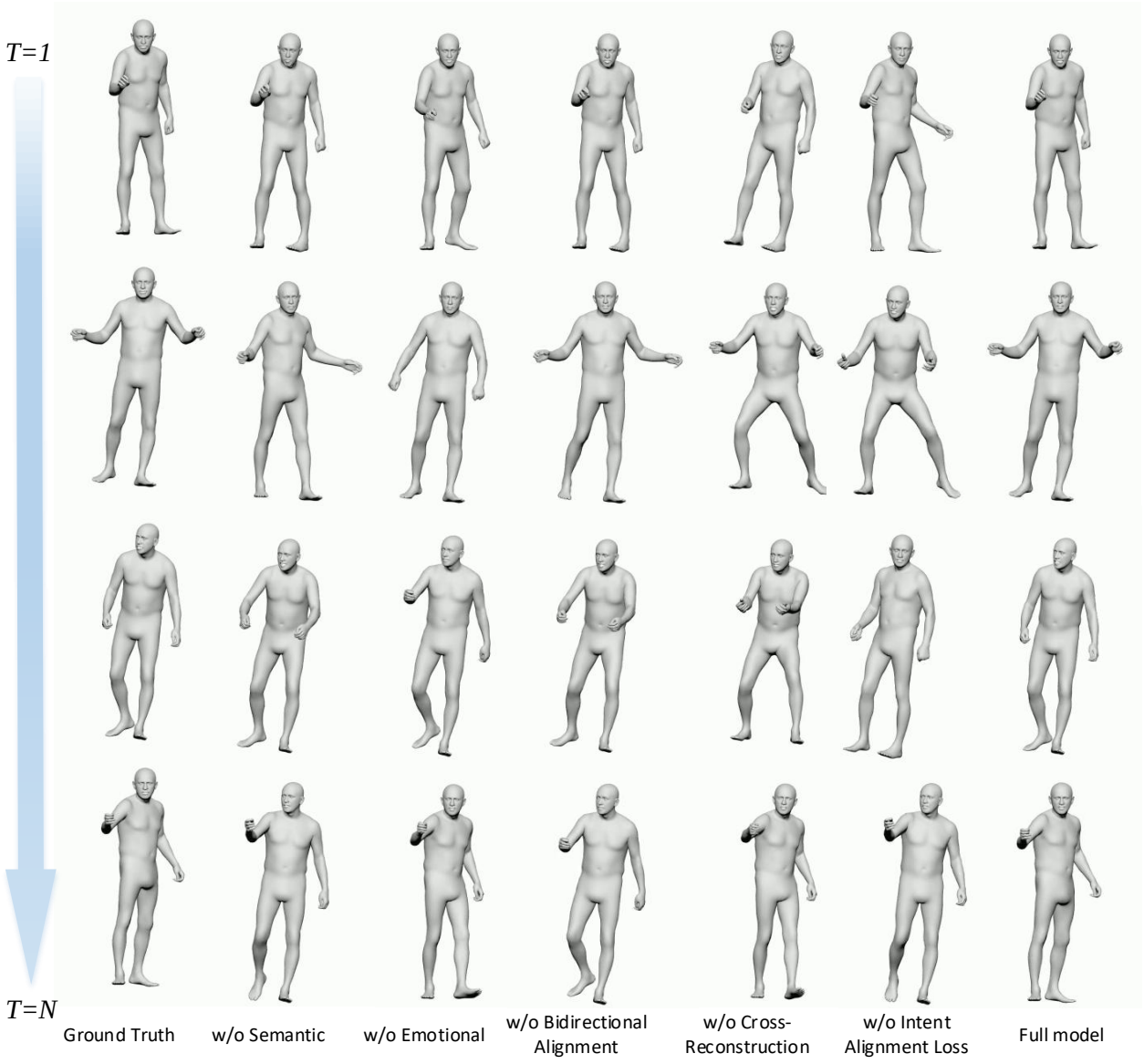


Fig. 2. Qualitative results of the ablation study. Comparison between Ground Truth, five ablated variants, and our full IACG model. The snapshots from top to bottom illustrate the motion evolution over time. The results demonstrate that removing key components (e.g., Semantic/Emotional streams or Alignment modules) degrades the synchrony and expressiveness of the generated gestures, whereas our full model maintains high fidelity to the reference motion.

that fail to convey specific meanings or align with the linguistic content of the speech.

- **w/o Emotional Stream:** Conversely, the variant without the emotional stream (Col. 3) produces *flat and constrained motions*. The gestures lack the necessary amplitude and dynamic range, failing to reflect the intensity and expressiveness required by the emotional audio input.

**Necessity of Bidirectional Alignment (Col. 4).** The visual results in Col. 4 demonstrate that simply extracting features is insufficient; they must be aligned. Without the Bidirectional Alignment module, the generated motions appear *stiff and uncoordinated*. The lack of interaction between the two streams leads to distinct discontinuities, where the emotional style does not naturally blend with the semantic gestures.

**Effectiveness of Optimization Constraints (Cols. 5 & 6).** Columns 5 and 6 illustrate the subtle but critical role of the Cross-Reconstruction and Intent Alignment Loss. Removing these constraints leads to a degradation in *synchronization and precision*. While the overall pose structure is maintained, the gestures are less decisive and occasionally drift out of sync with the audio beat compared to the full model.

**Full Model Superiority (Col. 7).** Finally, our full IACG model (rightmost column) synthesizes the most natural results. It successfully combines the semantic accuracy of the Ground Truth with expressive emotional dynamics, producing smooth, coherent, and high-fidelity gestures that outperform all ablated

TABLE I
INFERENCE TIME COMPARISON FOR GENERATING A ∼10-SECOND VIDEO

| Methods | Inference Time (s) |
|---|---|
| AMUSE | ∼ 37.15 |
| ProbTalk | ∼ 58.37 |
| DiffSHEG | ∼ 29.18 |
| EMAGE | ∼ 42.86 |
| MambaTalk | ∼ 17.26 |
| GestureLSM | ∼ **12.28** |
| Ours (IACG) | ∼ 15.57 |

baselines.

### G. 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 aims to reduce computational overhead while maintaining high-quality generation. We measured the inference time on an NVIDIA A100 GPU for generating a video of approximately 10 seconds, averaging over three runs, with results summarized in Table I. As shown, heavy diffusion-based models such as ProbTalk and EMAGE exhibit high latency, requiring 58.37 s and 42.86 s, respectively, to generate the same sequence. In contrast, our method (IACG) achieves an inference time of 15.57 s, representing a speedup of approximately 2.7×over EMAGE. Furthermore, our method outperforms MambaTalk (17.26 s) and is competitive with the ultra-lightweight GestureLSM (12.28 s). These results demonstrate that our framework strikes a favorable balance between generation quality and computational efficiency, making it highly suitable for practical applications.

### H. 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] H. Liu, Z. Zhu, N. Iwamoto, Y. Peng, Z. Li, Y. Zhou, E. Bozkurt, and B. Zheng, "Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis," in *European conference on computer vision*. Springer, 2022, pp. 612–630.

[2] H. Liu, Z. Zhu, G. Becherini, Y. Peng, M. Su, Y. Zhou, X. Zhe, N. Iwamoto, B. Zheng, and M. J. Black, "Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1144–1154.

[3] Z. Xu, Y. Lin, H. Han, S. Yang, R. Li, Y. Zhang, and X. Li, "Mambatalk: Efficient holistic gesture synthesis with selective state space models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 20 055–20 080, 2024.

[4] X. Qi, C. Liu, L. Li, J. Hou, H. Xin, and X. Yu, "Emotiongesture: Audio-driven diverse emotional co-speech 3d gesture generation," *IEEE Transactions on Multimedia*, vol. 26, pp. 10 420–10 430, 2024.

[5] H. Yi, H. Liang, Y. Liu, Q. Cao, Y. Wen, T. Bolkart, D. Tao, and M. J. Black, "Generating holistic 3d human motion from speech," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 469–480.

[6] Y. Yoon, B. Cha, J.-H. Lee, M. Jang, J. Lee, J. Kim, and G. Lee, "Speech gesture generation from the trimodal context of text, audio, and speaker identity," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–16, 2020.

[7] X. Liu, Q. Wu, H. Zhou, Y. Xu, R. Qian, X. Lin, X. Zhou, W. Wu, B. Dai, and B. Zhou, "Learning hierarchical cross-modal association for co-speech gesture generation," in *Proceedings of the IEEE/CVF CVPR*, 2022, pp. 10 462–10 472.

[8] 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. 13 401–13 412.

[9] K. Chhatre, N. Athanasiou, G. Becherini, C. Peters, M. J. Black, T. Bolkart *et al.*, "Emotional speech-driven 3d body animation via disentangled latent diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1942–1953.

[10] P. Liu, L. Song, J. Huang, H. Liu, and C. Xu, "Gesturelsm: Latent shortcut based co-speech gesture generation with spatial-temporal modeling," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2025, pp. 10 929–10 939.

[11] Y. Liu, Q. Cao, Y. Wen, H. Jiang, and C. Ding, "Towards variable and coordinated holistic co-speech motion generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1566–1576.

[12] J. Chen, Y. Liu, J. Wang, A. Zeng, Y. Li, and Q. Chen, "Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7352–7361.

[13] S. Alexanderson, R. Nagy, J. Beskow, and G. E. Henter, "Listen, denoise, action! audio-driven motion synthesis with diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–20, 2023.

[14] T. Ao, Z. Zhang, and L. Liu, "Gesturediffuclip: Gesture diffusion model with clip latents," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–18, 2023.

[15] R. BT, "Methodology for the subjective assessment of the quality of television pictures," *International Telecommunication Union*, vol. 4, p. 19, 2002.