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 details (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 $z \in \mathbb{R}^{1\times512}$. For training, we use the AdamW optimizer with a learning rate of $1e^{-4}$ and a batch size of 128. The VAE stage is trained for 6000 epochs, while the diffusion stage 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 $\mathbf{C} \in \mathbb{R}^{512}$ from captions and audio features $\mathbf{A} \in \mathbb{R}^{T\times1133}$ 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 The captioning component is pretrained on the text-to-motion datasets HumanML3D and KIT [17]. To better match the general motion-language space, we convert the gesture features into a commonly adopted human motion format [12] $x \in \mathbb{R}^{T\times659} \rightarrow \mathbb{R}^{T\times263}$ 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 controls: speech–gesture synchrony on BEAT and text–motion semantic alignment on HumanML3D, without directly reflecting the multimodal coordination, which 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], the FGD is calculated based on latent features extracted by a

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]. 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.	[caption]

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	$1.165^{\pm.000}$	$0.043^{\pm.000}$	-	-	-	-	$6.205^{\pm.043}$	$5.512^{\pm.114}$	$0.140^{\pm.008}$
Ours-Reg.	$1.201^{\pm.017}$	$0.038^{\pm.001}$	$2.302^{\pm.061}$	$1.910^{\pm.004}$	$12.781^{\pm.044}$	$1.260^{\pm.063}$	$6.872^{\pm.058}$	$5.303^{\pm.107}$	$0.102^{\pm.010}$
Ours-Dyn.	$1.189^{\pm.013}$	$0.038^{\pm.000}$	$2.866^{\pm.106}$	$1.943^{\pm.037}$	$14.471^{\pm.110}$	$1.404^{\pm.049}$	$6.955^{\pm.044}$	$5.440^{\pm.114}$	$0.095^{\pm.006}$
Ours-Hie.	$1.190^{\pm.015}$	$0.039^{\pm.001}$	$3.173^{\pm.123}$	$1.327^{\pm.049}$	$10.861^{\pm.066}$	$1.118^{\pm.061}$	$6.814^{\pm.056}$	$5.558^{\pm.126}$	$0.100^{\pm.008}$

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

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

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. 3.2.1), 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.

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 mo-

tions 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.

9.3. Incorporating Diverse Contexts

This work primarily leverages *gesture-descriptive captions* as the context to provide explicit semantic guidance for customized and coordinated gesture generation. Beyond this, incorporating multiple contextual inputs, such as *speech transcripts* or *conceptual information*, could provide richer supervision and further enhance the expressiveness and controllability of the generated gestures, representing a promising direction for future work.

10. User Study Details

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 clear criteria and a standardized layout to all participants to ensure consistency. 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.

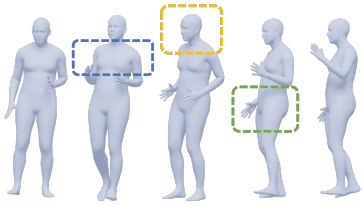
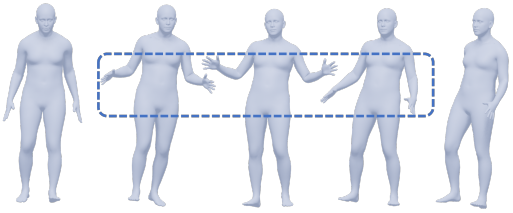
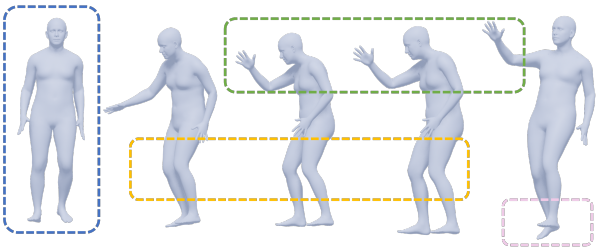
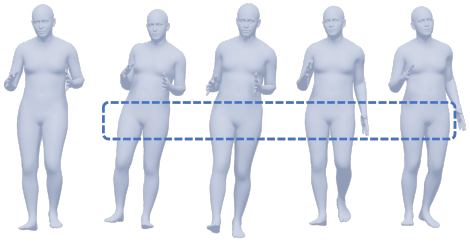
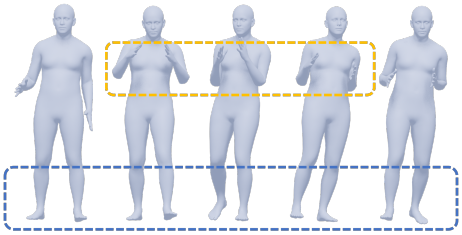
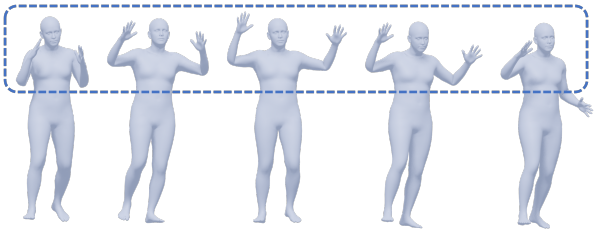
[Prompt]	*[Gesture]*	*[Caption]*
Explain the motion illustrated in <motion> using language.		*"A man rotates at the shoulders, then rolls his neck and finally rotates at his hips."*
Describe the action being represented by <motion> using text.		*"A person swings both arms back and forth while they're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

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.

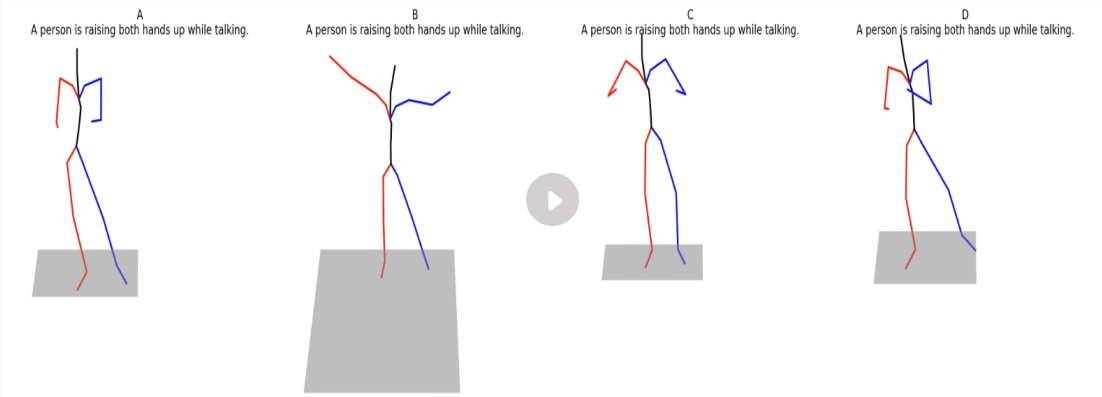
In the questionnaire, you will watch 10 short videos, each of which contains four virtual character gesture performances generated for the same audio and text description. Please evaluate each video based on the following three aspects:

- **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	○	○	○	○
Audio Synchrony	○	○	○	○
Text Matching	○	○	○	○

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