SemGest: A Multimodal Feature Space Alignment and Fusion Framework for Semantic-aware Co-speech Gesture Generation

Yo-Hsin Fang
RIKEN
Kyoto, Japan
yo-hsin.fang@a.riken.jp

Vijay John
RIKEN
Kyoto, Japan
vijay.john@riken.jp

Yasutomo Kawanishi
RIKEN
Kyoto, Japan
yasutomo.kawanishi@riken.jp

Abstract

This paper addresses the challenge of 3D co-speech gesture generation, aiming to generate body gestures that align with spoken content. Existing methods leverage multimodal features, such as speech and transcripts, to improve the expressiveness of generated gestures. However, generating gestures that express the semantic meaning of the speech remains challenging. To address this limitation, we propose SemGest, a framework featuring a semantic-to-gesture alignment mechanism and a feature fusion module that effectively integrates speech features and semantic features extracted from the transcribed text. A diffusion-based model is then conditioned on the fused features to generate realistic and semantic-aware co-speech gestures. By aligning semantic and gesture spaces and adaptively fusing speech and semantic features, the resulting feature space is more robust, aiding in the conditional generation process. We perform a detailed experimental analysis, demonstrating the advantages of our proposed framework over the baseline algorithms in generating vivid co-speech gestures. Our experimental results demonstrate the superiority of the proposed framework. Furthermore, ablation studies also validate the effectiveness of the proposed semantic-to-gesture alignment and feature fusion mechanisms in the proposed framework.

CCS Concepts

• **Computing methodologies** → *Image and video acquisition*; Motion processing.

Keywords

Pose generation; Multimodal; Diffusion; Deep learning

ACM Reference Format:
Yo-Hsin Fang, Vijay John, and Yasutomo Kawanishi. 2025. SemGest: A Multimodal Feature Space Alignment and Fusion Framework for Semantic-aware Co-speech Gesture Generation. In *Proceedings of the International Workshop on Generation and Evaluation of Non-verbal Behaviour for Embodied Agents (GENEA '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3746268.3759433

1 Introduction

Co-speech gesture generation is an important task that generates gestures corresponding to a given speech, which aids the speaker in effectively communicating their intent and emotions. Generating speech-synchronized or co-speech gestures has broad applications in areas such as robotics, AR/VR, and game development [9].

To generate co-speech gestures, previous studies have incorporated various multimodal inputs apart from speech, such as text [21, 37–39], speaker identity [21, 23, 37], and emotion [9, 21, 25]. In these studies, the smoothness and naturalness of synthetic gestures are improved, but the generated gestures are only correlated to the speech rhythm without being sensitive to the underlying semantics. Owing to this limitation, generating semantic-aware gestures, including iconic, metaphoric, and deictic gestures [4], is challenging. This can be primarily attributed to speech being the dominant modality [39], which hinders the ability of the system to effectively leverage the information from other modalities to generate the semantic-aware gestures. While early rule-based approaches [5, 6] relying on predefined correspondences between vocabularies and gestures can address this limitation, they typically generate deterministic results, limiting their flexibility and applicability. In recent years, with advancements in deep learning, various approaches [3, 18] attempt to address these issues using conditional generative models and predefined gesture classes to enhance the semantic awareness. However, the generated gestures remain constrained by the predefined gesture class. Other deep-learning-based methods have leveraged semantic features extracted from off-the-shelf Language Models (LMs) [16] or Vision-Language Models (VLMs) [35, 39] to learn a joint space of gestures and semantics. However, the learned joint space does not

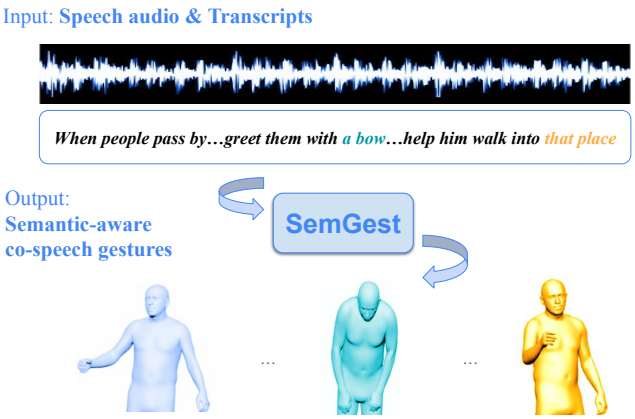


Figure 1: Goal. Given speech and transcripts, SemGest can generate upper-body co-speech gestures that express semantic meanings.

GENEA '25, Dublin, Ireland.

ACM ISBN 979-8-4007-2050-5/2025/10
https://doi.org/10.1145/3746268.3759433

Table 1: We compare the incorporated modalities and the fusion mechanisms between existing works and the proposed method.

Method	Input Modalities	Fusion Strategy
Trimodal [37]	Speech, Text, Identity	Concatenation of all encoded modalities
CaMN [21]	Speech, Text, Identity, Emotion, Facial blendweights	Concatenation of encoded modalities
DiffGesture [40]	Speech and initial poses	Concatenation of speech and initial pose as Transformer's input
LivelySpeaker [39]	Speech, Text, Identity	Stage 1: Gesture generation from text Stage 2: Concatenate time-aligned corrupted gestures, audio, and identity as denoiser's input
TalkSHOW [36]	Speech, Identity	Concatenation of MFCC and (one-hot) identity
EMAGE [20]	Speech, Text, Identity	Attention
MMoFusion [31]	Speech, Text, Identity, Emotion	Progressive fusion with masked style matrix
SemGest (Ours)	Speech, Text, Identity	Dual-branch cross-attention (See Fig. 3)

effectively bridge the distribution gap between the semantic and gesture feature space because of the many-to-many mapping between the input semantic-gesture features and the output gestures. Previous studies [21, 37] in co-speech gesture generation typically rely on concatenation or self-attention to integrate multimodal features, especially speech and text. These approaches overlook the inherent correlation between the two input modalities, leading to redundant or even conflicting information. These limitations in generating semantic-aware co-speech gestures from speech underscore the need to balance the influence of speech and semantic information.

To address these challenges, this paper proposes SemGest, a novel diffusion-based framework that leverages both speech and semantic features to guide the co-speech gesture generation. The proposed framework contains two key stages: semantic-aware feature extraction and conditional gesture generation. In the semantic-aware feature extraction stage, we first introduce a **semantic-to-gesture alignment** mechanism, utilizing a transformer-encoder-based model. The transformer-based model maps the CLIP [27] text embeddings to a pre-trained gesture space. The resulting embedding yields a semantic-gesture joint space aligning semantic and gesture features closely. This makes the semantic features carry gesture-relevant information that can guide the generation process. Apart from the alignment mechanism, we also propose a **feature fusion** mechanism in the semantic-aware feature extraction stage that employs a dual-branch cross-attention mechanism to integrate speech and semantic features. The attention mechanism comprises speech-to-semantic and semantic-to-speech cross-attention layers that learn the correlation between speech and semantic features. The attention mechanism additionally contains self-attention layers with an aggregation token to extract the final representation of speech and semantics to aid the conditional generation. Using the attention layers, the feature fusion mechanism effectively integrates speech and semantics and reduces redundant information, resulting in a robust latent representation that guides the following conditional gesture generation process. Using the semantic-to-gesture alignment and feature fusion mechanisms, our proposed framework generates realistic and semantic-aware co-speech gestures. Comparative experiments demonstrate the advantage of the proposed methods over the baseline methods. The ablation studies further validate the effectiveness of the semantic-aware feature extraction and feature fusion mechanisms. The main contributions of this paper are summarized as follows:

- We introduce a semantic-to-gesture alignment mechanism, which learns a robust semantic-motion space where gesture-relevant semantic features are obtained.
- We propose a feature fusion mechanism that progressively and effectively integrates speech and semantic features, resulting in a robust feature representation.
- Our framework achieves state-of-the-art performance on co-speech gesture generation using skeletal human body representation.

2 Related Work

2.1 Co-speech gesture generation

Co-speech gesture generation has been an active research topic for several years. Early studies [5, 6] relied on rule-based approaches to produce deterministic results. To tackle the amount of varying motions, recent advances in literature have shifted the focus towards data-driven methods based on models such as GANs [1, 14, 26], VAEs [12, 17], VQ-VAEs [7, 20, 34, 36], and Mamba [11]. Yoon *et al*. [37] proposed an LSTM-based model conditioned on temporal synchronized multimodal context to model upper-body gestures. However, this approach is limited to controlling personal styles. Liu *et al*. [21] introduced a cascaded architecture that incorporates additional features such as emotion, facial blendshape weights to synthesize more expressive upper-body gestures. Recently, diffusion-based methods [8, 9, 24, 33, 35, 40] have gained popularity because of their ability to model complex data distributions. Compared with these methods, which only generate upper-body gestures, Yi *et al*. [36] proposed a framework that generates 3D holistic co-speech gestures by separately modeling different body parts, with each body part having an independent correlation with speech. Liu *et al*. [20] adopted a cross-attention mechanism between speech and content while leveraging masked modeling to generate holistic co-speech motions. Chen *et al*. [7] introduced a prompt-based, data-augmented approach to enable synergistic and out-of-domain holistic motion generation conditioned on speech and user prompts. These works incorporate modalities beyond audio and motion to generate co-speech gestures, primarily relying on concatenation or cross-attention to integrate multimodal information (See Tab. 1). Wang *et al*. [31] proposed a progressive fusion strategy by leveraging concatenation, attention, and masked style matrix to integrate multimodal features, aiming to reduce the unnecessary features and noise. Compared with

the existing research, the proposed multimodal feature fusion module is based on the dual-branch cross-attention strategy to provide a robust latent representation that guides the conditional co-speech gesture generation.

2.2 Semantic-aware co-speech gesture generation

Although multimodal inputs are incorporated in co-speech gesture generation, the dominant speech rhythm overshadows other modality features to generate gestures. This results in the degradation of the generated gestures with limited semantic awareness and the ineffective utilization of available modalities. Several studies have been conducted to address this issue. Liang *et al*. [18] first mined the beat and semantic information from speech and then leveraged a semantic prompter to model semantic co-speech gestures. Zhi *et al*. [39] adopted a two-stage framework, consisting of the semantic-aware generator (SAG) and rhythm-aware generator (RAG). SAG utilizes CLIP [27] to construct a joint space for text and motion. Subsequently, a diffusion-based RAG models the distribution conditioned on speech rhythm. The method proposed by Liu *et al*. [19] first learned a joint space using consistency loss to enhance the semantic correspondence between speech and motion and trained a weakly supervised detector to identify salient postures, defined by large movements with rich semantic information, to enforce space alignment. Compared to the aforementioned literature, in our proposed framework, we introduce a semantic-to-gesture alignment mechanism that maps the semantic space to a pre-trained gesture space to align the two latent spaces.

3 Proposed Framework

Given an input audio sequence $A^{1:N} = \{\mathbf{a}_1, \ldots, \mathbf{a}_N\}$, the proposed framework $\mathcal{G}$ generates upper-body co-speech gestures $P^{1:N} = \mathbf{p}_1, \ldots, \mathbf{p}_N$ (N pose frames), where each $\mathbf{p}_i$ comprises J joints in 3D space. Apart from speech, the proposed framework is conditioned on the transcribed text and speaker identity, represented as C. Following previous studies [20, 21], the first M pose frames $\{\mathbf{p}_1, \ldots, \mathbf{p}_M\}$ serve as the initial seed to guide the generation of subsequent N frames, where $M \ll N$. Hence, the overall objective of the model is formulated as:

$$\arg\min_{\mathcal{G}} = \|P - \mathcal{G}(A^{1:N}, C, p^{1:M})\|. \tag{1}$$

The proposed framework SemGest is shown in Fig. 2. We next present the details of the feature extraction mechanism followed by conditional co-speech gesture generation mechanism in Sec. 3.2.

3.1 Sematic-aware feature extraction

In semantic-aware feature extraction, we first use two encoders to extract unimodal features from speech and transcribed text, respectively. Next, a dual-branch cross-attention feature fusion module is formulated to integrate the multimodal features, producing semantic-aware features.

Speech feature extraction. Given the input audio, we adopt a transformer-based encoder and decoder to obtain the speech latent representation. Specifically, we employ the Audio Spectrogram Transformer [13], which leverages the power of ViT [10], to embed the spectrogram to a latent vector $\mathbf{z}_A = E_A(A) \in \mathbb{R}^{1\times d}$. The decoder is then tasked with reconstructing the filterbanks from the latent vector. To preserve crucial speech information, we additionally enforce speech power reconstruction, which serves as an indicator of the gesture beats. Furthermore, to exploit the emotion cues in the speaker's speech for the co-speech gesture generation, we also train an emotion classifier that predicts emotion using the latent vector $\mathbf{z}_A$. For training, the module employs a combined loss consisting of the weighted sum of reconstruction loss of the filterbank and speech power reconstruction loss, and emotion classification cross-entropy.

Semantic-to-gesture alignment. We leverage the semantics in the underlying transcripts to generate semantic-aware co-speech gestures. Similar to previous studies [30, 39], we employ the text embeddings derived from pre-trained CLIP [27] (ViT-B/32). We propose a semantic-to-gesture projection module consisting of three transformer-encoder layers to construct a semantic-gesture joint space and extract gesture-relevant semantic features. Specifically, an encoder-decoder-based Transformer model is trained for gesture reconstruction. Subsequently, the proposed semantic-to-gesture projection module projects the CLIP text space onto the gesture space, aligning the CLIP text embeddings with the corresponding gesture latent vectors using the synchronized time stamps. This alignment process is supervised by the MSE loss between gesture-relevant semantic features and the gesture latent vectors.

Feature fusion. Recognizing the importance of modeling the correlation between speech and semantic, we propose a feature fusion module that adaptively aggregates the speech and semantic features (Sec Fig. 3). Inspired by dual-branch architectures [22, 32], our module first utilizes speech-to-semantic and semantic-to-speech cross-attention to model the correlation between the two input modalities. Subsequently, the two intermediate features are concatenated along the temporal axis with a special token appended at the beginning (similar to the class token in ViT [10]), resulting in an intermediate feature representation with shape $\mathbb{R}^{(1+1+N)\times d}$. This sequence is processed by a transformer encoder, and the special token's value is considered as the final feature representation $\mathbf{z}_{feat}$ for conditional co-speech gesture generation.

3.2 Conditional co-speech gesture generation

The architecture of the conditional generation module is based on the latent diffusion model [28], which applies forward diffusion and denoising on the gesture latent space. The gesture prior model reconstructs the input pose sequence. Sebsequently, the latent diffusion model uses a forward and backward diffusion process to learn the underlying distribution and generate co-speech gestures. Finally, a spatial-temporal self-attention module smooths the resulting gestures.

Gesture prior model. The gesture prior model adopts a transformer-based encoder-decoder architecture. The encoder $\mathcal{E}_P$ maps a pose sequence $p^{1:N}$ and positional encoding to a latent representation $\mathbf{z}_P$. The decoder $\mathcal{D}$, initialized by the seed pose, reconstructs the pose sequence from the query vector and the memory vector $\mathbf{z}_P$. Similar to the fusion module, a special token is appended at the beginning of the pose sequence to summarize the multi-frame sequence into a single-frame gesture embedding, $\mathbf{z}_P \in \mathbb{R}^{1\times d}$. To enhance the correlation between gesture beat and speech audio beat, we introduce an additional cross-attention layer that incorporates speech onset information.

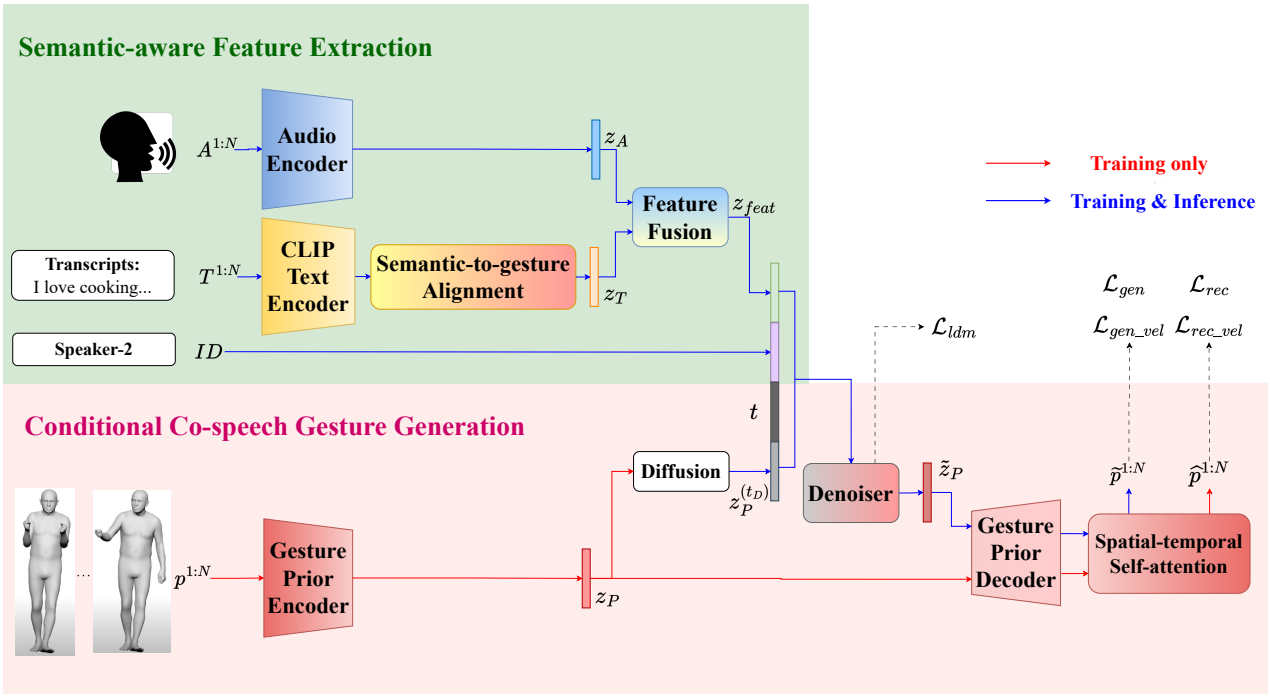


Figure 2: SemGest consists of two stages: semantic-aware feature extraction and conditional co-speech gesture generation. In the first stage, speech features are extracted from our pre-trained audio encoder, and gesture-relevant semantic features is produced from the proposed semantic-to-gesture alignment, which maps the CLIP text embeddings to a pre-trained gesture space. Next, a feature fusion module with cross-attention layers is used to fuse the speech and semantic features and embed them to a latent representation. In the second stage, the latent embedding is concatenated with noisy gesture latents (extracted by gesture prior encoder), timestamps, and speaker identity, and processed by denoiser Δ through reverse diffusion. Finally, the gesture decoder and the spatial-temporal attention module reconstruct or generate gestures $\widehat{p}^{1:N}$ from latent gestures z_P or denoised representation $\widetilde{z}_P$, respectively.

Forward diffusion process. To obtain noisy latent representation, fixed variance and linearly scaled noise scheduler are utilized to progressively add Gaussian noise to the gesture latent representation $\mathbf{z}_P$ over D diffusion timestamps:

$$q(\mathbf{z}_P^{(t_d)}|\mathbf{z}_P^{(0)}) = \mathcal{N}(\mathbf{z}_P^{(t_d)}; \sqrt{\bar{\alpha}_{t_d}}\mathbf{z}_P^{(0)}, (1-\bar{\alpha}_{t_d})\mathbf{I}), \tag{2}$$

where $\alpha_{t_d} = 1-\beta_{t_d}$, $\bar{\alpha}_{t_d} = \prod_{s=1}^{t_d} \alpha_s$. Following [15], α_{t_d} is a notation and β_{t_d} represents diffusion process variance.

Conditional denoising model. Starting from the noisy latent representation $\mathbf{z}_P^{(t_d)} \sim \mathcal{N}(0, I)$, the denoiser Δ, implemented as a transformer encoder, progressively refines the noisy latent vector and reconstructs the original gesture latent $\mathbf{z}_P^{(0)}$. To generate the gestures, the concatenated noisy gesture latent $\mathbf{z}_P^{(t_d)}$, positional-encoded timestep $PE(t) \in \mathbb{R}^d$, fused feature $\mathbf{z}_{feat}$, speaker identity s, and seed poses $p^{1:M}$ are given as input to the denoiser. This is represented as,

$$\mathbf{z}_P^{(t_d-1)} = \Delta([\mathbf{z}_P^{(t_d)}, PE(t_d), \mathbf{z}_{feat}, s, p^{1:M}]). \tag{3}$$

The denoiser output represents the noise between $\mathbf{z}_P^{(t_d)}$, and $\mathbf{z}_P^{(t_d-1)}$. Based on the noise prediction, a DDIM scheduler [29] is utilized to convert $\mathbf{z}_P^{(t_d-1)}$ back to $\mathbf{z}_P^{(t_d)}$.

Spatial-temporal self-attention. Inspired by [2], we utilize the spatial-temporal self-attention module to learn the spatial and temporal dependency and generate smooth gesture sequences.

Training. The audio encoder and the semantic-to-gesture projection layer are frozen, while the other modules are optimized jointly. In particular, our training process involves three primary stages: gesture reconstruction, noise modeling, and conditional gesture generation. Firstly, we train the gesture encoder, decoder, and the spatial-temporal self-attention module mainly using Huber loss (Eq. (4)) to minimize the discrepancy between the ground truth and the reconstructed gesture. Subsequently, a gesture embedding $\widehat{\mathbf{z}}_P$, extracted from the gesture encoder with gradient calculation disabled, undergoes the forward diffusion process and is transformed into a noisy embedding $\mathbf{z}_P^{(D)}$. The denoiser is trained with the Mean Squared Error to predict the noise $\widehat{\delta}^{t_d}$ added during the foward diffusion process (δ^{t_d}). In the final step, we employ the generation loss (Eq. (6)) to minimize the discrepancy the ground truth and the generated gesture. To improve robustness in the generation, a random noise is fed to the denoiser, resulting in a fully denoised latent $\widetilde{\mathbf{z}}_m$. This latent embedding is then processed by the gesture decoder and the spatial-temporal self-attention module to generate co-speech gestures $\widetilde{\mathbf{p}}^{1:N}$. We also incorporate velocity losses (Eq. (7)) as additional

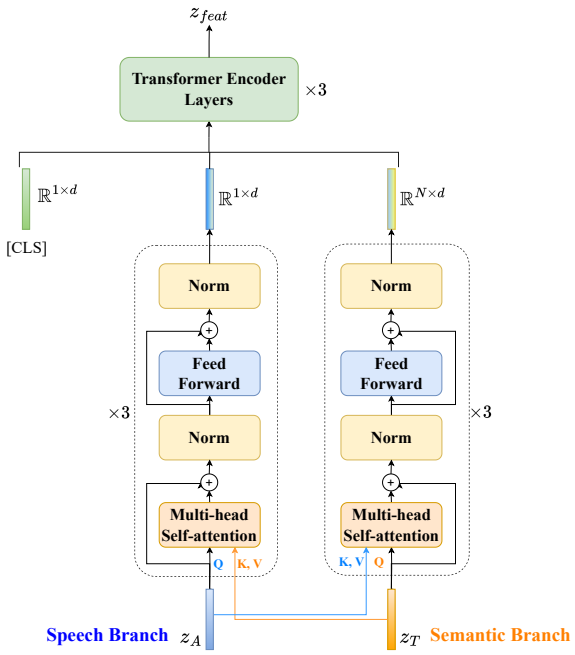


Figure 3: Details of fusion module. The fusion module integrates the speech feature and the semantic feature from the joint space. It first employs cross-attentions to learn the correlation between the two modalities, and then a transformer encoder is leveraged to transform the sequence of features into a feature representation, which serves as the condition for the generation process.

loss terms to regularize the reconstructed and generated gestures. From empirical results, we observe that there are redundant motion beats in specific joints. Hence, we emphasize the reconstruction and generation loss on the corresponding joints, denoted as $\mathcal{L}_{joint\_rec}$ and $\mathcal{L}_{joint\_gen}$ and similarly implemented by Huber loss. Details of these loss terms are explained in detail in the supplementary materials.

$$\mathcal{L}_{rec} = \mathcal{L}_{Huber}(p^{1:N}, \widehat{p}^{1:N}) \tag{4}$$

$$\mathcal{L}_{ldm} = \mathcal{L}_{MSE}(\delta^{t_d}, \widehat{\delta}^{t_d}) \tag{5}$$

$$\mathcal{L}_{gen} = \mathcal{L}_{Huber}(p^{1:N}, \widetilde{p}^{1:N}) \tag{6}$$

$$\mathcal{L}_{rec\_vel} = \mathcal{L}_{Huber}(\dot{p}^{1:N}, \dot{\widehat{p}}^{1:N}); \ \mathcal{L}_{gen\_vel} = \mathcal{L}_{Huber}(\dot{p}^{1:N}, \dot{\widetilde{p}}^{1:N}) \tag{7}$$

The overall training objective is the weighted sum of $\mathcal{L}_{rec}$, $\mathcal{L}_{ldm}$, $\mathcal{L}_{gen}$, $\mathcal{L}_{rec\_vel}$, $\mathcal{L}_{gen\_vel}$, $\mathcal{L}_{joint\_rec}$, and $\mathcal{L}_{joint\_gen}$.

Inference In the inference phase, similar to the last forward pass in training, a random noise $\mathbf{z}_P^{(t_D)} \sim \mathcal{N}(0, I)$ is sampled and iteratively refined by the denoiser. The fully denoised latent $\widetilde{\mathbf{z}}_P$ is fed to the gesture decoder and the spatial-temporal self-attention module for gesture generation.

4 Implementation Details

Dataset and human body representation. We conduct experiments on BEAT dataset [21]. In particular, we use version v0.2.1, which employs a 78-joint skeletal human body representation, capturing the Euler angle of joints in a manner that is invariant to body shape. Conventionally, we only consider the upper body joints, resulting in $J = 47$. The pose sequences are downsampled into 15 FPS and divided into chunks of 34 frames. While there are other 3D gesture generation datasets [38], these datasets lack detailed emotion annotation cues to train the audio encoder in the proposed method. Hence, we only conduct experiments using BEAT dataset.

Audio preprocessing. Speech sequences are segmented into approximately 2-second intervals, which are temporally aligned with the pose chunks. Each audio chunk is converted into a spectrogram with 128 mel-frequency bins, using a 25ms Hamming window and a 10ms frameshift. To enhance the robustness, we also inject noise for filterbank reconstruction.

Architecture. Except for the audio encoder that leverages AST and the motion decoder that adopts the transformer decoder, all the other modules are implemented by the transformer encoder. The spatial-temporal self-attention module contains a single layer, while all the other transformer-based models consist of 3 layers. The hidden space of all the modules is set to 512.

Denoiser. Following [9], we use 100 diffusion steps in training while 50 in inference. β ranges from 8.5×10^{-4} to 1.2×10^{-2}.

5 Experiments

5.1 Experiment settings

Evaluation metrics To quantitatively evaluate the generation quality, we adopted the following metrics, which are used in [21]:

- **Fréchet gesture distance (FGD)** compares the distribution between the synthetic gesture and ground truth motion on the feature space. The motion features are extracted by a pre-trained autoencoder.
- **Diversity (Div)** measures the L1 variance of the generated gesture.
- **Beat alignment (BA)** computes the synchrony between the audio beats and the motion kinematic beats in the generated gesture.
- **Semantic-relevant gesture recall (SRGR)** uses the semantic score, provided by the BEAT dataset, as the weight for the Probability of Correct Keypoint (PCK) between synthesized and ground truth motion.

Compared with the other metrics that only evaluate certain aspects of the generated gesture, we consider FGD as the main metric as it evaluates the overall motion feature distribution.

Baseline methods We evaluated the performance of our model against CaMN [21], Trimodal [37], LivelySpeaker [39], and DiffGesture [40]. To evaluate the performance of LivelySpeaker, we considered two settings: Rhythm-Aware Generator (RAG) only and the full framework (Semantic-Aware Generator and RAG). We utilized the officially released checkpoints of CaMN and LivelySpeaker, and the Trimodal checkpoint re-implemented by [21]. As DiffGesture is originally trained on TED Gesture [37, 38] and TED expressive [23], we retrained the official scripts on the BEAT dataset.

Table 2: Quantitative evaluation on BEAT dataset. We report the experimental result. Under the two settings, our frameworks significantly improve the FGD, indicating the distribution of the synthesized gestures is closer to that of the ground truth. Our framework also achieves the highest BA and SRGR.

Method	w/o post-processing				w/ post-processing			
	FGD ↓	Div	BA ↑	SRGR ↑	FGD ↓	Div	BA ↑	SRGR ↑
Ground truth	-	1710.17	0.89	-	-	1710.17	0.89	-
Trimodal [37]	372.44	1334.84	0.70	0.12	367.80	1335.17	0.70	0.12
CaMN [21]	265.52	2042.10	0.80	0.16	259.88	2083.16	0.80	0.16
DiffGesture [40]	215.66	3987.37	0.92	0.18	195.21	3806.84	0.92	0.19
LivelySpeaker (RAG) [39]	158.27	667.75	0.91	0.20	168.50	675.53	0.91	0.20
LivelySpeaker (full; RAG+SAG) [39]	169.54	475.38	0.92	0.21	180.87	482.28	0.91	0.21
SemGest (Ours)	**82.65**	1212.82	**0.93**	**0.24**	**83.42**	1212.96	**0.93**	**0.24**

Table 3: Ablation study on different components in the proposed method.

Method	w/o post-processing				w/ post-processing			
	FGD ↓	Div	BA ↑	SRGR ↑	FGD ↓	Div	BA ↑	SRGR ↑
Full	**82.652**	1212.824	**0.930**	0.243	**83.419**	1212.958	0.929	0.243
w/o Semantic projection module	169.540	1460.758	0.929	0.240	162.143	1460.762	0.929	0.240
Concatenation	177.421	1224.985	0.929	**0.251**	170.021	1224.873	0.928	0.252
Speech feature only	142.728	2995.257	**0.930**	0.244	141.945	2994.030	**0.931**	0.244
Semantic feature only	130.930	2044.934	0.929	**0.251**	130.393	2043.692	0.930	**0.251**
Replace AST with TCN	136.100		0.929					

For a fair comparison, both our framework and the baseline models were trained on Speaker-2, 4, 6, 8 from BEAT dataset [21], consisting of 16 hours of speech. The length of the seed pose M is set to 4 frames. During testing, we generated the whole pose sequence (around 1 minute long) chunk-by-chunk, leveraging the previous M chunks to generate the current pose chunk.

To mitigate the jittering artifacts presented in the gestures generated by certain methods, we applied Kalman smoothing as a post-processing step. We report the performance both before and after applying this smoothing technique.

5.2 Quantitative evaluation

Table 2 presents the performance of the quantitative evaluation. Our model outperforms all the other baselines in terms of FGD, indicating the generated gestures exhibit a distribution closer to that of the ground truth. Additionally, our model achieves the best score on both BA and SRGR. This demonstrates a balance between rhythm-aware and semantic-aware gesture generation. A comparison of the two LivelySpeaker settings reveals that the Semantic-Aware Generator (SAG) improves the SRGR, highlighting the importance of the semantic-motion joint space. Sharing similar objectives with SAG, our model achieves a higher SRGR, demonstrating the effectiveness of the proposed semantic-to-gesture alignment mechanism. Although DiffGesture exhibits a higher Div compared to the ground truth, we observe that it generates unnatural motions occasionally, leading to an abnormal Div score. Similarly, Trimodal and CaMN also produce unnatural poses, characterized by a lack of large movements and poor adherence to speech rhythm. These factors contribute to their poor performance on FGD and BA.

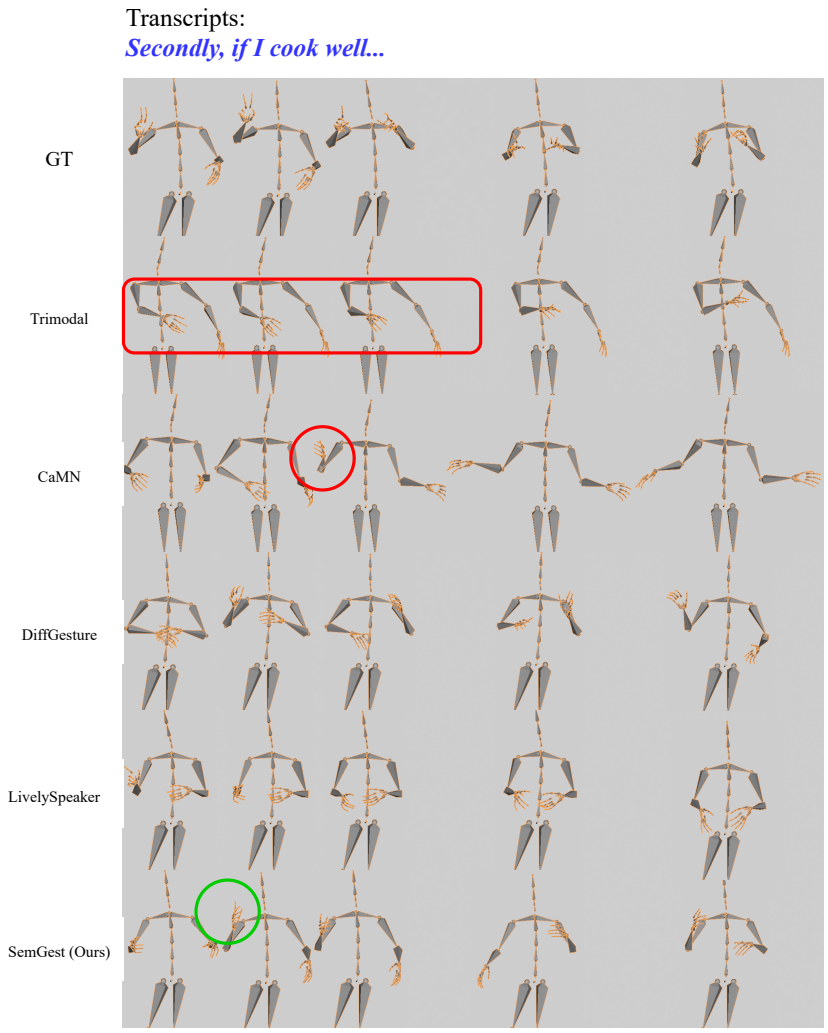
5.3 Qualitative evaluation

In Fig. 4, we visualize and compare the gestures generated by SemGest against baseline methods. It demonstrates that SemGest is able to generate semantic-aware gestures. For instance, SemGest demonstrates gesture *"two"*, which is semantically aligned with the saying *"secondly"*. SemGest can also generate gestures that align with speech rhythm. When saying *"gun fire"* with an astonished tone, the generated hands display a blow that is synchronized with the audio beat. This demonstrates SemGest's ability to balance semantic-aware gestures and rhythm-aware gestures. Besides the proposed method, LivelySpeaker [39] achieves good results in generating high-quality gestures but shows a preference for small movements. In contrast, Trimodal [37] and CaMN [21] tend to produce unnatural gestures, while DiffGesture [40] exhibits redundant gestures and lacks smooth transitions. Please refer to the supplementary materials for the video clips.

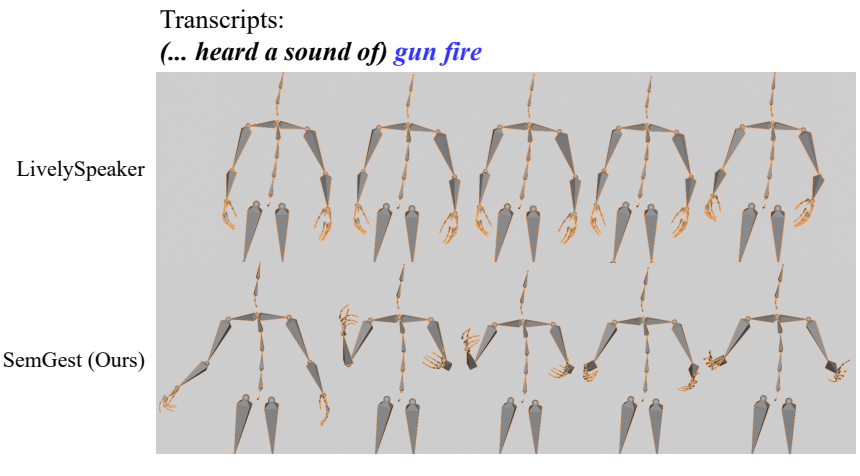
5.4 Ablation study

Semantic-to-gesture alignment. We ablated the semantic projection module by replacing the semantic features with CLIP text embeddings. As shown in Tab. 3, our full framework outperforms the ablated model in terms of FGD and SRGR. This validates the effectiveness of extracting gesture-relevant semantic information.

Feature fusion mechanism. We replaced our proposed dual-branch cross-attention mechanism with a simple concatenation of speech and semantic features along the temporal axis. The results, presented in Tab. 3, show that our full framework outperforms on FGD and achieves comparable BA and SRGR scores. This underscores the



(a)



(b)

Figure 4: Qualitative evaluation. We visualize ground thruth (GT) gestures the those generated by SemGest and baselines. (a) As shown in the green circle, SemGest exhibit *"two"* when saying *"secondly"*. We observe artifacts in Trimodal [37] and CaMN [21], highlighted in red. (b) Compared to LivelySpeaker [39], SemGest generates diverse and rhythm-synchronized gestures.

Table 4: Quantitative evaluation on BEATv2.0.1.

Method	FGD ↓	Div	BA ↑	SRGR ↑
EMAGE [20]	0.37	13.23	0.75	0.33
SynTalker [7]	**0.33**	12.98	0.74	**0.34**
Ours	0.72	9.20	**0.78**	**0.34**

effectiveness of the proposed feature fusion mechanism in improving the quality of the generated co-speech gesture.

The impact of speech and semantic modality. In this experiment, we only incorporated a single modality to evaluate its contribution to the generated co-speech gesture. The result is shown in Tab. 3. It is interesting to note that both the speech-only and text-only models outperform the early fusion approach, highlighting the importance of proper feature integration. Furthermore, the overall performance of the semantic-only model is better than the speech-only model, emphasizing the crucial role of semantic feature extraction in generating co-speech gestures.

Ablation of AST. We conducted an experiment in which we replaced the AST with the TCN network in EMAGE [20] and evaluated the performance using the FGD score. The result, reported in Tab. 3, demonstrates the effectiveness of AST in generating gestures that better align with the ground truth distribution and are rhythm-synchronized.

6 Discussion and Future Work

SMPL-X human body representation. We report a quantitative evaluation result on the BEATv2.0.1, which adopts SMPL-X human body representation, in Tab. 4. While our method achieves higher FGD compared with baselines, the generated gestures are visually plausible. Further studies and experiments are needed, and we will focus on this in our future work. Detailed analysis is provided in the supplementary materials.

Holistic co-speech gesture generation. Our model can only generate upper-body gestures. In our future work we will focus on extending the method to generate holistic gestures that include the lower-body and facial expressions.

Post-processing. In our future work, we will investigate methods to improve the smoothness and remove the post-processing step.

7 Conclusion

We present SemGest, a framework for co-speech gesture generation. To generate semantic-aware gestures, we devise a semantic-to-gesture alignment scheme to extract gesture-relevant semantic features from transcripts. Our proposed feature fusion mechanism models the correlation and adaptively integrates speech and semantic features, leading to a robust representation for diffusion-based model to generate expressive co-speech gestures. Extensive experiments demonstrate the superiority and effectiveness of our model.

Acknowledgments

This work was partly supported by Japan Society for the Promotion of Science (JSPS) KAKENHI JP24H00733.

References

[1] Chaitanya Ahuja, Dong Won Lee, Yukiko I Nakano, and Louis-Philippe Morency. 2020. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, 248–265.

[2] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. 2021. A spatio-temporal transformer for 3D human motion prediction. In *Proceedings of the 2021 International Conference on 3D Vision*. IEEE, 565–574.

[3] Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. 2021. Speech2AffectiveGestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2027–2036.

[4] Justine Cassell. 1998. *Computer Vision for Human–Machine Interaction: A Framework for Gesture Generation and Interpretation*. Cambridge university press. https://api.semanticscholar.org/CorpusID:61011414

[5] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*. 413–420.

[6] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. BEAT: the behavior expression animation toolkit. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*. 477–486.

[7] Bohong Chen, Yumeng Li, Yao-Xiang Ding, Tianjia Shao, and Kun Zhou. 2024. Enabling Synergistic Full-Body Control in Prompt-Based Co-Speech Motion Generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*.

[8] Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. 2024. DiffSHEG: A diffusion-based approach for real-time speech-driven holistic 3D expression and gesture generation. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7352–7361.

[9] Kiran Chhatre, Nikos Athanasiou, Giorgio Becherini, Christopher Peters, Michael J Black, Timo Bolkart, et al. 2024. Emotional speech-driven 3D body animation via disentangled latent diffusion. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1942–1953.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv* abs/2010.11929 (2020). https://api.semanticscholar.org/CorpusID:225039882

[11] Chencan Fu, Yabiao Wang, Jiangning Zhang, Zhengkai Jiang, Xiaofeng Mao, Jiafu Wu, Weijian Cao, Chengjie Wang, Yanhao Ge, and Yong Liu. 2024. MambaGesture: Enhancing Co-Speech Gesture Generation with Mamba and Disentangled Multi-Modality Fusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*.

[12] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F Troje, and Marc-André Carbonneau. 2023. ZeroEGGS: Zero-shot Example-based Gesture Generation from Speech. In *Computer Graphics Forum*, Vol. 42. Wiley Online Library, 206–216.

[13] Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. In *Proceedings of the Interspeech 2021*. 571–575. doi:10.21437/Interspeech.2021-698

[14] Ikhsanul Habibie, Mohamed Elgharib, Kripasindhu Sarkar, Ahsan Abdullah, Simbarashe Nyatsanga, Michael Neff, and Christian Theobalt. 2022. A motion matching-based framework for controllable gesture synthesis from speech. In *ACM SIGGRAPH 2022 conference proceedings*. 1–9.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[16] Taras Kucherenko, Patrik Jonell, Sanne Van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 242–250.

[17] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. 2021. Audio2Gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision*. 11293–11302.

[18] Yuanzhi Liang, Qianyu Feng, Linchao Zhu, Li Hu, Pan Pan, and Yi Yang. 2022. SEEG: Semantic energized co-speech gesture generation. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10473–10482.

[19] Fengqi Liu, Hexiang Wang, Jingyu Gong, Ran Yi, Qianyu Zhou, Xuequan Lu, Jiangbo Lu, and Lizhuang Ma. 2024. Emphasizing Semantic Consistency of Salient Posture for Speech-Driven Gesture Generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*.

[20] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. 2024.

EMAGE: Towards Unified Holistic Co-Speech Gesture Generation via Expressive Masked Audio Gesture Modeling. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1144–1154.

[21] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022. BEAT: A Large-Scale Semantic and Emotional Multi-Modal Dataset for Conversational Gestures Synthesis. *arXiv preprint arXiv:2203.05297* (2022).

[22] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (2023).

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

[24] Xiaofeng Mao, Zhengkai Jiang, Qilin Wang, Chencan Fu, Jiangning Zhang, Jiafu Wu, Yabiao Wang, Chengjie Wang, Wei Li, and Mingmin Chi. 2024. MDT-A2G: Exploring Masked Diffusion Transformers for Co-Speech Gesture Generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*.

[25] Xingqun Qi, Chen Liu, Lincheng Li, Jie Hou, Haoran Xin, and Xin Yu. 2024. EmotionGesture: Audio-driven diverse emotional co-speech 3D gesture generation. *IEEE Transactions on Multimedia* (2024).

[26] Xingqun Qi, Jiahao Pan, Peng Li, Ruibin Yuan, Xiaowei Chi, Mengfei Li, Wenhan Luo, Wei Xue, Shanghang Zhang, Qifeng Liu, et al. 2024. Weakly-Supervised Emotion Transition Learning for Diverse 3D Co-speech Gesture Generation. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10424–10434.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:231591445

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

[29] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising Diffusion Implicit Models. *arXiv:2010.02502* (October 2020). https://arxiv.org/abs/2010.02502

[30] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. 2022. MotionCLIP: Exposing human motion generation to clip space. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*. Springer, 358–374.

[31] Sen Wang, Jiangning Zhang, Weijian Cao, Xiaobin Hu, Moran Li, Xiaozhong Ji, Xin Tan, Mengtian Li, Zhifeng Xie, Chengjie Wang, and Lizhuang Ma. 2024. MMoFusion: Multi-modal Co-Speech Motion Generation with Diffusion Model. *ArXiv* abs/2403.02905 (2024). https://api.semanticscholar.org/CorpusID:268247486

[32] Hongxia Xie, Ming-Xian Lee, Tzu-Jui Chen, Hung-Jen Chen, Hou-I Liu, Hong-Han Shuai, and Wen-Huang Cheng. 2023. Most important person-guided dual-branch cross-patch attention for group affect recognition. In *Proceedings of the 19th IEEE/CVF International Conference on Computer Vision*. 20598–20608.

[33] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. 2023. DiffuseStyleGesture: Stylized audio-driven co-speech gesture generation with diffusion models. *arXiv preprint arXiv:2305.04919* (2023).

[34] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, and Haolin Zhuang. 2023. QPGesture: Quantization-based and phase-guided motion matching for natural speech-driven gesture generation. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2321–2330.

[35] Sicheng Yang, Zunnan Xu, Haiwei Xue, Yongkang Cheng, Shaoli Huang, Mingming Gong, and Zhiyong Wu. 2024. Freetalker: Controllable speech and text-driven gesture generation based on diffusion models for enhanced speaker naturalness. In *Proceedings of the 2024 International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 7945–7949.

[36] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. 2023. Generating holistic 3D human motion from speech. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 469–480.

[37] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics* 39 (2020), 1 – 16. https://api.semanticscholar.org/CorpusID:221507915

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

[39] Yihao Zhi, Xiaodong Cun, Xuelin Chen, Xi Shen, Wen Guo, Shaoli Huang, and Shenghua Gao. 2023. LivelySpeaker: Towards Semantic-Aware Co-Speech Gesture Generation. In *Proceedings of the 19th IEEE/CVF International Conference on Computer Vision*. 20807–20817.

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

Supplemental Materials

A Supplementary Video

We provide the results compared to baseline models [21, 37, 39, 40]. Specifically, we present:

- **A comparison with baseline models.** We provide video clips of Fig. 4a.
- Ablation comparison of different modules.
- Gesture generation on various emotion labels.
- A comparison of results with and without the post-processing step.

B Training Loss

As mentioned in Sec. 3.2, additional loss terms Eq. (8), Eq. (9), Eq. (10), and Eq. (11) are adopted to regulate reconstructing and generating gestures in specific joints. In particular, the loss terms are implemented as follows:

$$\mathcal{L}_{joint\_rec} = \mathcal{L}_{Huber}(p_J^{1:N}, \widehat{p}_J^{1:N}) \tag{8}$$

$$\mathcal{L}_{joint\_gen} = \mathcal{L}_{Huber}(p_J^{1:N}, \widetilde{p}_J^{1:N}), \tag{9}$$

$$\mathcal{L}_{joint\_rec\_vel} = \mathcal{L}_{Huber}(\dot{p}_J^{1:N}, \dot{\widehat{p}}_J^{1:N}) \tag{10}$$

$$\mathcal{L}_{joint\_gen\_vel} = \mathcal{L}_{Huber}(\dot{p}_J^{1:N}, \dot{\widetilde{p}}_J^{1:N}) \tag{11}$$

where J represents the specific joints, *e.g.* elbows.

We scale all the loss terms to balance contribution, and the total training loss $\mathcal{L}_{total}$ is:

$$\begin{aligned}\mathcal{L}_{total} = \mathcal{L}_{rec} + \mathcal{L}_{gen} + \mathcal{L}_{rec\_vel} + \mathcal{L}_{gen\_vel} + \mathcal{L}_{ldm} \\ + \lambda_{joint}\mathcal{L}_{joint\_rec} \\ + \lambda_{joint}\mathcal{L}_{joint\_gen} \\ + \lambda_{joint}\mathcal{L}_{joint\_rec\_vel} \\ + \lambda_{joint}\mathcal{L}_{joint\_gen\_vel},\end{aligned} \tag{12}$$

where λ_{joint} is set to 100.

C Evaluation Settings

To compute the FGD score, an autoencoder $\mathcal{M}$ is leveraged to extract the gesture feature. We use the official checkpoint provided by [21], which is trained on Speaker-2, 4, 6, 8, aligned with our training data.

D Analysis of Semantic-to-Gesture Alignment

To obtain the semantic-to-gesture alignment module, a gesture prior encoder consisting of 3-layer Transformer encoder is trained to map the CLIP [27] text embedding space to a pre-trained gesture latent space. Fig. 8 illustrates the training and evaluation pipeline. We conduct quantitative evaluation between applying cosine similarity loss (denoted as CosSimLoss) and MSE as the alignment loss $\mathcal{L}_{align}$ and present the results in Tab. 5. Training the semantic-to-gesture alignment module with MSE loss results in a lower FGD score, indicating the distribution of the generated gesture is closer to the distribution of the ground truth gestures. Additionally, we visualize the joint space of semantic features and gesture latents in Fig. 5. It is obvious that semantic features generated by the one trained with MSE are closer to the gesture embedding. Hence, we adopt MSE as the alignment loss.

E Analysis of Gesture Emotion Recognition

To assess the gesture emotion expressiveness, we train a gesture emotion classifier on the gesture features extracted by the autoencoder $\mathcal{M}$. As illustrated in Fig. 6, chunk-level gesture embeddings are extracted from $\mathcal{M}$, then the concatenation of all the gesture embeddings is fed to the gesture emotion classifier to predict the clip-level emotion label. In Tab. 6, we present the classification result of our framework, baseline methods, and the ablated models. Despite the lower-than-expected accuracy of our method, it is noteworthy that the speech-only model outperforms the full framework and the semantic-only model, indicating the effectiveness of the emotion classification loss used in training the audio encoder. This finding also suggests that transcripts may contain less explicit emotional information and implies the need for a more refined feature fusion mechanism to support both semantical and emotional co-speech gesture generation. Another observation is that the emotion accuracy of ground truth gestures is 65%, indicating room for improvement in gesture emotion recognition. To benefit future studies, we present the confusion matrix of the proposed method in Fig. 7.

F Supplementary Video: SMPL-X Data

BEATv2.0.1 adopts SMPL-X human body representation. It contains 55 joints, including both upper-body and lower-body joints.

Table 5: Quantitative evaluation on the semantic-to-gesture alignment. The autoencoder reported in this tableis trained on all of the 30-speaker data, while the one used in Table 2 and Table 3 of the main paper is trained on speaker-2, 4, 6, 8 only.

Loss	FGD ↓	BA ↑	SRGR ↑
CosSimLoss	1231.46	0.93	0.24
MSE	996.34	0.92	0.18

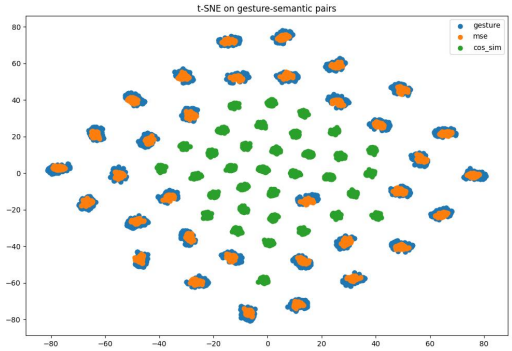


Figure 5: Visualization of the semantic-gesture joint space. We visualize the latent embeddings of the testing set. The blue dots represent the GT gesture embeddings, while the orange dots and the green dots represent the outputs of semantic-to-gesture alignment module trained with MSE and cosine similarity loss, respectively.

Table 6: Evaluation of gesture emotion recognition.

Method	Emotion Accuracy ↑
Ground truth	65.63
Trimodal [37]	28.13
CaMN [21]	48.44
DiffGesture [40]	54.69
LivelySpeaker (RAG) [39]	46.88
LivelySpeaker (full) [39]	51.56
SemGest (full)	45.31
SemGest (w/o semantic projection)	42.19
SemGest (concatenation)	39.06
SemGest (speech-only)	51.56
SemGest (semantic-only)	43.75

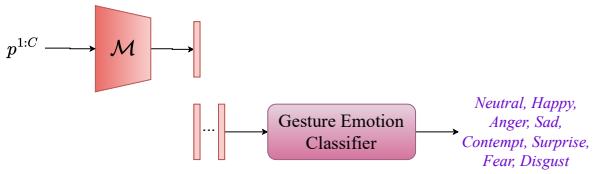


Figure 6: The framework of gesture emotion classifier.

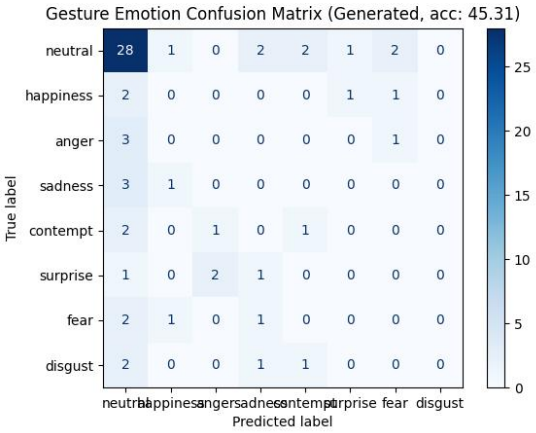


Figure 7: Gesture emotion recognition confusion matrix of the proposed method.

Following [7, 20], we downsample the pose sequences into 30 FPS, divide the sequence into chunks of 64 frames, and train the proposed framework on the same training set. As SemGest can only produce upper-body gestures while the baseline models [7, 20] generate holistic gesture, we use the identical lower-body gesture and facial expression in quantitative evaluation and visualization. The visualization video clips compared to baseline models [7, 20] are provided in the supplementary videos.

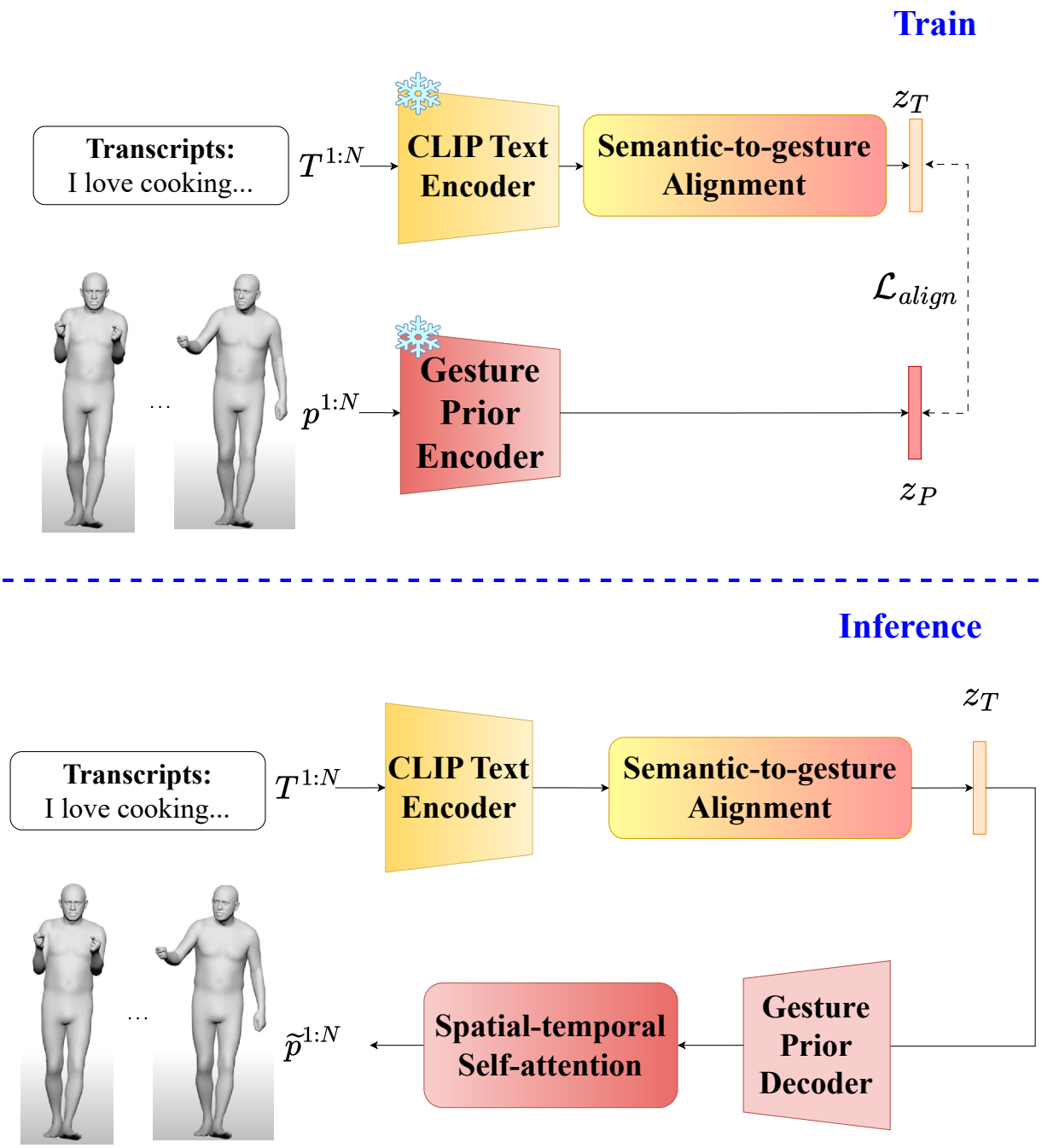


Figure 8: The training and evaluation pipeline of semantic-to-gesture alignment.