

CoordSpeaker: Exploiting Gesture Captioning for Coordinated Caption-Empowered Co-Speech Gesture Generation

Fengyi Fang Sicheng Yang Wenming Yang
Tsinghua University

Abstract

*Co-speech gesture generation has significantly advanced human-computer interaction, yet speaker movements remain constrained due to the omission of text-driven non-spontaneous gestures (e.g., bowing while talking). Existing methods face two key challenges: 1) the semantic prior gap due to the lack of descriptive text annotations in gesture datasets, and 2) the difficulty in achieving coordinated multimodal control over gesture generation. To address these challenges, this paper introduces **CoordSpeaker**, a comprehensive framework that enables coordinated caption-empowered co-speech gesture synthesis. Our approach first bridges the semantic prior gap through a novel gesture captioning framework, leveraging a motion-language model to generate descriptive captions at multiple granularities. Building upon this, we propose a conditional latent diffusion model with unified cross-dataset motion representation and a hierarchically controlled denoiser to achieve highly controlled, coordinated gesture generation. CoordSpeaker pioneers the first exploration of gesture understanding and captioning to tackle the semantic gap in gesture generation while offering a novel perspective of bidirectional gesture-text mapping. Extensive experiments demonstrate that our method produces high-quality gestures that are both rhythmically synchronized with speeches and semantically coherent with arbitrary captions, achieving superior performance with higher efficiency compared to existing approaches.*

1. Introduction

Gesture synthesis has garnered significant interest due to its broad applications in human-computer interaction, such as virtual reality [1], games [55], and digital avatars [11, 24]. To enhance the diversity and controllability of gesture generation, various modalities have been explored, including speech audio [43–45], text transcripts [24, 29, 53], emotion [32, 33], style [3, 10, 45], and speaker identity [46]. Among these, two critical aspects are highly emphasized: speech synchronization and semantic correlation. Though

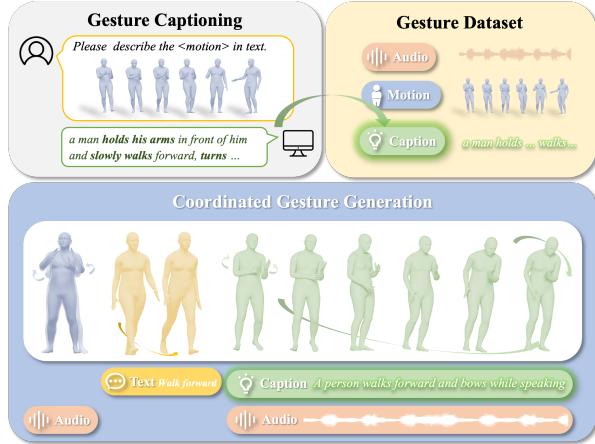


Figure 1. **CoordSpeaker** exploits *gesture captioning* to enable customized *coordinated speaker gesture generation*, producing both co-speech spontaneous gestures and caption-driven non-spontaneous motions. In a speech scenario, our method allows the speaker to naturally walk forward and bow while speaking, seamlessly delivering a closing gesture.

advances in deep neural networks [8, 24, 43] have significantly enhanced generation quality and diversity, current methods primarily focus on co-speech spontaneous gestures, while neglecting text-driven non-spontaneous gestures [47]. This constrains the natural full-body movement of speakers, which is particularly critical in digital avatars applications like public speaking or gaming. For instance, a game NPC may pace or lift an object while speaking, whereas a virtual teacher would naturally deliver lectures (spontaneous gesture) and point to key content on the screen as needed (non-spontaneous gesture). Thus, achieving coordinated gesture generation that ensures rhythmic synchronization for spontaneous gestures and semantic coherence for non-spontaneous gestures is essential.

Coordinated gesture generation remains challenging due to two key factors: 1) **Semantic Prior Gap**: Existing gesture datasets [10, 21, 23, 24] lack direct descriptive text annotations. While speech audio is naturally available, semantic cues are typically inferred through speech tran-

scriptions, which contain only limited and indirect guidance, resulting in weak semantic associations. Moreover, manually annotating gesture semantics is prohibitively expensive, further exacerbating the challenge. Gestures inherently possess richer semantic properties, which require a more direct and structured approach to extract and utilize. Although some methods attempt to alleviate this issue by incorporating additional motion datasets for joint training [47] or motion-text alignment pretraining [4], they still struggle with the inherent semantic gap. The former fails to achieve joint control, while the latter introduces additional training and inference costs.

2) Coordinated Multimodal Control Challenge: A fundamental challenge in generation domain lies in the effective coordination of heterogeneous multimodal conditions [26, 42, 51], particularly when they impose distinct or conflicting control objectives. Unlike naturally paired modalities (e.g., speech-text transcription), combining indirectly related ones (e.g., audio and description) requires more carefully balancing conflicts and complementarities. Previous methods simply concatenate multimodal conditions [4, 43, 47], treating them merely as signals of varying intensities rather than modeling their interactions, typically leading to inharmonious control. Given these challenges, it is critical to bridge the semantic gap and develop a coordinated gesture generation method that enables harmonious multimodal control with minimal cost.

To tackle these challenges, we propose *CoordSpeaker* (Fig. 1), a novel coordinated gesture generation approach that achieves both rhythmically synchronized and semantically consistent gesture synthesis while maintaining high generation quality and computational efficiency. Firstly, to mitigate the semantic prior gap, we introduce a gesture captioning framework, utilizing a motion-language model to generate descriptive gesture captions and a multi-granular captioning mechanism to enable precise semantic guidance. Secondly, to address the coordinated multimodal control challenge, we develop a coordinated gesture generation model learning a unified latent motion representation for cross-dataset modeling and leveraging a hierarchically controlled denoiser to ensure coordinated condition injection and efficient generation. To the best of our knowledge, this work represents the first exploration of gesture captioning as a solution to semantic prior gap, facilitating coordination between speech and semantics in gesture generation, meanwhile offering a novel perspective for bidirectional gesture-text mapping. Our contributions are as follows:

- We propose *CoordSpeaker*, a coordinated gesture generation approach producing both co-speech spontaneous gestures and caption-driven non-spontaneous motions.
- We present the first gesture captioning framework to bridge the semantic prior gap of gesture data and ensure precise multi-granular caption alignment.
- We introduce a hierarchical conditional latent diffusion

model enabling efficient and coordinated multimodal controlled gesture generation.

- Extensive experiments show our method effectively produces semantically coherent, rhythmically synchronized gestures while maintaining high fidelity and efficiency.

2. Related Work

2.1. Gesture Synthesis

Gesture synthesis is a complex task that aims to generate gesture motion sequences from multimodal inputs [27]. Early approaches employed RNNs [23, 48] and Transformers [29, 53] for gesture sequence modeling, but suffered from limited temporal modeling capacity and computational challenges, respectively. Diffusion models [2, 40, 52] have gained prominence in motion generation due to powerful generative capabilities. However, operating directly in motion space leads to high computational costs and training instability. In comparison, latent diffusion models (LDMs) [35] enable powerful generative capabilities while maintaining computational efficiency through low-dimensional latent space operations. Although current co-speech gesture generation methods have achieved significant progress in speech synchronization [5, 49, 54], semantic correlation is still limited by the lack of descriptive text annotations for gesture data [23, 24]. Yang et al. [47] propose a co-speech and text-driven approach, but it still suffers from semantic annotation gaps and does not support coordinated generation conditioned on joint signals. Chen et al. [4] leverage prompt-motion alignment pre-training to generate implicit text labels while introducing additional training and inference costs. In contrast to existing methods, we propose a coordinated gesture generation approach, which enables both semantic correlation and rhythmic consistency while maintaining low annotation and computational costs.

2.2. Motion-Text Translation

Human motion exhibits semantic coupling similar to natural language and is often viewed as a form of body language [17]. Previous research has explored various text-related motion tasks, including text-to-motion generation [12, 40], motion-to-text captioning [13, 17], and unified motion-language modeling [17, 39]. Recent text-to-motion works [7, 9] leverage pre-trained language models to extract semantics for motion control. Motion captioning aims to describe human motion using natural language, with early approaches relying on statistical models and RNNs to learn motion-language mappings [38]. More recently, bidirectional motion-text translation has gained attention. TM2T [13] pioneered this direction through tokenization, unified motion-language models [17, 18, 25] have emerged by fusing language data with large-scale motion models. Gestures inherently possess natural semantic properties, while no ef-

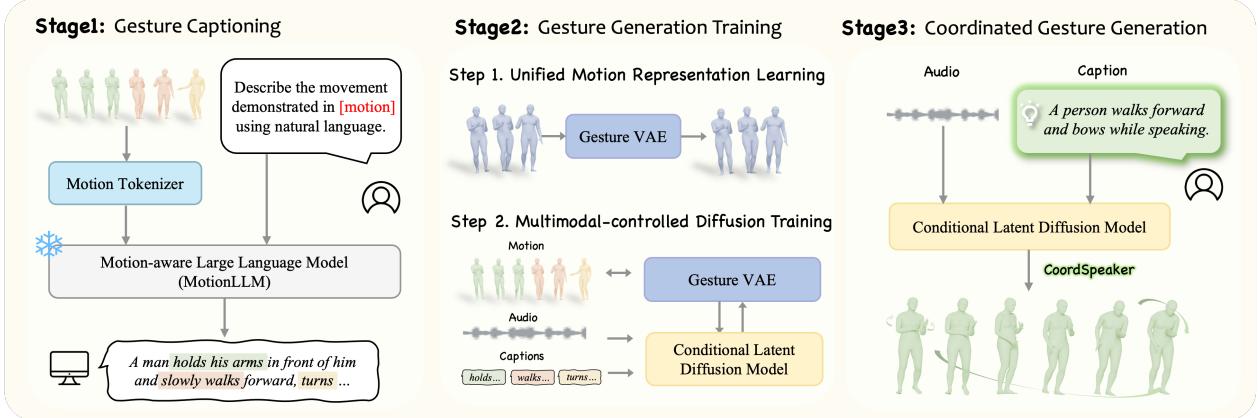


Figure 2. **Overview of CoordSpeaker.** We first introduce Gesture Captioning (Sec. 3.1) to bridge the semantic prior gap of gesture data, generating descriptive, multi-granular gesture captions at low cost. Subsequently, we propose a Coordinated Gesture Generation Model (Sec. 3.2) enabling harmonious coordination over heterogeneous multi-modal and multi-scale conditions. Our model can generate both rhythmically-synchronous and semantically-coherent gestures with high quality and superior efficiency.

fort has been made to explore gesture understanding and captioning. To address this gap, this paper first introduces a gesture captioning framework that bridges the lack of gesture captions and enables fine-grained semantic control over non-spontaneous gestures.

3. Method

In this section, we present a comprehensive framework for coordinated caption-empowered co-speech gesture generation (Fig. 2, 3). We first introduce a Gesture Captioning Framework (Sec. 3.1) to bridge the semantic prior gap, leveraging a motion-language model combined with a multi-granular captioning mechanism to generate descriptive, precisely aligned gesture captions, enabling fine-grained semantic injection over gesture generation. Subsequently, we propose a Coordinated Gesture Generation Model (Sec. 3.2) comprising two key components: (1) a gesture VAE that learns a unified low-dimensional latent representation, enabling compact cross-dataset motion modeling, and (2) a conditional latent diffusion model with a hierarchically controlled denoiser ensuring coordinated multimodal-controlled gesture generation. Our model can generate both rhythmically-synchronous and semantically-coherent speaker gestures with high quality and efficiency.

3.1. Gesture Captioning

3.1.1. Motion-Language Modeling

Our gesture captioning framework (Fig. 3, Top) comprises two main components: a *motion tokenizer* and a *motion-aware large language model (MotionLLM)*. The motion tokenizer is built upon the VQ-VAE architecture used in [13, 50], consisting of an encoder \mathcal{E}_M and a decoder \mathcal{D}_M that generates discrete motion tokens. The motion-aware language model adopts a transformer-based architecture with a

unified text-motion vocabulary $\mathbf{V} = \{\mathbf{V}_t, \mathbf{V}_m\}$, enabling flexible joint modeling of text and motion within a single model. To better match the general motion-language space, the gesture sequence is first converted into a unified motion representation space (detailed in Sec. 3.2.1), then projected to a 22-joint feature subset. Specifically, given an M -frame gesture motion sequence $m^{1:M} = \{x^i\}_{i=1}^M$, the motion tokenizer encodes and quantizes it into a discrete motion token sequence $s^{1:L} = \{s^i\}_{i=1}^L$. A carefully designed *prompt template* is then tokenized into text tokens $w^{1:N} = \{w^i\}_{i=1}^N$. Subsequently, the discrete motion and text tokens are mixed and jointly fed into the motion-aware language model to generate the corresponding gesture caption $\hat{w}^{1:L} = \{\hat{w}^i\}_{i=1}^L$. In practice, we leverage MotionGPT [17] for motion-language modeling and freeze it during inference. Notably, the captions are generated and cached offline, incurring no additional overhead. This framework produces direct and descriptive captions for gesture data, efficiently addressing the semantic prior gap, enabling precise semantic control over gesture generation. See supplementary material (Sec. 6.2) for more details.

3.1.2. Multi-Granular Captioning

Precise semantic control via gesture captions is challenging due to temporal dynamics and varying semantic granularity. To address this, we propose a *multi-granular captioning* mechanism that enables fine-grained caption alignment across temporal and semantic scales. We first introduce a *hierarchical* caption manner integrating both local and global gesture captions. Given a gesture sequence, we segment it as $m^{1:M} = \{m^{i:i+K-1}\}_{i=1}^{M-K+1}$, where K denotes the segment length. Each segment is processed by our gesture captioning framework independently to generate *local* captions $\mathbf{C}_{\text{local}} = \{w^i\}_{i=1}^L$. In contrast, *global* caption is formed by concatenating local captions with separator

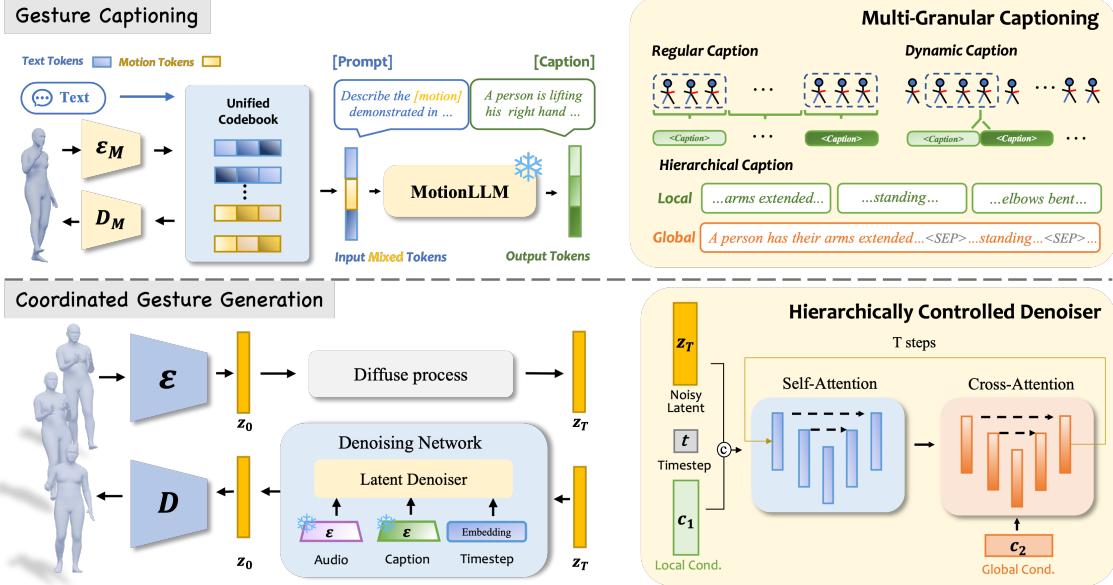


Figure 3. Model overview. **(Top) Gesture Captioning Framework:** A motion tokenizer and a motion-aware language model (MotionLLM) generate descriptive gesture [captions] from predefined [prompt] and [motion] inputs, addressing the semantic prior gap efficiently. A multi-granular captioning mechanism further enhances multi-scale semantic alignment via three strategies: Regular, Dynamic, and Hierarchical. **(Bottom) Coordinated Gesture Generation Model:** A gesture VAE first learns a unified latent motion space for cross-dataset modeling. A conditional latent diffusion model with a hierarchically controlled denoiser enables efficient and coordinated gesture generation via hierarchical multimodal condition injection.

tokens $\mathbf{C}_{\text{global}} = \{\mathbf{C}_{\text{local}}^1 <\text{SEP}> \mathbf{C}_{\text{local}}^2 <\text{SEP}> \dots \mathbf{C}_{\text{local}}^{M/K}\}$. While local captions provide fine-grained semantic supervision at the segment level, a global caption summarizes the overall gesture semantics across the entire sequence. To instantiate this mechanism, we explore three captioning strategies (Fig.3, Top): (1) *Regular Caption* applies uniform temporal segmentation to generate local captions for each fixed-length gesture segment, offering precise temporal alignment. (2) *Dynamic Caption* randomly samples segments during training, and introduces stochastic sampling and caption mixing for local captions, enhancing the robustness to varying temporal patterns and flexible caption combinations. (3) *Hierarchical Caption* combines both local and global captions to incorporate complementary fine-to-coarse semantic cues. This multi-granular captioning mechanism ensures precise caption alignment and enables flexible multi-scale semantic control over gesture generation.

3.2. Coordinated Gesture Generation

3.2.1. Unified Motion Representation

To leverage additional human motion semantic priors, we incorporate the human motion dataset HumanML3D [12] and introduce a *unified motion representation*. The various data formats are first converted into a unified feature format and then fit into a representative latent space via our motion encoder \mathcal{E} (Fig. 3). Specifically, data in different formats

is first converted into SMPL-X [30] axis-angle representation. After scaling and initial orientation adjustment, 3D joint positions are obtained through SMPL-X forward computation. Subsequently, following [12, 47], the i -th motion frame is represented as $x^i = \{\dot{r}^a, \dot{r}^x, \dot{r}^z, r^y, \mathbf{j}^p, \mathbf{j}^v, \mathbf{j}^r, \mathbf{c}^f\}$, where $\dot{r}^a, \dot{r}^x, \dot{r}^z, r^y$ denote the root joint's angular velocity, linear velocities and height, $\mathbf{j}^p, \mathbf{j}^v, \mathbf{j}^r$ represent joint positions, velocities and rotations, and \mathbf{c}^f indicates foot contact. We employ 55 joints to accommodate gesture data better. Consequently, each motion frame is represented by a 659-dimensional feature vector, denoted as $x \in \mathbb{R}^{T \times 659}$, where T is the sequence length.

3.2.2. Gesture VAE

A transformer-based [31] VAE model (Fig. 3) is adopted to encode motion representation into a compact and informative latent space, which consists of an encoder \mathcal{E} and a decoder \mathcal{D} , both enhanced with long skip connections to preserve motion details. Specifically, the motion sequence $x^{1:L}$ is first encoded into a latent vector $z \in \mathbb{R}^{n \times d}$ through the encoder \mathcal{E} , where d denotes the latent dimension. The encoder takes frame-wise motion features and learnable distribution tokens as input, producing Gaussian distribution parameters μ and σ for the motion latent space. These parameters are used to reparameterize [19] the latent vector through the standard VAE sampling process. For motion decoding, \mathcal{D} employs a cross-attention [41] mechanism. It

takes zero motion tokens as queries and the latent vector z as key and value, generating the reconstructed motion sequence $\hat{x}^{1:L}$. The VAE is trained by minimizing a combination of Mean Squared Error (MSE) for reconstruction accuracy and Kullback-Leibler (KL) divergence to regularize the encoded latent distribution $q(z|x^{1:L}) = \mathcal{N}(z; \mu_{\mathcal{E}}, \sigma_{\mathcal{E}}^2)$ toward a standard gaussian distribution:

$$\mathcal{L}_{\text{VAE}} = \|x^{1:L} - \hat{x}^{1:L}\|_2^2 + \beta \text{KL}(q(z|x^{1:L}) \| p(z)) \quad (1)$$

where β balances the reconstruction and regularization terms. This latent representation enables efficient gesture synthesis while maintaining motion fidelity and diversity.

3.2.3. Conditional Latent Diffusion Model

We introduce a *conditional latent diffusion model* with a hierarchically controlled denoiser (Fig. 3, Bottom) to generate high-quality, coordinated gestures in learned latent space.

Diffuse Process. The diffuse process in latent space follows a Markov chain that gradually adds Gaussian noise to the latent vector $z \in \mathbb{R}^{n \times d}$. For a noising step $t \in [1, T]$, the forward process is defined as:

$$q(z_t|z_{t-1}) = \mathcal{N}(\sqrt{\alpha_t} z_{t-1}, (1 - \alpha_t)\mathbf{I}), \quad (2)$$

$\alpha_t \in (0, 1)$ is constant variance schedule, $z_T \sim \mathcal{N}(0, \mathbf{I})$.

Hierarchically Controlled Denoiser. We propose a hierarchically controlled denoiser ϵ_{θ} to facilitate coordinated multimodal control over gesture generation. The denoiser employs a transformer-based endoder-decoder architecture to predict and remove noise iteratively. Starting from random noise z_T , the denoiser gradually recovers the latent vector \hat{z}_0 , which is then decoded into gesture motion through \mathcal{D} . Specifically, given the caption embedding \mathbf{C} and audio embedding \mathbf{A} , the denoising process is controlled by *hierarchical two-stage condition injection* (Fig. 3): First, the *local* condition $\mathbf{c}_1 = \{\mathbf{C}_{\text{local}}, \mathbf{A}\}$ is concatenated with the noised latent and timestep embeddings before being fed into the denoiser encoder \mathcal{E}_d performing self-attention, ensuring precise semantic and rhythmic synchronization with the local gesture segments. Second, the *global* condition $\mathbf{c}_2 = \{\mathbf{C}_{\text{global}}\}$ is incorporated into the denoiser decoder \mathcal{D}_d through cross-attention, enhancing both the overall gesture coherence and semantic relevance with high-level context. Overall, the reverse process is defined as follows:

$$h = \mathcal{E}_d(\text{concat}(z_t, t, \mathbf{c}_1)), \quad \epsilon_{\theta}(z_t, t, \mathbf{c}) = \mathcal{D}_d(h, \mathbf{c}_2), \quad (3)$$

where t is the timestep embedding, and the predicted noise ϵ_{θ} is used to recover \hat{z}_{t-1} . This hierarchical architecture aligns with our multi-granular captions effectively coordinates different modalities and scales, ensuring balanced contributions for heterogeneous conditions.

Classifier-free Guidance. To enhance the quality and controllability of generated gestures, we adopt classifier-free guidance [15]. During training, 10% of the condition inputs are randomly masked [36] to learn both conditional and unconditional distributions. During inference, noise prediction is computed as a weighted combination of outputs with different guidance:

$$\begin{aligned} \epsilon_{\theta}^s(z_t, t, \mathbf{c}) = & s_1 \epsilon_{\theta}(z_t, t, \mathbf{c} = \{\emptyset, \mathbf{A}\}) \\ & + s_2 \epsilon_{\theta}(z_t, t, \mathbf{c} = \{\mathbf{C}, \emptyset\}) \\ & + (1 - s_1 - s_2) \epsilon_{\theta}(z_t, t, \mathbf{c} = \{\emptyset, \emptyset\}), \end{aligned} \quad (4)$$

where s_1, s_2 are guidance scales for audio \mathbf{A} and caption \mathbf{C} , respectively.

Training and Inference. The latent diffusion model is trained with a ℓ_2 objective [16]:

$$\mathcal{L}_{\text{Diff}} = \mathbb{E}_{\epsilon, t} [\|\epsilon - \epsilon_{\theta}(z_t, t, \mathbf{c})\|_2^2], \quad (5)$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $z_0 = \mathcal{E}(x^{1:L})$. During inference, our model employs DDIM sampler [37] with 50 denoising steps to predict the latent \hat{z}_0 , then decodes it to motion sequence $\hat{x}^{1:L}$ using a forward pass through decoder \mathcal{D} .

4. Experiments

In this section, we evaluate our method through both quantitative and qualitative analyses in four aspects: (1) **Coordinated Gesture Generation**. Assessing the model's ability to generate speech-synchronized, semantically-relevant gestures under joint speech and caption control, compared to state-of-the-art baselines. (2) **Text-Driven Motion Generation**. Comparing with state-of-the-art text-to-motion methods to evaluate semantic understanding and non-spontaneous gesture generation. (3) **Gesture Captioning**. Evaluating the quality, human-alignment, and diversity of generated captions marks the first approach to bridging the semantic gap in gesture datasets. (4) **Ablation Studies**. Analyzing the impact of key components in our method.

4.1. Experimental Setup

Datasets We jointly train on the audio-to-gesture dataset BEAT [23] and the text-to-motion dataset HumanML3D [12]. BEAT offers 76 hours of speech-gesture data. We utilize four English speakers' gestures following [23]. HumanML3D contains 14,616 motions with 44,970 text descriptions, providing rich semantic priors. For coordinated generation, missing text in BEAT is generated via our captioning framework, while missing audio in HumanML3D is set to zero following [47].

Metrics The evaluation is conducted from three perspectives: (1) Reconstruction Quality: Motion smooth-

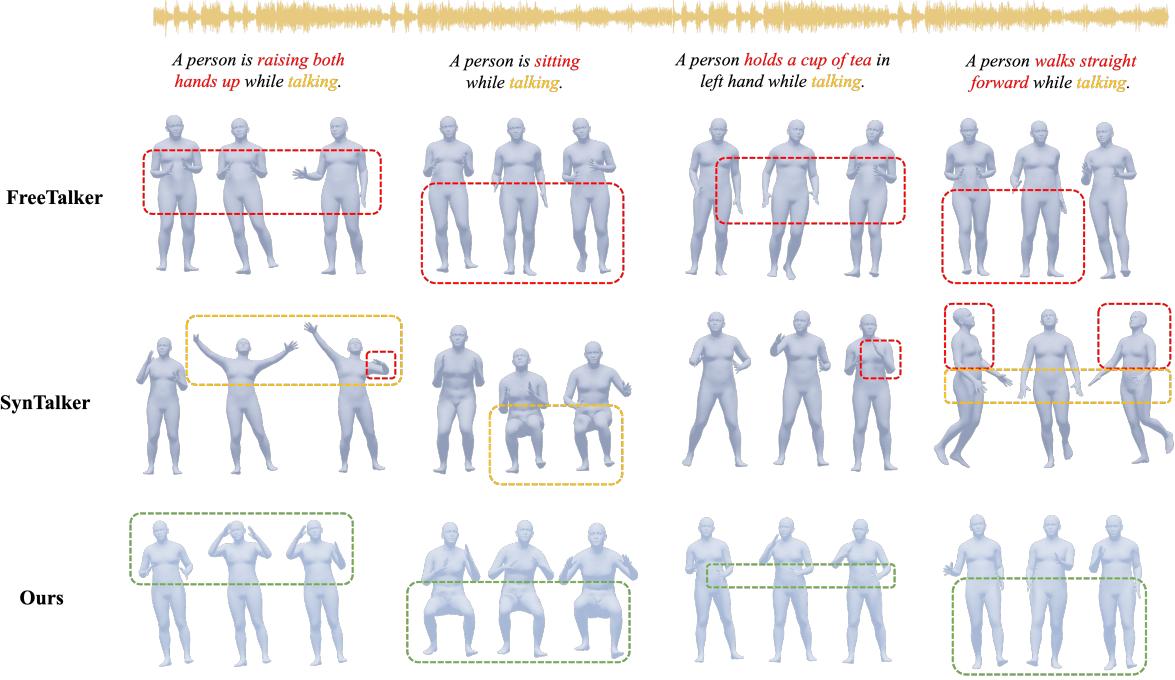


Figure 4. Qualitative comparison of coordinated gesture generation. Red boxes highlight semantic inconsistencies, yellow boxes indicate unnatural motions, and green boxes denote well-coordinated natural gestures. More results are in supplementary material (Sec. 8).

Table 1. Quantitative results of baseline comparisons and ablation studies. Metrics are reported with 95% confidence interval over 20 runs. ‘→’ denotes the closer to the real motion the better. We report BC $\times 10^{-1}$ and Top-1 R-Precision.

Methods	Reconstruction		Audio-to-Gesture			Text-to-Motion			
	Jerk→	Accel.→	FGD↓	BC↑	L1Div↑	FID↓	MM-Dist↓	Div→	R-Precision↑
GT	$1.165 \pm .000$	$0.043 \pm .000$	-	-	-	-	$6.205 \pm .043$	$5.512 \pm .114$	$0.140 \pm .008$
FreeTalker	$0.611 \pm .013$	$0.030 \pm .000$	$2.101 \pm .026$	$1.147 \pm .028$	$11.332 \pm .025$	$0.761 \pm .048$	$6.737 \pm .051$	$5.396 \pm .127$	$0.102 \pm .008$
Ours	$1.190 \pm .015$	$0.039 \pm .001$	$3.173 \pm .123$	$1.327 \pm .049$	$10.861 \pm .066$	$1.118 \pm .061$	$6.814 \pm .056$	$5.558 \pm .126$	$0.100 \pm .008$
Ours-w/o hed.	$1.201 \pm .017$	$0.038 \pm .001$	$2.302 \pm .061$	$1.910 \pm .004$	$12.781 \pm .044$	$1.260 \pm .063$	$6.872 \pm .058$	$5.303 \pm .107$	$0.102 \pm .010$
Ours-w/o mgc.	$1.239 \pm .013$	$0.039 \pm .000$	$3.123 \pm .139$	$2.256 \pm .045$	$15.363 \pm .103$	$2.568 \pm .099$	$7.031 \pm .044$	$5.447 \pm .150$	$0.082 \pm .006$
Ours-w/o mo.	$1.005 \pm .014$	$0.032 \pm .000$	$2.654 \pm .041$	$2.627 \pm .043$	$19.600 \pm .089$	$3.911 \pm .163$	$7.664 \pm .050$	$4.070 \pm .117$	$0.043 \pm .004$

ness and naturalness are measured using jerk and acceleration [20]. (2) Audio-to-Gesture: Gesture realism is assessed via FGD [48], diversity by the mean L1 distance between gestures [24], and speech-motion synchronization by Beat Consistency (BC) [22]. (3) Text-to-Motion: Frechet Inception Distance (FID) [7] and Diversity (DIV) [11] evaluate realism and diversity, while motion-retrieval precision (R Precision) and Multimodal Distance (MM-Dist) assess motion-text alignment [7]. Implementation details are in supplementary material (Sec. 6).

4.2. Coordinated Gesture Generation

Qualitative Comparison We first evaluate the performance of our method in generating coordinated, speech-synchronized, semantically-relevant gestures. As depicted

in Fig. 4, we compare with the only two works [4, 47] that addressed related coordinated generation tasks. Following [4], results are presented with a *calm* audio input and four distinct text captions. Results show that our method generates well-coordinated gestures that are both rhythmically aligned and semantically consistent, while achieving more natural appearances. In contrast, FreeTalker [47] fails to produce semantic motions under all settings, focusing solely on speech-driven gestures. While SynTalker [4] captures partial semantic relevance, it struggles to jointly coordinate multimodal conditions, leading to control conflicts (e.g., losing co-speech gestures in “*walks while talking*”), inconsistent details (e.g., incorrect hand poses for “*raising*” and “*holding*”), and unnatural stiffness (e.g., incorrect turning in “*walks straight forward*”). Ablations (Sec. 4.5) fur-

Table 2. User preference win rates (%) show that our results are perceived as more realistic and controllable, outperforming previous work [4] by 4.0%, 5.5%, and 1.5% in naturalness ($p < 0.001$), synchrony ($p < 0.05$), and matching ($p < 0.001$), respectively. All results are reported with 95% confidence intervals.

Methods	Naturalness	Synchrony	Matching
FreeTalker	18.0 ± 5.32	24.0 ± 5.92	15.5 ± 5.02
SynTalker	32.0 ± 6.46	26.5 ± 6.12	31.5 ± 6.44
Ours-w/o cap.	14.0 ± 4.81	17.5 ± 5.27	20.0 ± 5.54
Ours	36.0 ± 6.65	32.0 ± 6.46	33.0 ± 6.52

ther demonstrate that our superior semantic understanding stems from gesture captioning and enhanced multimodal coordination from proposed hierarchically controlled denoiser. More results are in supplement (Sec. 8 and video).

Quantitative Results Since only FreeTalker [47] and SynTalker [4] share a similar research scope with our work, and SynTalker does not provide a quantitative evaluation method under multimodal conditioning, we reproduce FreeTalker as primary baseline in this comparison, and present additional single-condition comparison with SynTalker in Sec. 4.3. As shown in Table 1, our method achieves comparable or superior results across all evaluation metrics, surpassing FreeTalker in reconstruction quality, beat consistency (BC 0.180 \uparrow), and diversity (Div 0.162 \uparrow), while achieving better coordination and substantially faster inference (Sec. 4.5). Note that due to existing quantitative metrics solely focus on single-modality factors, our aim is to achieve balance performance across multimodal metrics. More details are in supplement (Sec. 6.4).

Perceptual Study A user study was conducted to assess gesture quality under joint speech and caption control. 20 participants were recruited to evaluate 10 pairs of 9-second results based on: (i) *Naturalness*: realism and naturalness of generated gestures; (ii) *Synchrony*: synchronization with speech; (iii) *Matching*: alignment with text captions. Participants viewed video clips from different models and selected the best one for each aspect. Table 2 shows our proposed model outperforms others. A chi-square test further confirms the significant differences across methods in all aspects (Naturalness: $\chi^2=27.20$, $p=0.000005$; Synchrony: $\chi^2=8.68$, $p=0.034$; Matching: $\chi^2=17.72$, $p=0.0005$). More details are in supplementary material (Sec. 10).

4.3. Text-Driven Motion Generation

We further compare our method with SOTA text-to-motion approaches to validate its capacity in capturing semantics and generating non-spontaneous motions. For fair comparison, we report results on the HumanML3D test set using

only text condition. Table 3 shows our method achieves advanced performance compared to SOTA models, attaining the best text alignment (MM-Dist 3.584) and the second-best generation fidelity (FID 0.405). This underscores the effectiveness of our hierarchically controlled denoiser in integrating fine-grained semantics. Notably, our approach significantly outperforms the similar synergistic method SynTalker, highlighting its superiority in both coordinated generation and enhanced semantic control.

4.4. Gesture Captioning

Qualitative Results As shown in Fig. 5a, the proposed captioning framework demonstrates practical capabilities in generating descriptive gesture annotations. Our method can accurately describe both fine-grained hand movements (e.g., “*moving both hands...*”) and coarse-grained full-body actions (e.g., “*moves their waist...*”). Furthermore, the proposed approach proves advantageous in capturing continuous composite actions within extended time windows (e.g., “*a person [motion 1], then [motion 2] as [motion 3]*”). This demonstrates the effectiveness of our method in bridging the semantic prior gap in gesture datasets. Nevertheless, we also observe that the model encounters challenges in perceiving temporal order in longer gesture sequences. More results and discussion of limitations are provided in supplementary material (Sec. 8, 9).

Quantitative Captioning Evaluation Quantitative evaluation is inherently challenging due to the lack of Ground-Truth captions, precisely the “semantic prior gap” we aim to address. To compensate, we construct a 200-sample *expert-annotated set* and evaluate from three aspects: (1) motion-semantic relevance (MotionClipScore [39]), (2) language quality (via GPT [28]), (3) reference alignment (BLEU, ROUGE). As shown in Fig. 5b, our model performs comparably to expert annotations in both relevance and quality, with slightly better fluency (4.380 vs. 4.375), comparable ClipScore (0.690 vs. 0.709), and reasonable alignment for the open-ended task. Combined with qualitative results (Fig. 5a), these further confirm the effectiveness and robustness of our captioning framework.

4.5. Ablation Studies

Ablation studies are conducted to evaluate the impact of different components. Qualitative results are shown in Fig. 5c, while quantitative results are presented in Table. 1.

Multi-Granular Gesture Caption Fig. 5c (middle) shows that removing gesture captions results in only basic co-speech gestures, lacking meaningful non-spontaneous movements. Table 1 further demonstrates that removing multi-granular captioning (-w/o mgc.) significantly reduces semantic relevance, as evidenced by an increased MM-Dist

Table 3. Comparison with state-of-the-art methods on HumanML3D test set. Metrics follow the standard protocol in [7] and are reported with 95% confidence interval over 20 runs. ‘→’ denotes the closer to the real motion the better.

Methods	FID↓	MM-Dist↓	Div→	R-Precision↑		
				Top-1	Top-2	Top-3
GT	0.001 ^{±.001}	3.378 ^{±.007}	10.471 ^{±.083}	0.490 ^{±.003}	0.682 ^{±.003}	0.783 ^{±.003}
MDM [40]	1.390 ^{±.088}	4.599 ^{±.037}	10.704 ^{±.066}	0.363 ^{±.007}	0.553 ^{±.008}	0.662 ^{±.007}
T2M-GPT [50]	0.564 ^{±.012}	3.867 ^{±.008}	10.558 ^{±.083}	0.433 ^{±.003}	0.615 ^{±.002}	0.716 ^{±.003}
MLD [7]	0.963 ^{±.029}	3.898 ^{±.012}	10.401 ^{±.096}	0.429 ^{±.003}	0.613 ^{±.003}	0.717 ^{±.002}
MoMask [14]	0.222 ^{±.007}	3.620 ^{±.011}	10.621 ^{±.096}	0.461 ^{±.002}	0.657 ^{±.003}	0.760 ^{±.002}
SynTalker [4]	4.385 ^{±.034}	4.499 ^{±.012}	9.374 ^{±.073}	0.375 ^{±.003}	0.564 ^{±.003}	0.681 ^{±.002}
Ours	0.405 ^{±.012}	3.584 ^{±.012}	9.109 ^{±.235}	0.424 ^{±.003}	0.601 ^{±.003}	0.702 ^{±.003}

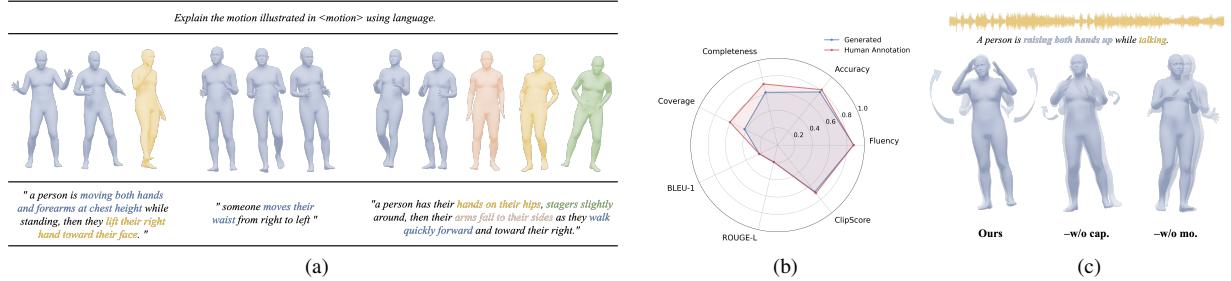


Figure 5. Visualization results. (a) Gesture captioning examples. Our captioning framework can effectively describe both overall motion patterns and fine-grained details. More results are in supplementary material (Sec. 8). (b) Quantitative captioning evaluation. Our model performs comparably to human annotations. (c) Qualitative ablation study. Results are generated using audio and single caption.

and a decreased R-Precision. These highlight the crucial role of the multi-granular captioning in providing precise semantic guidance. A detailed comparison of the three captioning strategies (*Reg.*, *Dyn.*, *Hie.*) is provided in supplementary material (Sec. 7); the Hierarchical strategy with Regular local captions is adopted, as it achieves the best overall balance across all metrics.

Hierarchically Controlled Denoiser Table 1 shows that removing the hierarchically controlled denoiser (-*w/o hcd*) degrades reconstruction quality and weakens condition coordination, resulting in a stronger bias toward co-speech gestures (BC 0.583 ↑) while losing semantic relevance (MM-Dist 0.058 ↑). This underscores the importance of the hierarchical denoiser for multimodal coordination.

Unified Cross-Dataset Motion Representation Ablating unified motion representation (Fig. 5c (right)) limits the semantic awareness beyond typical co-speech movements, leading to incomplete execution of intended movements like raising hands. Results in Table 1 (-*w/o mo.*) further confirm this degradation, showing a substantial drop in semantic relevance (MM-Dist 0.850 ↑ and R-Precision 0.057 ↓).

Inference Time Long inference time is a major bottleneck in diffusion models. We evaluate efficiency using Av-

erage Inference Time per Sentence (AITS) [7] on a single NVIDIA RTX 3090 GPU, with batch size one and excluding model loading time. Our method achieves an AITS of $0.842 \pm .002$ s, **over 6x faster** than FreeTalker ($6.632 \pm .044$ s) and SynTalker ($5.804 \pm .044$ s). This significant speedup can be attributed to our efficient hierarchical denoiser and optimized latent diffusion process, making our approach more practical for real-world applications.

5. Conclusion

In this study, we propose CoordSpeaker, a coordinated gesture generation approach designed to generate both co-speech spontaneous gesture and caption-driven non-spontaneous motion, under joint audio-caption control. We present the first Gesture Captioning Framework to bridge semantic prior gap by generating descriptive, well-aligned captions for gesture data. Building upon this, we propose a Coordinated Gesture Generation Model, leveraging a hierarchically controlled denoiser for efficient and coordinated gesture generation. Our approach pioneers gesture captioning and explores bidirectional gesture-text mapping. Extensive experiments demonstrate that CoordSpeaker produces high-quality gestures with enhanced semantic coherence and rhythmic synchronization while significantly improving efficiency, surpassing existing methods.

References

- [1] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International conference on 3D vision (3DV)*, pages 719–728. IEEE, 2019. 1
- [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. 1
- [4] 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, 6, 7, 8, 1
- [5] 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
- [6] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. 1
- [7] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 2, 6, 8
- [8] Qingrong Cheng, Xu Li, and Xinghui Fu. Siggesture: Generalized co-speech gesture synthesis via semantic injection with large-scale pre-training diffusion models. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 1
- [9] Wenzun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. In *European Conference on Computer Vision*, pages 390–408. Springer, 2024. 2
- [10] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F Troje, and Marc-André Carbonneau. Zeroeggs: Zero-shot example-based gesture generation from speech. In *Computer Graphics Forum*, pages 206–216. Wiley Online Library, 2023. 1
- [11] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5152–5161, 2022. 1, 6
- [12] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 2, 4, 5, 1
- [13] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pages 580–597. Springer, 2022. 2, 3
- [14] 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. 8
- [15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5
- [16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 5
- [17] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 1
- [18] Biao Jiang, Xin Chen, Chi Zhang, Fukun Yin, Zhuoyuan Li, Gang Yu, and Jiayuan Fan. Motionchain: Conversational motion controllers via multimodal prompts. In *European Conference on Computer Vision*, pages 54–74. Springer, 2024. 2
- [19] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013. 4
- [20] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 97–104, 2019. 6
- [21] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha Srinivasa, and Yaser Sheikh. Talking with hands 16.2m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 763–772, 2019. 1
- [22] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13401–13412, 2021. 6
- [23] 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, 2, 5, 3
- [24] 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, 6, 3

- [25] Mingshuang Luo, Ruibing Hou, Zhuo Li, Hong Chang, Zimo Liu, Yaowei Wang, and Shiguang Shan. M³gpt: An advanced multimodal, multitask framework for motion comprehension and generation. *arXiv preprint arXiv:2405.16273*, 2024. 2
- [26] Muhammad Hamza Mughal, Rishabh Dabral, Ikhсанул Habibie, Lucia Donatelli, Marc Habermann, and Christian Theobalt. Convofusion: Multi-modal conversational diffusion for co-speech gesture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1388–1398, 2024. 2
- [27] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. A comprehensive review of data-driven co-speech gesture generation. In *Computer Graphics Forum*, pages 569–596. Wiley Online Library, 2023. 2
- [28] OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence, 2024. Accessed: 2025-05-16. 7
- [29] Kunkun Pang, Dafei Qin, Yingruo Fan, Julian Habekost, Takaaki Shiratori, Junichi Yamagishi, and Taku Komura. Bodyformer: Semantics-guided 3d body gesture synthesis with transformer. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. 1, 2
- [30] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 4
- [31] Mathis Petrovich, Michael J Black, and Güл Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021. 4
- [32] Xingqun Qi, Chen Liu, Lincheng Li, Jie Hou, Haoran Xin, and Xin Yu. Emotiongesture: Audio-driven diverse emotional co-speech 3d gesture generation. *IEEE Transactions on Multimedia*, 2024. 1
- [33] Xingqun Qi, Jiahao Pan, Peng Li, Ruibin Yuan, Xiaowei Chi, Mengfei Li, Wenhan Luo, Wei Xue, Shanghang Zhang, Qifeng Liu, et al. Weakly-supervised emotion transition learning for diverse 3d co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10424–10434, 2024. 1
- [34] 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. 1
- [35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2
- [36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 5
- [37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5
- [38] Wataru Takano and Yoshihiko Nakamura. Statistical mutual conversion between whole body motion primitives and linguistic sentences for human motions. *The International Journal of Robotics Research*, 34(10):1314–1328, 2015. 2
- [39] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022. 2, 7
- [40] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2022. 2, 8
- [41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4
- [42] 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. 2
- [43] Zunnan Xu, Yukang Lin, Haonan Han, Sicheng Yang, Ronghui Li, Yachao Zhang, and Xiu Li. Mambatalk: Efficient holistic gesture synthesis with selective state space models. *Advances in Neural Information Processing Systems*, 37:20055–20080, 2025. 1, 2
- [44] Sicheng Yang, Zilin Wang, Zhiyong Wu, Minglei Li, Zhen-song 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.
- [45] 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. 1
- [46] Sicheng Yang, Haiwei Xue, Zhensong Zhang, Minglei Li, Zhiyong Wu, Xiaofei Wu, Songcen Xu, and Zonghong Dai. The diffusestylegesture+ entry to the genea challenge 2023. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 779–785, 2023. 1
- [47] Sicheng Yang, Zunnan Xu, Haiwei Xue, Yongkang Cheng, Shaoli Huang, Mingming Gong, and Zhiyong Wu. Freetalker: Controllable speech and text-driven gesture generation based on diffusion models for enhanced speaker naturalness. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7945–7949. IEEE, 2024. 1, 2, 4, 5, 6, 7
- [48] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geeyuk Lee. Speech gesture generation from the trimodal context of text, audio, and

- speaker identity. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020. 2, 6
- [49] Fan Zhang, Naye Ji, Fuxing Gao, and Yongping Li. Diffmofusion: Speech-driven gesture synthesis using denoising diffusion model. In *International Conference on Multimedia Modeling*, pages 231–242. Springer, 2023. 2
- [50] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023. 3, 8
- [51] Juntao Zhang, Yuehuai Liu, Yu-Wing Tai, and Chi-Keung Tang. C3net: Compound conditioned controlnet for multi-modal content generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26886–26895, 2024. 2
- [52] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondifuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6):4115–4128, 2024. 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. 1, 2
- [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
- [55] Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2430–2449, 2023. 1

CoordSpeaker: Exploiting Gesture Captioning for Coordinated Caption-Empowered Co-Speech Gesture Generation

Supplementary Material

In the supplementary material, we provide more implementation details (Sec. 6), additional experimental results (Sec. 7), more visual results (Sec. 8), discussion on limitations and future work (Sec. 9), and user study ethical considerations (Sec. 10) of the proposed CoordSpeaker.

6. Implementation Details

6.1. Network Details

Our transformer-based VAE and denoiser ϵ_θ are both composed of an encoder and a decoder, each containing 9 layers and 4 attention heads with GELU activation and residual connections. The latent dimension is set to $z \in \mathbb{R}^{1 \times 512}$. For training, we use the AdamW optimizer with a learning rate of $1e^{-4}$ and a batch size of 128. The VAE is trained for 6000 epochs, while the diffusion model is trained for 2000 epochs. We use 1000 diffusion steps for training and 50 steps for inference. During training, the noise variance β_t linearly scaled from 8.5×10^{-4} to 0.012. In the VAE stage, the KL loss weight β is set to $1e^{-4}$. During inference, for classifier-free guidance, the audio and caption guidance scales are set to $s_1 = 7$ and $s_2 = 0.75$ by default to balance contributions.

For condition embedding, we employ CLIP text encoder [34] and WavLM encoder [6] to extract semantic features $C \in \mathbb{R}^{512}$ from captions and audio features $A \in \mathbb{R}^{T \times 1133}$ from speech respectively. Both semantic and audio embeddings are projected through a linear layer into a 512-dimensional space before being fed into the denoiser.

6.2. Additional Details of Gesture Captioning

Prompt Template Table 4 presents a collection of prompt templates employed in our gesture captioning framework. These carefully curated templates are randomly sampled multiple times and paired with different gesture segments to generate diverse gesture captions. Templates are inspired by recent advances in motion-language modeling [17].

Motion Representation Alignment To better match the general motion-language space, we convert the gesture features into a commonly adopted human motion format [12] $x \in \mathbb{R}^{T \times 659} \rightarrow \mathbb{R}^{T \times 263}$ by retaining 22 key joints before performing captioning inference. Nevertheless, this conversion may omit some fine-grained finger-motion details. We provide more discussions in the following section (Sec. 9).

Caption Quality Control Given the differences in granularity and distribution between full-body motion and co-speech gestures, MotionLLM, pretrained on coarser human motion data, occasionally generate overly brief or action-oriented descriptions, such as interpreting exaggerated speaker movements as “The person is boxing”. To mitigate this, we implement a quality control mechanism that filters out captions with fewer than 5 words and their corresponding gesture segments, which are considered to lack clear non-spontaneous motion and cannot provide sufficient semantic guidance. This mechanism ensures that each retained segment is paired with a semantically rich caption as an effective training prior. In addition, global captions further complement local ones by providing broader contextual semantics. This strategy ensures caption quality and reduces the risk of ambiguous guidance during generation.

6.3. Dataset Details

To balance the data distribution between datasets during training, a weighted random sampling strategy is employed for the dataloader. Following [47], all motion sequences are resampled to 20 FPS and either truncated or padded to 180 frames. For the HumanML3D dataset, only sequences with lengths between 40 and 180 frames are utilized. Data is split into training, validation, and testing sets in an 8:1:1 ratio.

6.4. Evaluation Metrics

Coordination Evaluation Protocol Due to the absence of a unified multimodal benchmark, we follow standard practice [4, 47] and report Audio-to-Gesture and Text-to-Motion metrics on BEAT and HumanML3D in Sec. 4.2, respectively, to ensure fair comparison. However, this separation forces existing quantitative metrics to evaluate only single-modality factors: speech–gesture synchrony on BEAT and text–motion semantic alignment on HumanML3D, without directly reflecting the multimodal coordination that is critical to the joint generation task. Consequently, in Table 1 we focus on balanced performance across Audio-to-Gesture and Text-to-Motion metrics, as strong multimodal coordination inherently requires trade-offs between separate tasks. We believe that developing a unified multimodal benchmark would substantially benefit the coordination evaluation of this field, and our captioning framework may help facilitate its construction.

Fréchet Gesture Distance Following prior work [24], our FGD is calculated based on latent features extracted by

Table 4. Examples of prompt templates used in our gesture captioning framework.

Task	Input	Output
Gesture-to-Text	Give me a summary of the motion being displayed in [motion] using words.	
	Explain the motion illustrated in [motion] using language.	
	Describe the action being represented by [motion] using text.	
	What kind of action is being demonstrated in [motion]?	
	Describe the movement demonstrated in [motion] in words.	
	Generate a sentence that explains the action in [motion].	[caption]
	Please describe the movement depicted in [motion] using natural language.	
	Provide a description of the motion being displayed in [motion] using language.	
	Give me a brief summary of the movement depicted in [motion].	
	Describe the movement demonstrated in [motion] using natural language.	

Table 5. Ablation studies on multi-granular captioning strategies. “Reg.” denotes the Regular Caption strategy, “Dyn.” denotes the Dynamic Caption strategy, and “Hie.” denotes the Hierarchical Caption strategy. Each metric is reported under the 95% confidence interval from 20 times running. We report $BC \times 10^{-1}$ and Top-1 R-Precision.

Methods	Reconstruction		Audio-to-Gesture			Text-to-Motion			
	Jerk→	Accel.→	FGD↓	BC↑	L1Div↑	FID↓	MM-Dist↓	Div→	R-Precision↑
GT	1.165 ^{±.000}	0.043 ^{±.000}	-	-	-	-	6.205 ^{±.043}	5.512 ^{±.114}	0.140 ^{±.008}
Ours-Reg.	1.201 ^{±.017}	0.038 ^{±.001}	2.302 ^{±.061}	1.910 ^{±.004}	12.781 ^{±.044}	1.260 ^{±.063}	6.872 ^{±.058}	5.303 ^{±.107}	0.102 ^{±.010}
Ours-Dyn.	1.189 ^{±.013}	0.038 ^{±.000}	2.866 ^{±.106}	1.943 ^{±.037}	14.471 ^{±.110}	1.404 ^{±.049}	6.955 ^{±.044}	5.440 ^{±.114}	0.095 ^{±.006}
Ours-Hie.	1.190 ^{±.015}	0.039 ^{±.001}	3.173 ^{±.123}	1.327 ^{±.049}	10.861 ^{±.066}	1.118 ^{±.061}	6.814 ^{±.056}	5.558 ^{±.126}	0.100 ^{±.008}

a pre-trained autoencoder. Specifically, FGD is computed as:

$$FGD(g, \hat{g}) = \|\mu_r - \mu_g\|_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}) \quad (6)$$

where μ_r and Σ_r represent the mean and covariance matrix of the latent features z_r extracted from real gestures g , while μ_g and Σ_g correspond to those of generated gestures \hat{g} . A lower FGD indicates a better quality of generated gestures. To extract these latent features for our proposed unified motion representation (Sec. 3.1), we train a Full CNN-based autoencoder consisting of 4-layer convolutional encoders and decoders. Each convolutional layer is followed by a LeakyReLU activation function. The latent dimension is set to 240. This autoencoder is trained on the 4 English speakers of the BEAT dataset for 1000 epochs using the same training configuration as our VAE stage.

7. Additional Experimental Results

7.1. Comparison of Multi-Granular Captioning

Table 5 further compares the performance of different captioning strategies. Among these, the hierarchical strategy (*Ours-Hie.*) achieves the optimal balance across all metrics, making it well-suited for coordinated gesture generation. It yields better semantic relevance (lower MM-Dist at 6.814 vs. 6.872 and 6.955), improved motion quality (lower

FID by 11.3% and 25.6%), while maintaining comparable audio synchronization and motion diversity. Additionally, the dynamic strategy (*Ours-Dyn.*) exhibits advantages in co-speech gesture synchronization (BC: 1.943) and diversity (L1Div: 14.471). This may be attributed to its adaptive sampling mechanism, which introduces more rhythmic variation during training. Overall, these results suggest that the multi-granular captioning mechanism effectively supports multimodal coordination, enabling both fine-grained semantic control and rhythmically natural gestures.

8. More Visual Results

8.1. More Coordinated Generation Results

As shown in Fig. 6, we provide more coordinated generation results using *calm* audio and different text captions. These results further confirm the effectiveness of our proposed method in generating both co-speech spontaneous gestures and caption-driven non-spontaneous motions under joint speech-caption control. For dynamic results, please see our demo video appendix.

8.2. More Gesture Captioning Results

We present additional gesture captioning results in Fig. 7, further demonstrating the effectiveness of our approach in accurately mapping gestures to text. As highlighted in

colorful boxes, the model effectively captures both fine-grained hand movements and coarse-grained full-body motions while describing complex, continuous actions.

9. Limitations and Future Work

9.1. Enhancing Gesture Understanding

While gesture captioning effectively bridges the semantic gap in gesture generation, it still faces challenges in temporal consistency, occasionally leading to misordered actions in longer sequences. Our multi-granular captioning mechanism mitigates this: fine-grained local captions reduce the burden of describing long sequences, while global captions provide complementary long-range semantic context. Future improvements may stem from enhancing the temporal modeling capacity of motion-language models.

In addition, since MotionLLM occasionally produces coarse action-oriented descriptions, constructing broader gesture–caption benchmarks and fine-tuning a dedicated GestureLLM could further enhance the perception of fine-grained gesture semantics. We aim to expand our expert-annotated set in future work to support this direction. Moreover, incorporating audio or text transcripts into caption generation offers a promising avenue for producing more expressive gesture captions, which could further enhance coordinated gesture generation.

9.2. Fine-grained Gesture Representations

The proposed coordinated gesture generation framework could benefit from more refined motion representations. Given our primary focus on coordinated gestures and full-body movement synthesis, we adopt the BEAT dataset [23], which provides sufficient data for this purpose. However, integrating datasets with more precise head and finger motion, such as BEAT2 [24], could facilitate more holistic gesture generation. A key challenge that lies here is bridging the additional semantic gap for finer-grained facial expressions and finger movements, potentially requiring more detailed annotated datasets or a more powerful gesture-language model in future research.

10. Ethical Considerations in User Study

This section elaborates on the details of our user study protocol and participant demographics. The study recruited participants between 18 and 40 years of age, all possessing a minimum of an undergraduate degree to ensure a qualified assessment. Fig. 8 illustrates the interface of our evaluation platform, which presents a standardized template layout to all participants. To maintain data quality and ensure thorough evaluation, we implemented a response time threshold: any trial completed in less than 100 seconds was deemed insufficient for proper assessment and subsequently excluded from our analysis.

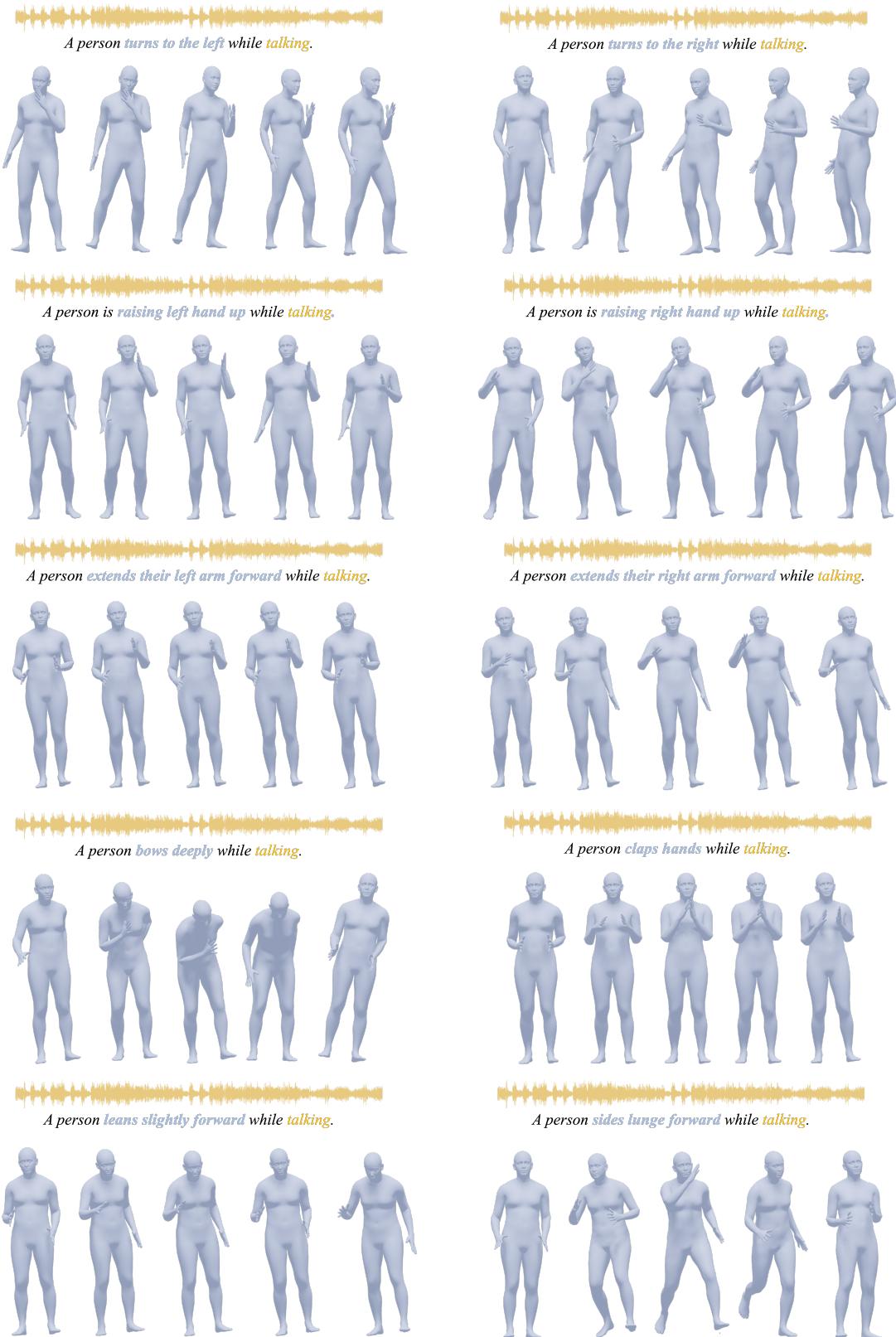


Figure 6. More visual results of coordinated gesture generation.

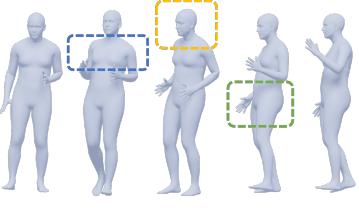
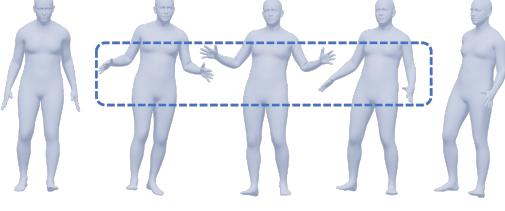
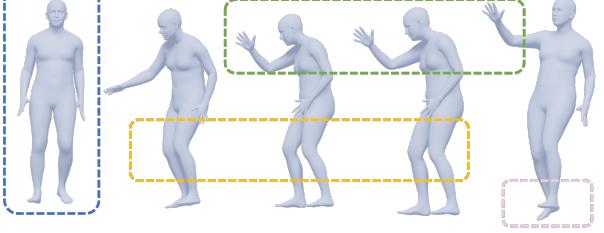
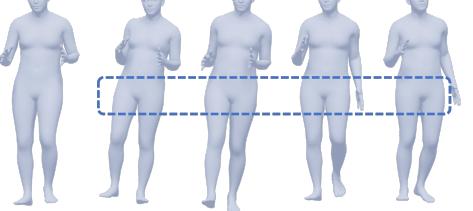
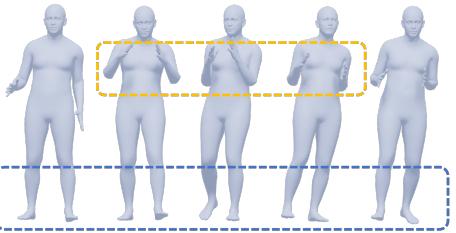
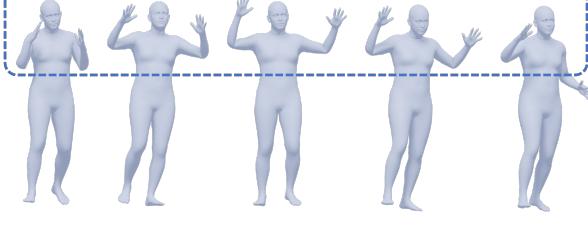
[Prompt]	[Gesture]	[Caption]
<p>Explain the motion illustrated in <gesture> using language.</p>		<p>"A man rotates at the shoulders, then rolls his neck and finally rotates at his hips."</p>
<p>Describe the action being represented by <gesture> using text.</p>		<p>"A person swings both arms back and forth while they're bent at the elbow."</p>
<p>Describe the motion displayed in <gesture> using natural language.</p>		<p>"A person stands with his hands by his sides face down and his right leg bent, and he moves his hand up and down as if tapping on a surface before alternating his feet."</p>
<p>Please describe the movement depicted in <gesture> using natural language.</p>		<p>"A man is moving his hips from right to left and movin his arms in rythm with his hips."</p>
<p>Explain the movement illustrated in <gesture> using text.</p>		<p>"A person stands with feet shoulder width ... lifting his arms out to shoulder level."</p>
<p>What does the <gesture> communicate? Please describe it in language.</p>		<p>"A person lifts both hands above their head, in an arcing motion, as if they were throwing a ball back onto the court."</p>

Figure 7. More gesture captioning results. **Colorful** boxes highlight the precise mapping between gestures and textual captions.

Gesture Video Evaluation

Dear Participants,

Thank you for participating in this survey. This study aims to evaluate the quality and performance of different gesture videos.

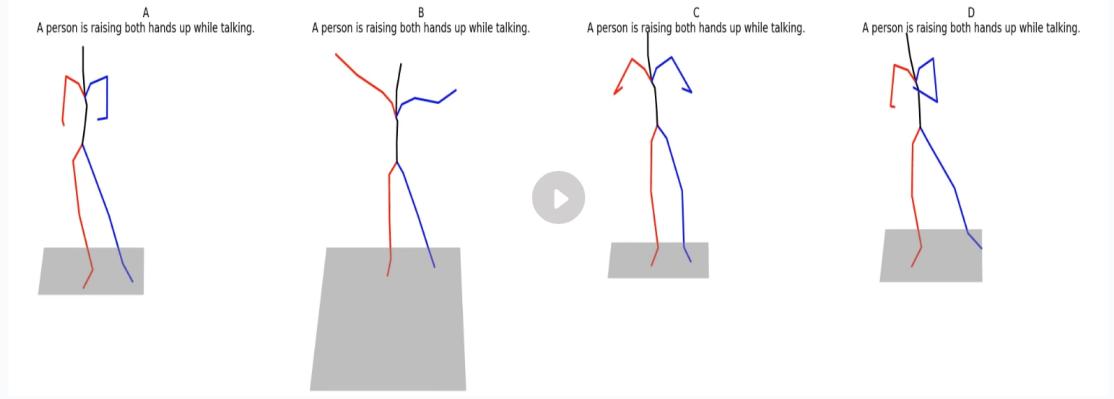
In the questionnaire, you will watch 10 short videos, each of which contains four virtual character gesture performances generated for the same audio and text description. Please evaluate each video based on the following three aspects:

- **Naturalness:** Is the gesture smooth and natural, and does it conform to people's daily movement habits?
- **Audio Synchrony:** Is the rhythm of gesture and voice coordinated, and can it accurately match the pauses, stresses, and other features of voice?
- **Text Matching:** Does the gesture accurately reflect the description of the text title and enhance semantic understanding?

Evaluation method: Based on the above three dimensions, please select the group of four videos in each group that best meets the dimension.

Please choose based on your intuition and subjective feelings, without considering too much objective criteria. Your feedback is of great value to our research, and thank you for your support!

Q1 Please watch the video and answer the following questions (Text Caption: "A person is raising both hands up while talking.")



Q1 For the following evaluation attributes, select the video that you think best matches.

	Video A	Video B	Video C	Video D
Naturalness	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Audio Synchrony	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Text Matching	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 8. The screenshots of the user study website for participants.