SemGes: Semantics-aware Co-Speech Gesture Generation using Semantic Coherence and Relevance Learning

Lanmiao Liu[1,2,3] Esam Ghaleb[1,2] Aslı Özyürek[1,2] and Zerrin Yumak[3]
[1]Max Planck Institute for Psycholinguistics [2]Donders Institute for Brain Cognition and Behaviour [3]Utrecht University
{lanmiao.liu, esam.ghaleb, asli.ozyurek}@mpi.nl z.yumak@uu.nl

Abstract

Creating a virtual avatar with semantically coherent gestures that are aligned with speech is a challenging task. Existing gesture generation research mainly focused on generating rhythmic beat gestures, neglecting the semantic context of the gestures. In this paper, we propose a novel approach for semantic grounding in co-speech gesture generation that integrates semantic information at both fine-grained and global levels. Our approach starts with learning the motion prior through a vector-quantized variational autoencoder. Built on this model, a second-stage module is applied to automatically generate gestures from speech, text-based semantics and speaker identity that ensures consistency between the semantic relevance of generated gestures and co-occurring speech semantics through semantic coherence and relevance modules. Experimental results demonstrate that our approach enhances the realism and coherence of semantic gestures. Extensive experiments and user studies show that our method outperforms state-of-the-art approaches across two benchmarks in co-speech gesture generation in both objective and subjective metrics. The qualitative results of our model, code, dataset and pre-trained models can be viewed at https://semgesture.github.io/.

1. Introduction

Human language is inherently multimodal, with gestures and speech complementing each other to convey pragmatic and semantic information [21, 35]. Co-speech gestures are non-verbal cues that are uniquely related to co-occurring speech, pragmatically, semantically, and temporally. For example, *representational iconic gestures* that visually express the semantic content of speech and interact with spoken language [13, 14, 19, 21, 41]. A long-standing goal in Computer Vision is to create digital humans that use non-verbal cues in sync with speech. Gesture generation—synthesizing movements from co-occurring speech, masked motion, or speaker identity—has advanced to enhance AI agents' expressiveness and realism [29]. However, much of the focus has gone into generating rhythmic beat gestures with limited semantic information, leaving representational gestures that convey semantic messages (e.g., iconic) less explored [29, 40].

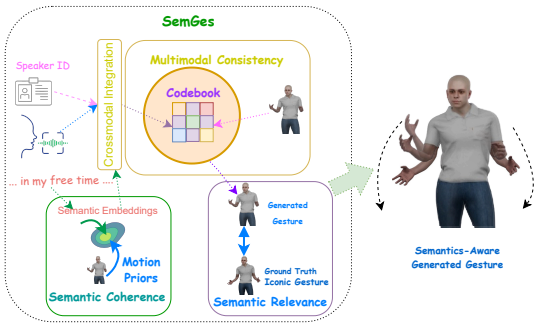

Figure 1. *SemGes* integrates audio, text-based semantics, and speaker identity to produce both contextually relevant (discourse-level) and fine-grained (local) gestures. A semantic coherence module aligns text and motion embeddings. The multimodal consistency loss synchronizes the quantized multimodal representations to match the quantized learned motion features for final speech-driven semantics-aware gesture generation. The semantic relevance loss selectively emphasizes gestures with semantic annotations.

Generating spontaneous and semantically rich gestures from speech comes with multiple challenges. First, it requires capturing global discourse-level information and local fine-grained details (e.g., salient words) to generate speech-driven gestures that reflect the intended meaning and align with speech temporally and semantically. Second, existing methods often generate repetitive and short sequences that do not span the full range of expressive motions required for natural communication. To leverage semantics when generating gestures, researchers have attempted to align motion with speech representations at a global level, e.g., by leveraging pre-

trained semantic representations such as CLIP [58] or focusing on semantically important keywords [6, 57]. Nonetheless, they often fail to (i) unify global and local semantic modelling within a single framework and (ii) exploit the relevance of the semantic information in guiding gesture generation [27]. At the same time, raw audio features and speaker identity are relevant to the timing and style of gestures. In this paper, we address these limitations by integrating speech, speech semantics, and gesturing style, exploiting semantic information at different levels.

Specifically, we propose a two-stage framework, namely, SemGes, that integrates speech, text-based semantics, and speaker identity into a unified gesture-generation model (see Figure 1). In stage 1, we build motion prior of holistic gestures (*i.e.*, body and hands) by training a vector-quantized variational autoencoder (VQ-VAE) to learn an efficient, compositional motion latent space. This stage results in a robust motion encoder & decoder and quantized codebooks that can reconstruct naturalistic gestures while allowing the reuse of learned codebook entries. Stage 2 leverages the learned motion priors to drive gesture synthesis by fusing three modalities using a cross-modal Transformer encoder: (i) text-based semantics, (ii) raw-audio speech features, and (iii) speaker identity for style consistency. We impose a *semantic coherence* loss that aligns text-based embeddings with the VQ-VAE motion latent space and a *semantic relevance* loss that emphasises representational gestures (*e.g.* iconic and metaphoric gestures). A multimodal consistency objective ensures the fused multimodal representations are compatible with the learned motion codebooks, enabling the generation of gestures that are both semantically rich and visually natural. Finally, we introduce a simple but effective long-sequence inference strategy that smoothly combines overlapping motion clips for extended durations. To summarize our contributions,

- We introduce a novel framework, *SemGes*, that first learns a robust VQ-VAE motion prior for body and hand gestures, and then generates gestures driven by fused speech audio, text-based semantics, and speaker identity in a cross-modal transformer.
- Our method jointly captures discourse-level context via a semantic coherence loss and fine-grained representational gestures (e.g., iconic, metaphoric) via a semantic relevance loss.
- We propose an overlap-and-combine inference algorithm that maintains smooth continuity over extended durations.
- Extensive experiments on two benchmarks, namely, the BEAT [27] and TED Expressive [33] datasets show that our method outperforms recent baselines in both objective metrics (e.g., Fréchet Gesture Distance (FGD), diversity, semantic alignment) and user judgment of generated gestures.

2. Related Work

Data-driven Co-Speech Gesture Generation. Current gesture generation approaches are based on generative deep neural networks. These approaches use advanced models such as Transformers [28], Generative Adversarial Networks [23], Normalizing Flows [18, 30], Vector Quantized Variational Autoencoder(VQ-VAE) [16] and Denoising Diffusion Probabilistic Models [47]. In addition, researchers have explored the impact of different model inputs on the naturalness and appropriateness of generated gestures. Various modal inputs have been used, such as text [15], audio [50, 59], image [32, 39], and speaking style [3]. For a comprehensive survey, we refer to Nyatsanga et al. [40]. Although there have been significant improvements in this field, current methods fall short in generating semantically grounded gestures at a fine-grained level. In other words, while the generated motions look convincing at first glance, they do not match well with the meaning of the text, or they mostly focus on beat-type gestures.

Semantics-aware Co-Speech Gesture Generation. A group of work focused on semantics-aware gesture generation where the semantic information is handled in two ways: global semantics and local semantics. Methods that focus on global semantic information [10, 22, 58] align gestures with text or audio, but they fall short in generating gesture types matching the semantic context, such as iconic, metaphoric and deictic gestures. To capture a wider range of semantic gestures, works like [5, 6, 27] adopt local semantic-aware modelling by integrating the semantic salient words to the neural network. However, these approaches often fail to ensure that the generated gestures align with both the broader audio or textual context and a combination of global and local semantics. Liang et al. [26], Voß and Kopp [49] incorporate both global and local semantics, however, they require extensive annotations. Recently, Zhang et al. [57] employed a generative retrieval framework based on LLMs to address the sparsity problem in datasets with semantic gestures. However, they do not explicitly model the different types of gestures [27] or gesture phases [12] grounded in linguistic research. Moreover, there is still not enough understanding of the impact of different annotations and fine-grained semantics.

Substantial research [6, 28, 57, 58] focused on two-stage latent space generative modelling to overcome the limitations of co-speech gesture generation and to generate more naturalistic and diverse gestures.

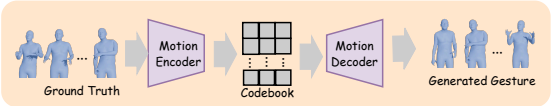

Figure 2. We pre-train two VQ-VAEs by reconstructing body and hand motions with a dedicated codebook for each.

These approaches first learn a latent space and then model gestures probabilistically, effectively integrating the strengths of different methods in different stages. Liu et al. [28], Zhang et al. [57] capture complex dependencies in the latent space using VQ-VAE while Zhi et al. [58] employs CLIP [46, 56] to align text and motion embeddings. Ao et al. [6] introduces a diffusion-based model that leverages semantic awareness, while Liu et al. [28] utilizes a transformer-based approach to generate holistic body gestures. In contrast with two-stage generative modeling approaches, end-to-end methods such as [5, 59], are prone to jittering artifacts, especially in hand-motion generation.

Our model contributes to the research line on semantics-aware co-speech gesture generation by taking into account both global and local semantics. Inspired by the previous work, we employ a two-stage latent space generative modelling for high-quality motion representation. We learn semantic coherence between text and gestures globally with cosine similarity. Moreover, our model takes into account the semantic relevancy of gesture types with minimally required annotations. In contrast with other semantic learning models, we focus on annotations with different gesture types embedded in linguistic research. Our work is closest to CAMN [27] in that sense; however, CAMN does not include semantic coherence learning by aligning text and gestures' latent space globally.

3. Methodology

We propose a two-stage approach that generates co-speech gestures by grounding them in raw speech, text-based semantics, and speaker identity. In Section 3.1, we introduce a VQ-VAE encoder-decoder that learns a robust motion prior. Section 3.2 details our gesture synthesis and inference pipeline based on speech, semantics, and identity.

Problem formulation. Our goal is to generate hand gestures $\boldsymbol{G}^h = (g_1^h, \ldots, g_T^h) \in \mathbb{R}^{T\times J}$ and body gestures $\boldsymbol{G}^b = (g_1^b, \ldots, g_T^b) \in \mathbb{R}^{T\times J}$, where T is the number of time steps and J the number of joints (e.g., 38 for hands, 9 for body). Each motion vector g_t^h or g_t^b is encoded in a Rot6D representation, capturing joint rotations at time t.

To model human motion of body and hands, we first learn a motion generator $\mathcal{M}_g$ (Stage 1), which synthesizes a plausible motion sequence:

$$\arg\min_{\mathcal{M}_g} \left\| \boldsymbol{G} - \mathcal{M}_g(g_1, \ldots, g_T) \right\|. \tag{1}$$

Next, we condition on (i) the raw input audio $\boldsymbol{A} = (a_1, \ldots, a_T)$, (ii) the speaker identity embedding I, and (iii) the text-based semantic embeddings of the speech $\boldsymbol{S} = (s_1, \ldots, s_T)$. Our second-stage model $\mathcal{M}_{a,s,i}$ uses these inputs to generate a latent sequence that the motion generator $\mathcal{M}_g$ then decodes into naturalistic gestures:

$$\arg\min_{\mathcal{M}_{a,s,i}} \left\| \boldsymbol{G} - \mathcal{M}_g(\mathcal{M}_{a,s,i}(\boldsymbol{A}, \boldsymbol{S}, I)) \right\|. \tag{2}$$

3.1. Stage 1: Learning Efficient Codebooks & Compositional Motion Priors

Realistic co-speech gestures require modelling the sequential motion of both body and hand joints. Rather than learning a single representation for the entire body, we adopt a compositional approach, using a discrete codebook of learned representations specific to each part (hands & body). Any gesture motion can then be represented by selecting appropriate codebook entries. Following [28, 48, 53], we employ a VQ-VAE architecture (see Fig. 2) with encoder $\mathcal{E}_m$ and decoder $\mathcal{D}_m$. Given hand motion $\boldsymbol{G}^h \in \mathbb{R}^{T\times J}$ and body motion $\boldsymbol{G}^b \in \mathbb{R}^{T\times J}$, the encoder produces latent vectors $\hat{z}^h$ and $\hat{z}^b$, which are quantized by selecting the nearest entries in the codebooks. Formally,

$$\mathbf{q}(\hat{\boldsymbol{z}}) = \arg\min_{z^i \in \mathcal{Z}} \left\| \hat{z}^j - z^i \right\|, \tag{3}$$

where z^i are the learned codebook entries, and $\hat{z}^j$ denotes an element of the latent vector for either hand or body. We train the VQ-VAE via a straight-through gradient estimator, minimizing:

$$\begin{aligned} \mathcal{L}_{\text{VQ-VAE}} = {} & \|\mathbf{g} - \hat{\mathbf{g}}\|^2 + \|\dot{\mathbf{g}} - \dot{\hat{\mathbf{g}}}\|^2 + \|\ddot{\mathbf{g}} - \ddot{\hat{\mathbf{g}}}\|^2 \\ & + \|\text{sg}[\mathbf{E}(\mathbf{g})] - \mathbf{q}(\hat{\boldsymbol{z}})\|^2 + \|\mathbf{E}(\mathbf{g}) - \text{sg}[\mathbf{q}(\hat{\boldsymbol{z}})]\|^2, \end{aligned} \tag{4}$$

where the first three terms reconstruct joint positions, velocities, and accelerations, and the last two terms implement the VQ-VAE commitment loss [48].

By the end of this stage, we have motion (m) encoder ($\mathcal{E}_m$), decoder ($\mathcal{D}_m$) and codebooks ($Quant^m(\cdot)$) for hands and body. In the next section (Section 3.2), we show how this discretized motion of hands and body guides speech, semantics and speaker identity-driven generation to produce realized co-speech gestures.

3.2. Stage 2: Speech and Identity Driven Semantic Gesture Generator

This stage focuses on generating gestures conditioned on three inputs: speech embeddings, text-based seman-

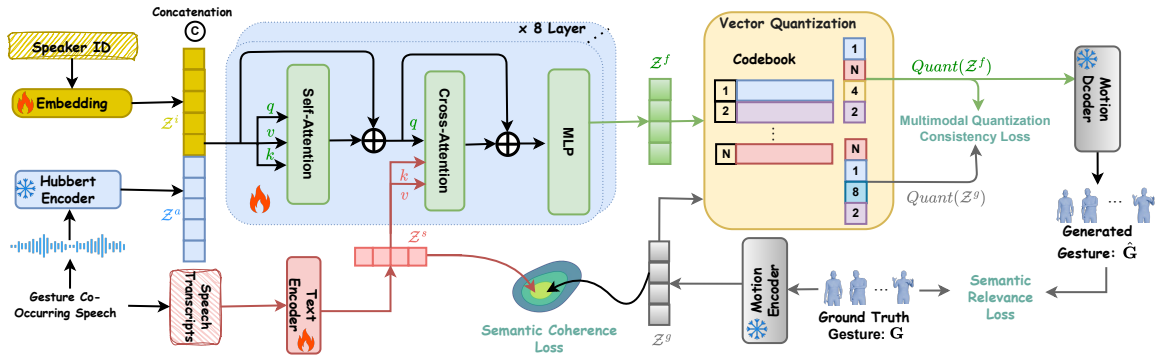

Figure 3. SemGes employs three training pathways: (1) Global semantic coherence, which minimizes latent disparities between gesture and text encoders; (2) Multimodal Quantization learning, where integrated multimodal representation codes are aligned with quantized motion to decode them into hand and body movements; and (3) Semantic relevance learning, which emphasizes semantic gestures.

tic embeddings, and speaker identity. As illustrated in Figure 3, the second-stage architecture has three main modules, which we elaborate on in the following subsections.

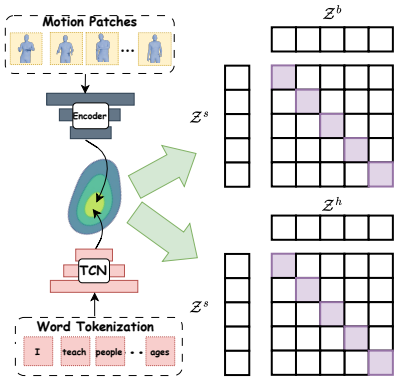

Figure 4. Semantic Coherence Embedding Learning.

3.2.1. Semantic Coherence Embedding Learning

To align text-based semantics with motion embeddings, we introduce a shared embedding space for both motion priors and speech transcripts. Specifically, we embed word tokens using a pre-trained FastText model [8], then feed these embeddings into a trainable text-based semantic encoder $\mathcal{E}_s$. At the same time, we use the *pre-trained motion encoder* $\mathcal{E}_m$ from Stage 1 to encode ground-truth gesture sequences. Thus, for a batch of paired (gesture, transcript) samples, we get:

$$\mathcal{Z}^S = \mathcal{E}_s(S), \quad \mathcal{Z}^h = \mathcal{E}_m^h(G^h), \quad \mathcal{Z}^b = \mathcal{E}_m^b(G^b), \tag{5}$$

where S is the tokenized speech transcript, and G^h and G^b correspond to the ground-truth hand gesture sequence and body gesture sequence, respectively. $\mathcal{Z}^h$ and $\mathcal{Z}^b$ represent the hand and body ground-truth motion encodings from Stage 1, and $\mathcal{Z}^s$ denotes the text-based semantic encoder output.

Semantic Coherence Loss. We maximize the similarity of correct (gesture, transcript) pairs and minimize it for mismatched pairs, enforcing *semantic coherence*. This aligns gestures and textual semantics in a common space while keeping $\mathcal{E}_m$ frozen, as illustrated in Figure 4. We impose the semantic coherence constraint separately on both hand and body movements to align gestures with transcripts in the shared embedding space. Specifically, we introduce two distinct cosine similarity losses: one between the text encoder output and the hand motion latent encoding and another between the text encoder output and the body motion latent encoding. Formally, we minimize:

$$\mathcal{L}_{\text{semantic-coherence}} = 1 - \cos(\mathcal{Z}^h, \mathcal{Z}^s) + 1 - \cos(\mathcal{Z}^b, \mathcal{Z}^s), \tag{6}$$

where the function $\cos(\cdot, \cdot)$ measures cosine similarity.

3.2.2. Crossmodal Integration

SemGes supports multi-modal inputs in the second training stage by combining audio features and speaker identity with semantic text embeddings using a Transformer encoder with self and cross-attention layers (see Figure 3). We begin by extracting HuBERT features [20] from raw speech (keeping the HuBERT encoder frozen). We concatenate audio features $\mathcal{Z}^a$ and speaker embeddings $\mathcal{Z}^i$, resuling in $\mathcal{Z}^r$, which we feed into a *self-attention layer*.

Next, we use a *cross-attention layer* that takes $\mathcal{Z}^r$ as the query and the motion-aligned text-based semantic features $\mathcal{Z}^s$ as the key-value pair. The final hidden representation $\mathcal{Z}^f$ serves as the *multimodal latent code*

that drives gesture synthesis when passed to our vector quantization and VQ-VAE-based motion decoder, which is learned in our first stage (see the yellow box in Figure 3).

Multimodal Quantization Consistency Loss. SemGes quantizes the multimodal latent code using separate hand and body codebooks. To align this code with the ground-truth motion latent codes, we apply independent quantization consistency losses for each component. Specifically, the quantization loss is defined as:

$$\mathcal{L}_{\text{quantization}} = \left\|Quant^h(\mathcal{Z}^f) - Quant^h(\mathcal{Z}^h)\right\|^2 + \left\|Quant^b(\mathcal{Z}^f) - Quant^b(\mathcal{Z}^b)\right\|^2 \tag{7}$$

where $Quant^h(\cdot)$ and $Quant^b(\cdot)$ denote the quantization functions for the hand and body codebooks, respectively.

The multimodal quantization loss aligns the integrated latent code $\mathcal{Z}^f$ with the learned motion code, a critical step since gesture synthesis is obtained through the quantized multimodal representation. Specifically, $\mathcal{Z}^f$ is vector-quantized using separate hand and body codebooks before being decoded by their respective VQ decoders. This process ensures that both hand and body movements contribute effectively to the final output. Formally, the generated gestures are given by:

$$\hat{G} = \hat{G}^h \oplus \hat{G}^b = \mathcal{D}^h_m(Quant^h(\mathcal{Z}^f)) \oplus \mathcal{D}^b_m(Quant^b(\mathcal{Z}^f)), \tag{8}$$

where $\oplus$ denotes concatenation, jointly synthesizing hand and body motions (*i.e.* $\hat{G}$).

3.2.3. Gesture Semantic Relevance Loss

To prioritize the generation of semantically meaningful gestures (e.g., iconic, metaphoric, or deictic), which are less frequent than beat gestures, we introduce a semantic relevance loss. This loss emphasizes semantic annotations while preventing over-penalization of minor deviations. Formally, it is defined as:

$$\mathcal{L}_{\text{semantic-relevance}} = \mathbb{E}\Big[\lambda\, \Psi(\mathbf{G} - \hat{\mathbf{G}})\Big], \tag{9}$$

where λ is the annotation relevance factor, and $\Psi(\cdot)$ is a piecewise function that applies a quadratic penalty for small errors and a linear penalty for larger ones:

$$\Psi(\mathbf{G} - \hat{\mathbf{G}}) = \begin{cases} \frac{1}{2}(\mathbf{G} - \hat{\mathbf{G}})^2, & \text{if } |\mathbf{G} - \hat{\mathbf{G}}| < \alpha, \\ \alpha\Big(|\mathbf{G} - \hat{\mathbf{G}}| - \frac{1}{2}\alpha\Big), & \text{otherwise,} \end{cases} \tag{10}$$

with $\alpha = 0.01$.

Combined Objective Functions. Finally, the overall objective is:

$$\mathcal{L}_{\text{SemGes}} = \mathcal{L}_{\text{semantic-coherence}} + \mathcal{L}_{\text{semantic-relevance}} + \mathcal{L}_{\text{quantization}}, \tag{11}$$

Algorithm 1 Long Gesture Sequence Algorithm

Require: Audio $\mathcal{A}$, aligned speech transcript $\mathcal{S}$, and speaker ID $\mathcal{I}$; Pre-trained codebooks and motion decoder (Stage 1)
Ensure: Long-sequence gesture $\hat{M}$

```
Partition (A, S, I) into clips {(A_c, S_c, I_c)}_{c=1}^{C}
Compute latent representation: Z^f ← Encode(A, S, I)
Quantize: Z^e ← VectorQuantize(Z^f)
Decode initial clip: M̂_1 ← Dec(Z^e)
for each clip c = 2 to C do
    Set first 4 frames of M̂_c to the last 4 frames of M̂_{c−1}
    Generate remaining frames of M̂_c
end for
return M̂
```

which jointly optimizes the model to generate gestures that are semantically coherent at both global and fine-grained levels while remaining faithful to the Stage 1 motion prior.

3.3. Inference of Long Gesture Sequences

Generating long sequences of gestures is challenging due to the need to maintain coherence and smooth transitions. Our Long-Sequence Gesture Motion algorithm (Alg. 1) addresses these challenges by partitioning the input speech, transcript, and speaker identity into aligned clips. For each clip, a multimodal latent representation is computed using our cross-modal encoder, vector-quantized via the Stage 1 codebooks, and decoded into gesture motions. Overlapping 4-frame segments between clips provides continuity, resulting in extended, naturalistic gesture sequences.

4. Experimental Setup

4.1. Datasets

Our proposed methodology is evaluated on two benchmarks, namely, BEAT [27] and the TED expressive dataset [33]. **The BEAT dataset** consists of 76 hours of multimodal recordings, which include speech audio recordings, speech transcriptions, and, more importantly, motion data collected from 30 participants, leveraging Motion Capture (MOCAP) technology. The participants expressed emotions in eight distinct scenarios across four languages. The motion data contains joint rotation angles, which were designed for consistency across varying body sizes. **The TED Expressive dataset** [33] is segmented from TED Talk videos into smaller shots based on scene boundaries. Liu et al. [33] extracted each frame's 2D human pose using OpenPose BEAT [9]. Using these 2D pose priors, ExPose [43] was

Table 1. Comparison of SemGesGen with other methods on the BEAT and TED-Expressive datasets. For BEAT, we compare with CaMN [27], DiffGesture [59], LivelySpeaker [58], and DiffSheg [10]. The same methods are evaluated on TED-Expressive. SRGR is not applicable (denoted with –) for TED-Expressive as it does not contain annotations for semantic relevance of gestures.

BEAT					TED-Expressive			
Method	**FGD ↓**	**BC ↑**	**Diversity ↑**	**SRGR ↑**	**Method**	**FGD ↓**	**BC ↑**	**Diversity ↑**
CaMN [27]	8.510	0.797	206.789	0.231	CaMN [27]	7.673	0.642	156.236
DiffGesture [59]	9.632	0.876	210.678	0.106	DiffGesture [59]	9.326	0.662	119.889
LivelySpeaker [58]	13.378	0.891	214.946	0.229	LivelySpeaker [58]	8.145	0.691	119.231
DiffSheg [10]	6.623	**0.922**	257.674	0.250	DiffSheg [10]	8.457	**0.712**	108.972
SemGes (Ours)	**4.467**	0.453	**305.706**	**0.256**	SemGes (Ours)	**7.263**	0.671	**302.772**

employed to annotate the 3D upper body keypoints, including 13 upper body joints and 30 finger joints. Both datasets' training and validation samples are divided into 34-frame clips.

Cross-Validation. We evaluate our approach on the BEAT dataset, following the protocol in [27], where the data is randomly split into a 19:2:2 ratio for training, validation, and testing. Similarly, for the TED Expressive dataset, we adapt the protocol in [33], using a random split of 8:1:1 for training, validation, and testing.

Implementation Details. The details of the model architectures and training are provided in Section 2 of the Supplementary Materials.

4.2. State-of-the-Art Baselines

We compare SemGes against a set of representative state-of-the-art models that focus on semantic-driven gesture generation. The selected models achieved strong performance on the BEAT and TED-Expressive datasets, making them suitable for a fair comparison with our method. The selected models are as follows:

1. **Cascaded Motion Network(CaMN) [27]** is the current benchmark model for the BEAT dataset. CaMN is based on LSTMs and integrates multiple input modalities, including audio, text, facial expressions, and emotion. Additionally, like SemGes, it leverages semantic relevance annotations to enhance gesture generation.
2. **DiffSHEG [10]** is a state-of-the-art diffusion-based model for real-time speech-driven holistic gesture generation. It is conditioned on noisy motion, audio, and speaker ID. DiffSHEG introduces a Fast Out-painting-based Partial Autoregressive Sampling method to efficiently generate arbitrary-length sequences in real time.
3. **LivelySpeaker [58]** generates semantically and rhythmically aware co-speech gestures by leveraging an MLP-based diffusion model. The model conditions gesture generation on text, noised motion, speaker ID, and audio to enable text-driven gesture control while incorporating global semantics.
4. **DiffGes [59]** models the diffusion and denoising processes within the gesture domain, enabling the generation of high-fidelity, audio-driven gestures conditioned on both audio and gesture inputs. Several recent studies[6, 26] have also demonstrated strong performance in this area.

We exclude certain models from our comparison. For instance, SEEG [26] and [57] rely on additional data annotations (e.g., Semantic Prompt Gallery or ChatGPT-generated annotations) that are not uniformly available In addition, other works, such as Ao et al. [6], Pang et al. [42], Zhang et al. [57], are excluded from our analysis due to the inaccessibility of their codebase. Voß and Kopp [49] is omitted due to its high computational cost and the unavailability of annotations.Liu et al. [31, 34], Mughal et al. [36, 37], Ng et al. [38], Yi et al. [54] are excluded as they primarily focus on holistic gestures with face and mesh data, which fall outside the scope of this work. Similarly, Chhatre et al. [11], Qi et al. [44] are excluded, as their emphasis lies in emotion-driven gesture generation rather than the semantic aspects. Furthermore, Ahuja et al. [1, 2], Alexanderson et al. [4], Habibie et al. [17], Liu et al. [33], Sun et al. [45], Yang et al. [51], Ye et al. [52] are omitted due to their lack of relevance to semantic-driven gesture generation.

5. Quantitive Objective Evaluations

Evaluation Metrics. We employ four standard objective metrics for evaluating the quality of gesture generation, namely, Fréchet Gesture Distance (FGD) [55], Beat Consistency Score (BC) [25], Diversity [24], and Semantic-Relevant Gesture Recall (SRGR) [27].

FGD measures how the generated gestures resemble real motion distributions by embedding sequences into a latent space via a pre-trained autoencoder. In contrast, **BC** focuses on synchronization with speech, measuring the alignment between speech onsets (audio beats) and motion beats, which are identified as velocity minima in

Table 2. Ablation studies evaluating the contributions of key components in SemGes on the BEAT and TED-Expressive Datasets. For BEAT, performance is measured using FGD (lower is better), BC, Diversity, and SRGR, while for TED-Expressive, SRGR is not applicable (denoted as –).

BEAT					TED-Expressive			
Model Variants	**FGD ↓**	**BC ↑**	**Diversity ↑**	**SRGR ↑**	**Model Variants**	**FGD ↓**	**BC ↑**	**Diversity ↑**
Baseline (VQVAE)	10.348	0.564	198.568	0.176	Baseline (VQVAE)	10.682	0.612	114.692
w/o Semantic Coherence Module	8.053	0.556	249.550	0.180	w/o Semantic Coherence Module	7.924	0.623	109.256
w/o Semantic Relevance Module	7.549	**0.573**	245.319	0.195	w/o Semantic Relevance Module	–	–	–
w/ SpeechCLIP Encoder	6.787	0.468	289.621	0.245	w/ SpeechCLIP Encoder	7.341	0.605	245.680
SemGes (Ours)	**4.467**	0.453	**305.706**	**0.256**	SemGes (Ours)	**7.263**	**0.671**	**302.772**

upper-body joints (excluding fingers). Meanwhile, **Diversity** captures the variability of generated motions by computing the average $L1$ distance between pairs of N generated clips. Finally, **SRGR** assesses semantic relevance by determining how well generated gestures align with the annotated semantic gestures. Further details on the objective metrics are included in the Supplementary Materials (Section 1).

Comparisons with Other Models. Table 1 compares the performance of our approach against four baseline methods across four evaluation metrics. As highlighted in the table, SemGes outperforms the baselines in FGD, Diversity, and SRGR.

For the BEAT dataset, our approach achieves the highest SRGR, which we attribute to the exploitation of semantic relevance information in our training objectives. In addition, our approach shows a significant improvement in FGD and Diversity, indicating a closer alignment with the ground truth gesture distribution and a broader range of generated gestures compared to the second-best baselines. The performance on the Beat Consistency (BC) metric is lower for our method. This is expected given our focus on improving semantic awareness of the generated gestures rather than optimizing for strict temporal alignment between rhythmic beat gestures and speech. In addition, the BC metric can be sensitive to rapid, jittery movements; even minor motion artefacts may be mistakenly counted as additional beats, thereby increasing the BC score artificially- a phenomenon also observed in the diffusion-based baselines, as further illustrated in our supplementary video.

We evaluate how our model handles the trade-off between semantic and beat scores by testing the model on beat-dominant gestures (without semantic content). The results show a significantly higher Beat score (0.689) than the full dataset Beat score (0.453). This confirms rhythmic consistency in beat-focused contexts. We provide additional evaluation in the supplementary materials (Section 3) to show how the model handles difficult cases (such as noisy speech or misaligned speech).

Note that the TED Expressive dataset lacks annotations for gesture semantic relevance, so SRGR is not applicable, and the semantic relevance loss was omitted during training. Nevertheless, SemGes produces diverse, naturalistic gestures on TED Expressive, outperforming baselines in FGD and Diversity metrics.

Ablation Study. We evaluate the contributions of key components in SemGes through ablation experiments. First, we assess a baseline VQ-VAE model (Stage 1 only), which uses two stacked encoder-decoder blocks and an MLP. In this experiment, we test its ability to generate gestures, conditioned on audio, masked motion, and speaker identity. As shown in Table 2, this baseline underperforms compared to state-of-the-art methods (Table 1). As a result, we motivate our two-stage design where the VQ-VAE is reserved to learn the motion latent space and Stage 2 leverages speech and identity conditioning to generate gestures.

Next, we examine Stage 2 by removing its components: (i) the Semantic Coherence Loss, (ii) the Semantic Relevance Loss, and (iii) by replacing the HuBERT-based speech encoder with SpeechCLIP. Results in Table 2 show that removing either the Semantic Coherence or Relevance Loss degrades FGD, Diversity, and SRGR scores, highlighting their roles in aligning gesture representations with textual semantics and capturing semantic importance. In addition, replacing the speech encoder results in marginal gains. The semantic encoder is fixed as FastText, which we believe is sufficient to capture the necessary semantic information [7]. Overall, these results confirm the importance of each module in generating semantics-aware gestures.

6. Qualitative & Subjective Evaluations

Visualization Comparisons. Before presenting the subjective ratings of the generated gestures, Figure 5 provides a visual comparison between the ground truth, results from our approach and two baseline models. We use examples from the BEAT dataset. It is clear from the figure that our approach not only achieves better speech-gesture alignment but also produces gestures that are more naturalistic, diverse, and semantically aware. For example, while CaMN generates smooth movements, its gestures tend to be slower and less varied compared to

Figure 5. Comparisons with baselines and ground truth gestures. Compared to the baseline method, our approach generates gestures that are aligned with speech content (semantics). For instance, when the speaker says "remix", our method produces gestures where the character raises both hands to emphasize the word before gradually lowering them—a movement that other methods fail to achieve. Similarly, when uttering "first", our method generates a raised hand gesture, producing an *iconic gesture*.

our model. Additionally, the baseline methods show varying degrees of jitter—DiffGesture shows the highest jitter, followed by LivelySpeaker and DiffSheg, with CaMN displaying the least. Although CaMN includes semantic information, our approach strikes a more effective balance, generating gestures that align with actual motion, as shown also with the objective metrics. Based on these qualitative observations, our subsequent rating study focuses on evaluating gestures produced by the ground truth, our model, CaMN, and DiffSHEG.

User Ratings of Generated Gestures. We conducted a user study using 40-second video clips from the BEAT test set, featuring subjects narrating six topics. Thirty native English speakers from the United Kingdom and the United States participated, with an average age of 36 ± 20 years and a female-to-male ratio of approximately 2:1. Each participant evaluated 24 videos generated by the ground truth, CaMN, DiffSHEG, and our model over a study duration of that lasted on average 27 ± 5 minutes. For data quality, participants were required to pass attention verification questions, *i.e.*, correctly answering at least two out of four questions regarding the narration topic. Participants rated the videos on three criteria: naturalness, diversity, and alignment with speech content and timing on a scale from 1 to 5. The videos were presented in a randomized order to avoid bias. In Section 3 of the Supplementary Materials, we provide screenshots and more details on the user study and interface.

Figure 6 shows that ground-truth gestures received an average rating of 4 across all metrics, establishing an upper bound and validating the participant survey. Our model received the highest ratings among the generated gestures, significantly outperforming CaMN and DiffSHEG in naturalness, synchronization, and diversity (indicated by the in Figure 6). These results prove that our approach produces gestures that are more natural, better aligned with speech, and more diverse than those generated by SOTA baselines.

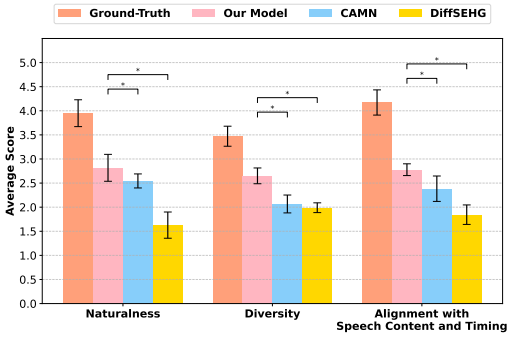

Figure 6. Average ratings of users for ground truth gestures and gestures generated through our approach, CAMN, and DiffSHEG. The bars illustrate the average user ratings across three metrics: naturalness, diversity, and alignment with speech content and timing. Statistical t-tests show that our approach received significantly higher ratings than CAMN and DiffSHEG, with $p < 0.05$.

7. Conclusion

We proposed SemGes, a novel two-stage approach to semantic grounding in co-speech gesture generation by integrating semantic information at both fine-grained and global levels. In the first stage, a motion prior generation module is trained using a vector-quantized variational autoencoder to produce realistic and smooth gesture motions. Building upon this model, the second stage generates gestures from speech, text-based semantics, and speaker identity while maintaining consistency between gesture semantics and co-occurring speech through semantic coherence and relevance modules. Subjective and objective evaluations show that our work achieves state-of-the-art performance across two public benchmarks, generating semantics-aware and diverse gestures. Future direction and limitations are discussed in Section 5 of the Supplementary Materials.

Acknowledgement

The project is funded by the Max Planck Society. We thank Sachit Misra for his invaluable assistance with rendering Avatar characters. We extend our gratitude to the members of the Multimodal Language Department at Max Planck Institute for Psycholinguistics for their feedback.

References

[1] Chaitanya Ahuja, Dong Won Lee, and Louis-Philippe Morency. Low-resource adaptation for personalized co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20566–20576, 2022. 6

[2] Chaitanya Ahuja, Pratik Joshi, Ryo Ishii, and Louis-Philippe Morency. Continual learning for personalized co-speech gesture generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20893–20903, 2023. 6

[3] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. Style-controllable speech-driven gesture synthesis using normalising flows. In *Computer Graphics Forum*, pages 487–496. Wiley Online Library, 2020. 2

[4] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–20, 2023. 6

[5] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics (TOG)*, 41(6): 1–19, 2022. 2, 3

[6] Tenglong Ao, Zeyi Zhang, and Libin Liu. Gesturediffuclip: Gesture diffusion model with clip latents. *ACM Transactions on Graphics (TOG)*, 42(4):1–18, 2023. 2, 3, 6

[7] Ben Athiwaratkun, Andrew Gordon Wilson, and Anima Anandkumar. Probabilistic fasttext for multi-sense word embeddings. *arXiv preprint arXiv:1806.02901*, 2018. 7

[8] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017. 4

[9] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 5

[10] Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation. In *CVPR*, 2024. 2, 6

[11] Kiran Chhatre, Nikos Athanasiou, Giorgio Becherini, Christopher Peters, Michael J Black, Timo 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*, pages 1942–1953, 2024. 6

[12] Ylva Ferstl, Michael Neff, and Rachel McDonnell. Adversarial gesture generation with realistic gesture phasing. *Computers & Graphics*, 89:117–130, 2020. 2

[13] Esam Ghaleb, Bulat Khaertdinov, Wim Pouw, Marlou Rasenberg, Judith Holler, Asli Ozyurek, and Raquel Fernández. Learning co-speech gesture representations in dialogue through contrastive learning: An intrinsic evaluation. In *Proceedings of the 26th International Conference on Multimodal Interaction*, pages 274–283, 2024. 1

[14] Esam Ghaleb, Bulat Khaertdinov, Aslı Özyürek, and Raquel Fernández. I see what you mean: Co-speech gestures for reference resolution in multimodal dialogue. In *Proceedings of the 63rd Conference of the Association for Computational Linguistics (ACL Findings)*, 2025. To appear. 1

[15] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 2

[16] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pages 580–597. Springer, 2022. 2

[17] Ikhsanul Habibie, Mohamed Elgharib, Kripasindhu Sarkar, Ahsan Abdullah, Simbarashe Nyatsanga, Michael Neff, and Christian Theobalt. A motion matching-based framework for controllable gesture synthesis from speech. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–9, 2022. 6

[18] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 2

[19] Judith Holler and Stephen C Levinson. Multimodal language processing in human communication. *Trends in Cognitive Sciences*, 23(8):639–652, 2019. 1

[20] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021. 4

[21] Adam Kendon. Gesture units, gesture phrases and speech. In *Gesture: Visible Action as Utterance*, chapter 7, page 108–126. Cambridge University Press, 2004. 1

[22] Taras Kucherenko, Patrik Jonell, Sanne Van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 international conference on multimodal interaction*, pages 242–250, 2020. 2

[23] Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1272–1279, 2022. 2

[24] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11293–11302, 2021. 6

[25] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 6

[26] Yuanzhi Liang, Qianyu Feng, Linchao Zhu, Li Hu, Pan Pan, and Yi Yang. Seeg: Semantic energized co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10473–10482, 2022. 2, 6

[27] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European Conference on Computer Vision*, pages 612–630. Springer, 2022. 2, 3, 5, 6

[28] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael 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*, pages 1144–1154, 2024. 2, 3

[29] Li Liu, Lufei Gao, Wentao Lei, Fengji Ma, Xiaotian Lin, and Jinting Wang. A survey on deep multi-modal learning for body language recognition and generation. *arXiv preprint arXiv:2308.08849*, 2023. 1

[30] Lanmiao Liu, Chuang Yu, Siyang Song, Zhidong Su, and Adriana Tapus. Human gesture recognition with a flow-based model for human robot interaction. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 548–551, 2023. 2

[31] Pinxin Liu, Luchuan Song, Junhua Huang, Haiyang Liu, and Chenliang Xu. Gesturelsm: Latent shortcut based co-speech gesture generation with spatial-temporal modeling. *arXiv preprint arXiv:2501.18898*, 2025. 6

[32] Xian Liu, Qianyi Wu, Hang Zhou, Yuanqi Du, Wayne Wu, Dahua Lin, and Ziwei Liu. Audio-driven co-speech gesture video generation. *Advances in Neural Information Processing Systems*, 35:21386–21399, 2022. 2

[33] 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. 2, 5, 6

[34] Yifei Liu, Qiong Cao, Yandong Wen, Huaiguang Jiang, and Changxing Ding. Towards variable and coordinated holistic co-speech motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1566–1576, 2024. 6

[35] David McNeill. Hand and mind. *Advances in Visual Semiotics*, 351, 1992. 1

[36] Muhammad Hamza Mughal, Rishabh Dabral, Ikhsanul Habibie, Lucia Donatelli, Marc Habermann, and Christian Theobalt. Convofusion: Multi-modal conversational diffusion for co-speech gesture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1388–1398, 2024. 6

[37] M Hamza Mughal, Rishabh Dabral, Merel CJ Scholman, Vera Demberg, and Christian Theobalt. Retrieving semantics from the deep: an rag solution for gesture synthesis. *arXiv preprint arXiv:2412.06786*, 2024. 6

[38] Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. From audio to photoreal embodiment: Synthesizing humans in conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1001–1010, 2024. 6

[39] Mang Ning, Mingxiao Li, Jianlin Su, Haozhe Jia, Lanmiao Liu, Martin Beneš, Wenshuo Chen, Albert Ali Salah, and Itir Onal Ertugrul. Dctdiff: Intriguing properties of image generative modeling in the dct space. *arXiv preprint arXiv:2412.15032*, 2024. 2

[40] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. A comprehensive review of data-driven co-speech gesture generation. In *Computer Graphics Forum*, pages 569–596. Wiley Online Library, 2023. 1, 2

[41] Aslı Özyürek. Hearing and seeing meaning in speech and gesture: Insights from brain and behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651):20130296, 2014. 1

[42] Kunkun Pang, Dafei Qin, Yingruo Fan, Julian Habekost, Takaaki Shiratori, Junichi Yamagishi, and Taku Komura. Bodyformer: Semantics-guided 3d body gesture synthesis with transformer. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. 6

[43] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 5

[44] Xingqun Qi, Jiahao Pan, Peng Li, Ruibin Yuan, Xiaowei Chi, Mengfei Li, Wenhan Luo, Wei Xue, Shanghang Zhang, Qifeng Liu, et al. Weakly-supervised emotion transition learning for diverse 3d co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10424–10434, 2024. 6

[45] Mingyang Sun, Mengchen Zhao, Yaqing Hou, Minglei Li, Huang Xu, Songcen Xu, and Jianye Hao. Co-speech gesture synthesis by reinforcement learning with contrastive pre-trained rewards. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2331–2340, 2023. 6

[46] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022. 3

[47] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[48] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[49] Hendric Voß and Stefan Kopp. Augmented co-speech gesture generation: Including form and meaning features to guide learning-based gesture synthesis. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2023. 2, 6

[50] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models. *arXiv preprint arXiv:2305.04919*, 2023. 2

[51] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, and Haolin Zhuang. Qpgesture: Quantization-based and phase-guided motion matching for natural speech-driven gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2321–2330, 2023. 6

[52] Sheng Ye, Yu-Hui Wen, Yanan Sun, Ying He, Ziyang Zhang, Yaoyuan Wang, Weihua He, and Yong-Jin Liu. Audio-driven stylized gesture generation with flow-based model. In *European Conference on Computer Vision*, pages 712–728. Springer, 2022. 6

[53] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *CVPR*, 2023. 3

[54] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 469–480, 2023. 6

[55] 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. 6

[56] Pengfei Zhang, Pinxin Liu, Hyeongwoo Kim, Pablo Garrido, and Bindita Chaudhuri. Kinmo: Kinematic-aware human motion understanding and generation. *arXiv preprint arXiv:2411.15472*, 2024. 3

[57] Zeyi Zhang, Tenglong Ao, Yuyao Zhang, Qingzhe Gao, Chuan Lin, Baoquan Chen, and Libin Liu. Semantic gesticulator: Semantics-aware co-speech gesture synthesis. *ACM Transactions on Graphics (TOG)*, 43(4):1–17, 2024. 2, 3, 6

[58] Yihao Zhi, Xiaodong Cun, Xuelin Chen, Xi Shen, Wen Guo, Shaoli Huang, and Shenghua Gao. Livelyspeaker: Towards semantic-aware co-speech gesture generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20807–20817, 2023. 2, 3, 6

[59] 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. 2, 3, 6