SemTalk: Holistic Co-speech Motion Generation with Frame-level Semantic Emphasis

Xiangyue Zhang[1,2*] Jianfang Li[2*] Jiaxu Zhang[1] Ziqiang Dang[2,3] Jianqiang Ren[2]
Liefeng Bo[2] Zhigang Tu[1†]
[1]Wuhan University [2]Tongyi Lab, Alibaba Group [3]Zhejiang University

Project page: https://xiangyue-zhang.github.io/SemTalk

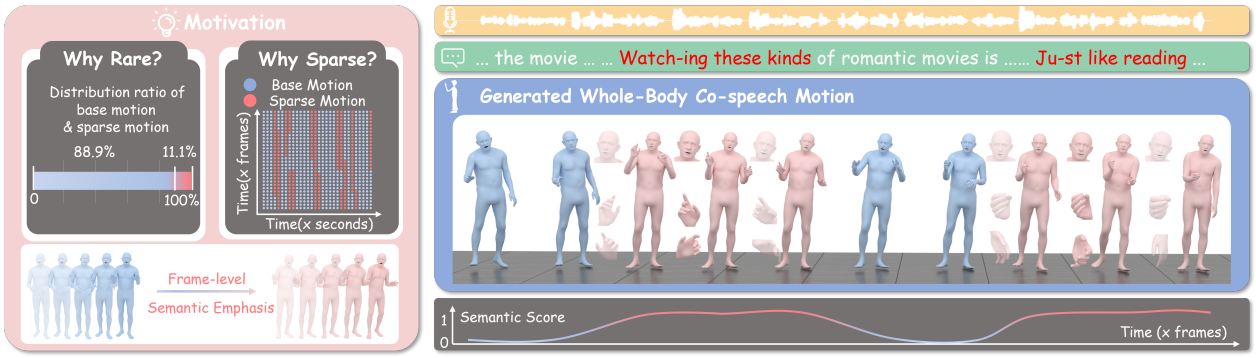

Figure 1. **On the left**, we analyze semantic labels from the BEAT2 dataset [31] and visualize frame-level motion, revealing that semantically relevant motions are rare and sparse, aligning with real-life observations. **On the right**, this observation drives the design of **SemTalk**, which establishes a rhythm-aligned base motion and dynamically emphasizes sparse semantic gestures at the frame-level. In this example, SemTalk amplifies expressiveness on words like "watching" and "just," enhancing gesture and torso movements. The semantic scores below are automatically generated by SemTalk to modulate semantic emphasis over time.

Abstract

A good co-speech motion generation cannot be achieved without a careful integration of common rhythmic motion and rare yet essential semantic motion. In this work, we propose SemTalk for holistic co-speech motion generation with frame-level semantic emphasis. Our key insight is to separately learn base motions and sparse motions, and then adaptively fuse them. In particular, coarse2fine cross-attention module and rhythmic consistency learning are explored to establish rhythm-related base motion, ensuring a coherent foundation that synchronizes gestures with the speech rhythm. Subsequently, semantic emphasis learning is designed to generate semantic-aware sparse motion, focusing on frame-level semantic cues. Finally, to integrate sparse motion into the base motion and generate semantic-emphasized co-speech gestures, we further leverage a learned semantic score for adaptive synthesis. Qualitative and quantitative comparisons on two public datasets demonstrate that our method outperforms the state-of-the-art, delivering high-quality co-speech motion with enhanced semantic richness over a stable base motion.

1. Introduction

Nonverbal communication, including body language, hand gestures, and facial expressions, is integral to human interactions. It enriches conversations with contextual cues and enhances understanding among participants [6, 14, 20, 24]. This aspect is particularly significant in holistic co-speech motion generation, where the challenge lies in synthesizing gestures that align with speech rhythm while also capturing the infrequent yet critical semantic gestures [25, 38].

Most existing methods [17, 30, 48] rely heavily on rhythm-related audio features as conditions for gesture generation. While these rhythm-based features successfully align gestures with the timing of speech, they often overshadow the sparse yet expressive semantic motion (see Fig. 1). As a result, the generated motions may lack the contextual depth necessary and nuanced expressiveness for natural interaction. Some methods try to address this by incorporating semantic information like emotion, style, and

* Equal contribution.
† Corresponding author.

content[10, 12, 23, 32]. However, the rhythm features tend to dominate, making the models difficult to capture sparse, semantically relevant gestures at the frame level. These rare but impactful gestures are often diluted or overlooked, highlighting the challenge of balancing rhythmic alignment with semantic expressiveness in co-speech motion generation.

In real-world human conversations, we have an observation that while most speech-related gestures are indeed rhythm-related, only a limited number of frames involve semantically emphasized gestures. This insight suggests that co-speech motions can be decomposed into two distinct components: (i) Rhythm-related base motion. These provide a continuous, coherent base motion aligned with the speech rhythm, reflecting the natural timing of speaking. (ii) Semantic-aware sparse motion: These occur infrequently but are essential for conveying specific meanings or emphasizing key points within the conversation.

Inspired by this observation, we propose a new framework SemTalk. SemTalk models the base motion and the sparse motion separately and then fuses them adaptively to generate high-fidelity co-speech motion. Specifically, we first focus on generating rhythm-related base motion by introducing *coarse2fine cross-attention module* and *rhythmic consistency learning*. We design a hierarchical *coarse2fine cross-attention module*, which progressively refines the base motion cues in a coarse-to-fine manner, starting from the face and moving through the hands, upper body, and lower body. This approach ensures consistent rhythmic transmission across all body parts, enhancing coherence base motion. Moreover, we propose a local-global *rhythmic consistency learning* approach, which enforces alignment at both the frame and sequence levels. Locally, a frame-level consistency loss ensures that each frame is precisely synchronized with its corresponding speech features, guaranteeing accurate temporal alignment. Globally, a sequence-level consistency loss sustains a coherent rhythmic flow across the entire motion sequence, preserving consistency throughout the generated gestures.

Furthermore, we introduce *semantic emphasis learning* approach, which focuses on generating semantic-aware sparse motion. This approach utilizes frame-level semantic cues from textual information, high-level speech features, and emotion to identify frames that require emphasis through a learned semantic score produced by a gating strategy, i.e., sem-gate. The sem-gate is designed to dynamically activate semantic motions at key frames through two weighting methods applied on the motion condition and the loss, respectively, and semantic label guidance, allowing the model to produce motion that enhances the motion with deeper semantic meaning and contextual relevance.

Finally, the base motion and sparse motion are integrated through *semantic score-based motion fusion*, which adaptively amplifies expressiveness by incorporating semantic-aware key frames into the rhythm-related base motion.

Our contributions are summarized below:

- We propose SemTalk, a novel framework for holistic co-speech motion generation that separately models rhythm-related base motion and semantic-aware sparse motion, adaptively integrating them via a learned semantic gate.
- We propose a hierarchical *coarse2fine cross-attention module* to refine base motion and a local-global *rhythmic consistency learning* to integrate latent face and hand features with rhythm-related priors, ensuring coherence and rhythmic consistency. We then propose *semantic emphasis learning* to generate semantic gestures at certain frames, enhancing semantic-aware sparse motion.
- Experimental results show that our model surpasses state-of-the-art methods qualitatively and quantitatively, achieving higher motion quality and richer semantics.

2. Related Work

Co-speech Gesture Generation. Co-speech gesture generation aims to produce gestures aligned with speech. Early rule-based methods [7, 19, 21, 22, 41] lacked variability, while deterministic models [5, 7, 29, 36, 46, 49] mapped speech directly to gestures. Probabilistic models, including GANs [1, 17, 40] and diffusion models [2, 10, 47, 54], introduced variability. Some methods incorporated semantic cues, such as HA2G [32] and SEEG [28], which used hierarchical networks and alignment techniques. SynTalker [8] employs prompt-based control but treats inputs as signal strengths rather than fully interpreting semantics. LivelySpeaker [53] combines rhythmic features and semantic cues using CLIP [39] but struggles to integrate gestures with rhythm and capture semantics consistently, moreover, it only provides global control, limiting fine-grained refinement. DisCo [29] disentangles content and rhythm but lacks explicit modeling of sparse semantic gestures. SemTalk addresses this by separately modeling rhythm-related base motion and semantic-aware sparse motion, integrating them adaptively through a learned semantic score.

Holistic Co-speech Motion Generation. Generating synchronized, expressive full-body motion from speech remains challenging, especially in coordinating the face, hands, and torso [9, 31, 34, 37, 48, 52]. Early methods introduced generative models to improve synchronization, but issues persisted. TalkSHOW [48] improved with VQ-VAE [42] cross-conditioning but handled facial expressions separately, causing fragmented outputs. DiffSHEG [9] and EMAGE [31] used separate encoders for expressions and gestures, but their unidirectional flow limited coherence. ProbTalk [33] leverages PQ-VAE [43] for improved body-facial synchronization but mainly relies on rhythmic cues, risking the loss of nuanced semantic gestures. Inspired by TM2D [15], which decomposes dance motion into music-related components, we separately model co-speech motion

into rhythm-related and semantic-aware motion.

3. Method

3.1. Preliminary on RVQ-VAE

Following [4, 16, 51], our approach uses a residual vector-quantized autoencoder (RVQ-VAE) to progressively capture complex body movements in a few players. To retain unique motion characteristics across body regions, we segment the body into four parts—face, upper body, hands, and lower body—each with a dedicated RVQ-VAE, following [3, 31]. This segmentation preserves each part's dynamics and prevents feature entanglement.

3.2. Overview

As shown in Figure 2, our SemTalk pipeline includes two main components: the Base Motion Blocks $f_r(\cdot)$ and the Sparse Motion Blocks $f_b(\cdot)$. Given rhythmic features γ_b, γ_h, a seed pose $\tilde{m}$, and a speaker ID id, the Base Motion Blocks generate rhythm-aligned codes q^b, forming the rhythmic foundation of the base motion:

$$f_r : (\gamma_b, \gamma_h, \tilde{m}, id; \theta_{f_r}) \rightarrow q^b, \tag{1}$$

where θ_{f_r} denotes the learnable parameters of the Base Motion Blocks. The Sparse Motion Blocks then take semantic features ϕ_l, ϕ_g, ϕ_e, along with γ_h, $\tilde{m}$ and id, to produce frame-level semantic codes q^s and semantic score ψ. ψ then triggers these codes only for semantically significant frames, producing a sparse motion representation:

$$f_s : (\phi_l, \phi_g, \phi_e, \tilde{m}, id; \theta_{f_s}) \rightarrow (q^s, \psi), \tag{2}$$

where θ_{f_s} represents the Sparse Motion Block parameters. Finally, the semantic emphasis mechanism $\mathcal{E}$ combines q^b and q^s, guided by ψ, to form the final motion codes q^m:

$$q^m = \mathcal{E}(q^b, q^s; \psi). \tag{3}$$

The motion decoder then uses q^m to generate the output m'.

3.3. Generating Rhythm-related Base Motion

The Base Motion Generation (Fig. 3 a) in SemTalk establishes a rhythmically aligned foundation by leveraging both rhythmic and speaker-specific features, enhancing the naturalness and personalization of generated motion.

Rhythmic Speech Encoding. To synchronize motion with speech, SemTalk incorporates rhythmic features: beats γ_b and HuBERT features γ_h. γ_b, derived from amplitude, short-time energy [11], and onset detection, mark key rhythmic points for aligning gestures with speech. Meanwhile, γ_h, extracted by the HuBERT encoder [18], captures high-level audio traits. In addition to rhythmic features γ,

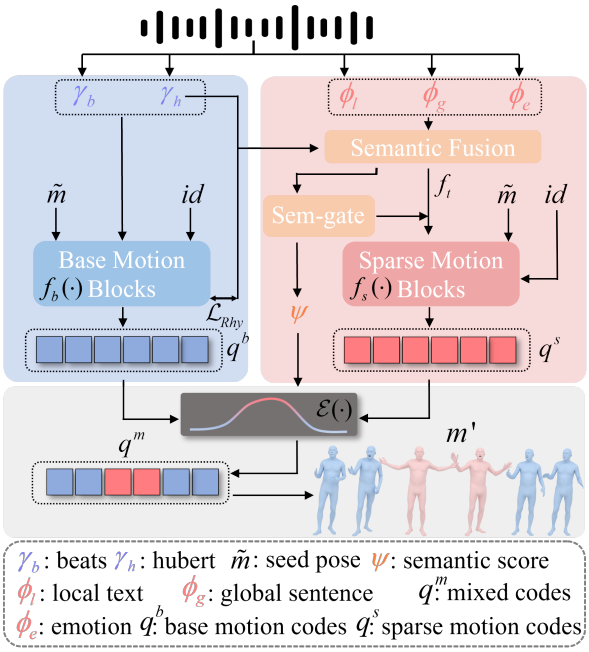

Figure 2. **An overview of the SemTalk pipeline.** SemTalk generates holistic co-speech motion by first constructing rhythm-aligned q^r in f_r, guided by rhythmic consistency loss L_{Rhy}. Meanwhile, f_s produce frame-level semantic codes q^s, activated selectively by the semantic score ψ. Finally, q^m is achieved by fusing q^r and q^s based on ψ, with motion decoder, yielding synchronized and contextually enriched motions.

SemTalk uses a seed pose $\tilde{m}$ and speaker identity id to generate a personalized, rhythm-aligned latent pose p. Then MLP-based Face Enhancement and Body Part-Aware modules utilize γ, p and id to obtain latent face f_e, hands f_h, upper body f_u and lower body f_l.

Coarse2Fine Cross-Attention Module. To facilitate the learning of base motion, we first proposed a transformer-based hierarchical *Coarse2Fine Cross-Attn Module* utilize f_e, f_h, f_u and f_l to obtain latent base motion f_b. The refinement begins with γ for f_e, which guides the rhythmic representation for f_h, followed by conditioning f_u and finally influencing f_l. Since mouth movements closely correspond to speech syllables with minimal delay, we use the face to guide hand motions, inspired by DiffSHEG [9]. As the upper and lower body movements are less directly driven by speech and instead reflect the natural swinging of the hands and torso, we adopt cascading guidance: hands influence the upper body, which in turn drives the lower body. This structured approach, moving from the face to the hands, upper body, and lower body, ensures smooth and coherent motion propagation across the entire body.

Rhythmic Consistency Learning. Inspired by CoG's use of InfoNCE loss [45] to synchronize facial expressions with audio cues, our approach adopts a similar philosophy of

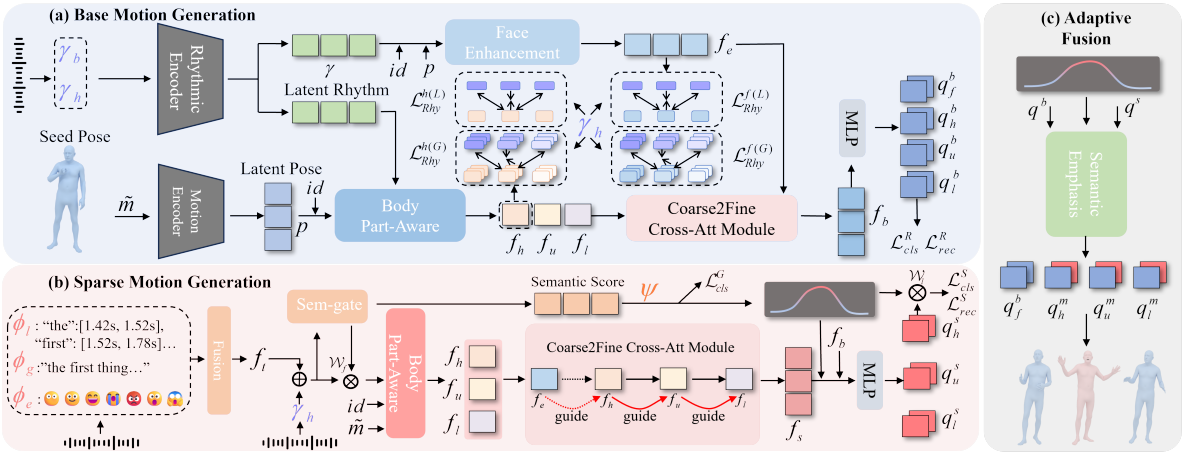

Figure 3. **Architecture of SemTalk.** SemTalk generates holistic co-speech motion in three stages. (a) Base Motion Generation uses *rhythmic consistency learning* to produce rhythm-aligned codes q^b, conditioned on rhythmic features γ_b, γ_h. (b) Sparse Motion Generation employs *semantic emphasis learning* to generate semantic codes q^s, activated by semantic score ψ. (c) Adaptively Fusion automatically combines q^b and q^s based on ψ to produce mixed codes q^m at frame level for rhythmically aligned and contextually rich motions.

aligning motion and speech rhythm. It can be defined as:

$$\mathcal{L}_{\text{Rhy}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\text{sim}\left(h\left(f_i\right),\gamma_h^i\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}\left(h\left(f_i\right),\gamma_h^j\right)/\tau\right)}, \tag{4}$$

where N denotes the number of frames(or the batch size), τ denotes the temperature hyperparameter, $h(\cdot)$ is the projection head for latent motion, f_i and γ_h^i are the latent motion and rhythmic features at frame (or sample) i, and $\text{sim}(\cdot)$ represents cosine similarity.

Unlike CoG, our approach fundamentally differs by incorporating separate local and global rhythmic consistency losses, which are applied to both latent face f_e and latent hands f_h, ensuring a more cohesive and synchronized representation across the entire motion sequence. This rhythmic consistency loss ensures that the motions are not only synchronized at the frame level but also maintain a consistent rhythmic flow across the entire sequence.

The local frame-level consistency loss $\mathcal{L}_{\text{Rhy}}^{(L)}$ aligns the motion features of each frame with the corresponding rhythmic cues γ_h. By leveraging HuBERT features γ_h instead of basic beat features γ_b, which only capture rhythmic pauses, we incorporate rich, high-level audio representations that enhance the model's ability to capture rhythm-related motion patterns and maintain temporal coherence.

The global sentence-level consistency loss $\mathcal{L}_{\text{Rhy}}^{(G)}$ is designed to ensure rhythmic coherence at a global level. Unlike local loss, $\mathcal{L}_{\text{Rhy}}^{(G)}$ reinforces rhythm consistency throughout the sequence, ensuring that the generated motion maintains smooth and rhythm-aligned throughout its duration.

By jointly minimizing $\mathcal{L}_{\text{Rhy}}^{(L)}$ and $\mathcal{L}_{\text{Rhy}}^{(G)}$, *rhythmic consistency learning* enables SemTalk to produce base motions that are rhythmically aligned and temporally cohesive, forming a solid rhythm-related base motion foundation.

3.4. Generating Semantic-aware Sparse Motion

The Sparse Motion Generation (Fig. 3 b) in SemTalk adds semantic-aware sparse motion to base motion by incorporating semantic cues drawn from speech content and emotional tone. By separating rhythm and semantics, this stage enhances motion generation by emphasizing contextually meaningful motion at key semantic moments.

Semantic Speech Encoding. To capture semantic cues in speech, similar to [10], *Semantic Emphasis Learning* combines frame-level text embeddings ϕ_l, sentence-level features ϕ_g from the CLIP model [39], and emotion features ϕ_e from the emotion2vec model [35]. These features form a comprehensive semantic representation f_t, together with audio feature γ_h, that reflects both the content and emotional undertones of speech, enabling SemTalk to activate motions that are sensitive to nuanced semantic cues.

Semantic Emphasis Learning. The process begins by generating f_t, combining local and global cues from text, speech, emotion embeddings and HuBERT features γ_h. Then, the sem-gate leverages multi-modal inputs to generate a semantic score, identifying frames that require enhanced semantic emphasis. The sem-gate in SemTalk refines keyframe motion by applying two forms of weighting methods $\mathcal{W}$: feature weighting $\mathcal{W}_f$ and loss weighting $\mathcal{W}_l$. Using f_t and γ_h, SemTalk computes a semantic score ψ, which dynamically scales feature weighting—filtering back semantic features f_t to activate frames with significant relevance, ensuring that the model emphasizes frames aligned with specific communicative intentions. Second, the loss weighting is applied by supervising ψ, with a classification loss $\mathcal{L}_{cls}^{G}$ based on semantic labels, further enhancing the model's ability to identify key frames. The two weight-

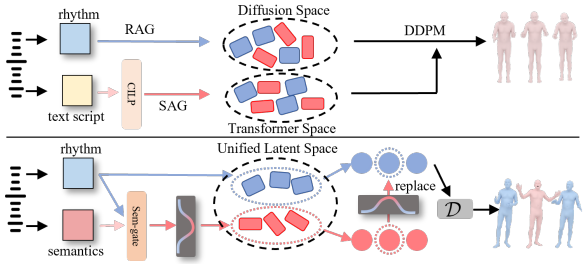

Figure 4. **Concept comparison with LivelySpeaker [53].** (Top) LivelySpeaker generates semantic gestures with CLIP embeddings in SAG and refines rhythm-related gestures separately using diffusion, causing potential jitter. (Bottom) SemTalk integrates text and speech, uses a semantic gate for fine-grained control, and unifies rhythm and semantics for smoother, more coherent motions.

ing methods allow SemTalk to selectively enhance semantic gestures while suppressing uninformative motion, leading to more expressive co-speech motion.

Once ψ is established, it modulates the integration of rhythm-aligned base motion f_b and sparse semantic motion f_s. Through alpha-blending, frames with high semantic relevance draw more from f_s, while others rely on f_b. The final motion codes q^s are computed as:

$$q^s = MLP(\psi f_s + (1-\psi) f_b), \tag{5}$$

To ensure cohesive propagation of semantic emphasis across body regions, we employ the *Coarse2Fine Cross-Attention Module*, similar to Sec. 3.3. In this stage, we focuses solely on body motion, excluding facial movements, as body gestures play a more critical role in conveying semantic meaning in co-speech interactions.

To foster diverse motion generation, SemTalk includes a code classification loss $\mathcal{L}_{cls}$ and a reconstruction loss $\mathcal{L}_{rec}$. These losses are specifically focused on frames with high semantic scores, guiding the model to prioritize the generation of sparse, meaningful gestures.

Discussion. Recently, LivelySpeaker [53] designs the Semantic-Aware Generator (SAG) and Rhythm-Aware Generator (RAG) for co-speech gesture generation, combining them through beat empowerment. While effective, key differences exist between LivelySpeaker and SemTalk, see Fig. 4. First, SAG generates gestures from text using CLIP embeddings, but bridging words and expressive gestures is challenging, causing jitter. SemTalk incorporates speech features (pitch, tone, emotion) alongside text and GT supervision for adaptive gestures. Second, LivelySpeaker applies global control, missing local semantic details, while SemTalk uses fine-grained, frame-level semantic control for subtle variations. Third, LivelySpeaker fuses SAG and RAG in separate latent spaces, leading to misalignment and inconsistencies. SemTalk jointly models rhythm and semantics in a unified framework, ensuring smoother transitions and coherence. We further compare SAG with our semantic gate in experiments.

3.5. Semantic Score-based motion fusion

The Adaptive Fusion stage (Fig. 3 c) in SemTalk seamlessly integrates semantic-aware sparse motion into the rhythmic-related base motion. By strategically enhancing frames based on their semantic importance, it maintains a smooth and natural motion flow across sequences. For each frame i, the semantic score ψ_i computed during the Sparse Motion Generation stage is compared to a threshold β. If $\psi_i > \beta$, the base motion's latent code q_i^r is replaced with the sparse semantic code q_i^s, effectively highlighting expressive gestures where they are most relevant; otherwise, $q_i = q_i^r$.

This selective replacement emphasizes semantically critical gestures while preserving the natural rhythmic base motion. By blending q^b and q^s based on semantic scores, SemTalk adapts to the expressive needs of the speech context while ensuring coherence. Additionally, the convolution structure of the RVQ-VAE decoder ensures smooth transitions between frames, preserving motion continuity.

4. Experiments

4.1. Experimental Setup

Datasets. For training and evaluation, we use two datasets: BEAT2 and SHOW. **BEAT2**, introduced in EMAGE [31], extends BEAT [30] with 76 hours of data from 30 speakers, standardized into a mesh representation with paired audio, text, and frame-level semantic labels. We follow [31] and use the BEAT2-standard subset with an 85%/7.5%/7.5% train/val/test split. **SHOW** [48] includes 26.9 hours of high-quality talk show videos with 3D body meshes at 30fps. Since it lacks frame-level semantic labels, we use the sem-gate from SemTalk, pre-trained on BEAT2, to generate them. Following [48], we select video clips longer than 10 seconds and split the data 80%/10%/10% for train/val/test.

Implementation Details. Our model is trained on a single NVIDIA A100 GPU for 200 epochs with a batch size of 64. We use RVQ-VAE [42], downscaling by 4. The residual quantization has 6 layers, a codebook size of 256 and a dropout rate of 0.2. We use five transformer layers to predict the last five layer codes. In *Base Motion Learning*, τ = 0.1; in *Sparse Motion Learning*, β = 0.5 empirically. The training uses ADAM with a 1e-4 learning rate. Following [31], we start with a 4-frame seed pose, gradually increasing masked frames from 0 to 40% over 120 epochs.

Metrics. We evaluate generated body gestures using FGD [50] to measure distributional alignment with GT, reflecting realism. DIV [26] quantifies gesture variation via the average L1 distance across clips. BC [27] assesses speech-motion synchrony. For facial expressions, we use MSE [47] to quantify positional differences and LVD [48] to measure discrepancies between GT and generated facial vertices.

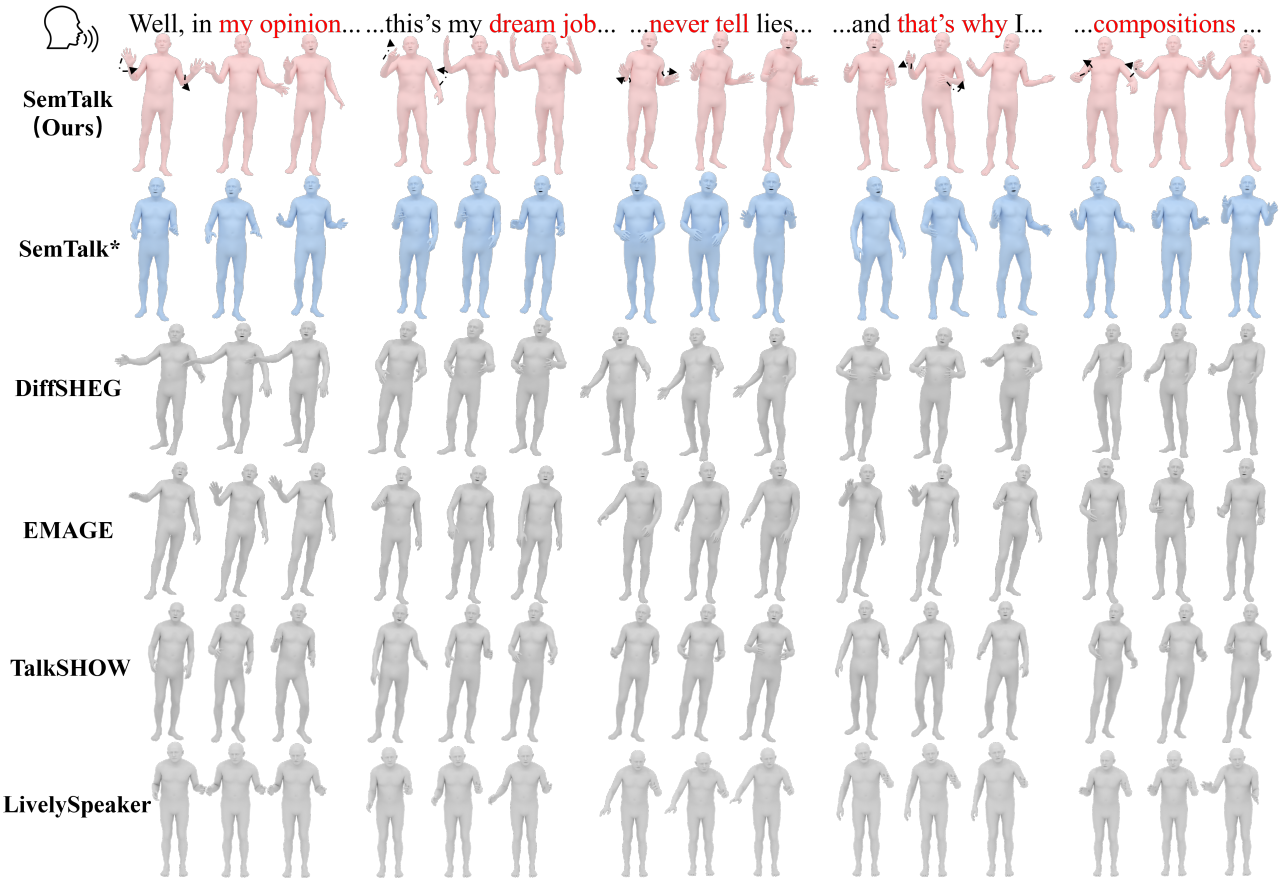

Figure 5. **Comparison on BEAT2 [31] Dataset.** SemTalk* refers to the model trained solely on the Base Motion Generation stage, capturing rhythmic alignment but lacking semantic gestures. In contrast, SemTalk successfully emphasized sparse yet vivid motions. For instance, when saying "my opinion," SemTalk generates a hand-raising gesture followed by an index finger extension for emphasis. Similarly, for "never tell," our model produces a clear, repeated gesture matching the rhythm, reinforcing the intended emphasis.

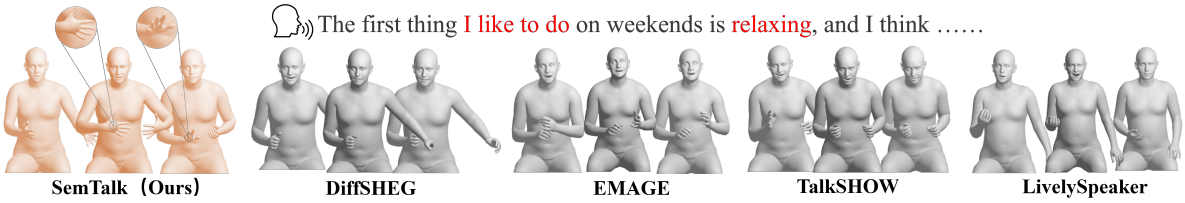

Figure 6. **Comparison on SHOW [48] Dataset.** Our method performs better in motion diversity and semantic richness.

4.2. Qualitative Results

Qualitative Comparisons. We encourage readers to watch our demo video for a clearer understanding of SemTalk's qualitative performance. Our method achieves superior speech-motion alignment, generating more realistic, diverse, and semantically consistent gestures than the baselines. As shown in Fig. 5, LivelySpeaker, TalkSHOW, EMAGE, and DiffSHEG exhibit jitter—EMAGE mainly in the legs and shoulders, while TalkSHOW affects the entire body. LivelySpeaker and DiffSHEG, which focus primarily on the upper body, produce slow and inconsistent motions, especially at speech clip boundaries. DiffSHEG improves gesture diversity over EMAGE and TalkSHOW, though EMAGE maintains greater naturalness. SemTalk surpasses all baselines in both realism and diversity. Compared to SemTalk*, SemTalk generates more expressive gestures, emphasizing key phrases (e.g., raising hands for "dream job" or pointing for "that is why"). While SemTalk* ensures rhythmic consistency, it lacks semantic expressiveness. By integrating frame-level semantic emphasis, SemTalk aligns motion with both rhythm and semantics, demonstrating the effectiveness of *rhythmic consistency learning* and *semantic emphasis learning*. In facial comparisons (Fig. 7), EMAGE shows minimal lip movement, while both DiffSHEG and

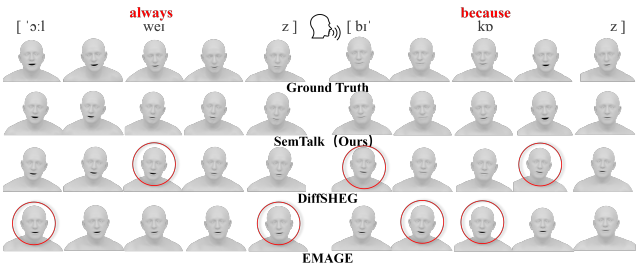

Figure 7. **Facial Comparison on the BEAT2 [31] Dataset.**

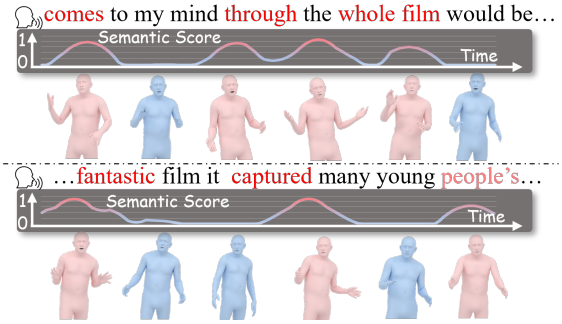

Figure 8. **Qualitative study on semantic score.** Semantic score aligns with keywords, influencing gesture intensity.

EMAGE reveal inconsistencies between lip motion and the rhythm of speech. In contrast, SemTalk produces smooth, natural transitions across syllables, resulting in realistic and expressive lips, significantly surpassing the baselines.

On the SHOW dataset (Fig. 6), SemTalk shows more agile gestures than all baselines, when applied to unseen data. Our method captures natural and contextually rich gestures, particularly in moments of emphasis such as "I like to do" and "relaxing," where our model produces lively hand and body movements that align with the speech content.

Semantic Score. Fig. 8 shows how semantic emphasis influences gesture intensity, with peaks in the semantic score aligning with keywords like "comes," "fantastic," and "captured." By extracting semantic scores from key frames, we track gesture emphasis trends. Furthermore, as shown in Fig. 9, SemTalk adapts to different emotional tones even when the text remains unchanged. This adaptability prevents overfitting to the text itself, allowing the model to generate gestures that vary according to the emotional delivery of the speech. The learned semantic score provides fine-grained, frame-level control, keeping gestures both rhythmically synchronized and semantically aligned in real time.

User Study. We conducted a user study with 10 video samples and 25 participants from diverse backgrounds, evaluating realism, semantic consistency, motion-speech synchrony, and diversity. Participants were required to rank shuffled videos across different methods. As shown in Fig. 10, our approach received dominant preferences across all metrics, especially in semantic consistency and realism.

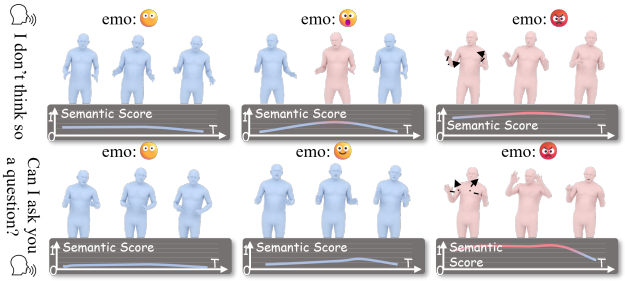

Figure 9. **Same words with different speech from the internet.** "emo" represents different emotional tones extracted from speech. SemTalk can generate different motions, even when the text script is the same, preventing overfitting to the text itself.

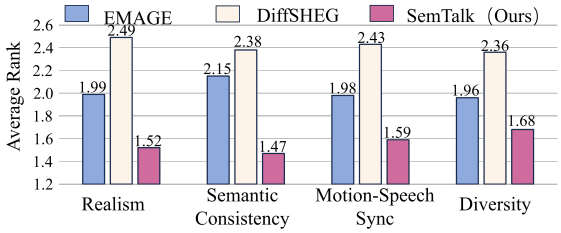

Figure 10. **Results of the user study.**

4.3. Quantitative Results

Comparison with Baselines. As shown in Tab. 1, SemTalk outperforms previous methods on BEAT2, achieving lower FGD, MSE, and LVD, indicating better distribution alignment and reduced motion errors. For fairness, we follow [31] and add a lower-body VQ-VAE to TalkSHOW, DiffSHEG, and SemTalk. Notably, SemTalk significantly reduces FGD, ensuring strong distribution matching. While TalkSHOW and EMAGE achieve competitive diversity (DIV) scores, SemTalk balances high semantic relevance with natural motion flow.

On the SHOW dataset, SemTalk excels with the lowest FGD, MSE, and the highest BC, indicating precise beat alignment with the audio and enhanced semantic consistency in generated motions. Although EMAGE exhibits high DIV, our model achieves comparable results while maintaining smooth, realistic motion free from jitter.

Sem-gate. Tab. 2 highlights the effectiveness of sem-gate. Without sem-gate, the model fails to emphasize key moments. Randomized semantic scores led to poor performance by preventing meaningful frame distinction. Introducing a learned sem-gate even (w/o $\mathcal{W}$) significantly improves semantic alignment and classification accuracy. Refinement is further enhanced through weighting strategies: feature weighting $\mathcal{W}_f$ enhances motion emphasis, while loss weighting $\mathcal{W}_l$ improves FGD and overall accuracy. These results suggest that weighting methods enhance the accuracy of the semantic score and help the model prioritize important frames. The best results come from applying two weighting methods together, where frames with stronger se-

Dataset	Method	FGD↓	BC↑	DIV↑	MSE↓	LVD↓
BEAT2	FaceFormer [13]	-	-	-	7.787	7.593
	CodeTalker [44]	-	-	-	8.026	7.766
	CaMN [30]	6.644	6.769	10.86	-	-
	DSG [47]	8.811	7.241	11.49	-	-
	LivelySpeaker [17]	11.80	6.659	11.28	-	-
	Habibie *et al.* [17]	9.040	7.716	8.213	8.614	8.043
	TalkSHOW [48]	6.209	6.947	**13.47**	7.791	7.771
	EMAGE [31]	5.512	7.724	13.06	7.680	7.556
	DiffSHEG [9]	8.986	7.142	11.91	7.665	8.673
	SemTalk (Ours)	**4.278**	**7.770**	12.91	**6.153**	**6.938**
SHOW	FaceFormer [13]	-	-	-	138.1	43.69
	CodeTalker [44]	-	-	-	140.7	45.84
	CaMN [30]	22.12	7.712	10.37	-	-
	DSG [47]	24.84	8.027	10.23	-	-
	LivelySpeaker [17]	32.17	7.844	10.14	-	-
	Habibie *et al.* [17]	27.22	8.209	8.541	145.6	47.35
	TalkSHOW [48]	24.43	8.249	10.98	139.6	45.17
	EMAGE [31]	22.12	8.280	**12.46**	136.1	42.44
	DiffSHEG [9]	24.87	8.061	10.79	139.0	45.77
	SemTalk (Ours)	**20.18**	**8.304**	11.36	**134.1**	**39.15**

Table 1. **Quantitative comparison with SOTA.** SemTalk consistently outperforms baselines across both the BEAT2 and SHOW datasets. Lower values are better for FMD, FGD, MSE, and LVD. Higher values are better for BC and DIV. We report FGD$\times 10^{-}1$, BC$\times 10^{-}1$, MSE$\times 10^{-}8$ and LVD$\times 10^{-}5$ for simplify.

mantic signals receive higher emphasis. We also compare sem-gate with LivelySpeaker's SAG [53]. We find that replacing the Sparse Motion stage with SAG and substituting motion using GT semantic labels led to poor performance. SAG relies only on text-motion alignment, ignoring emotional tone, making it more prone to overfitting the text. In contrast, our sem-gate applies GT supervision with two weighting methods, achieving more accurate and stable semantic motion.

Ablation Study on Components. We assess the impact of each component of our model on BEAT2 and present the results in Tab. 3, which reveals several key insights (more ablation results please see supplementary material) :

- **Rhythmic Consistency Learning (RC)** not only boosts performance on key metrics like FGD, LVD, and BC but also reduces the MSE, contributing to smoother and more realistic base motion.
- **Semantic Emphasis Learning (SE)** proves essential for selectively enhancing semantic-rich gestures. The inclusion of SE, as shown in rows with SE enabled, improves both diversity (DIV) and FGD, enabling the model to emphasize semantically relevant motions. SE demonstrates its effectiveness in focusing on frame-level semantic information, which contributes to the generation of lifelike gestures with enriched contextual meaning.
- **Coarse2Fine Cross-Attention Module (C2F)** effectively refines motion details, improving BC, FGD, and DIV. When combined with RVQ and RC, C2F achieves the best MSE and LVD, highlighting its role in enhancing motion realism and diversity hierarchically.

Method	FGD↓	BC↑	DIV↑	Acc(%)↑
w/o Sem-gate	4.893	7.702	12.42	-
SAG (LivelySpeaker [53])	4.618	7.682	12.45	-
Sem-gate (Random ψ)	4.634	7.700	12.44	50.07
Sem-gate (w/o $\mathcal{W}$)	4.495	7.633	12.26	72.32
Sem-gate (w/ $\mathcal{W}_f$)	4.408	7.679	12.28	78.52
Sem-gate (w/ $\mathcal{W}_l$)	4.366	**7.772**	11.94	77.83
Sem-gate (ours)	**4.278**	7.770	**12.91**	**82.76**

Table 2. **Ablation study on Sem-gate.** "Acc" denotes semantic classification performance on BEAT2. "w/o Sem-gate" means directly input f_t and γ_h without Sem-gate. "SAG (LivelySpeaker [53])" replaces the Sparse Motion Generation stage with LivelySpeaker's SAG method. "Random ψ" assigns frame-level scores randomly. "w/o $\mathcal{W}$" applies the semantic gate but excludes frame-level weighting. "w/ $\mathcal{W}_f$" applies feature weighting. "w/ $\mathcal{W}_l$" applies loss weighting. (as mentioned in Sec. 3.4). Sem-gate (ours) integrates both the semantic gate and frame-level weighting to enhance emphasis.

RC	SE	C2F	RVQ	FGD↓	BC↑	DIV↑	MSE↓	LVD↓
-	-	-	-	6.234	7.628	11.44	8.239	7.831
-	-	-	√	5.484	7.641	11.84	13.882	15.42
√	-	-	√	4.867	7.701	12.38	6.201	6.928
√	√	-	√	4.526	7.751	12.83	6.215	6.997
-	-	√	√	4.897	7.702	12.42	13.416	15.72
√	√	-	-	5.831	7.758	11.97	6.587	7.106
√	-	√	√	4.397	**7.776**	12.49	**6.100**	**6.898**
√	√	√	√	**4.278**	7.770	**12.91**	6.153	6.938

Table 3. **Ablation study on each key component.** "RC" denotes *rhythmic consistency learning*, "SE" denotes the *semantic emphasis learning*, and "C2F" denotes *Coarse2Fine Cross-Att Module*, "RVQ" denotes the RVQ-VAE.

- **RVQ-VAE (RVQ)** enhances the diversity and realism of generated motion. Though it slightly increases MSE and LVD, it notably improves FGD, leading to more natural motion generation compared to standard VQ-VAE.

5. Conclusion

We propose SemTalk, a novel approach for holistic co-speech motion generation with frame-level semantic emphasis. Our method addresses the integration of sparse yet expressive motion into foundational rhythm-related motion, which has received less attention in previous works. We develop a framework that separately learns rhythm-related base motion through *coarse2fine cross-attention module* and *rhythmic consistency learning*, while capturing semantic-aware motion through *Semantic Emphasis Learning*. These components are then adaptively fused based on a learned semantic score. Our approach has demonstrated state-of-the-art performance on two public datasets quantitatively and qualitatively. The qualitative results and user study show that our method can generate high-quality co-speech motion sequences that enhance frame-level semantics over robust base motions, reflecting the full spectrum of human expressiveness.

Acknowledgments. This work was supported by Alibaba Research Intern Program, the Young Scientists Fund of the National Natural Science Foundation of China No. 624B2110, the National Key Research and Development Program of China No. 2024YFC3015600, the Fundamental Research Funds for Central Universities No.2042023KF0180 & No.2042025KF0053. The numerical calculation is supported by super-computing system in Super-computing Center of Wuhan University and Tongyi Lab, Alibaba Group.

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. 2

[2] 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. 2

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

[4] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023. 3

[5] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 413–420, 1994. 2

[6] Justine Cassell, David McNeill, and Karl-Erik McCullough. Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & cognition*, 7(1):1–34, 1999. 1

[7] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. Beat: the behavior expression animation toolkit. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 477–486, 2001. 2

[8] Bohong Chen, Yumeng Li, Yao-Xiang Ding, Tianjia Shao, and Kun Zhou. Enabling synergistic full-body control in prompt-based co-speech motion generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6774–6783, 2024. 2

[9] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7352–7361, 2024. 2, 3, 8

[10] 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. 2, 4

[11] Selina Chu, Shrikanth Narayanan, and C-C Jay Kuo. Environmental sound recognition with time–frequency audio features. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1142–1158, 2009. 3

[12] Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. Emotional speech-driven animation with content-emotion disentanglement. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–13, 2023. 2

[13] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18770–18780, 2022. 8

[14] Susan Goldin-Meadow. The role of gesture in communication and thinking. *Trends in cognitive sciences*, 3(11):419–429, 1999. 1

[15] Kehong Gong, Dongze Lian, Heng Chang, Chuan Guo, Zihang Jiang, Xinxin Zuo, Michael Bi Mi, and Xinchao Wang. Tm2d: Bimodality driven 3d dance generation via music-text integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9942–9952, 2023. 2

[16] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 3

[17] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 101–108, 2021. 1, 2, 8

[18] 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. 3

[19] Chien-Ming Huang and Bilge Mutlu. Robot behavior toolkit: generating effective social behaviors for robots. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 25–32, 2012. 2

[20] Adam Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004. 1

[21] Michael Kipp. Gesture generation by imitation: From human behavior to computer character animation. Universal-Publishers, 2005. 2

[22] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. Towards a common

framework for multimodal generation: The behavior markup language. In *Intelligent Virtual Agents: 6th International Conference, IVA 2006, Marina Del Rey, CA, USA, August 21-23, 2006. Proceedings 6*, pages 205–217. Springer, 2006. 2

[23] 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

[24] Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. A large, crowdsourced evaluation of gesture generation systems on common data: The genea challenge 2020. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pages 11–21, 2021. 1

[25] Alex Lascarides and Matthew Stone. A formal semantic analysis of gesture. *Journal of Semantics*, 26(4):393–449, 2009. 1

[26] 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. 5

[27] 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. 5

[28] 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

[29] Haiyang Liu, Naoya Iwamoto, Zihao Zhu, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Disco: Disentangled implicit content and rhythm learning for diverse co-speech gestures synthesis. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3764–3773, 2022. 2

[30] 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. 1, 5, 8

[31] 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. 1, 2, 3, 5, 6, 7, 8

[32] 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

[33] 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. 2

[34] Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: Text-aligned whole-body motion generation. *arXiv preprint arXiv:2310.12978*, 2023. 2

[35] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*, 2023. 4

[36] Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics symposium on computer animation*, pages 25–35, 2013. 2

[37] 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. 2

[38] Aslı Özyürek, Roel M Willems, Sotaro Kita, and Peter Hagoort. On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of cognitive neuroscience*, 19(4):605–616, 2007. 1

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 4

[40] Manuel Rebol, Christian Gütl, and Krzysztof Pietroszek. Real-time gesture animation generation from speech for virtual human interaction. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–4, 2021. 2

[41] Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized talking head synthesis. *arXiv preprint arXiv:2301.03786*, 2(4):5, 2023. 2

[42] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2, 5

[43] Hanwei Wu and Markus Flierl. Learning product codebooks using vector-quantized autoencoders for image retrieval. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1–5. IEEE, 2019. 2

[44] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven

3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12780–12790, 2023. 8

[45] Zunnan Xu, Yachao Zhang, Sicheng Yang, Ronghui Li, and Xiu Li. Chain of generation: Multi-modal gesture synthesis via cascaded conditional control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6387–6395, 2024. 3

[46] Sicheng Yang, Zilin Wang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Qiaochu Huang, Lei Hao, Songcen Xu, Xiaofei Wu, Changpeng Yang, et al. Unifiedgesture: A unified gesture synthesis model for multiple skeletons. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1033–1044, 2023. 2

[47] 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, 5, 8

[48] 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. 1, 2, 5, 6, 8

[49] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4303–4309. IEEE, 2019. 2

[50] 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. 5

[51] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021. 3

[52] Jinsong Zhang, Minjie Zhu, Yuxiang Zhang, Zerong Zheng, Yebin Liu, and Kun Li. Speechact: Towards generating whole-body motion from speech. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 2

[53] 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, 5, 8

[54] 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