GestureLSM: Latent Shortcut based Co-Speech Gesture Generation with Spatial-Temporal Modeling

Supplementary Document

A. Overview

The supplementary document contains implementation details, metric details, additional experimental results and training analysis. For more visual results, **please see the demo videos**.

B. Implementation Details

In the construction of the RVQVAEs, the codebook is initialized uniformly, with each entry having a feature length of 128 and a total size of 1,024 per body region. The codebook updates occur solely during the quantization process, with resets following Contextual Gesture [4]. The RVQVAEs are trained for 30,000 iterations, with a learning rate of 2×10^{-4}. The GestureLSM model contains 3 layers of cross-attention for audio-gesture feature fusion and 8 layers of spatial-temporal attention blocks. The latent dimension is set to 256 with feed-forward size of 1024. During the second training stage for speech-to-gesture generation, the codebook remains frozen. We train the GestureLSM model for 1000 epochs. We utilize the Adam optimizer with a learning rate of 2×10^{-4}. All experiments are conducted on a single NVIDIA A100 GPU. We adopt a guidance dropout rate of 0.1 during training and a speech-conditioning ratio of 2 during generation.

C. Metric Details

Fréchet Gesture Distance (FGD) Fréchet Gesture Distance (FGD), introduced in [6], quantifies the similarity between the distributions of real and generated gestures, where a lower FGD signifies a closer match. Inspired by perceptual loss in image generation, FGD is computed using latent features extracted from a pretrained network:

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

where μ_r and Σ_r denote the mean and covariance of the latent feature distribution z_r derived from real gestures $\mathbf{g}$, while μ_g and Σ_g correspond to the statistics of the generated gestures $\hat{\mathbf{g}}$.

L1 Diversity L1 Diversity, proposed in [2], measures the variation across multiple gesture sequences, with higher values indicating greater diversity. The average L1 distance across N motion sequences is computed as:

$$\text{L1 div.} = \frac{1}{2N(N-1)} \sum_{t=1}^{N} \sum_{j=1}^{N} \left\| p_t^i - \hat{p}_t^j \right\|_1, \quad (2)$$

where p_t represents the joint positions at frame t. Diversity is evaluated on the complete test set. To ensure a focus on local motion, global translation is neutralized when computing joint positions.

Beat Constancy (BC) Beat Constancy (BC), as defined in [3], assesses the temporal alignment between gestures and audio rhythm. Higher BC values indicate stronger synchronization. Speech onsets are treated as audio beats, while motion beats correspond to local minima in the upper body joint velocity (excluding fingers). The alignment is determined using:

$$\mathrm{BC} = \frac{1}{g} \sum_{b_g \in g} \exp\left(-\frac{\min_{b_a \in a} \|b_g - b_a\|^2}{2\sigma^2}\right), \quad (3)$$

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

D. Additional Experimental Results

Feature Contributions. We assess feature variations: (1) w/o text: Exclude speech transcripts. (2) wavLM: Replace the CNN audio encoder with pretrained WavLM [1]. (3) concatenate: Use concatenation with an MLP for fusion instead of cross-attention. (4) addition: Element-wise addition of speech and gesture features. Tab. 1a shows cross-attention consistently outperforms other fusion methods, while WavLM provides no advantage.

Classifier Free Guidance. We evaluate guidance scale for conditional generation. We show their performance by the same number of sampling steps of 8. Tab.1b shows a guidance scale of 2 achieves the best performance.

Gesture Representation. We evaluate gesture quantization methods: (1) w/o quant: Directly use 6D-rotations of joints, (2) one quant: Single VQ quantizer for the whole body. (3) one residual: Single RVQ quantizer for the

whole body. (4) product quant: 2D quantizer based on ProbTalk [5]. Tab. 1c shows RVQ outperforms VQ and product quantization. Separating body regions further improves performance over holistic representations.

Sequential Design of Attention. We analyze the sequential design of the proposed two types od attentions. Tab.1d shows attention in spatial-temporal order present slightly improvement.

Skewness of Time Distribution. We further evaluate the skewness of the proposed beta schedule for time stamp distirbution. Tab. 1e shows with $\beta = 1.2$ and $\alpha = 2.0$ achieves the best performance. This indicate the emphasize and a more significant left skewness with an emphasis appraoching to 1 is important for the model learning.

E. Training Analysis

We analyze how the model's performance evolves during training. As shown in Fig. 1a, the model exhibits the highest beat constancy at the early stages of training. However, we observe that this corresponds to unnatural, exaggerated motion patterns in response to speech beats. As training progresses, beat constancy gradually decreases. Importantly, we argue that higher beat constancy is not necessarily better. For reference, the ground-truth gestures exhibit a beat constancy of 0.703, suggesting that aligning this metric with real human motion is a more meaningful target. Based on this, we propose evaluating beat constancy relative to the ground-truth rather than treating higher values as strictly superior.

For other metrics, we observe a clear upward trend in gesture diversity and a corresponding decrease in FGD as training progresses. These trends align with existing literature, and we maintain the standard evaluation approach for these metrics.

References

[1] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518, 2022. 1

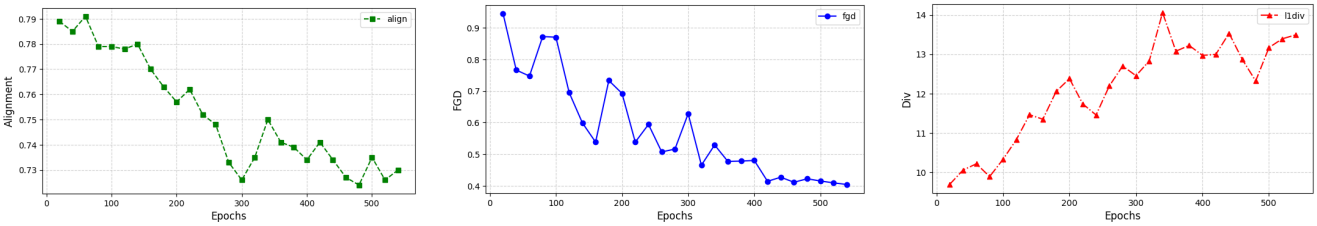
[2] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11293–11302, 2021. 1

[3] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. AI Choreographer: Music Conditioned 3D Dance Generation with AIST++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 1

[4] Pinxin Liu, Pengfei Zhang, Hyeongwoo Kim, Pablo Garrido, Ari Sharpio, and Kyle Olszewski. Contextual gesture: Co-speech gesture video generation through context-aware gesture representation, 2025. 1

[5] Yifei Liu, Qiong Cao, Yandong Wen, Huaiguang Jiang, and Changxing Ding. Towards variable and coordinated holistic co-speech motion generation. *arXiv preprint arXiv:2404.00368*, 2024. 2

[6] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity. *ACM TOG*, 39(6), 2020. 1



(a) Beat constancy over training. (b) Fréchet Gesture Distance (FGD) over training. (c) Gesture diversity over training.

Figure 1. Training dynamics of key evaluation metrics. (a) Beat constancy decreases, indicating a shift from overly rigid beat-following motions to more natural gestures. (b) FGD decreases, reflecting improved gesture realism. (c) Gesture diversity increases, suggesting a broader range of motion patterns learned by the model.

Table 1. Additional ablations of our method. We exam the speech feature, classifier free guidance scale, gesture representation, sequence order for the attention and the skewness for the sampling distribution. Bold indicates the best performance.

Features.	FGD↓	BC→	Div.↑
w/o text	4.323	0.743	13.17
w WavLM	4.567	**0.707**	13.23
concatenate	4.676	5.479	11.67
addition	6.012	6.234	13.11
cross-atten	**4.088**	0.714	**13.24**

(a) Speech Feature.

Scale.	FGD↓	BC→	Div.↑
1.0	4.215	0.741	12.79
1.5	4.141	0.730	13.26
2.0	**4.088**	0.714	13.24
2.5	4.124	0.714	13.61
3.0	4.157	**0.709**	**13.75**

(b) CFG Scale.

Represent.	FGD↓	BC→	Div.↑
w/o quant	8.727	0.612	**13.56**
one quant	6.343	0.734	13.42
one residual	5.256	0.755	13.35
product quant	4.412	0.737	13.41
Ours	**4.088**	**0.714**	13.24

(c) Gesture Motion Representation.

Order.	FGD↓	BC→	Div.↑
spatial-temporal	**4.088**	**0.714**	13.24
temporal-spatial	4.113	0.721	**13.34**

(d) Sequence Order.

Distribution.	FGD ↓	BC →	Div. ↑
α=2 β=1.2	**4.088**	0.714	13.24
α=2 β=1.0	4.123	**0.704**	13.44
α=2.2 β=1.4	4.362	0.754	13.65
α=1.8 β=1.4	4.341	0.743	**13.73**

(e) Skewness of the Distribution.