MambaTalk: Efficient Holistic Gesture Synthesis with Selective State Space Models

Zunnan Xu[⋆], **Yukang Lin**[*], **Haonan Han**[*], **Sicheng Yang**, **Ronghui Li**, **Yachao Zhang**[†], **Xiu Li**[†]
Shenzhen International Graduate School, Tsinghua University
University Town of Shenzhen, Nanshan District, Shenzhen, Guangdong, P.R. China

Abstract

Gesture synthesis is a vital realm of human-computer interaction, with wide-ranging applications across various fields like film, robotics, and virtual reality. Recent advancements have utilized the diffusion model to improve gesture synthesis. However, the high computational complexity of these techniques limits the application in reality. In this study, we explore the potential of state space models (SSMs). Direct application of SSMs in gesture synthesis encounters difficulties, which stem primarily from the diverse movement dynamics of various body parts. The generated gestures may also exhibit unnatural jittering issues. To address these, we implement a two-stage modeling strategy with discrete motion priors to enhance the quality of gestures. Built upon the selective scan mechanism, we introduce MambaTalk, which integrates hybrid fusion modules, local and global scans to refine latent space representations. Subjective and objective experiments demonstrate that our method surpasses the performance of state-of-the-art models. Our project is publicly available at https://kkakkkka.github.io/MambaTalk/.

1 Introduction

Gesture synthesis is a critical area of research in human-computer interaction (HCI), which has very broad application prospects, such as film, robotics, virtual reality, and digital human development [24]. The task is challenging due to the variable correlation between speech and gestures, as the same spoken content can elicit markedly different gestures among speakers. Meanwhile, the generated gestures should synchronize with the speaker's rhythm, emotional cues, and intentions [31, 1, 8, 42].

Recent works in co-speech gesture generation have shown great progress [11, 46, 67, 66, 2, 62]. By introducing new datasets [71] and more modalities [70, 33], previous work achieved end-to-end gesture generation based on RNN-based models [11, 46]. With the success of transformer in nature language processing [58] and video sequence modeling [26, 27, 75], recent works [8, 43, 55] leverage the power of attention mechanism to generate more expressive gestures that better synchronize with speech. By further combining emotional and style related features, EMoG [69] achieve better quality gesture generation. With the development in human recognition model [37], EMAGE [32] proposes a masked audio-gesture modeling strategy to enhance unified holistic gesture synthesis. Recently, with the development of diffusion model in generative tasks [39, 20, 40], the latest works [74, 2, 62, 7, 65] have applied the diffusion model to gesture synthesis, significantly improving the diversity of generated gesture. DiffuseStyleGesture [63] presents a diffusion model-based approach for generating diverse co-speech gestures by incorporating cross-local attention and self-attention mechanisms, and utilizing classifier-free guidance for style control. DiffuseStyleGesture+ [66] further considers the text modality as an additional input and utilizes channel concatenation to merge the text feature

[⋆]Work done as intern at Tecent.
[*]Equal Contribution.
[†]Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

with the audio feature. Deichler *et al.* [9] also incorporates the text modality as an additional input and employs contrastive learning to enhance the features. However, the exploration of generation for co-speech gesture sequences with low latency remains relatively uncharted, constraining its application in dynamic, interactive environments. RNN-based models often struggle with the long-term forgetting issue [54, 29], which impairs their ability to generate long sequences of gestures effectively. Additionally, these models may produce gestures that lack variability, tending towards an average representation [36]. Transformer-based models depend heavily on subtle positional encoding to capture the order of input elements [44, 50, 73]. Meanwhile, their computational complexity, which grows quadratically with the length of the input sequence, poses a challenge for generating long sequences of gestures. For the diffusion-based model, the intricate sampling strategy and iterative process lead to high computational expenses [48], which hinder their broad adoption in gesture generation scenarios.

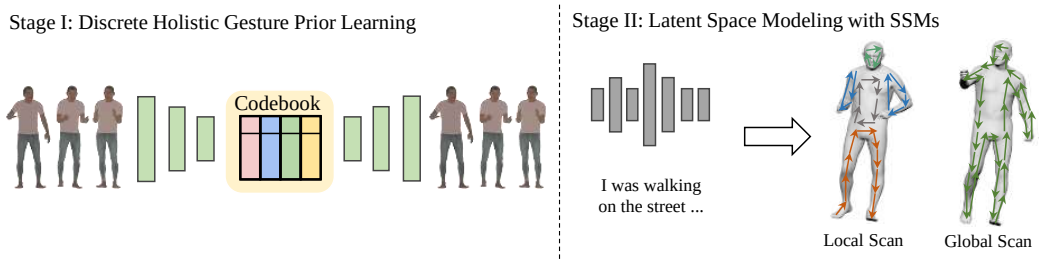

Figure 1: Our two-stage method for co-speech gesture generation with selective state space models. In the first stage, we construct discrete motion spaces to learn specific motion codes. In the second stage, we develop a speech-driven model of the latent space using selective scanning mechanisms.

State space models (SSMs) have recently shown significant potential in addressing challenges related to modeling sequences with low latency [12]. Inspired by continuous state space models from control systems and enhanced by HiPPO initialization [13], SSMs [16] show promise in addressing long-term forgetting issue. These advancements have been integrated into large-scale representation models [38, 41]. Some pioneering works have applied SSMs for tasks like language understanding [38, 41], content-based reasoning [12], and visual recognition [35, 57]. In our work, we further explore the potential of SSMs in co-speech gesture synthesis. We observe that directly applying the selective scan mechanism from Mamba [12] to gesture generation as a sequence modeling model would result in jittery outputs. To refine the generated gestures, we propose a two-stage modeling strategy. In the first training stage, we enhance the discrete motion priors derived from VQVAEs [53] by integrating velocity and acceleration losses. In the second stage, by utilizing motion priors from VQVAEs, we introduce individual learnable queries for different body parts, thereby alleviating the jittering issue. Meanwhile, considering that the direct application of Mamba encounters the challenge of limb movements across different body parts tending to average out, we propose a hybrid scanning approach in the second stage to enhance the motion representation in the latent space. Specifically, we refine the design of spatial and temporal modeling within latent spaces by introducing a global-to-local modeling strategy and integrating attention mechanisms along with a selective scanning approach into the framework's design. Considering the significant differences in deformation and movement patterns among different body parts [32], we propose local and global scan modules for refining the latent space representations of the movements across various body parts. These approaches enable dynamic interaction and iterative refinement of different body parts while maintaining low latency, leading to more diverse and rhythmic gestures. Our contributions can be summarized as below:

- We are the first to explore the potential of the selective scan mechanism for co-speech gesture synthesis, achieving a diverse and realistic range of facial and gesture animations.
- We introduce MambaTalk, an innovative framework that integrates hybrid scanning modules (e.g., local and global scan). The integration enhances the latent space representations for gesture synthesis, thereby refining the distinct movement patterns across various body parts.
- Extensive experiments and analyses demonstrate the effectiveness of our proposed method.

2 Related Work

2.1 Co-speech Gesture Generation

Co-speech gesture generation aims to automatically generate gestures based on speech input. Existing approaches can be broadly categorized into three groups: (i) Rule-based methods: These methods rely on pre-defined rules and gesture libraries to generate gestures based on speech features [23, 56]. While offering interpretable results, they require significant manual effort in creating gesture datasets and defining rules. (ii) Statistical models: These approaches leverage data-driven techniques to learn mapping rules between speech and gestures, often employing pre-defined gesture units [22, 25]. While overcoming the limitations of manual rule creation, these methods still rely on handcrafted features. (iii) Deep learning methods: Recent advancements in deep learning have enabled neural networks to capture the complex relationship between speech and gestures directly from raw multimodal data [70, 33, 68, 32]. This progress has established deep learning approaches, particularly recurrent neural networks (RNNs) [70, 33, 61], transformers [4, 45], and diffusion models [2, 74, 51, 72, 21, 6], as the prevailing paradigm for co-speech gesture generation. However, each of these models suffers from certain limitations that hinder their performance. RNNs inherently process sequences in a serial manner, where each timestep's computation depends on the output of the previous timestep. This limits their ability to efficiently handle long sequences and introduces cumulative latency. Meanwhile, RNNs lack inherent parallelism, further restricting their potential for high-speed computation. Transformers consider all positions within a sequence at every timestep, resulting in high computational complexity, especially for long sequences. While diffusion models significantly enhance the diversity of generated outputs, the sampling process is computationally expensive. To overcome these limitations, our method investigates the capacity of selective state space models in the field of gesture synthesis. To the best of our knowledge, we are the first to apply selective state space models to the task of gesture generation.

2.2 Selective State Space Models

State Space Models (SSMs) are a novel class of models recently integrated into deep learning for state space transformation [15, 10]. As foundational models evolve, various subquadratic-time architectures have emerged, including linear attention, gated convolution, recurrent models, and structured state space models (SSMs), aimed at mitigating the computational inefficiencies of Transformers when dealing with lengthy sequences. However, these advancements have yet to match the performance of attention mechanisms in critical modalities like language processing.

SSMs draw inspiration from continuous state space models in control systems and, when combined with HiPPO initialization [13], as seen in LSSL [16], show promise in tackling long-range dependency issues. However, the computational and memory demands of the state representation render LSSL impractical for real-world use. To address this, S4 [15] suggests normalizing the parameters into a diagonal structure. This has led to the emergence of various structured SSMs with diverse configurations, such as complex-diagonal structures [17, 14], multiple-input multiple-output (MIMO) support [49], diagonal-plus-low-rank decomposition [19], and selection mechanisms [12]. These models have been incorporated into large-scale representation models [38, 41].

These models primarily focus on the application of SSMs to long-range and sequential data like language and speech, for tasks such as language understanding [38, 41], content-based reasoning [12], and pixel-level 1-D image classification [15]. Recently, some pioneering work [35, 57, 59] have explored their application in visual recognition. We further demonstrate that by incorporating the selective scan mechanism from mamba [12] and the discrete motion priors from VQVAEs [53], our proposed MambaTalk is capable of matching the performance of existing popular holistic gesture synthesis models, highlighting the potential of MambaTalk as a powerful gesture synthesis model.

3 Method

We aim to synthesize sequential 3D co-speech gestures from speech signals (e.g., audio and text) using selective state space models. However, simply applying such a model to gesture synthesis leads to severe gesture jittering issues. We also found that maintaining performance is challenging due to the significant variations in movement patterns exhibited by different body parts. To overcome these challenges, we suggest modeling the gesture space using the acquired discrete motion patterns.

Subsequently, we propose to develop speech-conditioned selective state space models within this framework. This approach is designed to enhance the model's robustness against uncertainties that arise from cross-modal discrepancies. As shown in Figure 2, our framework consists of two stages: (i) modeling the discrete gestures and facial motion spaces (§3.2) and (ii) learning speech-conditioned selective state space models (§3.3) to generate 3D co-speech gestures.

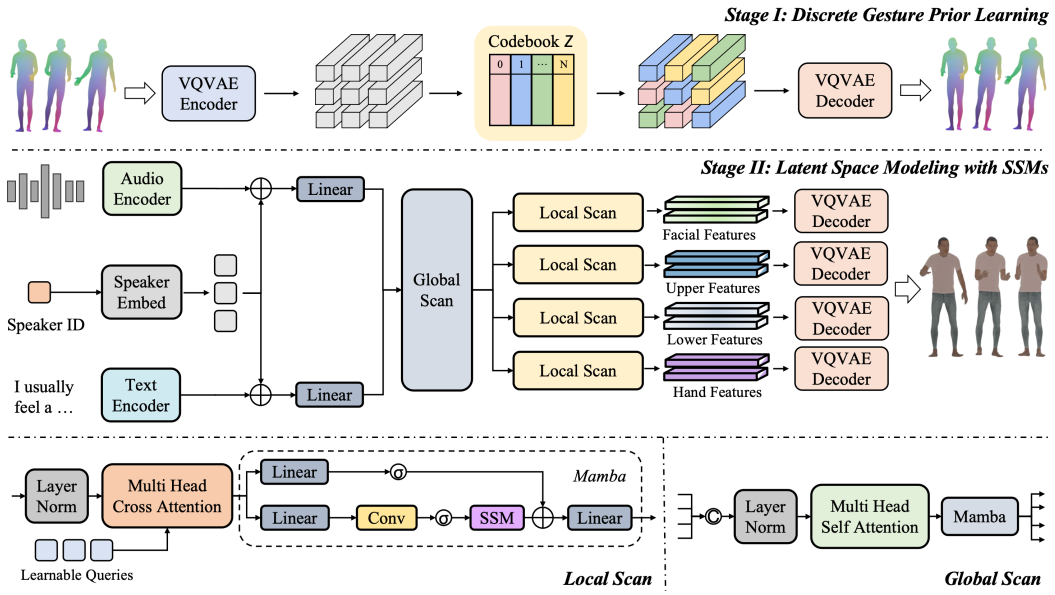

Figure 2: We propose a two-stage method for co-speech gesture generation. We first train multiple VQ-VAEs for face and different parts of body reconstruction. This step learns discrete motion priors through multiple codebooks. In the second stage, we train a speech-driven gesture generation model in the latent motion space with local and global scan modules.

3.1 Preliminaries

Selective State Spaces Model. In our approach, we adopt the Selective State Spaces model (Mamba [12]) that incorporates a selection mechanism and a scan module (S6). This model is designed to make sequence modeling, as it dynamically selects salient input segments for prediction, thereby enhancing its focus on pertinent information and improving overall performance. Unlike the traditional S4 model, which uses time-invariant matrices A, B, C, and scalar Δ, Mamba introduces selection mechanism that allows for the learning of these parameters from the input data using fully-connected layers. This adaptability enables model to better generalize and perform complex modeling tasks. Mamba operates by defining the state space with structured matrices that introduce specific constraints on the parameters, facilitating efficient computation and data storage. For each batch and each dimension, the model processes the input x_t, hidden state h_t, and output y_t at each time step t. We have $h_0 = \bar{B}_0 x_0$ when $t = 0$. When t > 0, the model's formulation is as follows:

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

where $\bar{A}_t$, $\bar{B}_t$, 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. With discretization, let Δ denote the sampling interval, $\exp(\Delta A)$ denote the matrix exponential, the transformation of the system's state over one time step can be represented as follows:

$$\begin{aligned} \bar{A} &= \exp(\Delta A), \\ \bar{B} &= (\Delta A)^{-1}(\exp(\Delta A) - I) \cdot \Delta B, \\ h_t &= \bar{A} h_{t-1} + \bar{B} x_t, \end{aligned} \tag{2}$$

where $(\Delta A)^{-1}$ denotes the inverse of matrix ΔA, I denotes the identity matrix. The scan module within Mamba is designed to capture temporal patterns and dependencies across multiple time steps by applying a set of trainable parameters or 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. By modifying the decoder's input and the range of features, we leverage Mamba to model local body features and global characteristics separately. These operations are learned during training and help the model to process sequential data.

3.2 Discrete Gestures and Facial Motion Spaces

To ensure visual realism in motion animations from speech signals, we learn extra motion priors to depict accurate movements and natural expressions. Building on this concept, we propose a method to represent the gesture motion space using multiple discrete codebooks.

Motion Quantization. Considering the substantial variations in deformation magnitude and periodicity among various body parts, our approach involves learning multiple codebooks tailored for the reconstruction of distinct body parts. For illustrative purposes, we detail the formulation of a single codebook. Denotes C as the dimensionality of each latent vector, N as the number of vectors in the codebook, for the codebook $\mathcal{Z} = \left\{\mathbf{z}_k \in \mathbb{R}^C\right\}_{k=1}^N$, we employ a set of allocated items $\{\mathbf{z}_k\}_{k\in\mathcal{S}}$ to represent the holistic gesture motion $\mathbf{M}_t$. Here, $\mathcal{S}$ represents the chosen index sets. The element-wise quantization function $Q(\cdot)$ maps each item $\hat{\mathbf{z}}_t$ in $\hat{\mathbf{Z}}$ to its closest match $\mathbf{z}_k$ in the codebook $\mathcal{Z}$:

$$\mathbf{Z_q} = Q(\hat{\mathbf{Z}}) := \underset{\mathbf{z}_k\in\mathcal{Z}}{\arg\min} \left\|\hat{\mathbf{z}}_t - \mathbf{z}_k\right\|_2, \tag{3}$$

where the codebook entries act as the foundational motion elements within the discrete motion space. To establish this, we follow [64, 32] to pre-train a CNN-based Vector Quantized-Variational Autoencoder (VQ-VAE), which comprises an encoder E, a decoder D, and a context-rich codebook $\mathcal{Z}$. This is done through the self-reconstruction of gesture motions.

The sequence of motions $\mathbf{M}_{1:T}$ is initially transformed into a temporal feature representation $\hat{Z} = E(\mathbf{M}_{1:T}) \in R^{T'\times H\times C}$, where H represents the count of gesture components, T' indicates the quantity of temporal units encoded (with $P = \frac{T}{T'}$ frames per unit). Subsequently, we derive the quantized motion sequence $Z_q \in R^{T'\times H\times C}$ by quantization function $Q(\cdot)$. This function Q maps each element in $\hat{Z}$ to its closest corresponding entry within the codebook $\mathcal{Z}$:

$$\mathbf{Z_q} = Q\left(E\left(\mathbf{M}_{1:T}\right)\right), \hat{\mathbf{M}}_{1:T} = D\left(\mathbf{Z_q}\right). \tag{4}$$

Training objectives. For the training of the quantized autoencoder, we employ motion-level losses to mitigate the jittering issue of generated gestures, along with two intermediate losses at the code level:

$$\begin{aligned}\mathcal{L}_{\text{VQ}} =& \mathcal{L}_{rec}(\mathbf{M}, \hat{\mathbf{M}}) + \mathcal{L}_{vel}(\mathbf{M}', \hat{\mathbf{M}}') + \mathcal{L}_{acc}(\mathbf{M}'', \hat{\mathbf{M}}'') \\ &+ \left\|\text{sg}(\hat{\mathbf{Z}}) - \mathbf{Z_q}\right\|_2^2 + \left\|\hat{\mathbf{Z}} - \text{sg}\left(\mathbf{Z_q}\right)\right\|_2^2,\end{aligned} \tag{5}$$

where $\mathbf{M}'$ and $\mathbf{M}''$ means the velocity and acceleration of motion, $sg(\cdot)$ denotes a stop-gradient operation, $\mathcal{L}_{\text{rec}}$ are Geodesic [52] loss and the last two terms are designed to refine the codebook entries. For facial motions, we utilize MSE loss for both velocity ($\mathcal{L}_{\text{vel}}$) and acceleration ($\mathcal{L}_{\text{acc}}$) loss. For body motions, we use L1 loss as $\mathcal{L}_{\text{vel}}$ and $\mathcal{L}_{\text{acc}}$. Additionally, for the foot contact loss, we employ MSE loss as the loss function. These terms work by minimizing the distance between the codebook Z and the embedded features $\hat{Z}$. Given that the quantization function (Equation 3) is non-differentiable, we utilize the straight-through gradient estimator [53] to propagate the gradients.

3.3 Speech-Driven Selective State Spaces Gesture Synthesis Model

Overall Framework. Utilizing the acquired discrete motion prior, we establish a cross-modal mapping from speech inputs to target motion codes, enabling the generation of realistic co-speech gesture motions. In our approach to speech-driven gesture synthesis, we utilize audio sequences $A = \{a_1, \ldots, a_N\}$ and text sequences $T = \{t_1, \ldots, t_N\}$ as inputs to guide the generation of co-speech gestures $G = \{g_1, \ldots, g_N\}$. Here, N signifies the total frame count, and $g_i \in R^{55\times 6+100+4+3}$

denotes 55 pose joints in Rot6D, R^{100} FLAME parameters, R^4 foot contact labels, R^3 global translations for the i-th frame. The gesture synthesis model, comprising audio encoders E_{A} and text encoders E_{T} and multiple selective state space models D_{B} for different parts of the body, is trained on the discrete motion space, conditioned on the speech, as shown in Figure 2.

Speech Feature Extraction. For audio feature extraction, two CNN-based audio feature extraction networks are employed to respectively extract amplitude and onset features. Specifically, we integrate amplitude and onset features along the channel dimension to obtain audio features $f_A = \{fa_1, ..., fa_N\}$. For processing speech input words, we employ pre-trained FastText [5] to obtain word embeddings, which are then refined by linear projections to produce text features $f_T = \{ft_1, ..., ft_N\}$. We further fuses features from the input modalities (e.g., audio and text features). The speaker embeddings s_{id} are first combined with audio and text features through additive operation. By concatenating feature vectors along the channel dimension, we further apply linear transformations to determine the weight factors, and then integrating the features through an element-wise summation. The process can be formalized as:

$$\begin{aligned} w_T &= \sigma(W_T \cdot [f_A + s_{id} \cdot \mathbf{1}, f_T + s_{id} \cdot \mathbf{1}]), \\ w_A &= \sigma(W_A \cdot [f_A + s_{id} \cdot \mathbf{1}, f_T + s_{id} \cdot \mathbf{1}]), \\ \bar{f}_T &= w_T \odot f_A + (1 - w_T) \odot f_T, \\ \bar{f}_A &= w_A \odot f_A + (1 - w_A) \odot f_T, \end{aligned} \tag{6}$$

where σ denotes the softmax operation, $\odot$ denotes the Hadamard product, and W_T and W_A represent linear mapping matrices used to adjust the dimensions of merged features. $\bar{f}_T$ and $\bar{f}_A$ denote the fused features.

Local and Global Scans. Recognizing the diverse deformations and motion patterns in various body parts, we propose using global scan module and multiple local scan modules to model the movements of different body parts (e.g., face, hand, upper and lower body) with fused multi-modal features from previous modules. By acquiring the speech features from audio and text encoders, we first improve the perception of motion patterns among them using a global scan module by combining the speech features along the sequence dimension. Subsequently, by utilizing self-attention mechanism($\mathcal{F}_{\text{MHSA}}$), we model the global information across different sequence tokens. These refined features are fed into Mamba to extract temporal perceptual information, which can be fomulized as below:

$$\begin{aligned} F_{speech} &= [\bar{f}_T, \bar{f}_A] \\ f_{mixed} &= \mathcal{F}_{\text{MHSA}}(F_{speech}), \\ f_{global} &= \text{Mamba}(f_{mixed}), \end{aligned} \tag{7}$$

where [] denotes the concatenation operation of $\bar{f}_T$ and $\bar{f}_A$ in the dimension of the sequence. To enhance the generalization of the model, we further incorporate additional learnable queries to foster the queries' ability to learn motion patterns. As shown in Figure 2, the queries (Q) are integrated with input global features through a multihead cross-attention mechanism ($\mathcal{F}_{\text{MHCA}}$). This allows queries to learn the most relevant information from the speech input. The process can be formally defined as follows:

$$f_{\text{face}} = \mathcal{F}_{\text{MHCA}}(Q_{\text{face}}, f_{global}), \tag{8}$$

where f_{face} denotes the feature of facial motion. This formulation can be applied similarly to obtain f_{hand}, f_{upper}, and f_{lower} by substituting the corresponding query vectors Q_{hand}, Q_{upper}, and Q_{lower} respectively. Utilizing the extracted perceptual features from various body parts, we proceed to employ Mamba to extract temporal features from the sequence, which can be formalized as belows:

$$F_{face} = \text{Mamba}(f_{face}), \tag{9}$$

where F_{face} corresponds to the temporal features of facial motion. The same approach is applied to generate temporal features for the hand, upper body, and lower body by entering their respective perceptual features f_{hand}, f_{upper}, and f_{lower} into the corresponding Mamba modules. The local latent features are then fed into their respective VQ-Decoders to produce the final motion predictions.

Training Objectives. The model's training objectives are composed of a composite loss function that harmonizes reconstruction and cross-entropy losses. This design aims to augment the accuracy of motion generation, encompassing the face, hands, upper, and lower body. The loss of latent reconstruction, represented by $L_{reclatent}$, is quantified using the Mean Squared Loss (MSELoss).

Here, z_i corresponds to the true latent vectors, while $\hat{z}_i$ are the vectors reconstructed by the model. The latent reconstruction loss is expressed as:

$$L_{reclatent} = \frac{1}{N}\sum_{i=1}^{N}\|z_i - \hat{z}_i\|^2, \tag{10}$$

where N denotes the number of frames. Concurrently, to encourage diversity in the generated motions, we optimize the cross-entropy loss for latent code class classification L_{cls}. Specifically, we employ Negative Log-Likelihood Loss (NLLLoss), where y_i represents the true class labels for each sample, and $\hat{y}_i$ denotes the model's predicted class labels. This loss is calculated as the negative sum of the logarithm of the predicted probabilities for the correct classes:

$$L_{cls} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}\log(\hat{y}_{ic}), \tag{11}$$

where N signifies the total number of frames, C is the total number of classes, and y_{ic} is a binary indicator of whether class c is the correct label for sample i. The total loss L is a weighted sum of the categorical and latent reconstruction losses, with α and β serving as balance hyper-parameters:

$$L = \alpha L_{cls} + \beta L_{reclatent}, \tag{12}$$

where $\alpha = 1$ and $\beta = 3$ for hands, upper and lower body motion. For facial motion, we set $\alpha = 0$ and $\beta = 3$. By optimizing the total loss, the model is trained to generate diverse gesture results.

4 Experiments

4.1 Experiments Setting

We train and evaluate on the BEATX-standard dataset proposed by [32]. BEATX-standard contains 60 hours of data with high finger quality for 25 speakers (12 female and 13 male). The dataset comprises 1762 sequences, each with an average duration of 65.66 seconds. Each sequence includes a response to a daily inquiry. We split datasets into 85%/7.5%/7.5% for the train/val/test set.

4.2 Implementation Details

We utilize the Adam optimizer with a learning rate of 2.5×10^{-4}. To maintain stability, we apply gradient norm clipping at a value of 0.99. In the construction of the VQVAEs, we employ a uniform initialization for the codebook, setting the codebook entries to feature lengths of 512 and establishing the codebook size at 256. The numerical distribution range for the codebook initialization is defined as $[-1/\text{codebook_size}, 1/\text{codebook_size})$. The codebook is solely updated during the first stage, and in the second stage of training for the speech-to-gesture mapping, the codebook remains frozen. The VQVAEs are trained for 200 epochs, with a learning rate of 2.5×10^{-4} for the first 195 epochs, which is then reduced to 2.5×10^{-5} for the final 5 epochs. During the second stage, the model is trained for 300 epochs. All experiments are conducted using one NVIDIA A100 GPU.

4.3 Metrics

To evaluate the realism of body gestures, we employ Fréchet Gesture Distance (FGD) [70] to measure the proximity of the distribution between the ground truth and generated gestures. Subsequently, Diversity [28] is quantified by computing the average L1 distance across multiple gesture clips. The synchronization between speech and motion is achieved using Beat Constancy (BC) [30]. For facial motions, we assess positional accuracy by calculating the vertex Mean Squared Error (MSE) [60]. Additionally, the difference between the ground truth and the generated facial vertices is measured using the vertex L1 difference (LVD) [68]. *More details about metrics and efficiency analysis are provided in the supplementary materials.*

4.4 Quantitative Results

As shown in Table 1, our method attains the highest BC, with a significant increase of 14.3% when compared to the previously top-performing method, which underscores the superior capability of

MambaTalk in discerning and correlating audio-motion beats. Furthermore, our method performs competitively with EMAGE in terms of FGD, and much better than others. This emphasizes the high quality and naturalness of our generated movements, showing the ability of MambaTalk to capture real motion dynamics. This also demonstrates the authenticity of our generated motions, affirming the successful capture of inherent motion characteristics. Some results are marked as "-" because these methods can not generate facial movements. Moreover, our method outperforms previous methods in terms of MSE and LVD, with substantial improvements of 36% and 14%, respectively. These two enhancements highlight the superior accuracy and fidelity of our method in capturing fine-grained details, affirming its efficacy in synthesizing realistic and authentic facial motions.

Table 1: Quantitative results on BEATX. FGD (Frechet Gesture Distance) multiplied by 10^{-1}, BC (Beat Constancy) multiplied by 10^{-1}, Diversity, MSE (Mean Squared Error) multiplied by 10^{-7}, and LVD (Learned Vector Distance) multiplied by 10^{-5}. The best results are in bold.

Methods	Venue	FGD ↓	BC ↑	Diversity ↑	MSE ↓	LVD ↓
Non-facial Gesture Synthesis						
S2G [11]	ICRA 2019	28.15	4.683	5.971	-	-
Trimodal [70]	TOG 2020	12.41	5.933	7.724	-	-
HA2G [34]	CVPR 2022	12.32	6.779	8.626	-	-
DisCo [31]	ACMMM 2022	9.417	6.439	9.912	-	-
CaMN [33]	ECCV 2022	6.644	6.769	10.86	-	-
DiffStyleGesture [63]	IJCAI 2023	8.811	7.241	11.49	-	-
Holistic Gesture Synthesis						
Habible *et al.* [18]	IVA 2021	9.040	7.716	8.21	8.614	8.043
TalkShow [68]	CVPR 2023	6.209	6.947	**13.47**	7.791	7.771
EMAGE [32]	CVPR 2024	5.512	7.724	13.06	7.680	7.556
MambaTalk (Ours)	-	**5.409**	**7.930**	13.07	**7.512**	**7.496**

4.5 Qualitative Analysis

User Study. We conducted a user study to assess the visual quality of the generated co-speech 3D gestures. For each method under comparison, we produced 10 gesture samples, which were then converted into video clips for evaluation by 39 participants. In each evaluation session, participants were presented with 20 seconds video clips generated by various models. They were instructed to assess the clips across the following dimensions: (i) naturalness, (ii) appropriateness, (iii) synchrony and (iv) smoothness. For naturalness, they evaluated the similarity of the generated gestures to those made by humans, paying attention to the authenticity and smoothness of the movements. In terms of appropriateness, they considered the alignment of the gestures with the spoken content, taking into account both the explicit meaning and the underlying semantics. For synchrony assessment, they examined the timing of the gestures in relation to the speech rhythm, audio, and facial expressions to ensure a harmonious and integrated performance. For smoothness, they assessed the gestures for any abrupt stops or unnatural jerks that might indicate a lack of fluidity in motion. We mainly compared two state-of-art methods with our proposed method (with and without VQVAE): CaMN [33], EMAGE [32], and the ground truth. As presented in Table 2, our method's average scores are higher than previous methods.

Visualization. As depicted in Figure 3, our approach yields gestures that exhibit enhanced rhythmic alignment and a more natural appearance. For instance, when conveying "we were", our method instructs the subject to hold both hands in front of the chest, a nuanced detail absent in both CaMN and EMAGE's outcomes, where either one or both arms hang down. Additionally, when representing "no place to", our method aligns with the ground truth by extending both arms upwards, whereas CaMN and EMAGE have their arms tucked in next to the body. In the case of "up", our generated result raises the right arm in alignment with the semantics of movement. In the context of "moving around" where our left and right arm swings may differ from the ground truth, the overall movement remains consistent.

Interestingly, for "sound of gunfire", a difficult semantic for the model to learn, our method still generates the result of the character's right hand clenched in a fist and the arm bent to indicate a tense

Table 2: User study results on naturalness (human likeness), appropriateness (the degree of consistency with the speech content), synchrony (the level of synchronization with the speech rhythm) and smoothness (the fluency of actions). The rating score range is 1-5, with 5 being the best. "Avg." denotes the average scores. ↑ indicates the higher the better.

Methods	Naturalness↑	Appropriateness↑	Synchrony↑	Smoothness↑	Avg.
CaMN [33]	3.08	3.34	3.25	3.50	3.29
EMAGE [32]	3.85	4.04	3.89	4.21	3.99
Ours$_{\text{w/o VQVAE}}$	1.24	1.24	1.18	1.29	1.24
Ours	4.04	4.00	3.91	4.35	4.08
Ground Truth	4.57	4.54	4.23	4.63	4.50

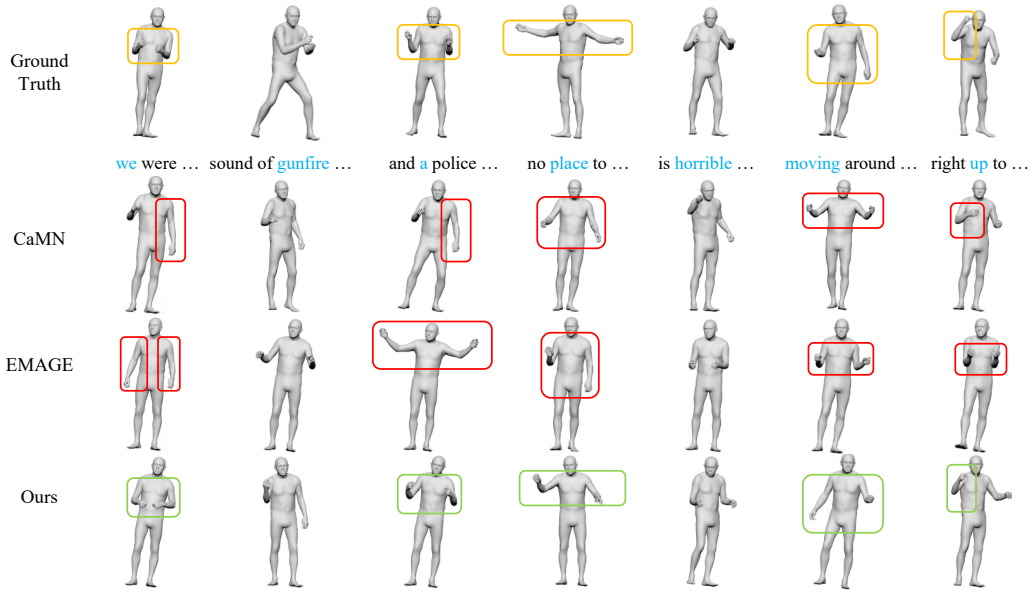

Figure 3: Visualization of the gestures generated by CaMN, EMAGE and our method. Unreasonable results are indicated by red boxes and reasonable ones by green boxes.

situation. For the emotion of fear expressed by "is horrible", the result of our method is similar to the ground truth, with the character's hands hanging down and face facing downward, which is a visual representation of the psychological state of panic and fear. In addition, as illustrated in Figure 3, our generated motions exhibit not only diverse characteristics, such as the range of motion and which hands to use, but also a high degree of consistency with the ground truth.

4.6 Ablation Study

Effect of VQVAEs. We confirm the significant role of the VQVAEs. As shown in Table 2 and Table 3, the integration of VQVAEs is essential for the functionality of our approach, contributing to the generation of gestures that exhibit smoother transitions and a more human-like quality. As demonstrated in Table 3, the removal of the VQVAEs ("−VQVAEs") from the model is also associated with performance decline, manifesting reduction in FGD, BC, Diversity, MSE and LVD.

Effect of Local Scan. We validate the effectiveness of the local scan. The ablation study is divided into two segments: (i) multi head cross attention and (ii) Mamba models for different part of bodys. As shown in Table 3, incorporating multi head cross attention enhances our method's capability to generate gestures with higher diversity and beat constancy. The incorporation of the Mamba from local scan generate gestures characterized by greater diversity and beat constancy. Concurrently, there is an observed improvement in the FGD of the generated gestures.

Table 3: Ablation study on different components of our proposed method. ↓ denotes the lower the better, and ↑ denotes the higher the better. FGD multiplied by 10^{-1}, BC multiplied by 10^{-1}, Diversity, MSE multiplied by 10^{-7}, and LVD multiplied by 10^{-5}.

Method	FGD ↓	BC ↑	Diversity ↑	MSE ↓	LVD ↓
Ours	5.409	7.930	13.072	0.751	7.496
− VQVAEs	12.051	7.447	8.462	1.316	9.235
− Local Scan ($\mathcal{F}_{\text{MHCA}}$)	7.554	7.139	11.473	0.805	7.553
− Local Scan (Mamba)	7.979	7.761	10.596	0.772	7.622
− Global Scan ($\mathcal{F}_{\text{MHSA}}$)	5.739	7.491	12.415	0.758	7.609
− Global Scan (Mamba)	6.267	7.823	12.047	0.825	7.707

Effect of Global Scan. We validate the effectiveness of the global scan, as listed in Table 3, the incorporation of global scan improves the overall performance of our method. For the multi-head self-attention module in global scan, the incorporation of multi-head self-attention acquires improvement of Diversity and a degradation for FGD. Additionally, the ablation results demonstrate that incorporating Mamba enhances the global scan's capability to generate gestures with higher diversity and beat constancy. The FGD of generated gestures is better at the same time.

Table 4: Ablation study on different audio encoders. ↓ denotes the lower the better, and ↑ denotes the higher the better. FGD multiplied by 10^{-1}, BC multiplied by 10^{-1}, Diversity, MSE multiplied by 10^{-7}, and LVD multiplied by 10^{-5}.

Method	FGD ↓	BC ↑	Diversity ↑	MSE ↓	LVD ↓
Ours	5.409	7.930	13.072	0.751	7.496
Whisper [47]	6.791	7.515	12.617	0.537	6.445
Wav2vec2 [3]	5.343	7.956	13.164	0.973	8.452

Effect of Different Audio Encoder. To validate the effectiveness of the audio encoder, we replace the CNN-based audio encoder with a pre-trained Whisper [47] and Vav2Vec2 [3], as listed in Table 4. Unlike CNN-based audio encoders that are randomly initialized and trained from scratch, when using Whisper and Wav2Vec2, we initialize the encoder using pre-trained weights and fix the parameters of the feature extractor. We observe a notable enhancement in facial generation when utilizing Whisper, however, the body generation results were subpar. In contrast, while Wav2Vec2 demonstrates some improvement in body generation, it results in a substantial decline in facial generation quality.

5 Conclusion

In this study, we propose a framework to employ the state space models in gesture synthesis. To alleviate the problem of jitter in gesture synthesis, we have implemented discrete motion priors, which enhance the effectiveness of the selective scan mechanism and lead to smoother results. We further incorporate the selective state space models with attention mechanisms to enhance the refinement of motion features in latent space. These modules capture the subtle movements and deformations of various body parts, thereby enhancing the overall quality of the generated gestures. By utilizing a linear time series modeling strategy with selective state space, our method achieves high-quality full body gesture generation with low latency.

Acknowledgements

This work was supported by the STI 2030-Major Projects under Grant 2021ZD0201404 and in part by the National Natural Science Foundation of China under Grant 62306165.

References

[1] Ao, T., Gao, Q., Lou, Y., Chen, B., Liu, L.: Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. ACM Transactions on Graphics (TOG) **41**(6), 1–19 (2022)

[2] Ao, T., Zhang, Z., Liu, L.: Gesturediffuclip: Gesture diffusion model with clip latents. arXiv preprint arXiv:2303.14613 (2023)

[3] Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)

[4] Bhattacharya, U., Rewkowski, N., Banerjee, A., Guhan, P., Bera, A., Manocha, D.: Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In: 2021 IEEE virtual reality and 3D user interfaces (VR). pp. 1–10. IEEE (2021)

[5] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the association for computational linguistics **5**, 135–146 (2017)

[6] Chemburkar, A., Lu, S., Feng, A.: Discrete diffusion for co-speech gesture synthesis. In: Companion Publication of the 25th International Conference on Multimodal Interaction. pp. 186–192 (2023)

[7] Chen, J., Liu, Y., Wang, J., Zeng, A., Li, Y., Chen, Q.: Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation. arXiv preprint arXiv:2401.04747 (2024)

[8] Chhatre, K., Daněček, R., Athanasiou, N., Becherini, G., Peters, C., Black, M.J., Bolkart, T.: Emotional speech-driven 3d body animation via disentangled latent diffusion. arXiv preprint arXiv:2312.04466 (2023)

[9] Deichler, A., Mehta, S., Alexanderson, S., Beskow, J.: Diffusion-based co-speech gesture generation using joint text and audio representation. In: Proceedings of the 25th International Conference on Multimodal Interaction. pp. 755–762 (2023)

[10] Fu, D.Y., Dao, T., Saab, K.K., Thomas, A.W., Rudra, A., Ré, C.: Hungry hungry hippos: Towards language modeling with state space models. arXiv preprint arXiv:2212.14052 (2022)

[11] Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J.: Learning individual styles of conversational gesture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3497–3506 (2019)

[12] Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)

[13] Gu, A., Dao, T., Ermon, S., Rudra, A., Ré, C.: Hippo: Recurrent memory with optimal polynomial projections. Advances in neural information processing systems **33**, 1474–1487 (2020)

[14] Gu, A., Goel, K., Gupta, A., Ré, C.: On the parameterization and initialization of diagonal state space models. Advances in Neural Information Processing Systems **35**, 35971–35983 (2022)

[15] Gu, A., Goel, K., Re, C.: Efficiently modeling long sequences with structured state spaces. In: International Conference on Learning Representations (2021)

[16] Gu, A., Johnson, I., Goel, K., Saab, K., Dao, T., Rudra, A., Ré, C.: Combining recurrent, convolutional, and continuous-time models with linear state space layers. Advances in neural information processing systems **34**, 572–585 (2021)

[17] Gupta, A., Gu, A., Berant, J.: Diagonal state spaces are as effective as structured state spaces. Advances in Neural Information Processing Systems **35**, 22982–22994 (2022)

[18] Habibie, I., Xu, W., Mehta, D., Liu, L., Seidel, H.P., Pons-Moll, G., Elgharib, M., Theobalt, C.: Learning speech-driven 3d conversational gestures from video. In: Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents. pp. 101–108 (2021)

[19] Hasani, R., Lechner, M., Wang, T.H., Chahine, M., Amini, A., Rus, D.: Liquid structural state-space models. In: The Eleventh International Conference on Learning Representations (2022)

[20] He, C., Shen, Y., Fang, C., Xiao, F., Tang, L., Zhang, Y., Zuo, W., Guo, Z., Li, X.: Diffusion models in low-level vision: A survey. arXiv preprint arXiv:2406.11138 (2024)

[21] Kim, G., Li, Y., Ko, H.: The ku-ispl entry to the genea challenge 2023-a diffusion model for co-speech gesture generation. In: Companion Publication of the 25th International Conference on Multimodal Interaction. pp. 220–227 (2023)

[22] Kipp, M., Neff, M., Kipp, K.H., Albrecht, I.: Towards natural gesture synthesis: Evaluating gesture units in a data-driven approach to gesture synthesis. In: Intelligent Virtual Agents: 7th International Conference, IVA 2007 Paris, France, September 17-19, 2007 Proceedings 7. pp. 15–28. Springer (2007)

[23] Kopp, S., Wachsmuth, I.: Synthesizing multimodal utterances for conversational agents. Computer animation and virtual worlds **15**(1), 39–52 (2004)

[24] Kucherenko, T., Jonell, P., Yoon, Y., Wolfert, P., Henter, G.E.: A large, crowdsourced evaluation of gesture generation systems on common data: The genea challenge 2020. In: 26th international conference on intelligent user interfaces. pp. 11–21 (2021)

[25] Levine, S., Krähenbühl, P., Thrun, S., Koltun, V.: Gesture controllers. In: ACM SIGGRAPH 2010 papers. pp. 1–11 (2010)

[26] Li, H., Cao, M., Cheng, X., Li, Y., Zhu, Z., Zou, Y.: G2l: Semantically aligned and uniform video grounding via geodesic and game theory. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12032–12042 (2023)

[27] Li, H., Cao, M., Cheng, X., Li, Y., Zhu, Z., Zou, Y.: Exploiting auxiliary caption for video grounding. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 18508–18516 (2024)

[28] Li, J., Kang, D., Pei, W., Zhe, X., Zhang, Y., He, Z., Bao, L.: Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11293–11302 (2021)

[29] Li, R., Zhang, Y., Zhang, Y., Zhang, H., Guo, J., Zhang, Y., Liu, Y., Li, X.: Lodge: A coarse to fine diffusion network for long dance generation guided by the characteristic dance primitives. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1524–1534 (2024)

[30] Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13401–13412 (2021)

[31] Liu, H., Iwamoto, N., Zhu, Z., Li, Z., Zhou, Y., Bozkurt, E., Zheng, B.: Disco: Disentangled implicit content and rhythm learning for diverse co-speech gestures synthesis. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 3764–3773 (2022)

[32] Liu, H., Zhu, Z., Becherini, G., Peng, Y., Su, M., Zhou, Y., Iwamoto, N., Zheng, B., Black, M.J.: Emage: Towards unified holistic co-speech gesture generation via masked audio gesture modeling. arXiv preprint arXiv:2401.00374 (2023)

[33] Liu, H., Zhu, Z., Iwamoto, N., Peng, Y., Li, Z., Zhou, Y., Bozkurt, E., Zheng, B.: Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII. pp. 612–630. Springer (2022)

[34] Liu, X., Wu, Q., Zhou, H., Xu, Y., Qian, R., Lin, X., Zhou, X., Wu, W., Dai, B., Zhou, B.: Learning hierarchical cross-modal association for co-speech gesture generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10462–10472 (2022)

[35] Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y.: Vmamba: Visual state space model. arXiv preprint arXiv:2401.10166 (2024)

[36] Lu, S., Yoon, Y., Feng, A.: Co-speech gesture synthesis using discrete gesture token learning. arXiv preprint arXiv:2303.12822 (2023)

[37] Lu, Y., Zhang, M., Lin, Y., Ma, A.J., Xie, X., Lai, J.: Improving pre-trained masked autoencoder via locality enhancement for person re-identification. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 509–521. Springer (2022)

[38] Ma, X., Zhou, C., Kong, X., He, J., Gui, L., Neubig, G., May, J., Zettlemoyer, L.: Mega: Moving average equipped gated attention. In: The Eleventh International Conference on Learning Representations (2022)

[39] Ma, Y., He, Y., Cun, X., Wang, X., Chen, S., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4117–4125 (2024)

[40] Ma, Y., Liu, H., Wang, H., Pan, H., He, Y., Yuan, J., Zeng, A., Cai, C., Shum, H.Y., Liu, W., et al.: Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. arXiv preprint arXiv:2406.01900 (2024)

[41] Mehta, H., Gupta, A., Cutkosky, A., Neyshabur, B.: Long range language modeling via gated state spaces. In: International Conference on Learning Representations (2023)

[42] Nyatsanga, S., Kucherenko, T., Ahuja, C., Henter, G.E., Neff, M.: A comprehensive review of data-driven co-speech gesture generation. In: Computer Graphics Forum. vol. 42, pp. 569–596. Wiley Online Library (2023)

[43] Pang, K., Qin, D., Fan, Y., Habekost, J., Shiratori, T., Yamagishi, J., Komura, T.: Bodyformer: Semantics-guided 3d body gesture synthesis with transformer. ACM Transactions on Graphics (TOG) **42**(4), 1–12 (2023)

[44] Press, O., Smith, N.A., Lewis, M.: Train short, test long: Attention with linear biases enables input length extrapolation. arXiv preprint arXiv:2108.12409 (2021)

[45] Qi, X., Liu, C., Li, L., Hou, J., Xin, H., Yu, X.: Emotiongesture: Audio-driven diverse emotional co-speech 3d gesture generation. arXiv preprint arXiv:2305.18891 (2023)

[46] Qian, S., Tu, Z., Zhi, Y., Liu, W., Gao, S.: Speech drives templates: Co-speech gesture synthesis with learned templates. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11077–11086 (2021)

[47] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International conference on machine learning. pp. 28492–28518. PMLR (2023)

[48] Ren, Y., Xia, X., Lu, Y., Zhang, J., Wu, J., Xie, P., Wang, X., Xiao, X.: Hyper-sd: Trajectory segmented consistency model for efficient image synthesis. arXiv preprint arXiv:2404.13686 (2024)

[49] Smith, J.T., Warrington, A., Linderman, S.: Simplified state space layers for sequence modeling. In: The Eleventh International Conference on Learning Representations (2022)

[50] Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)

[51] Tonoli, R.L., Marques, L.B.d.M., Ueda, L.H., Costa, P.D.P.: Gesture generation with diffusion models aided by speech activity information. In: Companion Publication of the 25th International Conference on Multimodal Interaction. pp. 193–199 (2023)

[52] Tykkälä, T., Audras, C., Comport, A.I.: Direct iterative closest point for real-time visual odometry. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). pp. 2050–2056. IEEE (2011)

[53] Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)

[54] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[55] Voß, H., Kopp, S.: Aq-gt: a temporally aligned and quantized gru-transformer for co-speech gesture synthesis. arXiv preprint arXiv:2305.01241 (2023)

[56] Wagner, P., Malisz, Z., Kopp, S.: Gesture and speech in interaction: An overview (2014)

[57] Wang, Z., Zheng, J.Q., Zhang, Y., Cui, G., Li, L.: Mamba-unet: Unet-like pure visual mamba for medical image segmentation. arXiv preprint arXiv:2402.05079 (2024)

[58] Wu*, X., Li*, H., Luo, Y., Cheng, X., Zhuang, X., Cao, M., Fu, K.: Uncertainty-aware sign language video retrieval with probability distribution modeling. ECCV 2024 (2024)

[59] Xiao, Y., Song, L., Huang, S., Wang, J., Song, S., Ge, Y., Li, X., Shan, Y.: Grootvl: Tree topology is all you need in state space model. arXiv preprint arXiv:2406.02395 (2024)

[60] Xing, J., Xia, M., Zhang, Y., Cun, X., Wang, J., Wong, T.T.: Codetalker: Speech-driven 3d facial animation with discrete motion prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12780–12790 (2023)

[61] Xu, Z., Zhang, Y., Yang, S., Li, R., Li, X.: Chain of generation: Multi-modal gesture synthesis via cascaded conditional control. arXiv preprint arXiv:2312.15900 (2023)

[62] Yang, S., Wang, Z., Wu, Z., Li, M., Zhang, Z., Huang, Q., Hao, L., Xu, S., Wu, X., Yang, C., et al.: Unifiedgesture: A unified gesture synthesis model for multiple skeletons. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 1033–1044 (2023)

[63] Yang, S., Wu, Z., Li, M., Zhang, Z., Hao, L., Bao, W., Cheng, M., Xiao, L.: Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models. arXiv preprint arXiv:2305.04919 (2023)

[64] Yang, S., Wu, Z., Li, M., Zhang, Z., Hao, L., Bao, W., Zhuang, H.: Qpgesture: Quantization-based and phase-guided motion matching for natural speech-driven gesture generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 2321–2330. IEEE (June 2023)

[65] Yang, S., Xu, Z., Xue, H., Cheng, Y., Huang, S., Gong, M., Wu, Z.: Freetalker: Controllable speech and text-driven gesture generation based on diffusion models for enhanced speaker naturalness. arXiv preprint arXiv:2401.03476 (2024)

[66] Yang, S., Xue, H., Zhang, Z., Li, M., Wu, Z., Wu, X., Xu, S., Dai, Z.: The diffusestylegesture+ entry to the genea challenge 2023. In: Proceedings of the 25th International Conference on Multimodal Interaction. pp. 779–785 (2023)

[67] Yazdian, P.J., Chen, M., Lim, A.: Gesture2vec: Clustering gestures using representation learning methods for co-speech gesture generation. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3100–3107. IEEE (2022)

[68] Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., Tao, D., Black, M.J.: Generating holistic 3d human motion from speech. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 469–480 (2023)

[69] Yin, L., Wang, Y., He, T., Liu, J., Zhao, W., Li, B., Jin, X., Lin, J.: Emog: Synthesizing emotive co-speech 3d gesture with diffusion model. arXiv preprint arXiv:2306.11496 (2023)

[70] Yoon, Y., Cha, B., Lee, J.H., Jang, M., Lee, J., Kim, J., Lee, G.: Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Transactions on Graphics (TOG) **39**(6), 1–16 (2020)

[71] Yoon, Y., Ko, W.R., Jang, M., Lee, J., Kim, J., Lee, G.: Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 4303–4309. IEEE (2019)

[72] Zhao, W., Hu, L., Zhang, S.: Diffugesture: Generating human gesture from two-person dialogue with diffusion models. In: Companion Publication of the 25th International Conference on Multimodal Interaction. pp. 179–185 (2023)

[73] Zhong, Z., Mi, Y., Huang, Y., Xu, J., Mu, G., Ding, S., Zhang, J., Guo, R., Wu, Y., Zhou, S.: Slerpface: Face template protection via spherical linear interpolation. arXiv preprint arXiv:2407.03043 (2024)

[74] Zhu, L., Liu, X., Liu, X., Qian, R., Liu, Z., Yu, L.: Taming diffusion models for audio-driven co-speech gesture generation. arXiv preprint arXiv:2303.09119 (2023)

[75] Zhuang, X., Li, H., Cheng, X., Zhu, Z., Xie, Y., Zou, Y.: Kdpror: A knowledge-decoupling probabilistic framework for video-text retrieval. ECCV 2024 (2024)

A Appendix / Supplemental material

A.1 More visualization results

Figure 4 presents the facial motion results generated by our method, showcasing the generation of facial expressions and movements with a high level of realism. Our method effectively synchronizes with the phonetic articulation of speech content, accurately reflecting the physical demands of pronunciation. For instance, when uttering "walking", "came" or "bus", our approach ensures that the mouth's movements, such as opening, correspond closely with the actual phonetic requirements. Other methods do not consistently achieve this level of accuracy in aligning with the phonetic and physical nuances of speech. Our method adeptly handles the subtleties of mouth closure and elongation required for sounds such as "in", closely aligning with the ground truth, whereas other approaches may exhibit inconsistencies in this regard. Moreover, in instances of silence, all methods, including ours, demonstrate a good capacity to learn and maintain the mouth's closed position, effectively reflecting the underlying patterns of speech and silence. Since CaMN does not specifically target the generation of facial movements, it results in a lack of variation in facial expressions throughout the process.

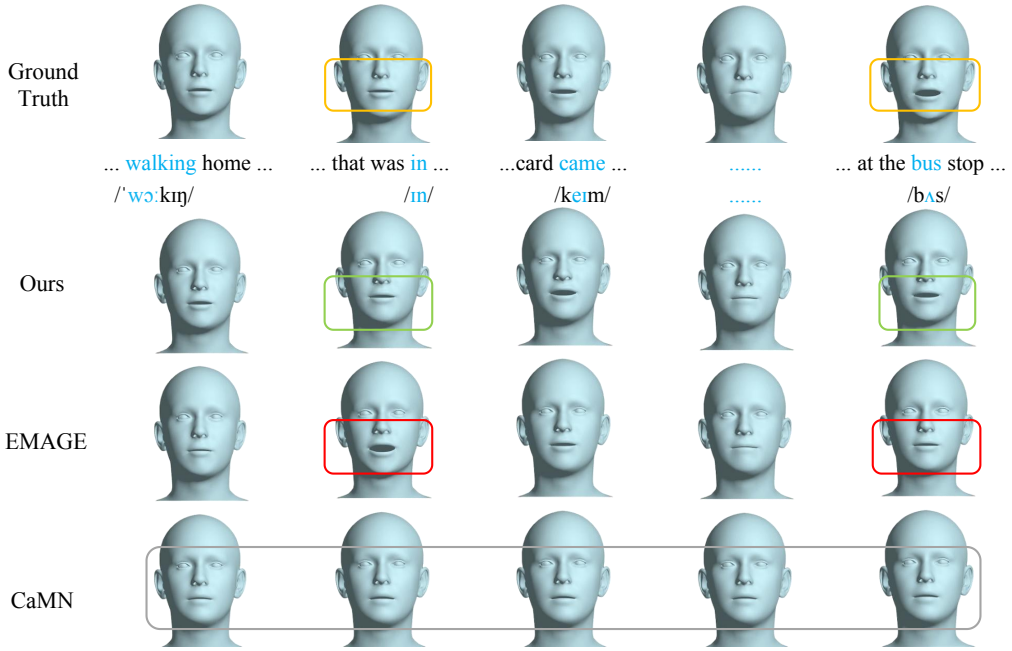

Figure 4: Visualization of the facial motions generated by CaMN, EMAGE and our method. Unreasonable results are indicated by red and gray boxes and reasonable ones by green boxes.

As illustrated in Figure 5, our methodology generates gestures that demonstrate improved rhythmic synchronization and a more lifelike appearance, effectively capturing the essence of the speaker's rhythmic patterns. For example, in the expression "actually" our method guides the individual to bring the hands inward in front of the chest, a subtle gesture not observed in the results produced by CaMN and EMAGE, where the arms are either hang limply at the sides or are splayed downward. Furthermore, in the depiction of "on the way back" our approach accurately reflects the ground truth by slightly bending down and raising one hand, while EMAGE cannot respond accurately to this and remains standing.

Furthermore, our approach accurately captures the semantic essence of movements. For instance, in response to the cue "hug" our method generates an inward-circling motion of the arms, aligning perfectly with the ground truth, which is a nuanced semantic element that other methodologies neglect. Similarly, in scenarios such as "so small", the result of our method is similar to the ground truth, with the character's hand moving inward. This attention to detail ensures semantic consistency, which is lacking in other approaches where actions are not aligned with the intended meaning.

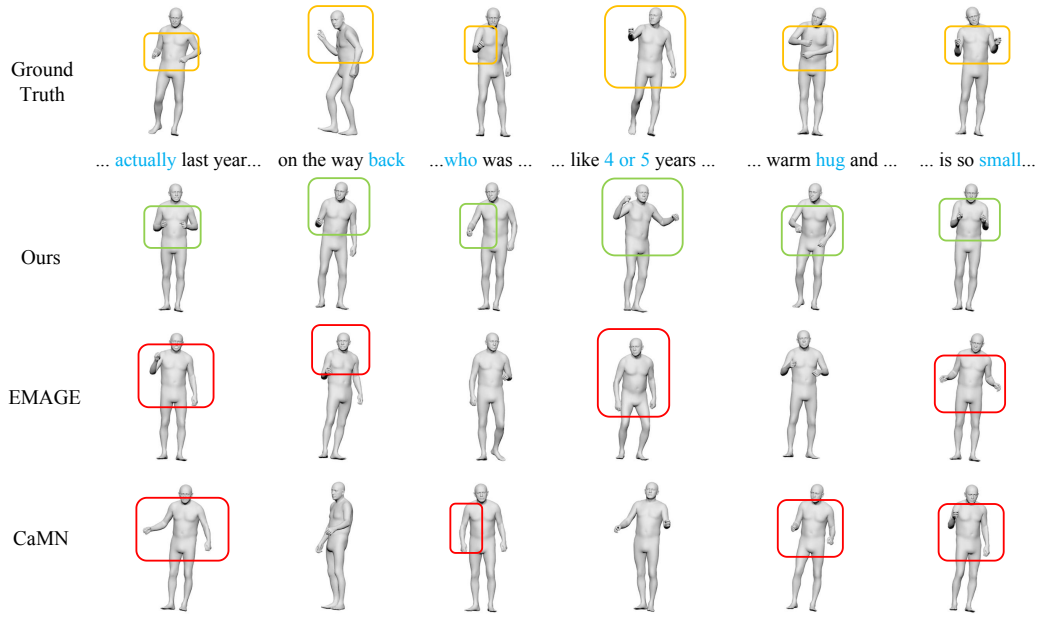

Figure 5: Visualization of the gestures generated by CaMN, EMAGE and our method. Unreasonable results are indicated by red boxes and reasonable ones by green boxes.

A.2 Efficiency Analysis

We leverage the linear computational complexity of Mamba and the sequence compression capability of VQVAE within our framework, which helps in reducing computational complexity. Although there are some specialized acceleration solutions, faster solutions are necessary because when the model is integrated into the system, there is not only a delay in generating gestures, but also delays in other modules. Therefore, to evaluate the efficiency of our pipeline, we conducted a series of measurements, focusing on the runtime of individual components. We measured the runtime of various components on the NVIDIA A100 GPU in our method over three runs and presented the average results in Table 5. We also compare our method's inference time with diffusion-based methods. The average computation time was determined based on the generation of a 31 second motion sequence, to showcase our model's low latency capabilities. The total inference time of our method is much slower than state-of-the-art diffusion-based method [63]. The result confirm that our pipeline is well-suited for applications requiring low latency gesture generation like interactive systems, where responsiveness is paramount.

Table 5: The time cost for generating one second (average) of gestures using the method's modules.

Modules	Run Time(s)
Diffusion-based method	
DiffStyleGesture [63]	0.64365 ± 0.0086
Our method	
Audio Encoders	0.00217 ± 0.0006
Text Encoders	0.00480 ± 0.0001
Local Scan	0.00219 ± 0.0004
Global Scan	0.00676 ± 0.0003
Face VQDecoder	0.00073 ± 0.0001
Hand VQDecoder	0.00077 ± 0.0001
Upper VQDecoder	0.00106 ± 0.0001
Lower VQDecoder	0.00068 ± 0.0001
Total Time	0.01917 ± 0.0018

A.3 Evaluation on BEAT dataset

To evaluate the generalisable benefit of our method, we conduct experiments on a large-scale multimodal dataset known as BEAT (Body-Expression-Audio-Text) [33]. This dataset encompasses 76 hours of multimodal data collected from 30 speakers engaging in conversations across four different languages while expressing eight distinct emotions. The dataset includes conversational gestures, facial expressions, emotional cues, and semantic content, along with annotations for audio, text, and speaker identity. To facilitate a fair comparison, we follow CaMN [33] and employ approximately 16 hours of speech data from English speakers. Furthermore, we implement the conventional approach of partitioning the dataset into distinct training, validation, and testing subsets, ensuring consistency with the data partitioning scheme utilized in prior research to uphold the integrity of the comparison.

To facilitate a fair comparison, we employ a total of N = 34 frame clips with a stride of 10 during the training process. The first four frames serve as seed poses, while the model is trained to generate the subsequent 30 poses, which collectively represent a duration of 2 seconds. Our models incorporate 47 joints from the BEAT dataset, comprising 38 hand joints and 9 body joints. As listed in Table 6, our method demonstrates a significant improvement compared to the CaMN (baseline), which also validates the generalisable benefits of our approach.

Table 6: Comparison with state-of-the-art method in the term of FGD, SRGR and BeatAlign. All methods are trained on BEAT datasets. $\downarrow$ denotes the lower the better while $\uparrow$ denotes the higher the better. The best results are in bold.

Methods	FGD $\downarrow$	SRGR $\uparrow$	BeatAlign $\uparrow$
Seq2Seq [71]	261.3	0.173	0.729
Speech2Gesture [11]	256.7	0.092	0.751
MultiContext [70]	176.2	0.195	0.776
Audio2Gesture [28]	223.8	0.097	0.766
CaMN [33]	123.7	0.239	0.783
TalkShow [68]	91.00	-	0.840
MambaTalk (ours)	**51.3**	**0.256**	**0.852**

A.4 Limitations

Currently, our approach to gesture synthesis involves using distinct modules to animate various body parts, which naturally introduces some latency. Developing a single, unified model capable of capturing the wide-ranging and intricate deformations and motion patterns characteristic of different body parts can be addressed in future research. This enhancement is anticipated to lower computational overhead and substantially reduce the processing time, thereby improving the real-time capabilities of the pipeline and ensuring a smoother and more responsive gesture generation system.

Meanwhile, exploring more robust audio representations or combining various types of pre-trained audio encoders could significantly enhance the quality of gesture generation. Our findings indicate that certain encoders, such as Whisper, are particularly effective for modeling facial movements, while others, like Wav2Vec2, are better suited for modeling body movements. This approach will further improve the overall performance of the method.

In addition, the issue of gesture diversity among speakers and across different cultures remains unaddressed. Addressing this gap is essential for improving the cross-cultural validity and expanding the applicability of gesture-based applications in diverse global contexts.

A.5 Pseudo Code

The local scanning procedure is illustrated in Algorithm 26. We attain local modeling of various body segments by individually processing actions within distinct regions. The approach to global scanning parallels this methodology; however, the key distinction lies in the simultaneous processing of motion representations across multiple body parts.

Algorithm 1 Local Scanning Process

Require: token sequence $\mathbf{T}_{l-1} : (\mathtt{B}, \mathtt{M}, \mathtt{D})$
Ensure: token sequence $\mathbf{T}_l : (\mathtt{B}, \mathtt{M}, \mathtt{D})$
/* model motions in different body regions separately $\mathbf{T}'_{l-1}$ */
$\mathbf{T}'_{l-1} : (\mathtt{B}, \mathtt{M}, \mathtt{D}) \leftarrow \mathbf{SelfAttention}(\mathbf{T}_{l-1})$
$\mathbf{z}_{\mathbf{face}} : (\mathtt{B}, \mathtt{M}, \mathtt{E}) \leftarrow \mathbf{Linear}^{\mathbf{face}}(\mathbf{T}'^{face}_{l-1})$
$\mathbf{z}_{\mathbf{upperbody}} : (\mathtt{B}, \mathtt{M}, \mathtt{E}) \leftarrow \mathbf{Linear}^{\mathbf{upperbody}}(\mathbf{T}'^{upperbody}_{l-1})$
$\mathbf{z}_{\mathbf{lowerbody}} : (\mathtt{B}, \mathtt{M}, \mathtt{E}) \leftarrow \mathbf{Linear}^{\mathbf{lowerbody}}(\mathbf{T}'^{lowerbody}_{l-1})$
$\mathbf{z}_{\mathbf{hand}} : (\mathtt{B}, \mathtt{M}, \mathtt{E}) \leftarrow \mathbf{Linear}^{\mathbf{hand}}(\mathbf{T}'^{hand}_{l-1})$
/* process with different parts of human body */
for o in {face, upperbody, lowerbody, hand} **do**
  $\mathbf{x}'_o : (\mathtt{B}, \mathtt{M}, \mathtt{E}) \leftarrow \mathbf{SiLU}(\mathbf{Conv1d}_o(\mathbf{x}))$
  $\mathbf{B}_o : (\mathtt{B}, \mathtt{M}, \mathtt{N}) \leftarrow \mathbf{Linear}^{\mathbf{B}}_o(\mathbf{x}'_o)$
  $\mathbf{C}_o : (\mathtt{B}, \mathtt{M}, \mathtt{N}) \leftarrow \mathbf{Linear}^{\mathbf{C}}_o(\mathbf{x}'_o)$
  /* softplus ensures positive $\boldsymbol{\Delta}_o$ */
  $\boldsymbol{\Delta}_o : (\mathtt{B}, \mathtt{M}, \mathtt{E}) \leftarrow \log(1 + \exp(\mathbf{Linear}^{\boldsymbol{\Delta}}_o(\mathbf{x}'_o) + \mathbf{Parameter}^{\boldsymbol{\Delta}}_o))$
  /* shape of $\mathbf{Parameter}^{\mathbf{A}}_o$ is $(\mathtt{E}, \mathtt{N})$ */
  $\overline{\mathbf{A}}_o : (\mathtt{B}, \mathtt{M}, \mathtt{E}, \mathtt{N}) \leftarrow \boldsymbol{\Delta}_o \bigotimes \mathbf{Parameter}^{\mathbf{A}}_o$
  $\overline{\mathbf{B}}_o : (\mathtt{B}, \mathtt{M}, \mathtt{E}, \mathtt{N}) \leftarrow \boldsymbol{\Delta}_o \bigotimes \mathbf{B}_o$
  $\mathbf{y}_o : (\mathtt{B}, \mathtt{M}, \mathtt{E}) \leftarrow \mathbf{SSM}(\overline{\mathbf{A}}_o, \overline{\mathbf{B}}_o, \mathbf{C}_o)(\mathbf{x}'_o)$
end for
/* get gated $\mathbf{y}_o$ */
$\mathbf{y}'_{\text{face}} : (\mathtt{B}, \mathtt{M}, \mathtt{E}) \leftarrow \mathbf{y}_{\text{face}} \bigodot \mathbf{SiLU}(\mathbf{z}_{\mathbf{face}})$
$\mathbf{y}'_{\text{upperbody}} : (\mathtt{B}, \mathtt{M}, \mathtt{E}) \leftarrow \mathbf{y}_{\text{upperbody}} \bigodot \mathbf{SiLU}(\mathbf{z}_{\mathbf{upperbody}})$
$\mathbf{y}'_{\text{lowerbody}} : (\mathtt{B}, \mathtt{M}, \mathtt{E}) \leftarrow \mathbf{y}_{\text{lowerbody}} \bigodot \mathbf{SiLU}(\mathbf{z}_{\mathbf{lowerbody}})$
$\mathbf{y}'_{\text{hand}} : (\mathtt{B}, \mathtt{M}, \mathtt{E}) \leftarrow \mathbf{y}_{\text{hand}} \bigodot \mathbf{SiLU}(\mathbf{z}_{\mathbf{hand}})$
/* residual connection */
$\mathbf{T}_l : (\mathtt{B}, \mathtt{M}, \mathtt{D}) \leftarrow \mathbf{Linear}^{\mathbf{T}}(\mathbf{y}'_{\mathbf{face}} + \mathbf{y}'_{\mathbf{upperbody}} + \mathbf{y}'_{\mathbf{lowerbody}} + \mathbf{y}'_{\mathbf{hand}}) + \mathbf{T}_{l-1}$
Return: $\mathbf{T}_l$

A.6 Evaluation Metrics

To evaluate the realism of body gestures, we employ Fréchet Gesture Distance (FGD)[70] to measure how close the distribution between the ground truth and generated body gestures is.

$$\mathrm{FGD}(\mathbf{g}, \hat{\mathbf{g}}) = \|\mu_r - \mu_g\|^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r \Sigma_g\right)^{1/2}\right), \tag{13}$$

where μ_r and Σ_r denote the mean and covariance of the latent feature distribution z_r for real human gestures g, while μ_g and Σ_g correspond to the mean and covariance of the latent feature distribution z_g for the synthesized gestures $\hat{g}$. We employ an encoder based on a Skeleton CNN (SKCNN) and a Full CNN-based decoder, constituting our autoencoder's pretrained network. This network is trained on both the BEATX-Standard and BEATX-Additional datasets. The preference for SKCNN over a Full CNN encoder stems from its superior performance in capturing gesture features, evidenced by a reduced reconstruction MSE loss of 0.095, as opposed to 0.103.

Subsequently, Diversity[28] is quantified by computing the average L1 distance across multiple body gesture clips. Higher Diversity signifies greater variance within the gesture clips. We compute the average L1 distance across various N motion clips using the following equation:

$$\text{Diversity } = \frac{1}{2N(N-1)} \sum_{t=1}^{N} \sum_{j=1}^{N} \left\| p_t^i - \hat{p}_t^j \right\|_1, \tag{14}$$

where p_t denotes the positions of joints in frame t. We assess diversity across the entire test dataset. Moreover, when calculating joint positions, translation is zeroed, indicating that L1 Diversity is exclusively concentrated on local motion dynamics.

The synchronization between the speech and motion is conducted using Beat Constancy (BC)[30]. BC indicates a more precise synchronization between the rhythm of gestures and the audio's beat. We define the onset of speech as the audio's beat and identify the local minima of the upper body

joints' velocity (excluding fingers) as the motion's beat. The synchronization between audio and gesture is determined using the following equation:

$$\mathrm{BC} = \frac{1}{g} \sum_{b_g \in g} \exp\left(-\frac{\min_{b_a \in a} \|b_g - b_a\|^2}{2\sigma^2} \right), \tag{15}$$

where g and a represent the sets of gesture beats and audio beats, respectively.

Turning focus to facial aspects, we gauge the positional accuracy through the calculation of vertex Mean Squared Error (MSE)[60]. This metric quantifies the average squared difference between the predicted facial landmarks and their corresponding ground truths, providing a clear indication of the facial model's accuracy:

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (f_i - \hat{f}_i)^2, \tag{16}$$

where n denotes the number of vertices, f_i represents the ground truth position of the i-th vertex, $\hat{f}_i$ denotes the predicted position of the i-th vertex. The sum is taken over all vertices to compute the average error.

Additionally, the disparity between the ground truth and the generated facial vertices is measured by the vertex L1 difference (LVD)[68], which measures the synchronization between speech and facial expression.

$$\mathrm{LVD} = \frac{1}{n} \sum_{i=1}^{n} \left\| f'_i - \hat{f}'_i \right\|_1, \tag{17}$$

where n denotes the number of vertices, f'_i represents the ground truth speed of the i-th vertex. $\hat{f}'_i$ denotes the speed of the i-th vertex in the generated facial expression. The sum is taken over all vertices to compute the average absolute difference.

A.7 Ethical Considerations in Crowdsourcing Research

This section provides additional details about the user study for qualitative analysis. For our user study, we have randomly selected ten videos generated by different methods, each containing 20-second video clips. For each participant, we paid compensation that exceeded the local average hourly wage.

The screenshot of our user study website is illustrated in the Figure 6, which displays the template layout presented to the participants. In addition to the main trials, participants were also subjected to several catch trials. These trials involved displaying Ground Truth videos and videos with distorted motion. Participants who failed to score the GT videos higher and the distorted motion videos lower were considered unresponsive or inattentive and their data was not included in the final evaluation.

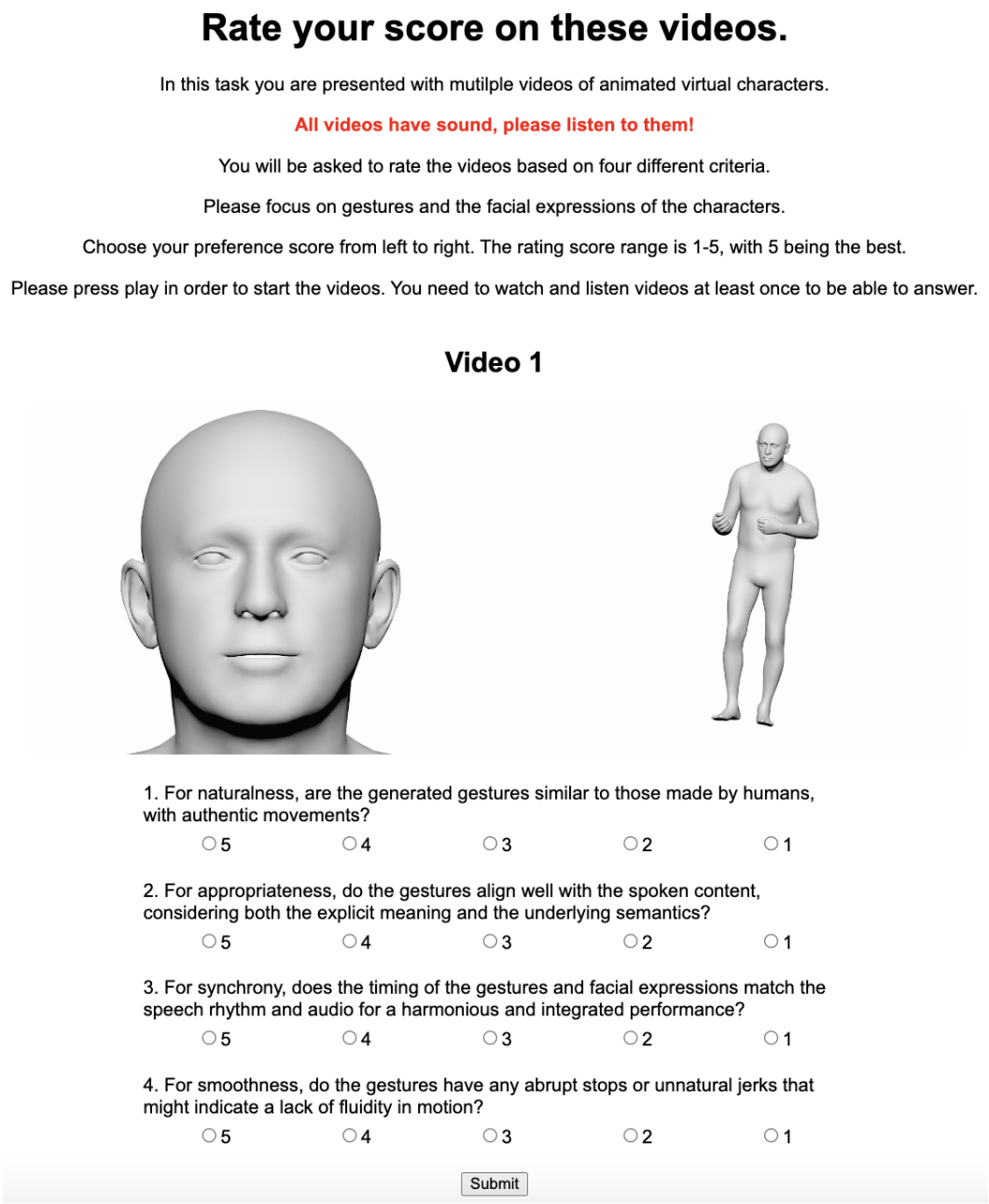

Figure 6: The screenshots of user study website for participants.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Overall, the abstract and introduction provide a concise yet comprehensive summary of the paper's objectives, methods, and findings, accurately reflecting its contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, we discuss the limitations in section A.4 in the Appendix part.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: By adhering to the principles mentioned in the Guidelines, we ensures that each theoretical result is underpinned by a full set of assumptions and complete, correct proofs, thus reinforcing the credibility and reliability of the paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: In Section 4. Experiments in the main text, we report all the experiments setting, implementation details and metrics, which disclose all the information needed to reproduce the main experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code is available.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In Section 4. Experiments in the main text, we report all the experiments setting and implementation details, facilitating readers' understanding of the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the error bars in Section A.2 Efficiency Analysis.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the computer resources in Section 4.2 Implementation Details, and show model runtimes in detail in Section A.2 Efficiency Analysis.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Yes, we do.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The work we perform makes no society impact. It is only an academic study.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cite the original owners of code, data and models.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Yes, we provide the details in appendix A.5 Ethical Considerations in Crowdsourcing Research.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Yes, we do.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.