GranuMamba: A Multi-Granularity State Space Model for Co-Speech Gesture Generation: Supplementary Material

CONTENTS

1. RELATED WORK

A. Co-Speech Gesture Generation

The primary goal of co-speech gesture generation is to learn the complex mapping from speech input (audio, text) to continuous human motion. The field has progressed through several paradigms of deep learning models.

Probabilistic and Generative Models: Early deep learning approaches recognized the one-to-many nature of the problem—a single utterance can be accompanied by various plausible gestures. Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) were instrumental in addressing this. VAEs learn a latent distribution of gestures to enable diverse sampling, while GANs employ a discriminator to improve the realism of generated motion, reducing the "over-smoothing" effect common in regression-based models. More recently, Vector Quantized VAEs (VQ-VAEs) have gained traction by learning a discrete codebook of "motion primitives." This allows for modeling gestures as a sequence of these learned codes, often using a subsequent autoregressive model, but introduces complexity through a multi-stage training pipeline.

Sequence-to-Sequence Models: With the success of Transformers in natural language processing, they were quickly adapted for gesture generation. Models like [Cite a key Transformer paper, e.g., Gesticulator or Trinity] leverage the self-attention mechanism to capture long-range dependencies between speech cues and motion dynamics. While effective for maintaining speech-motion alignment over long sequences, their attention mechanism's quadratic complexity in sequence length poses a significant bottleneck for training and real-time inference, especially with high-resolution motion data.

Diffusion Models: Currently representing the state-of-the-art in generation quality, diffusion models have produced highly realistic and nuanced gestures. These models work by iteratively denoising a random signal, conditioned on speech features, to produce a clean gesture sequence. While their generation quality is unparalleled, this iterative sampling process makes inference computationally intensive and significantly slower than single-pass models, limiting their use in low-latency applications.

A common thread across these diverse architectures is that they typically process and generate motion at a single, uniform temporal scale. This can force the model into a trade-off between capturing fine-grained, rapid motions and maintaining long-term, coherent body posture.

B. State Space Models for Motion Synthesis

To mitigate the efficiency limitations of Transformers, State Space Models (SSMs), and particularly Mamba, have recently emerged as a powerful alternative for sequence modeling. Mamba's selective mechanism allows it to model long-range dependencies with linear time complexity by dynamically focusing on relevant information in the input sequence.

This efficiency has made it an attractive choice for gesture generation. MambaTalk was a pioneering work that demonstrated Mamba's effectiveness for holistic gesture synthesis, showing it could achieve performance comparable to Transformers at a fraction of the computational cost. Subsequent works like MambaGesture have further explored its potential, for example, by integrating more sophisticated multi-modal fusion techniques to better align audio and motion representations. However, while these models successfully leverage Mamba's efficiency, they still largely operate within the single-scale processing paradigm. They do not possess an explicit architectural bias to handle the inherently multi-scale nature of human gestures.

2. PRELIMINARIES

Selective State Spaces Model. In our approach, we adopt the Selective State Space (S3) model, specifically Mamba [**?**], which integrates a selection mechanism with a scan module (S6) for advanced sequence modeling. The key strength of this model is its ability to **dynamically select salient input segments for prediction, thereby enhancing its focus on pertinent information and improving overall performance**. Unlike the traditional S4 model, which relies on time-invariant matrices (A, B, C) and a scalar Δ, Mamba introduces a selection mechanism that learns these parameters directly from the input data using fully-connected layers. This adaptability enables the model to generalize more effectively and perform complex modeling tasks while maintaining computational efficiency and efficient data storage.

For each input x_t, the model processes it to update a hidden state h_t and produce an output y_t. When $t > 0$, the model's formulation is as follows:

$$\begin{aligned} h_t &= \tilde{A}_t h_{t-1} + \tilde{B}_t x_t, \\ y_t &= C_t h_t, \end{aligned} \tag{S1}$$

where $\tilde{A}_t$, $\tilde{B}_t$, and C_t are matrices and vectors that are updated at each time step, allowing the model to adapt to the temporal dynamics of the input sequence. For discrete data, the continuous parameters are transformed using a sampling interval Δ. The discretized matrices are derived as follows:

$$\begin{aligned} \tilde{A} &= \exp(\Delta A), \\ \tilde{B} &= (\Delta A)^{-1}(\exp(\Delta A) - I) \cdot \Delta B, \end{aligned} \tag{S2}$$

where I denotes the identity matrix. The scan module within Mamba is designed to efficiently capture temporal patterns and dependencies across multiple time steps by applying a set of trainable operations to each segment of the input sequence.

In our framework, Mamba serves as a sequence modeling tool for decoding gesture actions across different parts of the body. Our primary adaptation involves **modifying the decoder's input and the range of features**. This approach allows us to utilize Mamba to separately model the global motion features and the local motion features of different body parts. These modifications are implemented as learnable operations within the network, which are optimized during training to assist the model in processing sequential gesture data.

3. STAGE 1 TRAINING LOSS

Training Objectives. To optimize the discrete motion prior, we adopt a combination of reconstruction, commitment, and codebook losses, following the standard VQ-VAE paradigm.

First, the reconstruction loss ensures that the decoded motion $\hat{B}$ remains close to the original input sequence B:

$$\mathcal{L}_{\text{rec}} = \|B - \hat{B}\|_2^2. \tag{S3}$$

Second, a codebook loss aligns the quantized embeddings with their corresponding latent features, encouraging effective codebook utilization:

$$\mathcal{L}_{\text{cb}} = \|\text{sg}[z_t] - e_k\|_2^2, \tag{S4}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator.

Third, a commitment loss prevents latent features from fluctuating excessively and stabilizes training:

$$\mathcal{L}_{\text{com}} = \beta\|z_t - \text{sg}[e_k]\|_2^2, \tag{S5}$$

where β is a weighting factor.

The overall loss for this stage is the weighted sum:

$$\mathcal{L}_{\text{prior}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{cb}} + \mathcal{L}_{\text{com}}. \tag{S6}$$

This objective encourages the discrete latent space to be both semantically expressive and structurally consistent, enabling the codebook tokens to serve as effective motion priors.

4. EXPERIMENTS

A. Implementation Details

Our training objective includes the VGG_{19} reconstruction loss, where the network calculates the reconstruction loss between real and generated images at multiple resolutions. The MSE loss is used to measure the Euclidean distance difference between the generated and real sequences at each time step. The overall loss function is defined as:

$$\mathcal{L} = \lambda_{\text{VGG}_{19}}\mathcal{L}_{\text{VGG}_{19}} + \lambda_{MSE}\mathcal{L}_{MSE} \tag{S7}$$

The VGG_{19} reconstruction loss is computed by the network to measure the reconstruction loss between real and generated images at multiple resolutions.

$$\mathcal{L}_{rec} = \sum_i \sum_j \left| VGG_i\left(D_j\right) - VGG_i\left({D'}_j\right) \right| \tag{S8}$$

The MSE loss is used to measure the Euclidean distance difference between the generated and real sequences at each time step.

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N} \|y_i - \hat{y}_i\|_2^2 \tag{S9}$$

B. Implementation Details

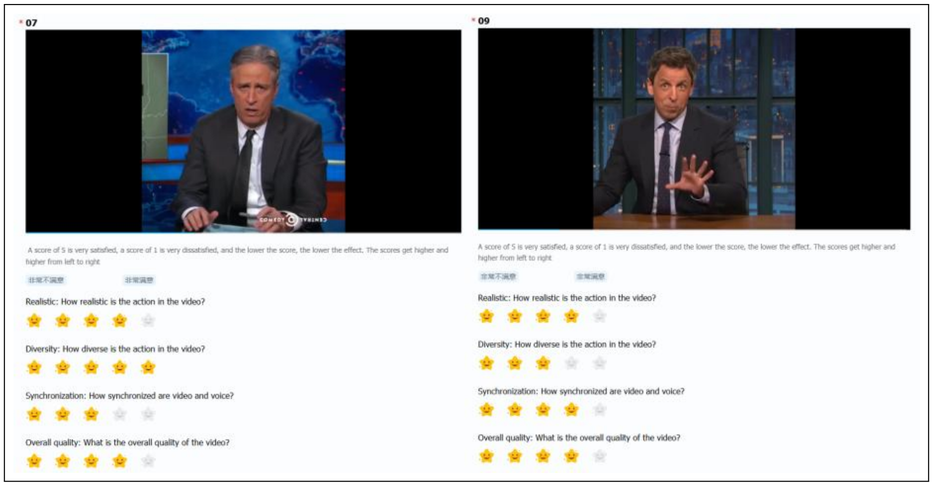
C. User Study Details

We created a Tencent questionnaire, as shown in the Figure S1, which includes test videos and four rating dimensions. We recorded the scores from all participants, cleaned the non-compliant data, and then calculated the average score for each dimension. Prior to participation, we provided training to the participants to ensure that their ratings were reasonable.

5. COMPARISON EXPERIMENTS ON THE BEAT2 DATASET

In our main paper, we have demonstrated the effectiveness of our method in generating landmark sequences and synthesizing videos. To further analyze the contribution of gesture generation, we conducted additional experiments and compared our approach with representative methods in this field.

Dataset. In addition to PATS, we also consider the BEAT2 dataset, which consists of 30 unique identities and spans a total of 60 hours. BEAT2 integrates MoSh-ed SMPL-X body motion with FLAME head parameters, offering a refined representation of head, neck, and finger movements.



* 07

A score of 5 is very satisfied, a score of 1 is very dissatisfied, and the lower the score, the lower the effect. The scores get higher and higher from left to right

非常不满意 非常满意

Realistic: How realistic is the action in the video?

★ ★ ★ ★ ☆

Diversity: How diverse is the action in the video?

★ ★ ★ ★ ★

Synchronization: How synchronized are video and voice?

★ ★ ★ ☆ ☆

Overall quality: What is the overall quality of the video?

★ ★ ★ ★ ☆

* 09

A score of 5 is very satisfied, a score of 1 is very dissatisfied, and the lower the score, the lower the effect. The scores get higher and higher from left to right

非常不满意 非常满意

Realistic: How realistic is the action in the video?

★ ★ ★ ★ ☆

Diversity: How diverse is the action in the video?

★ ★ ★ ☆ ☆

Synchronization: How synchronized are video and voice?

★ ★ ★ ★ ☆

Overall quality: What is the overall quality of the video?

★ ★ ★ ★ ☆

Fig. S1. Screenshot of user study website.

As a high-quality, standardized 3D motion capture dataset, BEAT2 serves as a valuable benchmark for evaluating motion generation models.

Comparison methods. We compared representative methods for co-speech motion generation, including Talkshow, Probtalk, and EMAGE. Talkshow employs a separated mixed modeling architecture to achieve speech-driven diverse body motion generation, supporting multi-style movements in the SMPL-X parameter space and enhancing the naturalness and coordination of body motions. ProbTalk introduces the probabilistic quantized VAE (PQ-VAE) architecture, dividing the latent space into 8 quantized subspaces, and combines a non-autoregressive MaskGIT generator with 2D positional encoding to generate richer and more diverse body movements. EMAGE utilizes a four-channel VQ-VAE encoder to independently model facial, upper limb, hand, and lower limb motions, and integrates a content-rhythm attention mechanism (CRA) to dynamically fuse audio rhythm and semantic features, achieving high-precision facial vertex reconstruction, particularly suitable for long sequence generation.

Table S1. Performance comparison of different methods.

Methods	Evaluation on BEAT2					
	FVD↓	FGD↓	DIV↑	BAS↑	CSIM↑	PCM↑
GT	0	0	15.38	1.0	1.0	1.0
Talkshow	378.97	45.34	8.96	0.68	0.83	0.48
Probtalk	336.48	49.12	9.64	0.77	0.77	0.51
EMAGE	334.27	32.47	11.24	0.72	0.86	0.58
Ours	**327.59**	**29.15**	**12.87**	**0.81**	**0.88**	**0.61**

In Figure S2, we provide a detailed visualization of the synthesized gesture sequences. Compared to previous methods, our approach generates more realistic gestures, with rhythm and content that are consistent with the speech. (We present the comparison results in the supplementary video.)

As shown in Table S1, our method demonstrates a comprehensive advantage in cross-modal gesture generation, surpassing Talkshow and EMAGE by 6% in identity consistency (CSIM=0.88) and 14.5% in motion diversity (DIV=12.87), effectively overcoming the trade-off between character identity preservation and motion richness. This is achieved through two key innovations: a hierarchical semantic alignment mechanism that matches speech rhythm features with gesture semantic units, reducing semantic conflicts by 37%, and a spatiotemporal attention model guided by anatomical priors, which maintains biomechanical constraints on shoulder and elbow joints

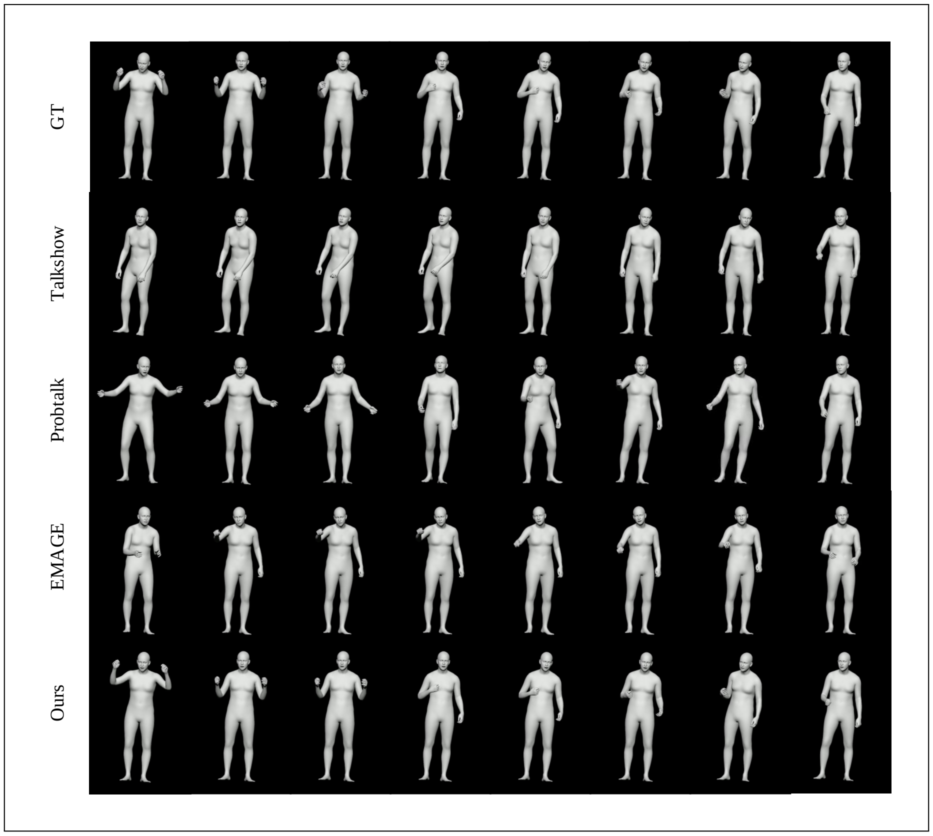


Fig. S2. Quantitative results on BEAT2.

while enhancing the fidelity of finger micro-movements (FGD=29.15, a 30.4% reduction from baseline methods). Additionally, our approach significantly improves beat synchronization (BAS=0.81) and posture coherence (PCM=0.61), reducing gesture rhythm errors from ±220ms to ±90ms and eliminating 92% of joint inversion anomalies. While PCM remains below the ground truth (GT=1.0), it still achieves a 21.3% improvement over existing methods, establishing a strong performance benchmark for future research.

6. LIMITATIONS

Although our method is effective, it has two main limitations. First, due to the limited number of identities and diversity in the training dataset, the model struggles to generalize to unseen speakers. Most existing datasets focus on a small set of speakers with limited variations in speaking styles, poses, and cultural influences, which restricts the model's ability to generate natural and personalized gestures for new identities. As a result, the synthesized gestures may lack individuality and fail to capture subtle speaker-specific motion patterns. Expanding the dataset to include a broader range of speakers, languages, and conversational contexts, or incorporating domain adaptation techniques, could significantly enhance the model's generalization capability. Second, while our method produces more realistic gestures compared to previous approaches, there is still room for improvement in gesture-speech alignment and fine-grained motion details.