SemConFlow: Semantic Grounding of Holistic Co-Speech Gesture Generation with Contrastive Flow-Matching

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. While the field of co-speech gesture generation has seen significant advances, producing holistic, semantically grounded gestures remains a challenge. Existing approaches rely on external semantic retrieval methods, which limit their generalisation capability due to dependency on predefined linguistic rules. Flow-matching-based methods produce promising results; however, the network is optimised using only semantically congruent samples without exposure to negative examples, leading to learning rhythmic gestures rather than sparse motion, such as iconic and metaphoric gestures. Furthermore, by modelling body parts in isolation, the majority of methods fail to maintain cross-modal consistency. We introduce a Contrastive Flow Matching-based co-speech gesture generation model that uses mismatched audio–text conditions as negatives, training the velocity field to follow the correct motion trajectory while repelling semantically incongruent trajectories. Our model ensures cross-modal coherence by embedding text, audio, and holistic motion into a composite latent space via cosine and contrastive objectives. Extensive experiments and a user study demonstrate that our proposed approach outperforms state-of-the-art methods on two datasets, BEAT2 and SHOW. Our project page is available at: https://marcos452.github.io/HoliticSemGes/

Keywords: co-speech gesture generation · contrastive-flow matching · holistic motion synthesis

1 Introduction

Human communication is multimodal: speech is produced and interpreted together with coordinated bodily visual signals, such as co-speech gestures, as part of a single composite message [12]. Importantly, combinations of visual cues (e.g., head, gestures, and facial signals) can change perceived communicative intent relative to each cue in isolation, indicating a compositional, cross-channel organisation of meaning [37]. Humans also integrate modalities by *congruency*:

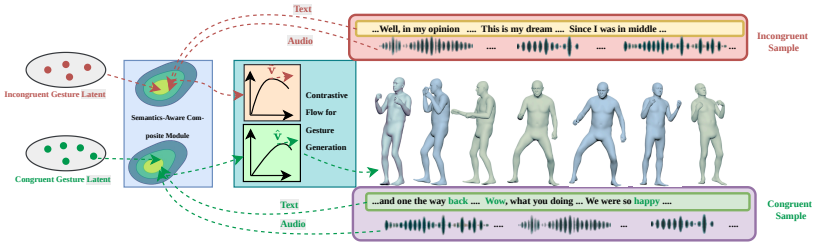

Fig. 1: Overall idea of our model. Audio and text transcripts are encoded into a shared semantic representation, where congruent (matching) and incongruent (mismatching) speech and gesture latents form contrastive pairs. Contrastive flow matching learns motion trajectories that align with the correct semantic context while diverging from mismatched samples, producing coherent full-body gestures.

content-aligned speech–gesture signals facilitate comprehension, whereas incongruent pairings can disrupt or mislead interpretation [15]. For generative modelling, this raises the bar beyond motion realism: generated motion must be semantically congruent, i.e., grounded in and aligned with the speech, and globally consistent across articulators (i.e., body parts).

Despite recent progress in holistic co-speech motion generation [4, 25, 26, 32, 44, 45, 48], two challenges remain. First, most holistic pipelines still perform *multimodal alignment and conditioning in a part-wise manner*: motion is tokenised into body-part latents, and speech and semantics embeddings are projected and injected separately for each region rather than being anchored to a single composite full-body representation. This factorised alignment limits cross-region coordination, so the model can satisfy semantic cues in one channel while others drift, leading to cross-articulator inconsistency (e.g., semantically informative hand motion paired with neutral or mismatched head and face behaviour). *Second, and most importantly,* the dominant generative paradigms such as Diffusion Models [1,4,30,48] and Flow Matching [24,26] optimise for distribution matching via distance-based (e.g., mean square error) or likelihood-style objectives. These objectives encourage motions that look realistic at first sight, but they lack the semantic grounding that links the motion with the *meaning* of the spoken content. Moreover, they lack explicit comparison against semantically *incongruent* alternatives, i.e., negative samples. Hence, they optimise to generate generic, beat-like gestures aligned with the rhythm of the speech, rather than sparse motion, such as iconic and metaphoric gestures that carry the intended semantics.

To address these two challenges, we present HolisticSemGes, an approach for high-fidelity holistic co-speech gesture generation with two complementary components. First, we introduce the **Semantics-Aware Composite Module (SACM)**. SACM explicitly shapes motion latents to be congruent with speech semantics by aligning speech-text embeddings with pre-quantised motion latents at the frame level. It uses a contrastive structure pulling matched speech–motion

pairs together, while pushing apart mismatched (semantically incongruent) pairs, whereas prior methods such as [22,45] neglect such incongruent relationships. In contrast with previous holistic work that encodes body parts separately [22], we treat them as a composite full-body representation to preserve the cross-channel organisation of meaning. Second, we propose a **Contrastive Flow Matching (CFM)** [36]- based co-speech gesture generation framework, different from previous work based on standard flow matching [26]. While standard flow matching [6] learns a conditional velocity field that transports noise to the target motion, it does not explicitly teach the model which motions are *semantically wrong* for a given utterance. CFM introduces mismatched audio–text conditions as *negative* context and trains the flow to align with the ground-truth trajectory as well as diverge from trajectories induced by incongruent semantics. This contrastive constraint makes the generation path from noise to motion more directionally stable and improves semantic grounding. Figure 1 presents the general idea of our approach. To summarise, our contributions are as follows:

- We introduce the *Semantics-Aware Composite Module* that learns a shared latent space across audio, text, and motion representations. This module improves cross-modal consistency by embedding the input modalities into a semantically aligned composite latent space.
- We propose a *Contrastive Flow Matching* based co-speech gesture generation framework, encouraging the latent flow to not only follow the correct motion trajectory but also avoid incorrect or mismatched speech-gesture directions. This leads to stable training, improved gesture diversity, and enhanced speech–gesture alignment.
- On two benchmark datasets, BEAT2 [22] and SHOW [40], we provide extensive objective and subjective evaluation and ablation studies showing state-of-the-art performances across several metrics.

2 Related Work

2.1 Holistic Co-Speech Gesture Generation

Early methods addressed individual body regions in isolation and relied on skeletal representations [9,17,27,41], which inherently prevent holistic semantic coherence. Recent work [4,11,22,26,40] have focused on holistic generation, including a combination of visual cues using mesh-based representations such as SMPL-X [33]. TalkSHOW [40] and EMAGE [22] train body part-specific VQ-VAEs [38] with autoregressive decoding, improving temporal coherence, yet learned codebooks are optimised for reconstruction fidelity alone with no semantic grounding. Diffusion-based methods such as DiffSHEG [4] achieve high sample diversity via audio-conditioned reverse diffusion, but require hundreds of inference steps and provide no mechanism for semantic guidance in the generation. Existing holistic approaches improve full-body realism [11,26], but they lack semantic grounding that explicitly aligns motion trajectories with the speech content.

2.2 Semantic Gesture Generation

Methods that focus on semantic grounding are mostly skeleton-based approaches and cannot generate holistic motion [21,23,41,47]. GestureDiffuCLIP [2] includes whole body and head movements in a holistic manner and constructs a shared audio–gesture embedding space via CLIP-inspired [34] contrastive pretraining. A key limitation of this approach is the decoupling of the alignment objective from the generative model. Contrastive pretraining shapes the conditioning space, but the generative objective reconstruction loss remains semantically agnostic, leaving semantically misaligned outputs unpunished during training. Semantic Gesticulator [46] and RAG-Gesture [29] address semantic grounding; however, they require an extra retrieval stage with predefined semantic guidance using linguistic insights. SemTalk [45] generates both rhythm and semantics-aware gestures; however, they are combined in a separate stage, leading to less realistic and diverse motion. In contrast, our method does not rely on an external retrieval stage or a separate pipeline for rhythm and semantics. Instead, it learns semantic grounding in an end-to-end manner with speech-text embeddings and semantics-guided contrastive flow-matching-based generation.

2.3 Generative Methods for Holistic Gesture Generation

Autoregressive Transformer decoders combined with (R)VQVAE discretisations are often used as the generative method [22,31,40,45]; however, they suffer from compounding errors over long sequences. Diffusion models [5, 48] achieve high fidelity and diversity, yet they require hundreds of denoising steps. Flow matching [26] learns a deterministic velocity field, enabling high-quality generation in a few steps more efficiently and with higher stability. MotionFlow [14] demonstrates its advantages for full-body motion synthesis. GestureLSM [26] introduces a flow matching-based co-speech gesture generation method; however, the network does not learn to punish semantically ungrounded motions. Our approach addresses this limitation using Contrastive Flow Matching (CFM), which embeds cosine similarity-based contrastive alignment into velocity field learning. This ensures that the deterministic path remains semantically faithful to the speech content throughout generation rather than converging to a statistically plausible but semantically averaged output.

3 Methodology

We propose HolisticSemGes, a novel approach for multimodal-driven holistic co-speech gesture generation over full-body SMPLX parameters. Sec 3.2 presents representation learning per-body-part Residual VQ-VAEs that compress full-body SMPL-X motion into compact latent sequences. Sec 3.3 introduce how to generate semantically coherent gestures, conditioned on speech acoustics and lexical content, with contrastive flow matching.

3.1 Problem Formulation

Our objective is to generate semantically grounded holistic gestures by jointly modelling multiple body parts conditioned on speech audio and textual semantics. We represent the gesture sequence for each part $r \in \mathcal{R}$, where $\mathcal{R} = \{\text{hand}, \text{upper}, \text{lower}, \text{face}\}$ as $\mathbf{G}^r = \{\mathbf{g}_i^r\}_{i=1}^L \in \mathbb{R}^{L \times J}$, where L denotes the number of time steps (i.e., sequence length) and J the number of joints within the corresponding region. To capture realistic motion dynamics, we first learn a motion generator $\mathcal{M}_g$ under a motion-aware representation learning framework. The learning objective of $\mathcal{M}_g$ is defined as follows:

$$\arg\min_{\mathcal{M}_g} \|\mathbf{G}^r - \mathcal{M}_g(\mathbf{G}^r)\| \tag{1}$$

Next, each motion stream undergoes encoding through a motion encoder $\mathcal{E}_m$, transforming kinematic data into intermediate feature representations $\mathbf{Z}^g$. Given multimodal inputs consisting of raw speech audio $\mathbf{A} = \{\mathbf{a}_i\}_{i=1}^L$, text-based semantic embeddings $\mathbf{S} = \{\mathbf{s}_i\}_{i=1}^L$, as well as mismatched (incongruent) speech audio $\hat{\mathbf{A}} = \{\hat{\mathbf{a}}_i\}_{i=1}^L$ and mismatched semantic embeddings $\hat{\mathbf{S}} = \{\hat{\mathbf{s}}_i\}_{i=1}^L$, our goal is to learn a multimodal generator $\mathcal{M}_m$ that maps $(\mathbf{A}, \mathbf{S}, \hat{\mathbf{A}}, \hat{\mathbf{S}})$ to a latent representation. The overall objective is defined as follows:

$$\arg\min_{\mathcal{M}_m} \left\| \mathbf{Z}^g - \mathcal{M}_m(\mathbf{A}, \hat{\mathbf{A}}, \mathbf{S}, \hat{\mathbf{S}}) \right\| \tag{2}$$

3.2 1st Stage: Motion Aware Representation Learning

We start with a motion prior learning stage that establishes fundamental movement patterns through the utilization of hierarchical RVQ-VAE with temporal compression inspired by [10, 22, 45]. This multi-scale tokenization approach uses residual vector quantizers (RVQ) [3, 18, 42] to discretize continuous gesture motions of hand, upper body, lower body and facial expressions into structured token sequences.

Given body-part-specific motion streams $\mathbf{G}^r$ for each body part r, we learn motion encoder $\mathcal{E}_m^r$ to extract latent representations. The encoded features are then quantised using region-specific codebooks $\mathbf{Q}^r$, and subsequently decoded by the motion decoder $\mathcal{D}_m^r$ to reconstruct the gesture sequence $\hat{\mathbf{G}}^r$ for each body part. The learning process optimises a composite objective function that balances reconstruction fidelity with quantisation stability across all body regions. We formulate the loss as:

$$\mathcal{L}_{total} = \mathcal{L}_{rec}(\mathbf{G}^r, \hat{\mathbf{G}}^r) + \mathcal{L}_{com}(\text{sg}[\mathbf{G}^r], \mathbf{Q}^r(\mathbf{Z}^g)) \tag{3}$$

where the first term denotes reconstruction loss between the generated gesture and the ground truth, the second term implements the VQ-VAE commitment loss [38].

3.3 2nd Stage: Generating Semantically Coherent and Grounded Gestures

The objective here is to condition motion generation on speech acoustics and text semantics, achieving meaningful, holistic gesture generation. We therefore apply efficient and robust mechanisms to map speech and text content to the latent motion space, learned in the first stage.

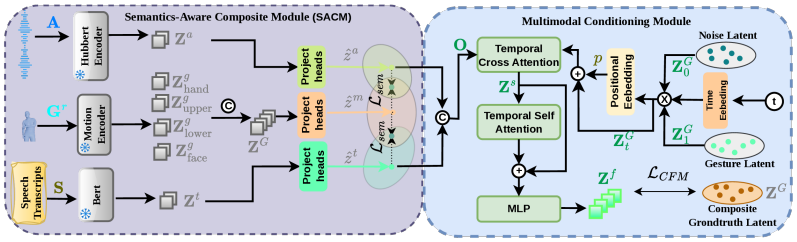

Fig. 2: Our architecture comprises two synergistic modules: (a) Semantics-Aware Composite Module (SACM), which anchors audio, textual, and motion features within an aligned semantic space to ensure cross-modal consistency; and (b) Multimodal Conditioning Module, which leverages a learned velocity field conditioned on multimodal priors and stochastic seed latents to synthesize expressive, kinematically coherent holistic gestures.

Semantics-Aware Composite Module (SACM). SACM aligns *text*, *audio*, and *holistic motion* in a shared semantic latent space (Fig. 2, left). Unlike part-wise alignment strategies that relate speech semantics to individual body regions in isolation [22, 45], we adopt a *body-compositional* view: we first aggregate all region latents from Stage 1 into a holistic motion representation, and then align this composite motion with language and acoustics using (1) a sequence-level cosine objective and (2) a CLIP-style contrastive objective.

Let $\mathbf{Z}_r^g \in \mathbb{R}^{L \times d_g}$ denote the Stage 1 latent sequence for region $r \in \mathcal{R}$, where $\mathcal{R} = \{\text{hand}, \text{upper}, \text{lower}, \text{face}\}$, d_g represent latent dimensions and L is the latent temporal length. We form a composite gesture latent by concatenation and scale normalisation:

$$\mathbf{Z}^G = \frac{1}{s}\,\text{Concat}\big(\mathbf{Z}^g_{\text{hand}}, \mathbf{Z}^g_{\text{upper}}, \mathbf{Z}^g_{\text{lower}}, \mathbf{Z}^g_{\text{face}}\big)\,, \tag{4}$$

where s is a constant normalisation factor. For semantics, a text encoder $\mathcal{T}$ (BERT [7]) maps transcripts to $\mathbf{Z}^t \in \mathbb{R}^{L \times d_t}$ and an audio encoder $\mathcal{A}$ (HuBERT [13]) maps waveforms to $\mathbf{Z}^a \in \mathbb{R}^{L \times d_a}$. Both modalities are temporally aligned to the motion latent length L. To compare heterogeneous modalities, we apply modality-specific projection heads Ψ^t, Ψ^a, Ψ^m into a common d-dimensional space followed by ℓ_2 normalization:

$$\hat{\mathbf{Z}}^t = \text{norm}(\Psi^t(\mathbf{Z}^t)), \quad \hat{\mathbf{Z}}^a = \text{norm}(\Psi^a(\mathbf{Z}^a)), \quad \hat{\mathbf{Z}}^m = \text{norm}(\Psi^m(\mathbf{Z}^G)). \tag{5}$$

To mitigate semantic drift between lexical and acoustic cues, we construct a fused (barycentric) target: $\bar{\mathbf{Z}} = \mathrm{norm}\big(\alpha\,\hat{\mathbf{Z}}^t + (1-\alpha)\,\hat{\mathbf{Z}}^a\big)$, where $\alpha \in [0,1]$.
(1) Sequence-level cosine alignment. Given a batch of size B and time index $i \in \{1,\dots,L\}$, we minimise the mean cosine distance over time and batch:

$$\mathcal{L}_{\cos} = 1 - \frac{1}{BK}\sum_{b=1}^{B}\sum_{l=1}^{L}\big(\bar{\mathbf{z}}_{b,l}\big)^{\top}\hat{\mathbf{z}}^m_{b,l}. \tag{6}$$

(2) CLIP-style InfoNCE alignment. To obtain batch-level discrimination with explicit negatives, we compute a single embedding per clip by temporal pooling (e.g., mean pooling) $\mathbf{u}^x_b = \mathrm{Pool}(\hat{\mathbf{Z}}^x_b) \in \mathbb{R}^d$ for $x \in \{m,a,t\}$. The (one-way) InfoNCE loss is:

$$\mathcal{L}_{\mathrm{NCE}}(\mathbf{u}^p,\mathbf{u}^q) = -\frac{1}{B}\sum_{b=1}^{B}\log\frac{\exp((\mathbf{u}^p_b)^{\top}\mathbf{u}^q_b/\tau)}{\sum_{l=1}^{B}\exp((\mathbf{u}^p_b)^{\top}\mathbf{u}^q_l/\tau)}, \tag{7}$$

where τ is a temperature parameter, and negatives are other samples in the mini-batch. We use a symmetric variant: $\mathcal{L}^{\mathrm{sym}}_{\mathrm{NCE}}(\mathbf{u}^p,\mathbf{u}^q) = \frac{1}{2}\big(\mathcal{L}_{\mathrm{NCE}}(\mathbf{u}^p,\mathbf{u}^q) + \mathcal{L}_{\mathrm{NCE}}(\mathbf{u}^q,\mathbf{u}^p)\big)$, and define:

$$\mathcal{L}_{\mathrm{clp}} = \tfrac{1}{2}\Big(\mathcal{L}^{\mathrm{sym}}_{\mathrm{NCE}}(\mathbf{u}^m,\mathbf{u}^a) + \mathcal{L}^{\mathrm{sym}}_{\mathrm{NCE}}(\mathbf{u}^m,\mathbf{u}^t)\Big). \tag{8}$$

The overall SACM objective is: $\mathcal{L}_{\mathrm{sem}} = \lambda_{\cos}\mathcal{L}_{\cos} + \lambda_{\mathrm{clp}}\mathcal{L}_{\mathrm{clp}}$. Overall, SACM anchors holistic motion latents in a joint speech-semantic manifold, improving cross-region coherence while preserving alignment with both lexical meaning and acoustic prominence.

Flow Matching Through Multimodal Conditioning. Given the aligned multimodal representations, we formulate gesture generation as learning a conditional velocity (flow) field that transports samples from a latent noise distribution to the motion manifold in the composite gesture-latent space (Fig. 2, right). We construct a multimodal conditioning sequence by concatenating audio and text embeddings along the channel dimension and projecting them to a unified conditioning space:

$$\mathbf{O} = \mathrm{Proj}_O\Big(\mathrm{Concat}(\hat{\mathbf{Z}}^a,\hat{\mathbf{Z}}^t)\Big) \in \mathbb{R}^{L\times d_O}. \tag{9}$$

Let $\mathbf{Z}^G_1 \in \mathbb{R}^{L\times d_G}$ denote the (Stage 1) composite gesture latent of the ground-truth motion and let $\mathbf{Z}^G_0 \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ denote a noise latent of the same shape. We sample a *flow time* $t \sim \mathcal{U}(0,1)$ (distinct from the text-modality superscript) and define the linear interpolant:

$$\mathbf{Z}^G_t = (1-t)\,\mathbf{Z}^G_0 + t\,\mathbf{Z}^G_1, \tag{10}$$

whose corresponding target velocity field is constant:

$$\hat{\mathbf{v}} = \frac{\partial \mathbf{Z}^G_t}{\partial t} = \mathbf{Z}^G_1 - \mathbf{Z}^G_0. \tag{11}$$

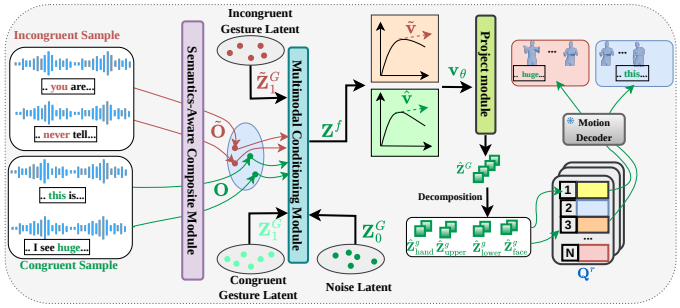

Fig. 3: Overview of the proposed semantic-aware contrastive flow matching framework for gesture generation. Audio and text transcripts are encoded into a shared semantic representation where congruent and incongruent gesture latents form contrastive pairs. Contrastive flow matching then learns motion trajectories aligned with the correct semantic context while diverging from mismatched inputs, producing coherent full-body gestures.

Before conditioning, we add a learned body-structure positional embedding $\mathbf{p}$ to the motion latent tokens to encode intra-frame body structure. We then apply a temporal cross-attention block ($TCAM$) to fuse the multimodal condition $\mathbf{O}$ with the noised motion latent:$\mathbf{Z}^s = \mathrm{TCAM}\big(\mathbf{Z}_t^G + \mathbf{p},\ \mathbf{O}\big)$, where $\mathbf{Z}_t^G + \mathbf{p}$ provides the queries and $\mathbf{O}$ provides keys/values, yielding conditioned motion features $\mathbf{Z}^s \in \mathbb{R}^{L\times d_s}$ (with d_s=1024 in our implementation). Next, temporal self-attention captures global motion context to promote temporal coherence, followed by a position-wise MLP:$\mathbf{Z}^f = \mathrm{MLP}(\mathrm{SelfAttn}(\mathbf{Z}^s))$. Finally, the flow network predicts the conditional velocity field $\mathbf{v}_\theta(\mathbf{Z}_t^G, t \mid \mathbf{O})$ from the fused features.

Contrastive Flow for Gesture Generation To encourage the conditional flow to remain sensitive to multimodal semantics (and reduce semantic averaging), we add a contrastive regularisation term inspired by contrastive conditional flow formulations [36]. In Fig. 3, during training, for each sample we construct an *incongruent* conditioning sequence $\tilde{\mathbf{O}}$ by mismatching audio–text pairs (e.g., via an in-batch random permutation of transcripts while keeping audio fixed, or vice versa).

Let $\mathbf{Z}_1^G$ denote the ground-truth composite gesture latent and $\mathbf{Z}_0^G \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ the noise latent. We sample $t \sim \mathcal{U}(0, 1)$ and define the interpolated latent $\mathbf{Z}_t^G$ as in Eq. 10. Under the straight-line path, the positive target velocity is (Eq. 11). To form a negative (incongruent) target, we sample a mismatched gesture latent $\tilde{\mathbf{Z}}_1^G$ from the example associated with $\tilde{\mathbf{O}}$, while reusing the same noise seed $\mathbf{Z}_0^G$ for comparability. The corresponding negative velocity is: $\tilde{\mathbf{v}} = \tilde{\mathbf{Z}}_1^G - \mathbf{Z}_0^G$.

Given the conditional velocity predictor $\mathbf{v}_\theta(\mathbf{Z}_t^G, t \mid \mathbf{O})$, we optimize the following contrastive flow-matching objective:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}\Big[\big\|\mathbf{v}_\theta(\mathbf{Z}_t^G, t \mid \mathbf{O}) - \hat{\mathbf{v}}\big\|_2^2 - \lambda \big\|\mathbf{v}_\theta(\mathbf{Z}_t^G, t \mid \mathbf{O}) - \tilde{\mathbf{v}}\big\|_2^2\Big], \qquad \lambda \in [0, 1), \tag{12}$$

where λ controls the strength of contrastive regularisation. Intuitively, the first term attracts the predicted velocity toward the congruent trajectory induced by $\mathbf{O}$, while the second term repels it from trajectories associated with the incongruent multimodal context $\tilde{\mathbf{O}}$.

Sampling and decoding. At inference time, we sample $\mathbf{Z}_0^G \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and integrate the learned conditional ODE $\frac{d\mathbf{Z}}{dt} = \mathbf{v}_\theta(\mathbf{Z}, t \mid \mathbf{O})$ to obtain a generated holistic latent $\hat{\mathbf{Z}}_1^G$. We apply a lightweight projection head to map the generated latent to a stable manifold compatible with the Stage 1 VQ codebooks:

$$\hat{\mathbf{Z}}^G = \text{Proj}\Big(\hat{\mathbf{Z}}_1^G\Big). \tag{13}$$

We then decompose the holistic latent back into region-specific latents (inverse of the concatenation order):

$$\hat{\mathbf{Z}}^G = \bigoplus_{r \in \mathcal{R}} \hat{\mathbf{Z}}_r^g, \quad \mathcal{R} = \{\text{hand}, \text{upper}, \text{lower}, \text{face}\}. \tag{14}$$

Each region latent is quantised by the corresponding VQ codebook $\mathbf{Q}^r$:

$$\mathbf{Z}_r^q = \text{Quant}\big(\hat{\mathbf{Z}}_r^g; \mathbf{Q}^r\big). \tag{15}$$

Finally, the motion decoder $\mathcal{D}_m$ reconstructs the full-body gesture sequence:

$$\hat{\mathbf{G}} = \mathcal{D}_m\big(\{\mathbf{Z}_r^q\}_{r \in \mathcal{R}}\big). \tag{16}$$

Training Objective The comprehensive training objective for the semantic gesture generation integrates contrastive flow matching loss and semantic alignment loss:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{CFM}} \mathcal{L}_{\text{CFM}} + \lambda_{\text{sem}} \mathcal{L}_{\text{sem}}. \tag{17}$$

The hyperparameters λ_{CFM} and λ_{sem} balance the contribution of optimization objective.

4 Experiment

4.1 Datasets

Our methodology is evaluated on two datasets, namely BEAT2 [22] and SHOW [40], objectively and subjectively. BEAT2 [22] presents a large-scale multimodal motion capture dataset comprising 60 hours of synchronised speech and motion data from 25 speakers (12 female, 13 male). The dataset is organised into

BEAT2-Standard (27 hours) and BEAT2-Additional (33 hours) subsets, with the latter capturing more spontaneous and natural gestural behaviours. BEAT2 significantly advances beyond its predecessor through refined mesh-based representation and MoSh++ [28] optimisation, which enhances body proportion accuracy, finger articulation precision, and facial expression fidelity. The dataset facilitates comprehensive multimodal learning with robust cross-modal alignment between speech audio and full-body motion coded in SMPL-X parameters [33]. The BEAT2-Standard subset [22] is split with an 85%/7.5%/7.5% train/validation/test division across all 25 speakers.

SHOW dataset [40] is a comprehensive audiovisual dataset featuring 26.9 hours of synchronized speech and 3D motion data from four speakers. The dataset provides high-quality 30fps SMPL-X [33] body meshes with holistic full-body, hand, and facial expressions paired with 22kHz audio. Each video undergoes rigorous processing through PyMAF-X [43] incorporating facial landmarks and shape alignment to ensure motion accuracy. The dataset emphasizes speech-gesture synchronization and rhythm correlation, segmented into 10-second clips totaling 12,019 sequences. Following established protocols [40], the data is split into 80%/10%/10% for the train/val/test division across all 4 speakers.

4.2 Experimental Setup

Implementation Details. All experiments are conducted on a single NVIDIA A100 GPU. For the BEAT2 dataset [22], our training follows a two-stage pipeline. In the first stage, we train four separate body-part-specific temporal Residual Vector Quantised VAEs (RVQVAE). This stage is optimized for 800 epochs (approximately 140 hours) with a batch size of 64 and a clip length of 64 frames. In the second stage, the semantic gesture generator is optimised for 1000 epochs (approximately 63 hours) using the Adam optimiser with a constant learning rate of 1×10^{-3}, a batch size of 128, and a clip length of 64. For the SHOW dataset [40], we extend the temporal window to a clip length of 196 frames by following the setup in [40]. The motion representation stage for SHOW is similar to the BEAT2 configuration, while the subsequent semantic generation stage is trained for 1000 epochs with a batch size of 128 and a slightly increased learning rate of 1.5×10^{-3} to accommodate the expanded temporal context.

Comparison Baselines. We conduct comprehensive comparisons against state-of-the-art holistic gesture generation methods such as Semtalk [39], GestureLSM [26], EMAGE [22], Talkshow [40] and RAGGesture [29] All methods mentioned above are trained from scratch using their original code base on the BEAT2 and SHOW dataset following the train/val/test splits as explained in Section 4.1 DataSets. Other approaches such as [2, 23, 46] are not included as they do not focus on holistic co-speech gesture synthesis.

4.3 Quantitative Results

Evaluation Metrics. We evaluate all methods using three complementary metrics. (1) **Fréchet Gesture Distance (FGD)** [41] measures the distributional

Dataset	Method	FGD↓	BC↑	Diversity↑
BEAT2(25 speakers)	SHOW [40]	6.204	0.652	82
	EMAGE [22]	5.521	0.692	88
	RAGGesture [29]	4.446	0.475	114
	GestureLSM [26]	4.268	0.525	112
	SemTalk [45]	4.264	0.727	116
	Ours	**2.247**	**0.780**	**120**
SHOW(4 speakers)	SHOW [40]	24.35	0.825	102
	EMAGE [22]	22.23	0.823	106
	RAGGesture [29]	23.54	0.812	94
	GestureLSM [26]	22.89	0.809	102
	SemTalk [45]	20.17	0.829	110
	Ours	**18.92**	**0.831**	**112**

Table 1: Quantitative comparison with state-of-the-art methods on the BEAT2 and SHOW datasets. ↓ and ↑ indicate whether lower or higher values are better, respectively. Our method achieves consistently strong performance across Fréchet Gesture Distance (FGD), Beat Consistency (BC) and Diversity metrics.

similarity between generated and ground-truth gestures in a learned latent feature space. Lower FGD indicates closer alignment with the real motion distribution and improved holistic realism. (2) **Beat Consistency (BC)** [20] evaluates speech–motion synchronization by measuring the alignment between kinematic motion peaks and corresponding acoustic onset patterns, reflecting rhythmic and temporal consistency. (3) **Diversity** [19] is calculated as the average pairwise Euclidean distance among generated samples in the test set, assessing motion variability and the stochastic richness of the generative model.

Quantitative Comparison Evaluation with Baselines. We provide comparison with the recent state-of-the-art methods listed above, on both BEAT2 and SHOW datasets. It is important to note that prior works [22, 26, 45] report results on BEAT2 under single-speaker settings, leaving multi-speaker generalization underexplored. We therefore evaluate on the full 25-speaker BEAT2 benchmark. As shown in Table 1, our approach achieves consistently strong performance across all metrics. On BEAT2(25 speakers), our method results in the lowest FGD value, indicating closer alignment with the ground-truth motion indicating improved motion realism. It further achieves the highest BC score, demonstrating improved temporal coordination between speech and gesture. Our approach also produces the most diverse gesture patterns without compromising motion realism. On the SHOW dataset (4 speakers), our model maintains better results for FGD, BC and Diversity metrics exceeding the performance of other methods. Taken together, these results suggest that our framework generalizes effectively across datasets and speaker identities, providing consistent improvements without sacrificing motion naturalness.

Ablation Study. Table 2 evaluates three design choices on BEAT2 and SHOW: (i) removing the Semantics-Aware Composite Alignment Module (SACM), (ii) restricting SACM to a single modality (text-only vs. audio-only), and (iii) replacing Contrastive Flow Matching with standard Flow matching (FM).

Dataset	Method	FGD↓	BC↑	Diversity↑
BEAT2	w/o SACM	4.125	0.592	102
	w/ SACM(text only)	3.216	0.736	112
	w/ SACM(audio only)	3.849	0.765	105
	w/ standard FM	4.269	0.776	111
	Ours	**2.247**	**0.780**	**120**
SHOW	w/o SACM	24.28	0.726	92
	w/ SACM(text only)	20.47	0.815	108
	w/ SACM(audio only)	22.85	0.821	98
	w/ standard FM	22.86	0.825	101
	Ours	**18.92**	**0.831**	**112**

Table 2: Ablation study on BEAT2 and SHOW datasets. We analyze the contribution of each component, including the Multimodal Semantics-Aware Composite Module (SACM) under different modality settings, and the Contrastive Flow Matching objective compared to Standard Flow Matching.

1. *SACM removal.* Removing SACM leads to a clear and consistent degradation across datasets, with the most impact on Beat Consistency and FGD metrics. This confirms that explicit semantic alignment is necessary for stable speech–motion coupling.
2. *Modality within SACM.* Text-only or audio-only alignment provides improved results in comparison to removing SACM, but each modality emphasizes different aspects: text-only alignment improves FGD and Diversity, whereas audio-only alignment tends to preserve temporal consistency measured through BC (while still improving the FGD and Diversity metrics). This supports the view that text and audio provide complementary supervision.
3. *Contrastive FM vs standard FM.* Replacing contrastive flow matching with standard FM degrades the FGD metric leading to decreased motion realism. Standard FM also have lower diversity and BS values with respect to the Contrastive Flow Matching. Overall, Contrastive Flow Matching consistently improves the overall results supporting our initial claims and contributions.

Overall, our complete model achieves the best balance on both datasets, simultaneously improving motion realism (lowest FGD), beat consistency (highest BC), and diversity.

4.4 Qualitative Results

Qualitative comparisons. Figure 4 compares our method against EMAGE, SemTalk, and GestureLSM on four semantically salient spans (discourse marker [8], self-reference [35], negation and causality [16]) As shown in the figure, our outputs show clearer *gesture stroke* realization that is consistent with the intended discourse function and the ground truth.

A discourse marker is a word or phrase that manages the flow and structure of communication, acting as a linguistic glue connecting sentences or words. For the discourse marker "Well" our model produces a synchronized beat gesture with a brief upward hand lift aligned with the stressed syllable, clearly

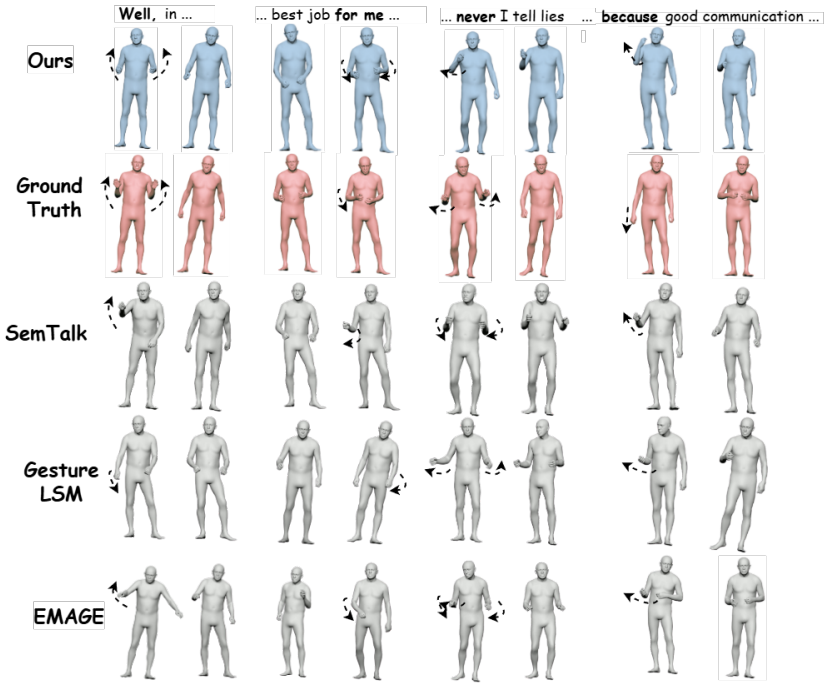

Fig. 4: Qualitative comparison of semantic gesture generation on BEAT2. Our model synthesizes context-grounded gestures that precisely map to linguistic semantics: discourse marker "Well", deictic pointing for "for me" rhythmic denial for "never" and iconic explanation for "because". In contrast, baselines typically collapse into semantically neutral, rhythm-driven arm swings, failing to capture the distinct communicative functions of each clause.

marking discourse initiation. In contrast, SemTalk maintains rhythmic motion without emphasis-specific articulation, while EMAGE exhibits neutral or loosely synchronized movements. For the phrase "for me" our model generates a clear deictic gesture by directing the hand toward the torso during "me" forming an explicit self-referential motion aligned with lexical emphasis. SemTalk shows mild prosodic emphasis but lacks precise referential targeting, whereas EMAGE remains largely rhythm-driven. GestureLSM displays minor frame-level jitter that reduces motion coherence. For the negation phrase "never tell lies", our model produces repeated forward index gestures synchronized with "never" reinforcing semantic negation, while baselines largely maintain neutral hand configurations. For the causal segment "because good communication" our framework generates a forward-expanding iconic gesture unfolding across the clause, visually supporting the explanatory structure. SemTalk shows clause-aligned motion but lacks progressive expansion, and other baselines exhibit temporal discontinuities near emphasis peaks. Overall, while SemTalk demonstrates rhythm-aligned motion with limited semantic sensitivity, our model consistently produces discourse-

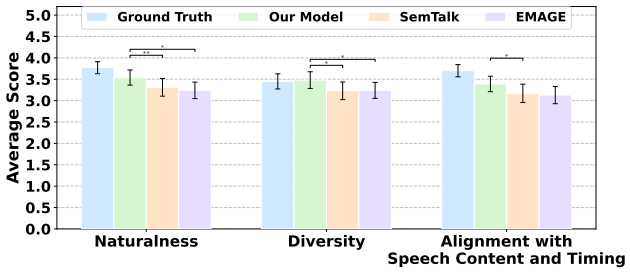

Fig. 5: User study evaluation results comparing Ground Truth, Our Model, SemTalk, and EMAGE across three criteria: Naturalness, Diversity, and Alignment with Speech Content and Timing. Error bars denote standard deviation across participants. Statistical significance is indicated by * ($p < 0.05$) and ** ($p < 0.01$).

aware beat gestures, deictic references, and clause-level iconic expansions while maintaining smooth inter-frame transitions without noticeable jitter.

User Study. We recruited 30 native English speakers from the UK and US, who evaluated 24 video sequences presented in randomized order using a five-point Likert scale across three dimensions: naturalness, diversity, and alignment with speech content and timing, including ground-truth motion, EMAGE, SemTalk, and our proposed model. Details on the user study are provided in the supplementary material. User study results, illustrated in Fig. 5, confirm that our model significantly outperforms SemTalk and EMAGE in perceptual quality. Specifically, our framework achieves higher scores in Naturalness and Alignment to Speech Content and Timing, indicating that the Semantic Coherence Module (SACM) and Contrastive Flow Matching (CFM) effectively capture nuanced speech-motion mappings. While Diversity remains similar across SemTalk and EMAGE, our approach surpasses these methods (as well the ground truth) without degrading motion realism and alignment to speech.

5 Conclusion

We presented HolisticSemGes, a semantic-aware framework for holistic co-speech gesture generation that integrates speech audio, transcripts, and motion latents within a unified semantic space through the Semantics-Aware Composite Module (SACM), enabling cross-articulator coordination across the whole body. In addition, we introduce Contrastive Flow Matching-based(CFM) co-speech gesture generation model, which guides motion trajectories by aligning with semantically congruent speech–motion pairs while diverging from incongruent alternatives. Comprehensive experiments demonstrate consistent improvements in motion realism, speech–motion synchronization, semantic relevance, and gesture diversity across multiple speaker identities, producing coherent, expressive, and semantically grounded gestures.

References

1. Alexanderson, S., Nagy, R., Beskow, J., Henter, G.E.: Listen, denoise, action! audio-driven motion synthesis with diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–20 (2023)
2. Ao, T., Zhang, Z., Liu, L.: Gesturediffuclip: Gesture diffusion model with clip latents. ACM Transactions on Graphics (TOG) **42**(4), 1–18 (2023)
3. Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Roblek, D., Teboul, O., Grangier, D., Tagliasacchi, M., et al.: Audiolm: a language modeling approach to audio generation. IEEE/ACM transactions on audio, speech, and language processing **31**, 2523–2533 (2023)
4. Chen, J., Liu, Y., Wang, J., Zeng, A., Li, Y., Chen, Q.: Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7352–7361 (2024)
5. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18000–18010 (2023)
6. Dao, Q., Phung, H., Nguyen, B., Tran, A.: Flow matching in latent space. arXiv preprint arXiv:2307.08698 (2023)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
8. Fraser, B.: What are discourse markers? Journal of pragmatics **31**(7), 931–952 (1999)
9. Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J.: Learning individual styles of conversational gesture. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3497–3506 (2019)
10. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1900–1910 (2024)
11. Guo, Z., Zhang, J.: Fasttalker: Jointly generating speech and conversational gestures from text. In: European Conference on Computer Vision. pp. 177–194. Springer (2025)
12. Holler, J., Levinson, S.C.: Multimodal language processing in human communication. Trends in Cognitive Sciences **23**(8), 639–652 (2019)
13. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: 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)
14. Hu, V.T., Yin, W., Ma, P., Chen, Y., Fernando, B., Asano, Y.M., Gavves, E., Mettes, P., Ommer, B., Snoek, C.G.: Motion flow matching for human motion synthesis and editing. In: Arxiv (2024)
15. Kelly, S.D., Özyürek, A., Maris, E.: Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. Psychological Science **21**(2), 260–267 (2010). https://doi.org/10.1177/0956797609357327, https://journals.sagepub.com/doi/abs/10.1177/0956797609357327

16. Konig, E., Siemund, P.: Causal and concessive clauses: Formal and semantic relations. Topics in English Linguistics **33**, 341–360 (2000)
17. Kucherenko, T., Jonell, P., Van Waveren, S., Henter, G.E., Alexandersson, S., Leite, I., Kjellström, H.: Gesticulator: A framework for semantically-aware speech-driven gesture generation. In: Proceedings of the 2020 international conference on multimodal interaction. pp. 242–250 (2020)
18. Lee, D., Kim, C., Kim, S., Cho, M., Han, W.S.: Autoregressive image generation using residual quantization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11523–11532 (2022)
19. Li, J., Kang, D., Pei, W., Zhe, X., Zhang, Y., He, Z., Bao, L.: Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11293–11302 (2021)
20. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13401–13412 (2021)
21. Liang, Y., Feng, Q., Zhu, L., Hu, L., Pan, P., Yang, Y.: Seeg: Semantic energized co-speech gesture generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10473–10482 (2022)
22. Liu, H., Zhu, Z., Becherini, G., Peng, Y., Su, M., Zhou, Y., Zhe, X., Iwamoto, N., Zheng, B., Black, M.J.: 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. pp. 1144–1154 (2024)
23. Liu, L., Ghaleb, E., Ozyurek, A., Yumak, Z.: Semges: Semantics-aware co-speech gesture generation using semantic coherence and relevance learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13963–13973 (2025)
24. Liu, L., Yu, C., Song, S., Su, Z., Tapus, A.: 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. pp. 548–551 (2023)
25. Liu, P., Liu, H., Song, L., Corso, J.J., Xu, C.: Intentional gesture: Deliver your intentions with gestures for speech. arXiv preprint arXiv:2505.15197 (2025)
26. Liu, P., Song, L., Huang, J., Liu, H., Xu, C.: Gesturelsm: Latent shortcut based co-speech gesture generation with spatial-temporal modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10929–10939 (October 2025)
27. Liu, P., Zhang, P., Kim, H., Garrido, P., Shapiro, A., Olszewski, K.: Contextual gesture: Co-speech gesture video generation through context-aware gesture representation. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 9803–9812 (2025)
28. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5442–5451 (2019)
29. Mughal, M.H., Dabral, R., Scholman, M.C., Demberg, V., Theobalt, C.: Retrieving semantics from the deep: an rag solution for gesture synthesis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16578–16588 (2025)
30. Ning, M., Li, M., Su, J., Jia, H., Liu, L., Beneš, M., Chen, W., Salah, A.A., Ertugrul, I.O.: Dctdiff: Intriguing properties of image generative modeling in the dct space. arXiv preprint arXiv:2412.15032 (2024)

31. Ning, M., Li, M., Zhang, L., Liu, L., Blaschko, M.B., Salah, A.A., Ertugrul, I.O.: Spectrum matching: a unified perspective for superior diffusability in latent diffusion. arXiv preprint arXiv:2603.14645 (2026)
32. Paar, F., Liu, L., Özyürek, A., Thill, S., Ghaleb, E.: Duogesture: Neuro-inspired and biomechanically informed dual-stream co-speech gesture generation. arXiv preprint arXiv:2605.26236 (2026)
33. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: 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. pp. 10975–10985 (2019)
34. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
35. Rogers, T.B., Kuiper, N.A., Kirker, W.S.: Self-reference and the encoding of personal information. Journal of personality and social psychology **35**(9), 677 (1977)
36. Stoica, G., Ramanujan, V., Fan, X., Farhadi, A., Krishna, R., Hoffman, J.: Contrastive flow matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1185–1194 (2025)
37. Trujillo, J.P., Holler, J.: Conversational facial signals combine into compositional meanings that change the interpretation of speaker intentions. Scientific Reports **14**(1), 2286 (2024). `https://doi.org/10.1038/s41598-024-52589-0`, `https://doi.org`
38. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
39. Yang, Q., Huang, L., Wang, K., Guan, J., He, S., Li, F., Zhou, H., Yu, L., Li, Y., Feng, H., et al.: Gesturehydra: Semantic co-speech gesture synthesis via hybrid modality diffusion transformer and cascaded-synchronized retrieval-augmented generation. arXiv preprint arXiv:2507.22731 (2025)
40. Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., Tao, D., Black, M.J.: Generating holistic 3d human motion from speech. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 469–480 (2023)
41. Yoon, Y., Cha, B., Lee, J.H., Jang, M., Lee, J., Kim, J., Lee, G.: Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Transactions on Graphics (TOG) **39**(6), 1–16 (2020)
42. Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., Tagliasacchi, M.: Soundstream: An end-to-end neural audio codec. IEEE/ACM Transactions on Audio, Speech, and Language Processing **30**, 495–507 (2021)
43. Zhang, H., Tian, Y., Zhang, Y., Li, M., An, L., Sun, Z., Liu, Y.: Pymaf-x: Towards well-aligned full-body model regression from monocular images. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(10), 12287–12303 (2023)
44. Zhang, P., Liu, P., Garrido, P., Kim, H., Chaudhuri, B.: Kinmo: Kinematic-aware human motion understanding and generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11187–11197 (2025)
45. Zhang, X., Li, J., Zhang, J., Dang, Z., Ren, J., Bo, L., Tu, Z.: Semtalk: Holistic co-speech motion generation with frame-level semantic emphasis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13761–13771 (2025)
46. Zhang, Z., Ao, T., Zhang, Y., Gao, Q., Lin, C., Chen, B., Liu, L.: Semantic gesticulator: Semantics-aware co-speech gesture synthesis. ACM Transactions on Graphics (TOG) **43**(4), 1–17 (2024)

47. Zhi, Y., Cun, X., Chen, X., Shen, X., Guo, W., Huang, S., Gao, S.: Livelyspeaker: Towards semantic-aware co-speech gesture generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 20807–20817 (2023)
48. Zhu, L., Liu, X., Liu, X., Qian, R., Liu, Z., Yu, L.: Taming diffusion models for audio-driven co-speech gesture generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10544–10553 (2023)