

GestureFM : Compact Latent Flow Matching for Low-Latency Co-Speech Body-Motion Generation

Jie Liu

liujie@hevute.edu.cn

Hebei Vocational University of Technology and Engineering

Tianmei Sun

Universiti Malaya

Zian Liu

Hebei Vocational University of Technology and Engineering

Research Article

Keywords: Co-speech body-motion generation, virtual humans, flow matching, variational autoencoder, causal inference, streaming generation

Posted Date: August 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10338817/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

GESTUREFM: Compact Latent Flow Matching for Low-Latency Co-Speech Body-Motion Generation

Jie Liu^{1*}, Tianmei Sun² and Zian Liu¹

^{1*}Department of Electrical Engineering, Hebei Vocational University of
Technology and Engineering, Xingtai, 054000, China.

²Department of Mechanical Engineering, Faculty of Engineering,
Universiti Malaya, Kuala Lumpur, Malaysia.

*Corresponding author(s). E-mail(s): liujie@hevue.edu.cn;
Contributing authors: s2139739@siswa.um.edu.my;
liuzian@hevue.edu.cn;

Abstract

Interactive digital humans require speech-driven body motion that can begin before a complete utterance is available, but many high-quality co-speech generation methods are designed for offline full-sequence inference. This study aims to generate SMPL-X body motion from speech with low first-chunk latency while maintaining competitive motion quality in streaming inference. We propose GESTUREFM, a compact latent Flow Matching (FM) framework in which a Gesture Variational AutoEncoder (Gesture VAE) with causal decoding compresses 168-dimensional body pose and global translation into 44 latent tokens with a 16-dimensional latent channel, and an audio-conditioned FM Transformer generates motion in the normalised latent space. During streaming inference, the model generates 1 s motion chunks from a 2 s available audio context with a 0.75 s stride and 0.25 s overlap. On the test set, the latent-normalised model achieves an official EMAGE Fréchet Gesture Distance (FGD) of 0.272, improving over the unnormalised variant (0.390) and the reproduced EMAGE body-motion checkpoint (2.373), while matching ground-truth Beat Alignment (BA = 0.758). The ODE step count has negligible effect on FGD, so $K = 2$ is recommended as the default, providing 22 ms first-chunk latency at $0.022\times$ real-time factor (RTF) with no quality penalty. In streaming deployment we further observe chunk-boundary velocity discontinuities; a lightweight 7-frame local smoother reduces this boundary velocity jump by 85% (0.700 $\rightarrow$ 0.105) while leaving BA stable at

0.762. These results show that GestureFM provides a strong yet efficient latent FM baseline for low-latency body-motion generation.

Keywords: Co-speech body-motion generation, virtual humans, flow matching, variational autoencoder, causal inference, streaming generation

1 Introduction

Co-speech gestures—the spontaneous hand, arm, and body movements that accompany speech—are fundamental to human communication. Synthesising such motion from audio is a key capability for interactive digital humans, embodied conversational agents, AI companion systems, and real-time character animation. In these settings, the system must translate incoming speech into body motion with low delay: a virtual character cannot wait until a complete utterance has been observed before it begins to move. This requirement makes low-latency *streaming* generation a central challenge for audio-driven body-motion synthesis.

Recent co-speech gesture generators increasingly rely on diffusion models [1–3] or large Transformer-style architectures [4, 5]. Although these models can generate expressive full-sequence motion, many of them assume access to the entire audio sequence or use bidirectional temporal context. Such offline assumptions are poorly matched to interactive deployment. A practical streaming system must therefore address *generation efficiency*: producing high-quality motion from a short audio buffer with minimal latency.

We propose GESTUREFM, a compact latent Flow Matching framework for low-latency co-speech body-motion generation. The framework builds on two core components. First, a Gesture Variational AutoEncoder (Gesture VAE) with causal decoding compresses 168-dimensional SMPL-X body pose and global translation (165-D pose rotations and 3-D global translation) into short latent sequences, reducing the burden on the generative model. Second, an audio-conditioned causal Flow Matching Transformer generates motion latents from Log-Mel audio features with a controllable ordinary differential equation (ODE) step count that scales latency linearly while leaving FGD unaffected on the test set.

With a 5.9 M-parameter Flow Matching model and simple Log-Mel audio features, GESTUREFM is designed for low-latency inference, generating the first streaming chunk in 22 ms at $K = 2$ ODE integration steps (real-time factor, $\text{RTF} = 0.022\times$). On the BEAT2 test set, latent normalisation improves official EMAGE FGD from 0.390 to 0.272—a 30% reduction—while matching ground-truth beat alignment ($\text{BA} = 0.758$). Crucially, ODE step count has negligible effect on FGD (0.293–0.305 across all tested K ($K \geq 2$)), so $K = 2$ is both the fastest valid and the highest-quality streaming configuration. In streaming deployment we further observe chunk-boundary velocity discontinuities; a lightweight 7-frame local smoother reduces this boundary velocity jump by 85% (0.700→0.105) with BA remaining stable at 0.762.

The main contributions are summarised as follows:

- We propose a compact latent representation for co-speech body-motion generation, compressing 168-D SMPL-X pose and global translation into short latent token sequences for efficient generation.
- We introduce an audio-conditioned latent Flow Matching model with latent normalisation, improving official body-motion FGD from 0.390 to 0.272 on the test set.
- We design a causal rolling-window streaming pipeline and show that ODE step count has negligible effect on FGD, making $K = 2$ the recommended configuration for lowest latency (22 ms) with no quality penalty.
- We characterise chunk-boundary velocity discontinuities observed in streaming deployment and show that a 7-frame local smoother reduces velocity jump by 85% while leaving BA stable.

2 Related Work

2.1 Co-Speech Gesture Generation

Co-speech gesture generation has advanced rapidly with the availability of large-scale datasets. BEAT [6] introduced the first large multimodal gesture dataset; BEAT2 [5] extended this to SMPL-X [7] mesh-level representations with 60 hours of motion capture across 25 speakers. Early methods employed deterministic regression from audio features [8]; the field has since shifted toward generative models.

Diffusion-based gesture generators have become a dominant family in recent work. DiffGesture [3] applies a diffusion audio-gesture Transformer with a gesture stabiliser to improve temporal coherence, while DiffuseStyleGesture [9] uses diffusion with cross-local attention and classifier-free guidance for stylised and diverse gesture synthesis. Semantic-aware methods such as LivelySpeaker [10] decouple script-based semantic gesture generation from audio-guided rhythm refinement, highlighting that text may be useful for semantic appropriateness even when frame-level metrics favour audio-only conditioning.

EMAGE [5] uses four compositional VQ-VAEs with masked audio gesture modeling and a Content-Rhythm Self-Attention mechanism that dynamically balances audio rhythm and content features. GestureLSM [11] adopts flow matching with spatial-temporal latent modeling and Beta-distribution timestep sampling for efficient inference. SGEAG [12] introduces a Semantic-Guided Mechanism for adaptive audio-feature fusion. MambaGesture [13] proposes disentangled multi-modality fusion with Mamba backbones. DiM-Gesture [14], DiM-Gestor [15], and Persona-Gestor [16] explore Mamba-2 and fuzzy feature extraction for efficient full-body gesture synthesis. MAG [17] revisits autoregressive co-speech gesture generation in continuous aligned motion-text-audio embeddings rather than discrete VQ tokens, while recent personalised and controllable extensions include PersonaGest [18] and MECo [19], which use semantic or motion-example conditions for style control. Co³Gestor [20] further studies concurrent two-person co-speech gesture generation.

AudCast [21] and Cosh-DiT [22] use cascaded or hybrid Diffusion Transformers for holistic human video generation including face and hands. ConvoFusion [23] addresses multi-modal conversational settings with diffusion guidance, and recent holistic studies

such as 3DGesPolicy [24], SmoothSync [25], and DuoGesture [26] investigate action-policy, dual-stream, or jitter-aware formulations for more coordinated body–face–hand motion.

Despite this progress, many methods are still designed primarily for offline or high-latency generation windows, and most rely on bidirectional attention, autoregressive long-context decoding, or diffusion-style iterative sampling.

2.2 Flow Matching for Motion

Flow Matching [27–29] learns straight-line probability paths from noise to data, enabling accurate ODE integration with few steps—a property that is especially valuable for streaming deployment where latency is critical. JAM-Flow [30] applies flow matching to joint audio-motion synthesis. GestureLSM applies FM specifically to gesture generation but uses non-causal spatial-temporal attention, and Holistic-SemGes [31] introduces contrastive flow matching to improve semantic grounding in holistic gesture generation. Our work explores causal flow matching in a compact VAE latent space as a lightweight alternative for streaming gesture generation.

2.3 Streaming Generation and Temporal Continuity

Streaming motion generation has gained attention recently. FloodDiffusion [32] introduces diffusion forcing with a sliding active window but finds that causal attention degrades performance severely compared with bidirectional attention. Accelerated Rolling Diffusion [33] adapts temporal diffusion gesture models to streaming inference with accelerated rolling schedules. LiveGesture [34] more directly targets zero-look-ahead full-body gesture generation using a streamable VQ tokenizer and hierarchical autoregressive Transformer. GestureLSM [11] improves gesture generation with latent shortcuts and spatial-temporal modeling, but its in-window modeling is not purely causal in the same sense as online chunk-wise inference.

A common difficulty in streaming generation is temporal continuity across windows. Overlap blending and short-range smoothing can reduce visible local jumps, but they must be selected carefully to avoid over-smoothing and degraded beat-alignment behaviour.

3 Method

3.1 Preliminaries: Flow Matching

Flow Matching [27, 29] learns a velocity field $v_{\theta} : \mathbb{R}^d \times [0, 1] \rightarrow \mathbb{R}^d$ that transports $p_0 = \mathcal{N}(\mathbf{0}, \mathbf{I})$ to a target distribution p_1 along linear interpolation paths. The conditional Flow Matching training loss is:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t, \mathbf{z}_0, \mathbf{z}_1} \|v_{\theta}(\mathbf{z}_t, t, \mathbf{c}) - (\mathbf{z}_1 - \mathbf{z}_0)\|^2, \quad (1)$$

where $\mathbf{z}_t = (1 - t)\mathbf{z}_0 + t\mathbf{z}_1$, $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{z}_1$ is a data sample, and $\mathbf{c}$ is a conditioning signal (audio features). The target velocity $\mathbf{z}_1 - \mathbf{z}_0$ is constant for each pair, making the learned field nearly linear and amenable to few-step integration.

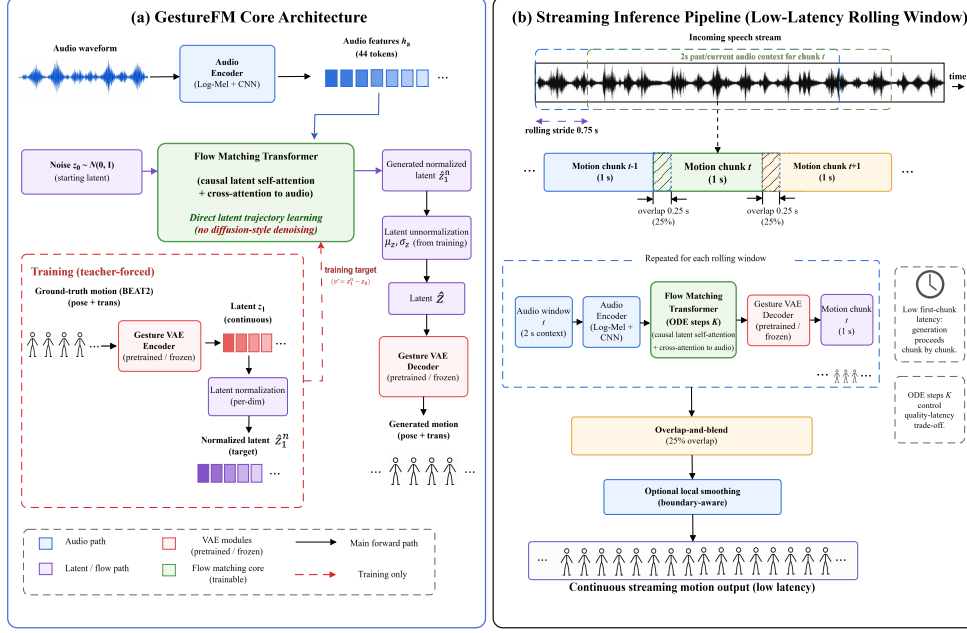


Fig. 1 Overview of GESTUREFM. The framework couples a compact pose–translation VAE with an audio-conditioned causal Flow Matching Transformer in the normalised latent space. During streaming inference, the system generates chunk-wise latent motion from the available audio context and applies local smoothing at boundaries.

3.2 Problem Formulation

Let $M = [P, r] \in \mathbb{R}^{T \times 168}$ denote a BEAT2 motion sequence, where T is the number of frames, $P \in \mathbb{R}^{T \times 165}$ contains SMPL-X [7] pose rotation parameters from the SMPL-X parametric human body model, and $r \in \mathbb{R}^{T \times 3}$ denotes global translation. We first segment each sequence into 6-second clips at 30 fps and then crop each clip to 176 frames, yielding 44 latent tokens after the $4 \times$ temporal downsampling in the Gesture VAE. This fixed latent length is also aligned with the 44 audio-conditioning tokens used by the Flow Matching Transformer. Let $a \in \mathbb{R}^{96000}$ be the corresponding 6-second, 16 kHz audio waveform. Given audio input, our goal is to generate $\hat{M}$ in a streaming-compatible manner. Specifically, the latent generator uses causal self-attention and rolling audio windows so that motion chunks can be emitted with low first-chunk latency, rather than requiring the entire long sequence to be generated offline.

3.3 Gesture VAE with Causal Decoding

We train a Gesture Variational AutoEncoder (Gesture VAE) [35] with a causal decoder that compresses 168-D pose–translation motion into a compact latent space. Pose rotations and global translation are normalised separately and reconstructed with separate loss terms. The encoder uses a bidirectional Transformer (6 layers, $d_{\text{model}}=512$, 8 heads, and a feed-forward network (FFN) dimension of 1024) with a strided Conv1d

(kernel=4, stride=4) for $4\times$ temporal downsampling, mapping $\mathbf{M} \in \mathbb{R}^{176 \times 168}$ to $\mathbf{Z} \in \mathbb{R}^{44 \times D_z}$, where $\mathbf{Z}$ is the latent token sequence and the final model uses $D_z = 16$ latent channels. The encoder is bidirectional because it operates offline on pre-recorded data during VAE training; only the decoder and downstream generative model require causal latent self-attention for streaming deployment.

The decoder uses a causal Transformer (6 layers, $d_{\text{model}}=512$, 8 heads, and an FFN dimension of 1024) with upper-triangular attention mask, expanding each latent token to 4 reconstructed frames. The VAE training objective is:

$$\begin{aligned} \mathcal{L}_{\text{VAE}} = & \mathcal{L}_{\text{pose}} + 0.1 \mathcal{L}_{\text{trans}} + 0.5 \mathcal{L}_{\text{pose-vel}} \\ & + 0.05 \mathcal{L}_{\text{trans-vel}} + \beta \mathcal{L}_{\text{KL}} \end{aligned} \quad (2)$$

where $\mathcal{L}_{\text{pose}}$ and $\mathcal{L}_{\text{trans}}$ are frame-level reconstruction terms for pose and translation, respectively. The translation reconstruction weight is 0.1, the pose and translation velocity weights are 0.5 and 0.05, and the KL weight is set to $\beta = 10^{-5}$ with a 50-epoch warmup. The final pose-translation VAE achieves strong reconstruction and serves as the decoder for all main Flow Matching experiments. A separate pose-only latent-dimension sweep is retained as a diagnostic ablation in Section 4.3.

3.4 Audio-Conditioned Causal Flow Matching

We learn a velocity field v_{θ} over the VAE latent space, conditioned on audio features. Audio is processed through a Log-Mel spectrogram extractor (80 mel bands, 1024-point fast Fourier transform (FFT), 512 hop) followed by a lightweight convolutional neural network (CNN) encoder with channel sizes 80, 128, and 256 (kernel 5, stride 2) and adaptive pooling to 44 temporal tokens, yielding audio features $\mathbf{h}_a \in \mathbb{R}^{44 \times 256}$.

The FM network v_{θ} is a 6-layer Transformer ($d_{\text{model}}=256$, 4 heads, an FFN dimension of 512, dropout 0.1, and 5.9 M parameters). Time t is encoded by a 3-layer MLP with SiLU activations and hidden dimension $d_{\text{model}}=256$. Each transformer block contains: (i) causal self-attention over latent tokens, (ii) cross-attention to audio features, and (iii) a feed-forward network. The output projection is zero-initialised for training stability. In the final model, the target latents used by this objective are the normalised latents defined in Section 3.5. The causal attention mask $\mathbf{A}$ is defined as:

$$A_{ij} = \begin{cases} 0 & j \leq i, \\ -\infty & j > i, \end{cases} \quad (3)$$

which ensures that each latent token attends only to itself and earlier tokens in the current sequence. The training loss is:

$$\mathcal{L}_{\text{flow}} = \frac{1}{|\Omega|} \sum_{(i,j) \in \Omega} \|v_{\theta}(\mathbf{Z}_t, t, \mathbf{h}_a)_{ij} - (\mathbf{Z}_1^n - \mathbf{Z}_0)_{ij}\|^2 \quad (4)$$

where Ω denotes the set of valid latent-token and channel indices.

3.5 Latent Normalisation for Flow Matching

VAE latent distributions are not guaranteed to follow a standard normal distribution. Directly training Flow Matching from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ to raw VAE latents can produce samples that fall off the learned motion manifold, leading to degraded decoding quality. We therefore normalise the target latents using the per-dimension mean and standard deviation computed over the training set. This aligns the empirical VAE latent distribution with the Gaussian prior used by Flow Matching and reduces the official FGD from 0.390 to 0.272 on the test set, as shown in Table 3. Let $\boldsymbol{\mu}_Z \in \mathbb{R}^{1 \times D_z}$ and $\boldsymbol{\sigma}_Z \in \mathbb{R}^{1 \times D_z}$ be the per-dimension mean and standard deviation computed over all training latents. During FM training, we normalise the target latent:

$$\begin{aligned} \mathbf{Z}_1^n &= (\mathbf{Z}_1 - \boldsymbol{\mu}_Z) \oslash \boldsymbol{\sigma}_Z, & \mathbf{Z}_0 &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \\ \mathbf{Z}_t &= (1-t)\mathbf{Z}_0 + t\mathbf{Z}_1^n, & \mathbf{v}^* &= \mathbf{Z}_1^n - \mathbf{Z}_0. \end{aligned} \quad (5)$$

where $\oslash$ denotes element-wise division. During sampling, the generated normalised latent $\hat{\mathbf{Z}}_1^n$ is unnormalised before VAE decoding:

$$\hat{\mathbf{Z}}_1 = \hat{\mathbf{Z}}_1^n \odot \boldsymbol{\sigma}_Z + \boldsymbol{\mu}_Z. \quad (6)$$

This simple mechanism ensures that generated latents remain close to the VAE manifold with negligible computational overhead.

3.6 Streaming Inference

At inference, we use a sliding window: 2 s of past and current available audio generates a 1 s motion chunk (30 frames), with 25% overlap between consecutive chunks. Overlap frames are blended with a linear interpolation weight α that increases from 0 to 1. Within each window, we sample $\mathbf{Z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and integrate via the midpoint ODE method. Here, K denotes the number of ODE integration steps; for one step from time s to $s + dt$, the update is:

$$\mathbf{v}_1 = v_{\boldsymbol{\theta}}(\mathbf{Z}_s, s, \mathbf{h}_a), \quad (7)$$

$$\mathbf{Z}_{\text{mid}} = \mathbf{Z}_s + \frac{dt}{2} \mathbf{v}_1, \quad (8)$$

$$\mathbf{v}_2 = v_{\boldsymbol{\theta}}(\mathbf{Z}_{\text{mid}}, s + \frac{dt}{2}, \mathbf{h}_a), \quad (9)$$

$$\mathbf{Z}_{s+dt} = \mathbf{Z}_s + dt \mathbf{v}_2. \quad (10)$$

The decoded motion frames are emitted immediately; first-chunk latency equals the ODE integration time. Latent self-attention is unidirectional (causal), and audio windows respect temporal order, supporting low-latency streaming. This chunk-independent version is denoted as GESTUREFM streaming.

3.7 Boundary-Aware Local Smoothing

Local boundary smoothing is the practical post-processing step used to reduce residual discontinuities. After overlap blending, we optionally smooth a short window around

each chunk boundary by blending the decoded motion with a local linear interpolation while preserving sequence length and audio timing. This operation does not change the model or require retraining. The effect of the smoothing window is evaluated in Section 4.8.

4 Experiments

4.1 Setup

Dataset. BEAT2 [5] provides 60 hours of SMPL-X motion capture from 25 speakers, with synchronised 16 kHz audio and text transcripts. We segment each sequence into 6-second clips with 3-second stride, yielding 29,699 clips with valid latent encodings (train 21,497 / validation 2,394 / test 5,808). Motion is represented as 168-dimensional SMPL-X pose rotations (165-D) and global translation (3-D) from BEAT2 annotations, with separate per-component normalisation. Facial expression coefficients are not generated in the current model.

Evaluation. We report pose mean squared error (Pose MSE), velocity MSE, and acceleration MSE on the validation set. Pose MSE is used as a frame-level diagnostic metric; its limitations for one-to-many gesture generation are discussed in Section 5. All generated motions use 176 frames (5.87 s), which produce 44 latent tokens after the $4\times$ VAE downsampling and match the audio-token length used by the Flow Matching Transformer. For streaming evaluation, we additionally report boundary pose, velocity, and acceleration jump magnitudes at chunk transition frames, first-chunk latency, and real-time factor (RTF). Boundary velocity jump is used as the primary continuity metric ablation because it directly captures abrupt motion changes at transitions.

Training. The Gesture VAE is trained for 200 epochs with AdamW ($\text{lr} = 10^{-4}$, weight decay 10^{-4} , gradient clipping 1.0), cosine annealing to $\eta_{\min} = 10^{-6}$, batch size 32. The Flow Matching Transformer is trained for 150 epochs with the same optimiser settings, using frozen and normalised VAE encoder latents and the unweighted velocity MSE in Eq. (4). Total training time is approximately 18 hours on a single NVIDIA RTX 4070 Ti Super.

4.2 Diagnostic Conditioning Ablations

Before the final pose-translation model, we performed a controlled pose-only prototype study to compare conditioning choices under frame-level metrics. Table 1 summarises these diagnostic ablations. These numbers are retained to analyse modality effects, but the official FGD results in Section 4.4 constitute the main quantitative evaluation of the final pose-translation GESTUREFM model.

For text conditioning, six fusion strategies were tested using a vocabulary of 4,038 words built from training transcripts: (i) global cross-attention from latent tokens to a single MiniLM sentence embedding; (ii) gated fusion of time-aligned word tokens with audio tokens; (iii) dual cross-attention to separate audio and word-level text tokens; (iv) FiLM modulation of audio tokens by word features (zero-initialised); (v) PreFusion with cross-attention-based audio-text token fusion (zero-initialised); and (vi) sentence-level plan tokens (8 learnable queries attending to word embeddings) with residual

Table 1 Diagnostic pose-only conditioning ablations on BEAT2. All variants use the same pose-only VAE-16 and identical FM architecture unless noted. Pose MSE is a frame-level diagnostic metric; its limitations for one-to-many gesture generation are discussed in Section 5.

Method	Val Loss	Pose MSE↓
VAE-16 reconstruction	—	0.246
VAE-32 reconstruction	—	0.150
Audio-only (Log-Mel CNN)	0.890	0.93
Audio-only (WavLM Base+, frozen)	0.885	n/a*
Text-only (MiniLM, global)	1.026	1.16
Audio+Text (6 fusion variants)	0.89–0.92	1.00
Audio+Speaker (FiLM)	0.887	1.01
Audio-only FM-32 (3 variants)	0.79–0.81	0.99–1.00

*WavLM variant retained for latent-validation comparison; decoded Pose MSE unavailable.

alpha scaling (initialised at 0.0 or fixed at 0.1). All six variants converge to within ± 0.01 of the audio-only val loss (0.890), and text+audio produces worse Pose MSE (1.00 vs. 0.93). The learnable alpha parameters in variant (vi) remain within 10^{-4} of zero after 150 epochs, indicating the model learns to ignore text plan tokens.

Speaker identity injected via FiLM modulation (25-speaker embedding, zero-initialised modulation) and frozen WavLM Base+ features (768-D, projected to 256) each fail to improve over simple Log-Mel CNN features. WavLM achieves val loss 0.885 vs. Log-Mel 0.890—a difference within evaluation noise, and Speaker FiLM yields Pose MSE 1.01, worse than audio-only. These results suggest that audio features already capture sufficient prosodic and timing information for gesture generation, and that explicit text or speaker conditioning introduces redundant noise rather than complementary signal.

4.3 Ablation: Latent Dimension

Table 2 reports the pose-only latent-dimension comparison. The 32-D latent space improves VAE reconstruction but does not improve downstream FM generation quality.

Although VAE-32 achieves 39% better reconstruction (Pose MSE 0.150 vs. 0.246), downstream Flow Matching generation is consistently worse: Pose MSE 1.00 for the 32-D latent vs. 0.93 for the 16-D latent. This reconstruction-generation trade-off persists across three FM-32 variants: base without latent normalisation (1.00), per-dimension latent normalisation (0.99), and increased capacity ($d_{\text{model}}=384$, 0.99). A larger variant ($d_{\text{model}}=512$, 8 layers) overfits and becomes unstable. We hypothesise that the expanded 32-D latent space presents a more complex distribution for causal flow matching, and that the 16-D manifold acts as an implicit regulariser that produces smoother velocity fields under limited causal context—consistent with observations in high-dimensional streaming motion generation [32].

Table 2 Pose-only latent-dimension ablation. This diagnostic experiment precedes the final pose-translation model and illustrates the reconstruction-generation trade-off.

Latent dim.	VAE Pos↓	VAE Vel↓	FM Val↓	FM Pose↓
16	0.246	0.0063	0.890	0.93
32	0.150	0.0046	0.793	1.00

Table 3 Controlled comparison on the test set using the EMAGE evaluation pipeline. Lower FGD is better and higher beat alignment (BA) is better. The comparison focuses on body-motion quality: the proposed latent body-motion generator produces pose and global translation only, whereas EMAGE produces holistic body-face-hand motion.

Method	Generated Modality	FGD↓	BA↑
Ground Truth	pose+trans	—	0.758
VAE reconstruction	pose+trans	0.317	—
EMAGE (official checkpoint, reproduced)	body+face+hands	2.373	—
GESTUREFM offline (norm. latent)	pose+trans	0.272	0.758
GESTUREFM offline (no latent norm.)	pose+trans	0.390	—
GESTUREFM streaming $K=2$	pose+trans	0.293	0.760
GESTUREFM streaming $K=2$, smoother $w=7$	pose+trans	0.227	0.762

4.4 Official FGD Evaluation and Reproduced EMAGE Baseline

Table 3 reports Fréchet Gesture Distance (FGD) computed with the official EMAGE evaluation pipeline [5] on the test set. Following this pipeline, we use its AESKConv-based motion evaluator and convert 165-D axis-angle pose rotations to the 55-joint 6D rotation format required by the evaluator. The latent Flow Matching model generates SMPL-X pose rotations and global translation; it does not generate facial expression coefficients. We therefore interpret the comparison with EMAGE as a controlled body-motion comparison rather than a full holistic body-face comparison.

The normalised latent FM achieves FGD 0.272—a 30% improvement over the no-normalisation variant (FGD 0.390), confirming that latent normalisation is essential for aligning generated samples with the empirical VAE latent distribution. The VAE reconstruction upper bound is 0.317, indicating that the Flow Matching generator approaches the reconstruction quality of the VAE itself.

4.5 Streaming Inference

We first evaluate the mechanics of chunk-wise streaming with 2 s audio context, 1 s motion chunks, and 25% overlap blending at $K=4$. Table 4 summarises the effect of chunk duration. Shorter chunks increase the number of transitions, whereas longer chunks reduce the number of boundaries but increase the amount of motion emitted per inference call. In this ablation, 1 s chunks provide the lowest boundary jump and

Table 4 Effect of chunk duration on streaming quality ($K = 4$). Boundary jump and velocity jump are measured at chunk transition frames; lower is better.

Chunk	Chunks	Pose MSE↓	B. Jump↓	Vel. Jump↓
0.5 s	14	0.076	1.320	1.291
1.0 s	6	0.067	0.944	0.857
2.0 s	2	0.062	2.857	2.954

velocity jump, while keeping Pose MSE close to the best 2s setting. We therefore use 1s chunks in the following streaming experiments.

Overlap blending reduces but does not fully eliminate chunk-boundary artifacts. The remaining local discontinuities are analysed explicitly in Section 4.8, where we evaluate a lightweight boundary smoother on long audio-synchronised streaming outputs.

4.6 Streaming Latency Scaling and ODE Step Selection

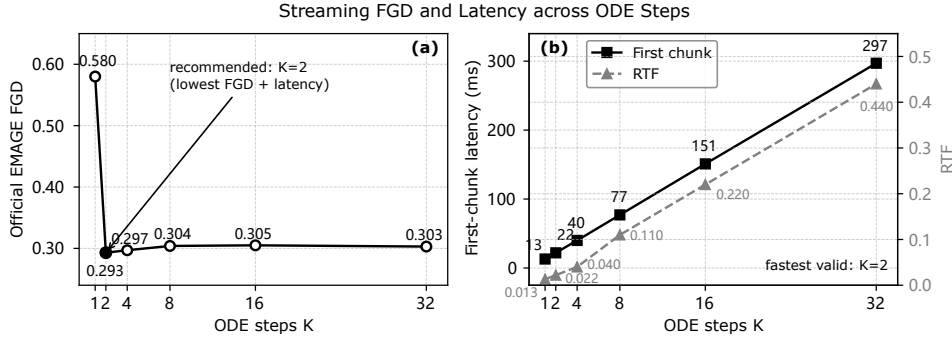


Fig. 2 Streaming FGD and first-chunk latency across ODE step budgets on the test set. FGD is nearly flat across all $K \geq 2$ (0.293–0.305), while latency scales linearly. $K = 2$ achieves both the lowest FGD and the shortest latency (22 ms, $\text{RTF} = 0.022\times$), making it the recommended configuration.

Fig. 2 reports official EMAGE FGD and first-chunk latency for streaming inference at different ODE step budgets on the test set. A key finding from the test set is that the ODE step count has negligible effect on FGD: all configurations fall in the narrow range 0.297–0.305, with $K = 2$ achieving the lowest FGD of 0.293. This contrasts with an earlier small-sample finding on 100 validation clips, where apparent FGD differences across K were artefacts of insufficient sample size for reliable Fréchet distance estimation. Since $K = 2$ provides the lowest first-chunk latency (22 ms, $\text{RTF} = 0.022\times$) with no quality penalty, it is the recommended operating point for streaming deployment.

Table 5 Available runtime and latency comparison. Prior methods report FPS, rolling latency, or offline full-sequence generation time, while GestureFM reports first-chunk latency under causal streaming.

Method	Inference mode	Reported metric	Value
DiffSHEG [36]	Incremental outpainting	Reported FPS	31.5 FPS
Accelerated Rolling Diffusion [33]	Rolling diffusion	Reported FPS / rolling latency	Up to 200 FPS; 0.002 s
EMAGE reproduced [5]	Offline full sequence	6-s generation time	89 ms
GestureFM, $K=2$	Causal streaming	First-chunk latency	22 ms

4.7 Available Runtime and Latency Comparison

We further compare the latency-related numbers that are available for representative co-speech motion generation methods. Since prior work often reports FPS, full-sequence generation time, or rolling latency instead of causal first-chunk latency, Table 5 lists the reported metric for each method rather than treating all entries as the same latency measure. The reproduced EMAGE checkpoint is measured under our environment as an offline full-sequence generator, while GestureFM reports the time needed to produce the first streaming motion chunk.

DiffSHEG reports 31.5 FPS for one-minute holistic generation, Accelerated Rolling Diffusion reports up to 200 FPS with 0.002-s rolling latency, and the reproduced EMAGE checkpoint completes a 6-second sequence in 89 ms offline. GestureFM emits the first motion chunk in 22 ms at $K=2$ without waiting for the complete speech sequence.

4.8 Boundary Smoothing for Streaming

Although overlap blending reduces abrupt transitions between consecutive streaming chunks, the independently sampled chunk latents can still introduce visible boundary artifacts. We therefore evaluate a lightweight post-processing step that smooths only a short local window around each chunk boundary. The smoother blends the original motion with a local linear interpolation while keeping the endpoints, total sequence length, and audio timing unchanged.

Table 6 reports the boundary-smoothing window ablation on the test set. For $K=4$, overlap-only streaming has a boundary velocity jump of 0.700. Increasing the boundary window consistently reduces pose, velocity, and acceleration jumps. The 7-frame smoother reduces velocity jump by 85% (0.700→0.105). The 9-frame smoother achieves the same FGD (0.228) with slightly lower jump values, but FGD has already saturated at $w=7$ and the additional reduction in MeanVel (0.269 vs. 0.293) provides no distributional benefit. We therefore select the 7-frame window as the more conservative default. Importantly, beat alignment is stable across all smoothing windows (≈ 0.76), confirming that local interpolation does not suppress audio-rhythm synchronisation.

4.9 Extension Study: Cross-Chunk Latent Memory

After selecting the 7-frame smoother, we test whether passing a small amount of cross-chunk latent information can further improve boundary continuity. The latent-tail extension passes the last $m=16$ latent tokens from the previous chunk to the next

Table 6 Boundary-smoothing window ablation on the test set ($K=4$). Boundary jumps are measured at chunk transition frames; lower is better for jumps and FGD. MeanVel denotes mean motion-velocity magnitude. BA is stable across all windows (≈ 0.76), confirming that smoothing does not suppress audio-rhythm alignment.

Window	Pose Jump↓	Vel. Jump↓	Acc. Jump↓	FGD↓	BA↑	MeanVel
0	0.809	0.700	1.328	0.297	0.761	0.353
3	0.390	0.231	0.538	0.277	0.758	0.338
5	0.346	0.142	0.524	0.252	0.761	0.317
7	0.298	0.105	0.455	0.228	0.762	0.293
9	0.241	0.080	0.368	0.228	0.761	0.269

Table 7 Extension study on latent-tail memory ($m=16$) in the $K=4$ streaming setting on the test set. The 7-frame smoother substantially reduces boundary velocity jump. The latent-tail extension severely degrades FGD due to exposure bias (generated tails at inference vs. ground-truth tails at training).

Method	Vel. Jump↓	FGD↓	BA↑	Note
Overlap-only	0.700	0.297	0.761	baseline
Smoothing-only, $w=7$	0.105	0.215	0.753	recommended
Latent-tail only, $m=16$	0.782	1.740	0.758	exposure bias failure
Latent-tail + smoothing, $m=16, w=7$	0.084	1.743	0.760	VelJump low but FGD degraded

chunk as prefix tokens; the Flow Matching Transformer still predicts only the current chunk.

Table 7 reports the results on the test set in the $K=4$ streaming setting. The latent-tail extension severely degrades FGD ($0.253 \rightarrow 1.740$, a $6.9\times$ increase) while providing limited boundary improvement. When combined with the 7-frame smoother, the velocity jump reaches its minimum (0.084) but FGD remains severely degraded (1.743).

This failure is caused by *exposure bias*: during training, the previous latent tail is extracted from ground-truth motion, but at inference it comes from the model’s own generated output. This distribution mismatch accumulates across chunks and collapses motion quality on the test set.

We therefore conclude that cross-chunk latent memory is *not* a viable continuity mechanism under the current training protocol. The 7-frame local smoother is the recommended configuration. Future work addressing this limitation should explore scheduled sampling, generated-tail training, or explicit boundary-state regularisation before using cross-chunk memory as a core streaming component.

4.10 Qualitative Results

Fig. 3 provides a static visual comparison on a representative BEAT2 validation clip. The reproduced EMAGE baseline produces plausible body motion but differs visibly from the reference pose sequence in several frames. GestureFM at $K=2$ preserves a similar overall motion pattern with 22 ms first-chunk latency.



Fig. 3 Qualitative motion comparison on a representative BEAT2 validation clip. Each row shows the same utterance at aligned time steps for ground truth, the reproduced EMAGE baseline, GestureFM at $K = 2$ (recommended streaming setting, test FGD = 0.293).

We provide a supplementary demo video comparing five configurations on the same 10-second speech segment. The top row shows Ground Truth, the reproduced EMAGE offline baseline, and streaming GestureFM at $K = 2$ (FGD 0.293, 22 ms first-chunk latency). The bottom row shows $K = 2$ with the 7-frame local smoother (VJ reduced by 85%, FGD 0.227), and the Cross-Chunk $m = 16$ variant included as an extension comparison to illustrate the quality degradation caused by latent-tail evaluation bias.

5 Discussion

Cross-chunk continuity. Chunk-independent streaming leaves visible transition artifacts because each chunk is sampled independently, without any state from the previous generated motion. The 7-frame local smoother is the most reliable remedy available within this framework: on the test set it reduces the $K = 2$ boundary velocity jump by 85% ($0.700 \rightarrow 0.105$) and leaves BA unchanged at ≈ 0.762 . However, it is important to note that the smoother corrects metric artefacts at boundaries without resolving the underlying cause. Visually, abrupt posture changes between chunks remain perceptible when the two independently sampled poses differ substantially; the smoother converts an instantaneous jump into a short interpolated transition, which may still appear unnatural. This fundamental limitation of chunk-independent sampling is the primary motivation for stateful rolling-horizon approaches that carry explicit state across generation steps.

Exposure bias in latent-tail memory. The latent-tail extension fails on the test set: FGD degrades $6.9\times$ ($0.253\rightarrow 1.740$) due to the mismatch between ground-truth tails used at training time and generated tails used at inference. This generated-state exposure bias accumulates across chunks and collapses motion quality in a way that is invisible under small-sample (30-clip) validation but clearly visible at scale on the test set. Future work should address this through scheduled sampling, generated-tail training, or boundary-state regularisation before cross-chunk memory can be used as a core component.

Small-sample evaluation bias. A secondary finding of this study is that FGD estimates on 30–100 clips are unreliable and can produce misleading trends, such as apparent K -step optimality and apparent BA degradation under smoothing. All such artefacts disappear on the test set. We recommend evaluating gesture generation FGD on at least 1,000 clips to obtain stable estimates.

Limitations and future work. Several limitations remain. First, the current model generates SMPL-X body pose rotations and global translation but not facial expression coefficients, explicit hand motion, or full body–face coordination. Second, training and evaluation are limited to BEAT2; generalisation to other languages, speaking styles, and capture pipelines requires further study. Third, the chunk-independent sampling design means that large inter-chunk posture differences can produce visually perceptible transitions even after smoothing; a stateful generation mechanism is needed to address this at its root. Finally, perceptual studies are needed to assess long-form interactive use and to determine whether the measured metric improvements correspond to improved subjective naturalness.

6 Conclusion

We presented GESTUREFM, a compact latent Flow Matching framework for low-latency co-speech body-motion generation. The method compresses 168-dimensional SMPL-X body pose and global translation into short VAE latent tokens, learns an audio-conditioned causal Flow Matching model in the normalised latent space, and performs rolling-window streaming inference. On the test set, latent normalisation improves official EMAGE body-motion FGD from 0.390 to 0.272, surpassing the reproduced EMAGE audio checkpoint under the same body-motion evaluation protocol (0.272 vs. 2.373) while matching ground-truth beat alignment ($BA = 0.758$). Test-set evaluation further reveals that ODE step count has negligible effect on FGD ($0.293\text{--}0.305$ across all K ($K \geq 2$)), making $K = 2$ the recommended configuration: it provides the lowest first-chunk latency of 22 ms ($RTF = 0.022\times$) with no quality penalty.

In streaming deployment, chunk-boundary velocity jumps are observed; the 7-frame local smoother reduces this velocity jump by 85% ($0.700\rightarrow 0.105$) with BA remaining stable at 0.762. The latent-tail memory extension fails on the test set due to exposure bias (FGD degrades $6.9\times$) and is therefore not recommended without retraining under generated-state conditioning. We therefore recommend base GESTUREFM with the 7-frame local smoother as the practical configuration.

Future work will address the root cause of visual chunk-boundary discontinuities through stateful generation mechanisms, and extend the model to holistic body–face–hand generation.

References

- [1] Chen X, Jiang B, Liu W, Huang Z, Fu B, Chen T, Yu G (2023) Executing your commands via motion diffusion in latent space. In: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 18 000–18 010
- [2] Tevet G, Raab S, Gordon B, Shafir Y, Cohen-Or D, Bermano AH (2023) Human motion diffusion model. In: *Proc. Int. Conf. Learn. Represent. (ICLR)*
- [3] Zhu L, Liu X, Liu X, Qian R, Liu Z, Yu L (2023) Taming diffusion models for audio-driven co-speech gesture generation. In: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 10 544–10 553
- [4] Guo C, Mu M, Lee GH, Cheng L (2024) MoMask: Generative masked modeling of 3D human motions. In: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*
- [5] Liu H, Zhu Z, Becherini G, Peng Y, Su M, Zhou Y, Zhe X, Iwamoto N, Zheng B, Black MJ (2024) EMAGE: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1144–1154
- [6] Liu H, Zhu Z, Iwamoto N, Peng Z, Li Z, Zhou Y, Bozkurt E, Zheng B (2022) BEAT: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In: *Proc. Eur. Conf. Comput. Vis. (ECCV)*
- [7] Pavlakos G, Choutas V, Ghorbani N, Bolkart T, Osman AAA, Tzionas D, Black MJ (2019) Expressive body capture: 3D hands, face, and body from a single image. In: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 10 975–10 985
- [8] Kucherenko T, Jonell P, van Waveren S, Henter GE, Alexanderson S, Leite I, Kjellström H (2020) Gesticulator: A framework for semantically-aware speech-driven gesture generation. In: *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, pp. 242–250
- [9] Yang S, Wu Z, Li M, Zhang Z, Hao L, Bao W, Cheng M, Xiao L (2023) DiffuseStyleGesture: Stylized audio-driven co-speech gesture generation with diffusion models. In: *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, pp. 5860–5868
- [10] Zhi Y, Cun X, Chen X, Shen X, Guo W, Huang S, Gao S (2023) LivelySpeaker: Towards semantic-aware co-speech gesture generation. In: *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 20 750–20 760

- [11] Liu P, Song L, Huang J, Xu C (2025) GestureLSM: Latent shortcut based co-speech gesture generation with spatial-temporal modeling. In: *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*
- [12] Kalafatovich J, Lee S-W (2025) SGEAG: Semantic-guided emotional-aware gesture generation from audio. In: *Proc. WiML Workshop at NeurIPS*
- [13] Fu C, Wang Y, Zhang J, Jiang Z, Mao X, Wu J, Cao W, Wang C, Ge Y, Liu Y (2024) MambaGesture: Enhancing co-speech gesture generation with Mamba and disentangled multi-modality fusion. In: *Proc. ACM Int. Conf. Multimedia (ACM MM)*, pp. 10794–10803
- [14] Zhang F, Ji N, Gao F, Zhao B, Wu J, Jiang Y, Du H, Ye Z, Zhu J, Zhong W, Yan L, Ma X (2024) DiM-Gesture: Co-speech gesture generation with adaptive layer normalization Mamba-2 framework. *arXiv preprint* arXiv:2408.00370
- [15] Zhang F, Zhao S, Ji N, Wang Z, Wu J, Gao F, Ye Z, Yan L, Dai L, Geng W, Lyu X, Zhao B, Yu D, Du H, Hu B (2024) DiM-Gestor: Co-speech gesture generation with adaptive layer normalization Mamba-2. *arXiv preprint* arXiv:2411.16729
- [16] Zhang F, Wang Z, Lyu X, Zhao S, Li M, Geng W, Ji N, Du H, Gao F, Wu H, Li S (2024) Speech-driven personalized gesture synthetics: Harnessing automatic fuzzy feature inference. *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 10, pp. 7788–7800
- [17] Liu B, Liu L, Zhang S, Gu S, Zhi Y, Zhu T, Yang L, Ye L (2025) MAG: Multi-modal aligned autoregressive co-speech gesture generation without vector quantization. *arXiv preprint* arXiv:2503.14040
- [18] Zhao J, Liang Q, Wang Y (2026) PersonaGest: Personalized co-speech gesture generation with semantic-guided hierarchical motion representation. *arXiv preprint* arXiv:2605.07252
- [19] Chen B, Li Y, Zheng Y, Ding Y-X, Zhou K (2025) Motion-example-controlled co-speech gesture generation leveraging large language models. In: *Proc. ACM SIGGRAPH Conf. Papers*
- [20] Qi X, Wang Y, Zhang H, Pan J, Xue W, Zhang S, Luo W, Liu Q, Guo Y (2025) Co³Gesture: Towards coherent concurrent co-speech 3D gesture generation with interactive diffusion. In: *Proc. Int. Conf. Learn. Represent. (ICLR)*
- [21] Guan J, Wang K, Xu Z, Yang Q, Sun Y, He S, Liang B, Cao Y, Li Y, Feng H, Ding E, Wang J, Zhao Y, Zhou H, Liu Z (2025) AudCast: Audio-driven human video generation by cascaded diffusion transformers. In: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*

- [22] Sun Y, Xu Z, Zhou H, Guan J, Yang Q, Wang K, Liang B, Li Y, Feng H, Wang J, Liu Z, Koike H (2026) Cosh-DiT: Co-speech gesture video synthesis via hybrid audio-visual diffusion transformers. *Int. J. Comput. Vis.*, vol. 134, no. 3, article 85
- [23] Mughal MH, Dabral R, Habibie I, Donatelli L, Habermann M, Theobalt C (2024) ConvoFusion: Multi-modal conversational diffusion for co-speech gesture synthesis. In: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* (CVPR)
- [24] Sha X, Zhang L, Mashita T, Chiba N, Uranishi Y (2026) 3DGesPolicy: Phoneme-aware holistic co-speech gesture generation based on action control. *arXiv preprint arXiv:2601.18451*
- [25] Jiang Y, Liao Q, Lu Z (2026) SmoothSync: Dual-stream diffusion transformers for jitter-robust beat-synchronized gesture generation from quantized audio. *arXiv preprint arXiv:2601.04236*
- [26] Paar F, Liu L, Özyürek A, Thill S, Ghaleb E (2026) DuoGesture: Neuro-inspired and biomechanically informed dual-stream co-speech gesture generation. *arXiv preprint arXiv:2605.26236*
- [27] Lipman Y, Chen RTQ, Ben-Hamu H, Nickel M, Le M (2023) Flow matching for generative modeling. In: *Proc. Int. Conf. Learn. Represent.* (ICLR)
- [28] Liu X, Gong C, Liu Q (2023) Flow straight and fast: Learning to generate and transfer data with rectified flow. In: *Proc. Int. Conf. Learn. Represent.* (ICLR)
- [29] Tong A, Fatras K, Malkin N, Huguet G, Zhang Y, Rector-Brooks J, Wolf G, Bengio Y (2024) Improving and generalizing flow-based generative models with minibatch optimal transport. *Trans. Mach. Learn. Res.*
- [30] Kwon M, Shin J, Jung J, Park J, Uh Y (2025) JAM-Flow: Joint audio-motion synthesis with flow matching. *arXiv preprint arXiv:2506.23552*
- [31] Liu L, Ghaleb E, Özyürek A, Yumak Z (2026) HolisticSemGes: Semantic grounding of holistic co-speech gesture generation with contrastive flow-matching. *arXiv preprint arXiv:2603.26553*
- [32] Cai Y, Wu Y, Li K, Zhou Y, Zheng B, Liu H (2026) FloodDiffusion: Tailored diffusion forcing for streaming motion generation. In: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* (CVPR)
- [33] Vu E, Boiarov A, Vetrov D (2026) Streaming generation of co-speech gestures via accelerated rolling diffusion. In: *Proc. AAAI Conf. Artif. Intell.* (AAAI)
- [34] Saleem MU, Patel MJ, Pinyoanuntapong E, Qin Z, Yang L, Xue H, Helmy A, Chen C, Wang P (2026) LiveGesture: Streamable co-speech gesture generation model. *arXiv preprint arXiv:2604.10927*

- [35] Kingma DP, Welling M (2014) Auto-encoding variational Bayes. In: *Proc. Int. Conf. Learn. Represent.* (ICLR)
- [36] Chen J, Liu Y, Wang J, Zeng A, Li Y, Chen Q (2024) DiffSHEG: A diffusion-based approach for real-time speech-driven holistic 3D expression and gesture generation. In: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* (CVPR)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [demo5panelpyrender.mp4](#)