FLOWTALKER: CAUSAL STREAMING AUDIO-DRIVEN 3D HEAD ANIMATION VIA FACTORIZED MOTION MODELING WITH RECTIFIED FLOW

Anonymous authors
Paper under double-blind review

ABSTRACT

Causally generating expressive 3D head animation as audio streams in is essential for real-time digital human interaction. However, most existing methods rely heavily on future audio context, leading to noticeable performance degradation when only limited or no look-ahead is available in real-time streaming inference. Current streaming-oriented approaches achieve low latency, yet their animation quality remains suboptimal, particularly in terms of expressive facial motion and long-term temporal consistency. We present **FlowTalker**, a framework for audio-driven 3D head animation under causal streaming inference. Pose, expression, and mouth motion evolve at different temporal scales and respond to distinct audio cues, but prior methods typically model them in a shared motion space. FlowTalker preserves these distinct motion dynamics by encoding them into separate token streams using spatially factorized vector quantization. We further introduce a Draft-then-Refine paradigm, in which a kinematics-aware autoregressive model predicts factorized motion tokens, and a selective rectified-flow module subsequently refines the expression and mouth latents to enhance expressive facial dynamics. Extensive experiments show that FlowTalker achieves state-of-the-art lip synchronization and facial motion quality under causal streaming inference, while showing the smallest performance degradation when transitioning from non-causal to causal settings. FlowTalker further enables high-fidelity photorealistic animation of 4D head avatars reconstructed from a single image.

1 INTRODUCTION

Audio-driven 3D head animation synthesis (Daněček et al., 2022; Thambiraja et al., 2023; Peng et al., 2023; Du et al., 2023; Zhuang et al., 2024; Lin et al., 2024; Nocentini et al., 2024; Kim et al., 2025; Wang et al., 2026) has become a key component of interactive digital human systems. The resulting 3D motion can directly drive photorealistic head avatars represented with explicit 3D geometry (Qian et al., 2024; Aneja et al., 2024; Taubner et al., 2025; He & Hoi, 2026; Wu et al., 2026; Xu et al., 2026). However, existing methods (Ma et al., 2024a; Sun et al., 2024; Chae-Yeon et al., 2025) are primarily designed for offline generation with full-utterance audio. Live interaction instead requires motion to be generated incrementally as speech arrives, without access to future audio. Such causal streaming capability is essential for telepresence and embodied conversational agents, where delayed responses harm presence and natural turn-taking. This responsiveness depends not only on inference speed, but also on how much audio context must be observed before generation can begin.

Reducing future audio context creates a fundamental latency-quality trade-off: shorter audio buffers enable earlier responses but discard future phonetic and prosodic cues needed for accurate facial motion prediction. Existing streaming-oriented methods (Han et al., 2025; Lee et al., 2025; Yang et al., 2025) reduce response latency, yet still struggle to preserve expressive facial dynamics and coherent head motion under causal streaming inference. To address this challenge, we introduce **FlowTalker**, a draft-then-refine framework for causal streaming audio-driven 3D head animation, as illustrated in Fig. 1. Fig. 2 contrasts non-causal inference protocols, including full-utterance and block-wise inference, with causal sliding-window inference, and quantifies the resulting latency-quality trade-off. FlowTalker achieves the best facial motion quality while showing the smallest performance degradation from non-causal to causal streaming inference.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

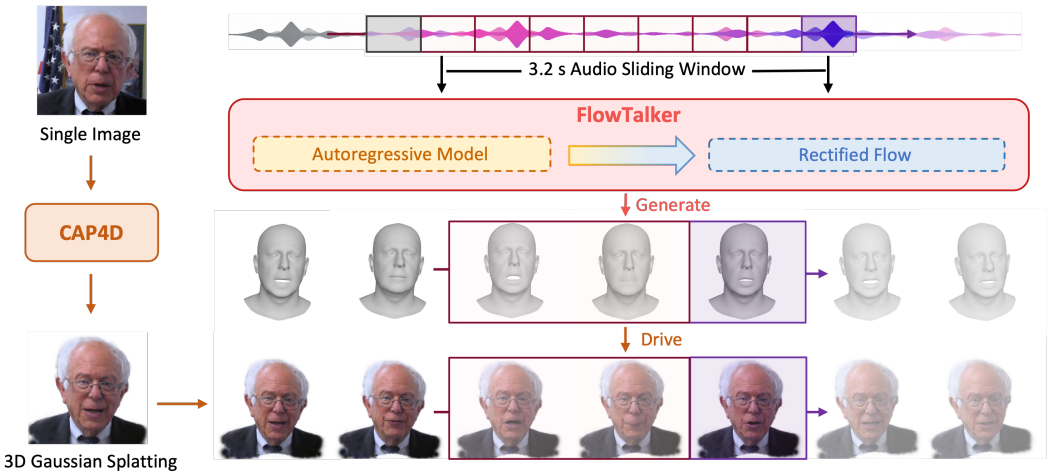


Figure 1: **Causal streaming audio-driven 3D head animation.** FlowTalker incrementally generates expressive full-head motion from audio and drives a photorealistic 3D avatar reconstructed from a single image (Taubner et al., 2025).

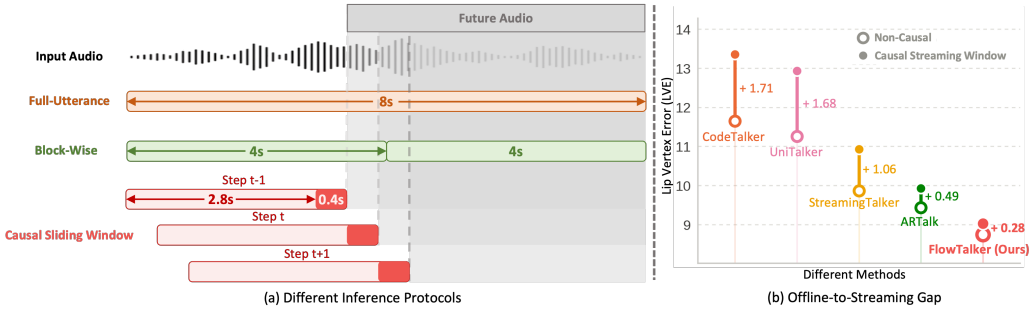


Figure 2: **Latency-quality trade-off.** (a) After initialization, causal sliding-window inference updates motion every 0.4 s using only a 3.2 s window of observed audio, whereas full-utterance and block-wise protocols require much more future audio. (b) LVE under non-causal and causal settings; the gap indicates quality degradation caused by causal conversion (LVE: lower is better).

Concretely, FlowTalker first learns a factorized motion representation motivated by the distinct temporal dynamics and audio dependencies of head pose, expression, and mouth motion. Pose follows broad rhythmic context, expression reflects affective and prosodic cues, and mouth motion tracks phonetic content. These distinct motion characteristics become particularly important under limited audio context. Although FLAME (Li et al., 2017) distinguishes global pose, expression, and jaw parameters, its expression coefficients span multiple facial regions, while existing generators often encode motion in a shared latent space (Xing et al., 2023; Sun et al., 2024; Chu et al., 2025). We therefore introduce a spatially factorized Vector Quantized Variational Autoencoder (Van Den Oord et al., 2017) with dedicated codebooks for pose, expression, and mouth motion. Each codebook captures the distinct motion patterns of pose, expression, and mouth motion, encoding each component as a separate token stream.

Building on this factorized motion representation, we propose a Draft-then-Refine pipeline. In the Draft stage, a kinematics-aware autoregressive transformer predicts the factorized motion tokens from the observed audio window and aligned motion history. To compose these motion patterns into coordinated full-head motion, masked self-attention organizes cross-stream conditioning from global motion to local articulation: pose provides context for expression, while both guide mouth prediction. Discrete drafting can, however, attenuate fine-grained facial variations. In the Refine stage, a selective rectified-flow module (Liu et al., 2022) therefore refines the expression and mouth embeddings in continuous space to recover expressive lip dynamics and micro-expressions, while keeping the drafted pose fixed to preserve the global motion trajectory. The refiner requires only a few integration steps for efficient streaming inference. Our main contributions can be summarized:

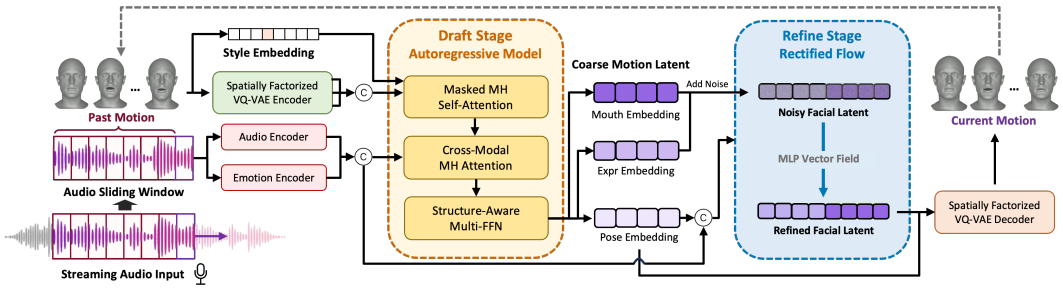


Figure 3: **Overview of FlowTalker.** Given a sliding window of audio and past motion, the autoregressive drafter predicts discrete pose, expression, and mouth tokens for the newly arrived audio, which are mapped to their codebook embeddings. A rectified-flow refiner then enhances the expression and mouth embeddings in continuous latent space, conditioned on audio, emotion, and pose. The VQ-VAE decoder maps the pose embedding and refined facial latents to FLAME parameters.

- We present **FlowTalker**, a framework for causal streaming audio-driven 3D head animation synthesis that generates highly expressive full-head motion with low response latency.
- We learn factorized motion representations for pose, expression, and mouth motion, and introduce a Draft-then-Refine framework for causal streaming generation. The autoregressive drafter coordinates these components through conditioning from global motion to local articulation, while selective continuous refinement restores fine-grained facial dynamics.
- Extensive experiments show that FlowTalker achieves state-of-the-art lip synchronization and facial motion quality with strong temporal coherence under causal streaming, and enables high-fidelity 3D photorealistic head-avatar animations.

2 Related Work

Audio-Driven 3D Facial Animation. Generating realistic 3D facial animation from audio has evolved from deterministic regression to generative modeling. Early seminal works such as VOCA (Cudeiro et al., 2019) and MeshTalk (Richard et al., 2021) pioneer audio-to-mesh mappings. However, these regression-based approaches often yield over-smoothed motion due to regression-to-the-mean. To capture long-term temporal dependencies, FaceFormer (Fan et al., 2022) introduces transformer-based (Vaswani et al., 2017) autoregressive (AR) motion prediction. Yet, its continuous regression still tends to average out subtle dynamics. CodeTalker (Xing et al., 2023), MemoryTalker (Kim et al., 2025), and ARTalk (Chu et al., 2025) address this by incorporating discrete motion priors, reformulating synthesis as a code-query process that yields more vivid facial motion. Recently, diffusion-based paradigms (Stan et al., 2023; Ma et al., 2024a; Sun et al., 2024; Xie et al., 2026) produce high-fidelity facial motion via iterative denoising, but repeated sampling can be costly for real-time use. To balance fidelity and efficiency, FlowTalker adopts a Draft-then-Refine design: a kinematics-aware AR model predicts factorized motion tokens, and a selective rectified-flow refiner enhances fine-grained facial dynamics in a few steps.

Streaming 3D Facial Animation. Most prior methods are designed for full-utterance or block-wise generation (Fan et al., 2022; Xing et al., 2023; Sun et al., 2024; Chu et al., 2025). Although their generators can run quickly, restricting them to causal streaming inference often degrades animation quality. Streaming-oriented methods reduce response latency through lightweight generation (Han et al., 2025; Lee et al., 2025; Yang et al., 2025; Chu et al., 2026). However, preserving rich full-head motion with only causal audio context remains challenging. FlowTalker addresses this through a factorized motion representation and a Draft-then-Refine motion generator.

3 Methods

We present FlowTalker, a draft-then-refine framework for causal streaming audio-driven 3D head animation synthesis, as illustrated in Fig. 3. At each streaming step, FlowTalker generates the full-head motions from a sliding audio window and aligned motion history, guided by an emotion em-

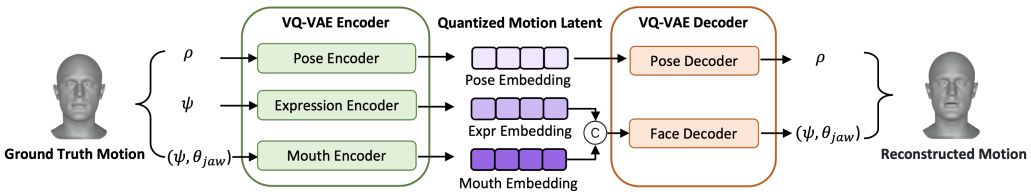


Figure 4: **Spatially factorized VQ-VAE.** Pose, expression, and mouth motion are encoded and quantized with separate codebooks. The pose decoder reconstructs ρ_t, while the face decoder jointly reconstructs $(\psi_t, \theta_{\text{jaw},t})$ from the expression and mouth codebook embeddings.

bedding extracted from the sliding audio window (Ma et al., 2024b) and a global speaker-style embedding (Sun et al., 2024). Built upon the factorized pose, expression, and mouth motion representations described in Sec. 3.2, FlowTalker introduces a Draft-then-Refine pipeline. A kinematics-aware autoregressive drafter first predicts discrete motion tokens (Sec. 3.3), followed by a selective rectified-flow refiner that enhances fine-grained facial dynamics (Sec. 3.4).

3.1 Preliminaries

We use FLAME (Li et al., 2017) to represent 3D head geometry and motion. FLAME parameterizes a head mesh by identity shape $\beta_{\text{shape}} \in \mathbb{R}^{100}$, facial expression $\psi \in \mathbb{R}^{50}$, and pose $\theta = (\rho, \theta_{\text{jaw}}) \in \mathbb{R}^6$, where $\rho, \theta_{\text{jaw}}$ denote global head and jaw rotations, respectively. The resulting mesh is $\mathbf{V} = M(\beta_{\text{shape}}, \psi, \rho, \theta_{\text{jaw}}) \in \mathbb{R}^{5023\times 3}$. With the dataset-provided identity shape fixed, FlowTalker predicts only the motion sequence $\{(\rho_t, \theta_{\text{jaw},t}, \psi_t)\}_{t=1}^{T}$.

3.2 Spatially Factorized Motion Representation

The distinct temporal dynamics and audio dependencies of head pose, expression, and mouth articulation motivate learning component-specific motion patterns, enabling the generator to model each component separately and selectively refine facial motion. This separation, however, cannot be achieved by simply partitioning FLAME parameters: although FLAME separately parameterizes global pose ρ, expression ψ, and jaw rotation θ_{jaw}, the PCA coefficients of ψ span multiple facial regions and cannot be cleanly divided into expression- and mouth-specific subsets. We therefore learn this factorization in latent space, using separate codebooks to encode pose, expression, and mouth motion as distinct token streams.

As shown in Fig. 4, three dedicated encoders map ρ_t, ψ_t, and $(\psi_t, \theta_{\text{jaw},t})$ to pose, expression, and mouth latents $\mathbf{z}_{e,b}^t$, where $b \in \{\text{pose}, \text{expr}, \text{mouth}\}$. Each latent is quantized with a branch-specific codebook $\mathcal{Q}_b$, yielding a discrete index I_b^t and its codebook embedding $\mathbf{z}_{q,b}^t = \mathcal{Q}_b(I_b^t)$. The two facial branches learn complementary representations: the expression branch captures holistic facial variation from ψ_t, while the mouth branch additionally incorporates jaw rotation to focus on mouth articulation. The pose decoder reconstructs ρ_t independently. Since expression and mouth motion both involve the FLAME expression coefficients ψ_t, a shared face decoder combines their codebook embeddings to reconstruct a unified set of facial parameters $(\psi_t, \theta_{\text{jaw},t})$.

The indices $\mathbf{I}^t = [I_{\text{pose}}^t, I_{\text{expr}}^t, I_{\text{mouth}}^t]$ define three discrete motion-token streams predicted autoregressively by the drafter. By selecting learned motion primitives rather than directly regressing continuous motion, this formulation mitigates over-smoothing under causal context while modeling dependencies across motion components. At inference, the predicted indices are mapped to codebook embeddings as structural priors for continuous refinement in Sec. 3.4.

We use a full-face vertex loss to ensure overall reconstruction fidelity, together with upper- and lower-face losses defined by FLAME masks $\mathcal{M}_{upper}$ and $\mathcal{M}_{lower}$ to provide region-specific supervision for the spatially factorized motion representation:

$$\begin{aligned}\mathcal{L}_{\text{vert}} = \lambda_{global}\|\hat{\mathbf{V}} - \bar{\mathbf{V}}\|^2 + \lambda_{up}\|(\hat{\mathbf{V}} - \bar{\mathbf{V}}) \odot \mathcal{M}_{upper}\|^2 \\ + \lambda_{low}\|(\hat{\mathbf{V}} - \bar{\mathbf{V}}) \odot \mathcal{M}_{lower}\|^2,\end{aligned} \tag{1}$$

where $\bar{\mathbf{V}}$ and $\hat{\mathbf{V}}$ are ground-truth and reconstructed vertices. We further impose a parameter loss on head pose and jaw rotation:

$$\mathcal{L}_{\text{param}} = \|\hat{\mathbf{r}} - \bar{\mathbf{r}}\|_2^2, \qquad \mathbf{r} = [\rho; \theta_{\text{jaw}}]. \tag{2}$$

The standard VQ-VAE loss over the three branches $\mathcal{B} = \{\text{pose}, \text{expr}, \text{mouth}\}$ is

$$\mathcal{L}_{\text{vq}} = \sum_{b \in \mathcal{B}} \left(\|\text{sg}[\mathbf{z}_{e,b}] - \mathbf{z}_{q,b}\|_2^2 + \beta_{\text{vq}} \|\mathbf{z}_{e,b} - \text{sg}[\mathbf{z}_{q,b}]\|_2^2 \right), \tag{3}$$

where $\mathbf{z}_{e,b}$ is the encoder latent, $\mathbf{z}_{q,b}$ is its codebook embedding, and $\text{sg}[\cdot]$ denotes stop-gradient.

3.3 Kinematics-Aware Autoregressive Motion Drafting

At streaming step k, the model receives a new audio segment a_k corresponding to C motion frames. It appends a_k to the sliding window and discards the oldest segment, maintaining W frames of audio context: $\mathcal{A}_{\text{win}}^k = [a_{k-L+1}, \ldots, a_k]$, where $L = W/C$ is the number of segments in the window. The preceding $L-1$ segments and their aligned generated motion provide the historical context for predicting the motion of a_k. A pretrained HuBERT encoder (Hsu et al., 2021) converts the audio waveform within each window into motion-aligned audio features $\mathbf{A}_{\text{win}}^k \in \mathbb{R}^{T_a \times D_{\text{aud}}}$. From the same window, emotion2vec (Ma et al., 2024b) extracts $\mathbf{E}_{\text{emo}} \in \mathbb{R}^{768}$. A global style embedding $\mathbf{E}_{\text{id}} \in \mathbb{R}^{128}$ from a pretrained style encoder provides speaker-style conditioning (Sun et al., 2024).

Masked self-attention is applied over the embedded pose, expression, and mouth token streams, with $\mathbf{E}_{\text{id}}$ injected via AdaIN (Huang & Belongie, 2017). Within each sliding window, tokens attend to all history segments and all frames of the current segment, while future segments are masked to preserve segment-level causality. The mask also imposes directed dependencies across the factorized token streams: expression attends to pose, while mouth attends to both pose and expression. This coarse-to-fine hierarchy preserves the spatial factorization during autoregressive generation while coordinating the components through directed conditioning to produce coherent full-head motion.

For cross-modal attention, audio and emotion features are projected into the drafter hidden space, with positional encodings added to the audio features and the emotion embedding prepended as a context token:

$$\tilde{\mathbf{A}}_{\text{win}} = \mathbf{A}_{\text{win}} W_{\text{aud}} + \mathbf{P}_{\text{aud}}, \qquad \tilde{\mathbf{E}}_{\text{emo}} = \mathbf{E}_{\text{emo}} W_{\text{emo}}, \qquad \mathcal{C} = [\tilde{\mathbf{E}}_{\text{emo}}; \tilde{\mathbf{A}}_{\text{win}}]. \tag{4}$$

The motion features then attend to $\mathcal{C}$ through multi-head cross-attention, followed by a structure-aware Multi-FFN with branch-specific prediction heads for pose, expression, and mouth. With the VQ-VAE frozen, the drafter is trained via teacher-forced next-token prediction:

$$\mathcal{L}_{\text{AR}} = \mathbb{E}\left[\sum_{t=1}^{T} \sum_{b \in \mathcal{B}} \text{CE}(\hat{P}_b^t, \bar{I}_b^t) \right], \tag{5}$$

where $\hat{P}_b^t$ and $\bar{I}_b^t$ are the predicted distribution and ground-truth token index for branch b.

3.4 Selective Rectified Flow Refinement

Autoregressive drafting promotes temporal consistency, but discrete motion tokens attenuate fine-grained lip motion and micro-expressions. We therefore refine $\mathbf{z}_{\text{face}}^t = [\mathbf{z}_{q,\text{expr}}^t; \mathbf{z}_{q,\text{mouth}}^t]$ in continuous space. We perturb the coarse facial latent (Meng et al., 2021):

$$\mathbf{z}_{\text{face},0}^t = \mathbf{z}_{\text{face}}^t + \sigma \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \tag{6}$$

and transport it to $\hat{\mathbf{z}}_{\text{face}}^t$ with a conditional rectified-flow ODE (Liu et al., 2022):

$$\frac{d\mathbf{z}}{d\tau} = v_\phi(\mathbf{z}, \tau, c_{\text{in}}^t, \tilde{\mathbf{E}}_{\text{emo}}), \quad \mathbf{z}(0) = \mathbf{z}_{\text{face},0}^t, \quad \hat{\mathbf{z}}_{\text{face}}^t = \mathbf{z}(1), \tag{7}$$

where $\tau \in [0, 1]$. The condition concatenates pose with projected window-level audio:

$$\mathbf{a}_{\text{ctx}} = W_{\text{flat}} \text{Flatten}(\tilde{\mathbf{A}}_{\text{win}}), \qquad c_{\text{in}}^t = [\mathbf{z}_{q,\text{pose}}^t; \mathbf{a}_{\text{ctx}}]. \tag{8}$$

The velocity field v_ϕ is a lightweight MLP modulated by $\tilde{\mathbf{E}}_{\text{emo}}$ through AdaIN.

The target $\mathbf{z}^t_{\text{face},1} = [\bar{\mathbf{z}}^t_{e,\text{expr}}; \bar{\mathbf{z}}^t_{e,\text{mouth}}]$ comes from the frozen VQ-VAE encoder before quantization. We sample the straight-line path

$$\mathbf{z}^t_\tau = (1-\tau)\mathbf{z}^t_{\text{face},0} + \tau \mathbf{z}^t_{\text{face},1}, \quad \tau \sim \mathcal{U}(0,1), \tag{9}$$

and optimize velocity matching:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{t,\tau,\boldsymbol{\epsilon}} \left[\left\| v_\phi(\mathbf{z}^t_\tau, \tau, c^t_{\text{in}}, \tilde{\mathbf{E}}_{\text{emo}}) - (\mathbf{z}^t_{\text{face},1} - \mathbf{z}^t_{\text{face},0}) \right\|^2 \right]. \tag{10}$$

For geometric regularization, we decode the one-step estimate $\tilde{\mathbf{z}}^t_{\text{face}} = \mathbf{z}^t_\tau + (1-\tau)v_\phi$ with the drafted pose and minimize

$$\mathcal{L}_{\text{mesh}} = \mathbb{E}_t \left[\left\| \tilde{\mathbf{V}}^t - \bar{\mathbf{V}}^t \right\|^2 \right]. \tag{11}$$

At inference, an 8-step Euler solver produces $\hat{\mathbf{z}}^t_{\text{face}}$.

3.5 Training Objectives

For stable optimization, training proceeds in three stages: factorized motion representation learning, autoregressive motion drafting, and selective motion refinement: (i) the spatially factorized VQ-VAE learns component-specific codebooks and reconstructs FLAME meshes using $\mathcal{L}_{\text{VQ-VAE}} = \mathcal{L}_{\text{vert}} + \lambda_{\text{param}}\mathcal{L}_{\text{param}} + \lambda_{\text{vq}}\mathcal{L}_{\text{vq}}$; (ii) the autoregressive drafter predicts pose, expression, and mouth token indices from the causal streaming context using $\mathcal{L}_{\text{AR}}$; and (iii) the rectified-flow refiner enhances the drafted facial latents to recover fine-grained facial dynamics using $\mathcal{L}_{\text{Refine}} = \mathcal{L}_{\text{flow}} + \lambda_{\text{mesh}}\mathcal{L}_{\text{mesh}}$.

4 Experiments

4.1 Setup

Datasets. We train FlowTalker on TFHP (Sun et al., 2024) using its official train/test split, comprising 1052 clips from 588 subjects with all mesh sequences tracked at 25 fps. To evaluate cross-dataset generalization, we test FlowTalker zero-shot on 160 VOCASET (Cudeiro et al., 2019) sequences from 4 subjects. We re-track their FLAME parameters at 25 fps to match TFHP training frame rate.

Metrics. All metrics are computed after removing global head pose from both predicted and ground-truth meshes, using FLAME masks of 254 lip and 884 upper-face vertices. LVE (Lip Vertex Error, mm): Following Richard et al. (2021), LVE averages the frame-wise maximum ℓ_2 error over corresponding lip vertices. FDD (Upper-Face Dynamic Deviation, $\times 10^{-5}$ m): Following Xing et al. (2023), FDD measures the difference in temporal variation of upper-face vertices between prediction and ground truth. MOD (Mouth Opening Distance, mm): Following Sun et al. (2024), MOD measures the mean absolute error in mouth-opening distance.

Baselines. We compare FlowTalker against a diverse set of representative audio-driven 3D head animation methods. Autoregressive baselines include FaceFormer (Fan et al., 2022), CodeTalker (Xing et al., 2023), and ARTalk (Chu et al., 2025), while DiffPoseTalk (Sun et al., 2024) represents diffusion-based generation. We also include UniTalker (Fan et al., 2024) as a large-scale pre-trained paradigm and StreamingTalker (Yang et al., 2025) as a streaming-oriented approach.

Comparison Settings. We use dataset-specific baseline settings. All baselines are retrained or fine-tuned on TFHP for fair comparison. On VOCASET, we use released VOCASET-trained models when available; otherwise, methods are evaluated zero-shot without VOCASET training or fine-tuning. We evaluate both datasets under two sequence-length settings: (i) Short (200 frames, 8s): typical conversational segments, used to evaluate lip sync and facial motion quality. (ii) Long (2000 frames, 80s): stress tests for temporal coherence and robustness to error accumulation.

Implementation Details. FlowTalker is implemented in PyTorch (Paszke et al., 2017). The spatially factorized VQ-VAE uses three codebooks (pose, expression, mouth), each with $K = 256$ entries and embedding dimension $D = 64$. The kinematics-aware Autoregressive (AR) model is a 4-layer transformer-based decoder with 4 attention heads. The rectified-flow refiner is a 6-layer MLP. FlowTalker is trained in three stages. All stages are optimized using AdamW (Loshchilov & Hutter, 2017) with an initial learning rate of 1×10^{-4} and cosine decay schedule. For Stage 1 (VQ-VAE),

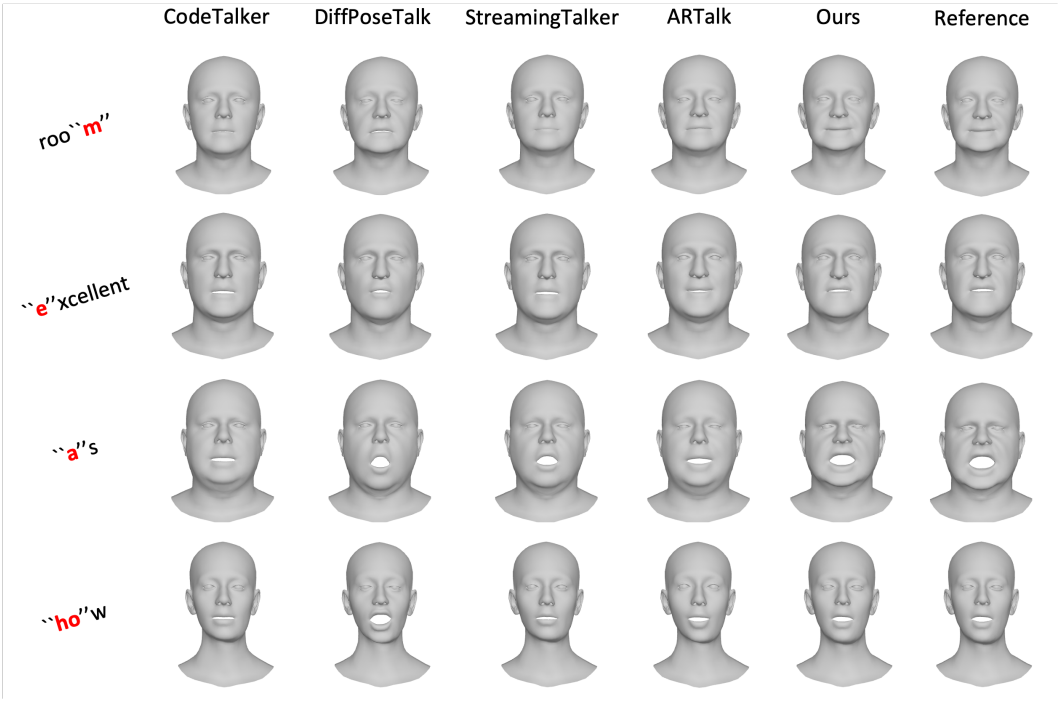


Figure 5: **Qualitative comparison on representative phoneme segments.** FlowTalker achieves the most accurate mouth articulation and natural facial expressions. Red letters denote target phoneme cues. The first three rows are from TFHP and the last from VOCASET. More results are provided in the supplementary material and videos.

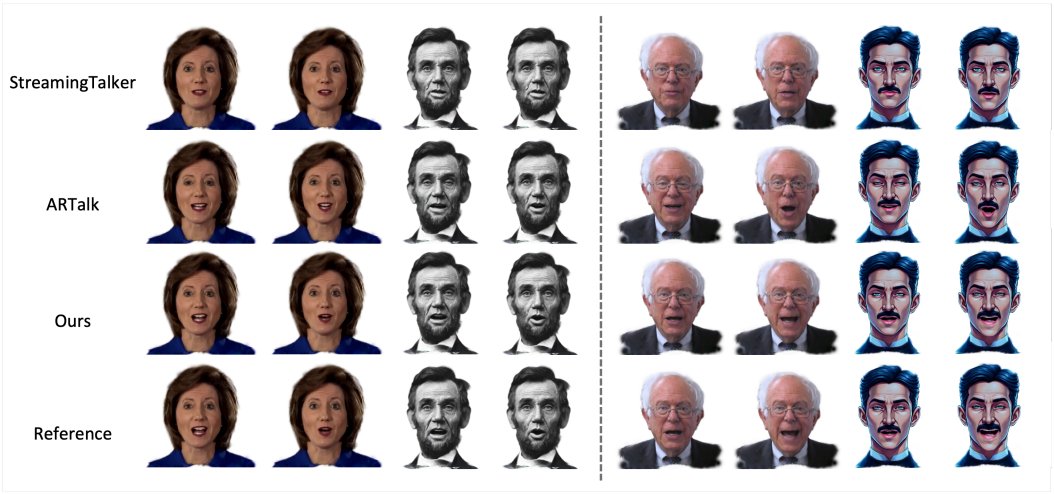


Figure 6: **Qualitative comparison on photorealistic avatar results.** We show two TFHP Short test cases separated by the dashed line, each rendered on two CAP4D avatars (Taubner et al., 2025) at synchronized timestamps. The reference uses ground-truth FLAME parameters. FlowTalker delivers the best visual quality among the compared methods.

we train for 80 epochs with a batch size of 64. For Stage 2 (AR drafting), we train for 40 epochs with a batch size of 16. For Stage 3 (flow refinement), we train for 40 epochs with a batch size of 128, and set the Gaussian perturbation scale to $\sigma = 0.2$. We use a $W = 3.2$ s sliding audio window (80 frames at 25 fps) with a stride of $C = 400$ ms (10 frames) per update. Streaming begins once the first full window is available, generating the initial motion sequence. Each subsequent update uses the current audio window and corresponding motion history to generate 10 new frames. All experiments run on a single NVIDIA RTX A6000 GPU. More details are provided in the supplementary material.

Table 1: **Quantitative comparison in the mesh domain.** Baselines follow their official inference protocols, while FlowTalker is evaluated with both full-utterance and causal sliding-window inference. * indicates zero-shot evaluation on VOCASET. [Key: **Best**, Second Best].

	TFHP (Sun et al., 2024)						VOCASET (Cudeiro et al., 2019)					
	Short			Long			Short			Long		
Method	**LVE↓**	**FDD↓**	**MOD↓**	**LVE↓**	**FDD↓**	**MOD↓**	**LVE↓**	**FDD↓**	**MOD↓**	**LVE↓**	**FDD↓**	**MOD↓**
FaceFormer (Fan et al., 2022)	12.69	22.05	2.71	15.31	25.74	3.59	8.21	15.48	1.83	10.04	17.71	2.58
CodeTalker (Xing et al., 2023)	11.65	20.36	2.45	13.57	22.86	3.09	7.72	15.25	1.77	9.46	17.59	2.29
DiffPoseTalk (Sun et al., 2024)*	10.04	19.81	2.27	11.38	21.28	2.66	9.01	14.94	1.88	10.09	16.57	2.23
UniTalker (Fan et al., 2024)	11.26	25.31	2.19	12.91	28.57	3.01	7.48	17.74	1.85	9.07	20.54	2.41
StreamingTalker (Yang et al., 2025)	9.87	19.72	1.89	11.07	20.54	2.25	7.35	14.76	1.72	8.52	15.95	2.16
ARTalk (Chu et al., 2025)*	9.44	18.13	1.82	10.06	19.34	2.14	7.52	15.28	1.76	8.48	16.43	2.11
FlowTalker (Sliding Window)*	9.03	17.52	1.79	9.85	18.67	1.88	7.14	14.29	1.67	7.83	15.38	1.84
FlowTalker (Full Utterance)*	**8.75**	**16.98**	**1.73**	**9.31**	**17.86**	**1.83**	**6.89**	**13.91**	**1.64**	**7.45**	**14.84**	**1.79**

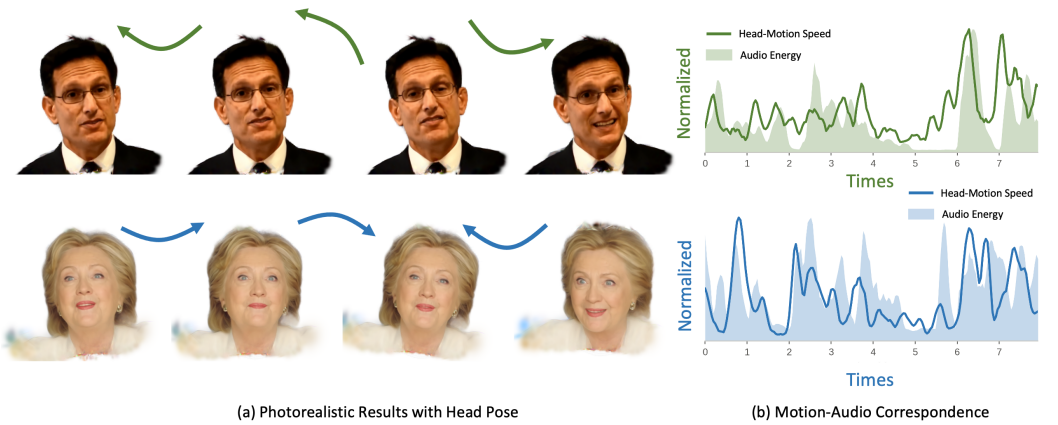


Figure 7: **Photorealistic avatar results with global head pose.** (a) Rendered frames from two TFHP Short test cases at different timestamps, with arrows indicating the direction of global head rotation. (b) Normalized head-motion speed and audio energy. Their temporal relationship reveals more active head motion during speech and reduced head motion during pauses.

4.2 Results

Quantitative Results. Tab. 1 reports results on TFHP and VOCASET under the Short and Long settings. Both the causal and full-utterance variants of FlowTalker outperform all baselines across all metrics, including in the zero-shot evaluation on VOCASET. The smaller performance drop from Short to Long sequences further demonstrates stronger long-horizon stability.

Qualitative Results. Figures 5 and 6 compare mesh and photorealistic results without global head pose. FlowTalker more accurately captures phoneme-specific articulation, including lip closure, rounding, and opening amplitude, while producing natural expressions and reducing over-smoothed facial motion. The photorealistic results show that these improvements remain consistent across different avatars. Please refer to the supplementary videos for dynamic comparisons.

Table 2: **User study results.** [Key: **Best**, Second Best].

Method	AV Sync	Expr Nat.	Temp Coh.	Emo Align.
UniTalker	3.2	3.0	3.6	2.8
DiffPoseTalk	3.6	3.9	2.9	3.2
StreamingTalker	3.9	3.7	4.0	3.9
ARTalk	4.2	4.1	4.2	3.8
FlowTalker	**4.8**	**4.7**	**4.7**	**4.6**

User Study. Thirty participants anonymously rated 20 test cases (10 mesh and 10 photorealistic avatar) on a five-point scale for Audio-Visual Synchronization, Expression Naturalness, Temporal Coherence, and Emotional Alignment. FlowTalker scores highest on all four criteria in Tab. 2.

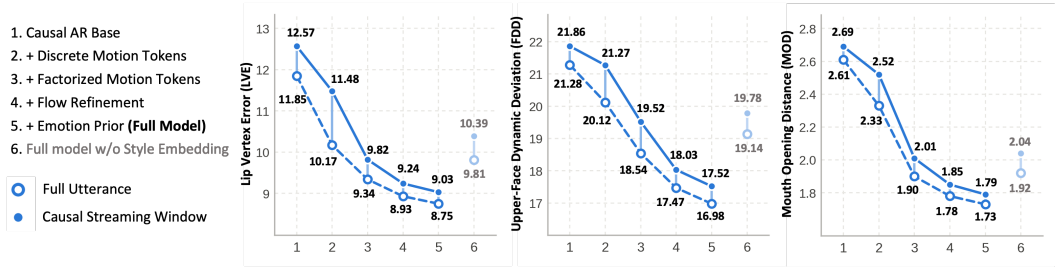


Figure 8: **Component ablation** on TFHP Short under full-utterance and causal sliding-window inference. From left to right, the plots report LVE, FDD, and MOD, where lower values are better.

Response Latency. After a one-time 3.2 s initialization, each update requires 400 ms of new audio and 23 ms of inference on an RTX A6000, resulting in approximately 423 ms steady-state response latency. Combined with its leading facial motion quality in Tab. 1 and smallest causal gap in Fig. 2(b), FlowTalker offers a favorable balance between motion quality and response latency.

Full-Head Motion. As shown in Fig. 7, FlowTalker predicts global head pose jointly with expression and mouth motion, producing natural head movements that become more active during speech and decrease during pauses, consistent with head-motion patterns observed in real human behavior.

4.3 Ablation Study

We evaluate all components of FlowTalker on TFHP Short under both full-utterance and causal sliding-window inference, as shown in Fig. 8. Unlike the system-level comparison in Fig. 2(b), which applies causal inference to baseline checkpoints trained under their original protocols, all FlowTalker variants, including the Causal AR Base, share the same causal sliding-window training protocol. The gap between the two curves measures each variant's sensitivity to removing future audio, while differences across variants isolate the contribution of individual components.

Factorized Motion Tokens. Replacing continuous FLAME regression with discrete motion tokens improves full-utterance inference, while yielding only limited gains under causal sliding-window inference. As a result, the gap between the two settings widens across all three metrics, with the LVE gap increasing from 0.72 to 1.31. Factorizing the motion tokens into pose, expression, and mouth streams improves overall performance and, importantly, substantially narrows this gap, reducing the LVE gap to 0.48; the FDD and MOD gaps also decrease. These results show that factorized motion modeling improves both generation quality and robustness under causal inference.

Flow refinement. Flow-based refinement improves all three metrics while narrowing the causal gap, most notably reducing the FDD gap from 0.98 to 0.56. This highlights its effectiveness in preserving facial dynamics under causal inference.

Conditioning Priors. Adding emotion conditioning further improves performance in both inference settings and yields the smallest gap across all three metrics: 0.28 in LVE, 0.54 in FDD, and 0.06 in MOD. Removing style conditioning, in contrast, substantially degrades performance in both settings, underscoring the importance of speaker-specific style information for motion generation.

5 Conclusion

In this paper, we present FlowTalker, a framework for causal streaming audio-driven 3D head animation synthesis. To improve motion quality with limited future context, we introduce a spatially factorized motion representation and a Draft-then-Refine paradigm. FlowTalker learns dedicated pose, expression, and mouth codebooks that capture component-specific motion patterns as separate token streams. Building on these factorized representations, a kinematics-aware autoregressive drafter generates coordinated motion through structured conditioning under causal context, while a selective rectified-flow refiner enhances fine-grained facial dynamics in continuous space. This discrete-to-continuous design enables high-quality full-head animation with low response latency, providing a practical foundation for interactive telepresence and live broadcasting.

AI USE STATEMENT

We used large language models solely for language editing to improve the clarity and fluency of the manuscript. We did not use these tools to generate research ideas, develop the methodology, design or conduct experiments, analyze data, or formulate scientific conclusions. All AI-assisted text was reviewed by the authors. We take full responsibility for the final content of this work, including its text, claims, and conclusions.

ETHICS STATEMENT

FlowTalker is intended to support human-computer interaction and digital human applications. Privacy and consent are important when training or deploying models on personalized audio-visual data. The collection and use of real-world recordings should be based on explicit consent and comply with applicable data-protection requirements. Like other talking-head generation methods, FlowTalker could be misused to create misleading or unauthorized synthetic media. We encourage responsible deployment practices, including content authentication through watermarking and reliable detection tools that distinguish synthetic content from authentic media, to help mitigate misuse. We also aim to support transparency and responsible research by providing generated examples and evaluation materials that may facilitate research on deepfake detection and media forensics.

REFERENCES

Shivangi Aneja, Artem Sevastopolsky, Tobias Kirschstein, Justus Thies, Angela Dai, and Matthias Nießner. Gaussianspeech: Audio-driven gaussian avatars. *arXiv preprint arXiv:2411.18675*, 2024.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Lee Chae-Yeon, Oh Hyun-Bin, Han EunGi, Kim Sung-Bin, Suekyeong Nam, and Tae-Hyun Oh. Perceptually accurate 3d talking head generation: New definitions, speech-mesh representation, and evaluation metrics. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21065–21074. IEEE, 2025.

Xuangeng Chu, Nabarun Goswami, Ziteng Cui, Hanqin Wang, and Tatsuya Harada. Artalk: Speech-driven 3d head animation via autoregressive model. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–9, 2025.

Xuangeng Chu, Yu Han, Wei Mao, and Shih-En Wei. Personalizing causal audio-driven facial motion via dynamic multi-modal retrieval. *arXiv preprint arXiv:2604.23692*, 2026.

Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian conference on computer vision*, pp. 87–103. Springer, 2016a.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pp. 251–263. Springer, 2016b.

Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10101–10111, 2019.

Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20311–20322, 2022.

Chenpeng Du, Qi Chen, Tianyu He, Xu Tan, Xie Chen, Kai Yu, Sheng Zhao, and Jiang Bian. Dae-talker: High fidelity speech-driven talking face generation with diffusion autoencoder. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4281–4289, 2023.

Sebastian Egger, Raimund Schatz, and Stefan Scherer. It takes two to tango-assessing the impact of delay on conversational interactivity on perceived speech quality. In *Interspeech*, pp. 1321–1324, 2010.

Xiangyu Fan, Jiaqi Li, Zhiqian Lin, Weiye Xiao, and Lei Yang. Unitalker: Scaling up audio-driven 3d facial animation through a unified model. In *European Conference on Computer Vision*, pp. 204–221. Springer, 2024.

Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18770–18780, 2022.

Zhen Han, Mattias Teye, Derek Yadgaroff, and Judith Bütepage. Tiny is not small enough: High quality, low-resource facial animation models through hybrid knowledge distillation. *ACM Transactions on Graphics (TOG)*, 44(4):1–18, 2025.

Yisheng He and Steven Hoi. Meshlam: Feed-forward one-shot animatable textured mesh avatar reconstruction. *arXiv preprint arXiv:2604.22865*, 2026.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510, 2017.

Hyung Kyu Kim, Sangmin Lee, and Hak Gu Kim. Memorytalker: Personalized speech-driven 3d facial animation via audio-guided stylization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11241–11251. IEEE, 2025.

Jiye Lee, Chenghui Li, Linh Tran, Shih-En Wei, Jason Saragih, Alexander Richard, Hanbyul Joo, and Shaojie Bai. Audio driven real-time facial animation for social telepresence. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

Yihong Lin, Zhaoxin Fan, Xianjia Wu, Lingyu Xiong, Liang Peng, Xiandong Li, Wenxiong Kang, Songju Lei, and Huang Xu. Glditalker: Speech-driven 3d facial animation with graph latent diffusion transformer. *arXiv preprint arXiv:2408.01826*, 2024.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Zhiyuan Ma, Xiangyu Zhu, Guojun Qi, Chen Qian, Zhaoxiang Zhang, and Zhen Lei. Diffspeaker: Speech-driven 3d facial animation with diffusion transformer. *arXiv preprint arXiv:2402.05712*, 2024a.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15747–15760, 2024b.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Federico Nocentini, Thomas Besnier, Claudio Ferrari, Sylvain Arguillere, Stefano Berretti, and Mohamed Daoudi. Scantalk: 3d talking heads from unregistered scans. In *European Conference on Computer Vision*, pp. 19–36. Springer, 2024.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Ziqiao Peng, Yihao Luo, Yue Shi, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 5292–5301, 2023.

Shenhan Qian. Vhap: Versatile head alignment with adaptive appearance priors, sep 2024. URL `https://github.com/ShenhanQian/VHAP`.

Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20299–20309, 2024.

Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1173–1182, 2021.

Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. In *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games*, pp. 1–11, 2023.

Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Lin, Jenny Sheng, Yu-Hui Wen, Minjing Yu, and Yong-jin Liu. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. *ACM Transactions on Graphics (ToG)*, 43(4):1–9, 2024.

Jennifer Tam, Elizbeth Carter, Sara Kiesler, and Jessica Hodgins. Video increases the perception of naturalness during remote interactions with latency. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pp. 2045–2050. 2012.

Felix Taubner, Ruihang Zhang, Mathieu Tuli, and David B Lindell. Cap4d: Creating animatable 4d portrait avatars with morphable multi-view diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5318–5330. IEEE Computer Society, 2025.

Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. Imitator: Personalized speech-driven 3d facial animation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 20621–20631, 2023.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Zhongju Wang, Zhenhong Sun, Beier Wang, Yifu Wang, Daoyi Dong, Huadong Mo, and Hongdong Li. 3dxtalker: Unifying identity, lip sync, emotion, and spatial dynamics in expressive 3d talking avatars. *arXiv preprint arXiv:2602.10516*, 2026.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Yue Wu, Xuanhong Chen, Yufan Wu, Wen Li, Yuxi Lu, and Kairui Feng. Fastavatar: Towards unified and fast 3d avatar reconstruction with large gaussian reconstruction transformers. In *International Conference on Learning Representations*, volume 2026, pp. 45230–45252, 2026.

Jiu-Cheng Xie, Jiwang Zheng, Yongkang Xia, Jian Xiong, Chi-Man Pun, Hao Gao, and Feng Xu. Ethead: Generating expressive 3d facial animation and head movement from speech. *arXiv preprint arXiv:2608.01605*, 2026.

Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. 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.

Haolan Xu, Keli Cheng, Lei Wang, Ning Bi, and Xiaoming Liu. Emotag: Emotion-aware talking head synthesis on gaussian splatting with few-shot personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Yifan Yang, Zhi Cen, Sida Peng, Xiangwei Chen, Yifu Deng, Xinyu Zhu, Fan Jia, Xiaowei Zhou, and Hujun Bao. Streamingtalker: Audio-driven 3d facial animation with autoregressive diffusion model. *arXiv preprint arXiv:2511.14223*, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3661–3670, 2021.

Yixiang Zhuang, Baoping Cheng, Yao Cheng, Yuntao Jin, Renshuai Liu, Chengyang Li, Xuan Cheng, Jing Liao, and Juncong Lin. Learn2talk: 3d talking face learns from 2d talking face. *IEEE Transactions on Visualization and Computer Graphics*, 31(9):5829–5841, 2024.

A Overview of Appendix

This appendix provides implementation details, controlled analyses, and additional experiments that complement the main paper. Sec. B defines the causal streaming and latency protocols, explains the causal evaluation of baselines, and studies the effects of window length, update stride, and audio encoder choice. Sec. C describes data preparation and the unified evaluation setup, including checkpoint selection, speaker conditioning, and the user study. Sec. D reports model and training details, including the FLAME region masks and loss weights. Sec. E motivates the factorized representation through component-wise audio coupling and analyzes learned token specialization, directed cross-stream generation, the drafter's training-inference gap, and individual loss components. Sec. F provides a controlled evaluation of photorealistic avatar results, while Sec. G describes additional visualizations and supplementary videos. Finally, Sec. H discusses limitations and future directions.

B Causal Streaming Protocol and Latency

B.1 Protocol

FlowTalker implements causal streaming through sliding-window inference. It processes 16 kHz audio and generates motion at 25 fps using a sliding window of $W = 80$ frames (3.2 s) with a stride of $C = 10$ frames (400 ms). After the initial audio window becomes available, the model generates the corresponding motion to initialize the stream. Each subsequent update advances the window by one stride and generates C new motion frames conditioned on the observed audio window and aligned motion history. No future audio beyond the current generation boundary is used, and previously generated motion remains unchanged.

The 400 ms update interval is consistent with delay ranges found acceptable in prior conversational studies (Egger et al., 2010; Tam et al., 2012), supporting this choice for streaming interaction. The stride ablation in Sec. B.4 further evaluates its quality-latency trade-off.

B.2 Latency Measurement

Response latency consists of audio acquisition and model inference, i.e., $T_{\text{response}} = T_{\text{acquisition}} + T_{\text{inference}}$. For full-utterance and block-wise methods, $T_{\text{acquisition}}$ is the duration of audio required before motion can be generated. FlowTalker begins generation once the first 3.2 s audio window is available; after this one-time initialization, each update requires only 400 ms of new audio, giving a steady-state acquisition time of $T_{\text{acquisition}} = 400$ ms.

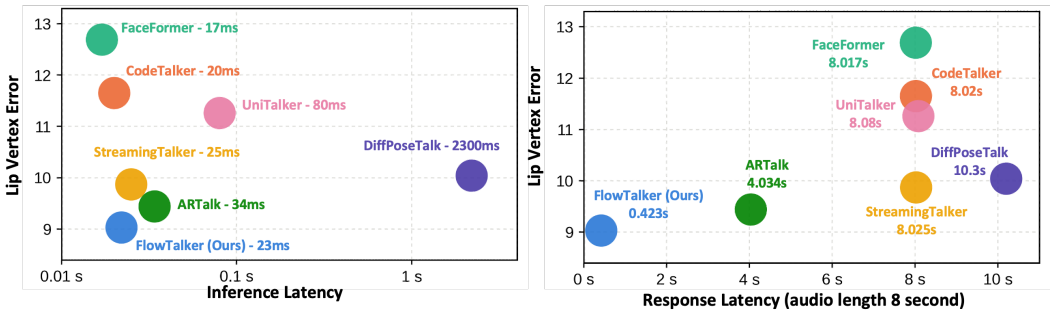


Figure 9: **Latency comparison on TFHP Short setting.** Left: LVE versus inference time for one generation step under each method's native protocol. Right: LVE versus response latency, including the required audio acquisition and inference time. Baselines follow their official protocols, while FlowTalker is measured per steady-state causal update after a one-time 3.2 s initialization.

Fig. 9 reports the inference and response latency on 8 s TFHP Short clips, measured on a single NVIDIA RTX A6000 GPU. For FlowTalker, one update takes $T_{\text{inference}} = 23$ ms, including HuBERT (Hsu et al., 2021), emotion2vec (Ma et al., 2024b), autoregressive drafting, rectified-flow refinement, and VQ-VAE decoding. We exclude one-time global style extraction and photoreal-

Table 3: **System-level robustness under causal streaming on TFHP Short.** Non-causal LVE is measured under the original full-context inference setting, whereas causal LVE is measured using causal sliding-window inference. Δ LVE measures the performance degradation from non-causal to causal inference (lower is better). [Key: **Best**].

Method	Non-causal LVE↓	Causal LVE↓	Δ LVE↓
FaceFormer (Fan et al., 2022)	12.69	14.14	+1.45
CodeTalker (Xing et al., 2023)	11.65	13.36	+1.71
UniTalker (Fan et al., 2024)	11.26	12.94	+1.68
DiffPoseTalk (Sun et al., 2024)	10.04	12.54	+2.50
StreamingTalker (Yang et al., 2025)	9.87	10.93	+1.06
ARTalk (Chu et al., 2025)	9.44	9.93	+0.49
FlowTalker (ours)	**8.75**	**9.03**	**+0.28**

istic rendering. Although several baselines have comparable inference times, their response latency is dominated by the 4-8 s of audio required by their official protocols; DiffPoseTalk additionally requires 2.3 s for diffusion sampling. FlowTalker's steady-state response latency is only $400 + 23 = 423$ ms, and its inference finishes well within each 400 ms update interval without accumulating delay.

B.3 Baselines under the Causal Protocol

Tab. 1 evaluates each baseline using its official inference protocol, all of which are non-causal and have access to future audio. For Fig. 2(b), we instead evaluate all methods under the unified causal streaming protocol ($W = 80, C = 10$). We use TFHP Short as a controlled setting, where each test clip contains 8 s of audio. For FlowTalker, the non-causal result is obtained by extending the input sliding window to cover the entire audio clip. This comparison evaluates whether existing methods can be directly deployed for causal streaming without retraining or modification. We therefore keep each baseline unchanged and restrict only the future audio available at inference. The resulting performance drop measures system-level robustness to causal deployment rather than architecture-specific effects. Retraining the baselines causally would constitute a different comparison, as the resulting models would no longer correspond to the original published methods. A controlled component-level analysis under a shared causal training protocol is provided in Fig. 8.

Although StreamingTalker (Yang et al., 2025) and DiffPoseTalk (Sun et al., 2024) generate motion frame-wise and block-wise, their released inference code encodes the entire 8 s audio clip before generation, making both methods non-causal. Under the causal evaluation, every audio encoder receives only the observed audio window. For DiffPoseTalk, we report causal-generation quality offline, as each 0.4 s update requires 2.3 s of inference, making real-time streaming infeasible.

As shown in Tab. 3, the causal protocol reduces steady-state audio acquisition to 0.4 s per update. FlowTalker achieves the best facial motion quality in both settings and the smallest performance degradation when transitioning to causal inference.

B.4 Update Stride and Window Length

The window length W controls how much past audio context is available at each update, while the stride C determines how much new audio must be accumulated before the next update and how many new motion frames are generated per step. Tab. 4 evaluates their effects by varying one parameter at a time while holding the other fixed.

To evaluate temporal consistency beyond frame-wise metrics, we introduce a sequence-level jitter error. We evaluate the 20 TFHP Short clips, each containing 200 frames at 25 fps. After removing global head pose, we define the second temporal difference of mesh vertex i as $\Delta^2\mathbf{v}_{t,i} = \mathbf{v}_{t+1,i} - 2\mathbf{v}_{t,i} + \mathbf{v}_{t-1,i}$ and compute

$$J_{\mathrm{err}}(\hat{\mathbf{V}}, \mathbf{V}) = \frac{10^3}{(T-2)N} \sum_{t=2}^{T-1} \sum_{i=1}^{N} \left\| \Delta^2\hat{\mathbf{v}}_{t,i} - \Delta^2\mathbf{v}_{t,i} \right\|_2 , \tag{12}$$

Table 4: **Sliding window-length and stride ablation on TFHP Short.** We vary the update stride and window length independently to evaluate their effects on facial motion quality and temporal stability.

Update stride (fixed $W = 80$)						*Window length* (fixed $C = 10$)					
C	Interval	LVE↓	FDD↓	MOD↓	Jitter↓	W	History	LVE↓	FDD↓	MOD↓	Jitter↓
1	40 ms	10.07	20.01	2.03	0.431	20	0.8 s	12.72	22.51	2.74	0.512
5	200 ms	9.36	18.65	1.91	0.164	40	1.6 s	10.16	19.24	2.17	0.198
10	400 ms	9.03	17.52	1.79	0.091	80	3.2 s	9.03	17.52	1.79	0.091
20	800 ms	8.89	17.31	1.77	0.076	160	6.4 s	8.86	17.15	1.76	0.069

Table 5: **Audio encoder ablation on TFHP Short.** We replace HuBERT with Wav2Vec 2.0 while keeping the remaining configuration fixed; time denotes encoding one audio window on an RTX A6000. [Key: **Best**].

Encoder	LVE↓	FDD↓	MOD↓	Encoding time (ms)
Wav2Vec 2.0 (Baevski et al., 2020)	9.05	17.56	**1.78**	8.7
HuBERT (Hsu et al., 2021)	**9.03**	**17.52**	1.79	8.5

where $\hat{\mathbf{V}}$ and $\mathbf{V}$ are the predicted and ground-truth sequences and $N = 5{,}023$; the factor 10^3 expresses the error in millimeters. For each clip, we average the second-order temporal error over all valid time steps and mesh vertices to obtain a single sequence-level score, and then average this score across the 20 clips. Because it measures second-order temporal differences relative to the ground truth, it does not penalize expressive motion that follows the reference dynamics. Instead, it penalizes deviations from the reference temporal evolution, including spurious oscillations. Lower values indicate better temporal fidelity.

Update stride. A shorter stride reduces steady-state response latency by requiring less new audio before each update, but increases computational cost by invoking the full generation pipeline more frequently. For example, reducing C from 10 to 1 increases the number of pipeline executions per second by $10\times$. Reducing C from 20 to 10 lowers the audio acquisition latency from 800 ms to 400 ms with only a minor change in quality, whereas smaller strides of 5 or 1 frame degrade all metrics and substantially increase jitter.

Window length. Longer sliding windows provide richer prosodic and co-articulatory context. Performance improves substantially as W increases from 20 to 80, while extending the window to $W = 160$ brings only marginal gains at higher per-update computational cost.

Chosen operating point. Based on the results in Tab. 4, we set $W = 80$ and $C = 10$, achieving a 400 ms update interval with near-saturated facial motion quality and stable temporal behavior.

B.5 Audio Encoder

The causal streaming protocol restricts the audio encoder to the observed audio within each sliding window, while allowing bidirectional attention within the window. To assess whether FlowTalker depends on a specific speech encoder, we replace HuBERT (Hsu et al., 2021) with Wav2Vec 2.0 (Baevski et al., 2020) while keeping the motion generator, streaming protocol, dataset, and training configuration unchanged.

As shown in Tab. 5, HuBERT and Wav2Vec 2.0 yield comparable facial motion quality and per-window encoding time, with only marginal differences across all metrics. This suggests that the performance of FlowTalker is not specific to HuBERT.

C Evaluation Protocol and Baseline Credibility

C.1 Data Preparation

TFHP. We use the official train/test split and all sequences are tracked at 25 fps. The Short protocol evaluates the first 200 frames (8 s) of each test clip. For the Long protocol, we select 10 TFHP sequences longer than 2,000 frames. Since TFHP is derived from HDTF (Zhang et al., 2021), we directly evaluate long sequences without concatenating separate clips.

VOCASET contains 480 sequences from 12 subjects at 60 fps, each lasting about 4 s. We use 160 sequences from 4 evaluation subjects and concatenate clips with their corresponding audio from the same subject to construct the Short (200-frame) and Long (2,000-frame) test sequences. VOCASET provides registered meshes rather than FLAME parameters. To obtain ground truth for our unified FLAME-based evaluation, we resample the videos to 25 fps, following ARTalk (Chu et al., 2025), and track them with VHAP (Qian, 2024) to obtain 100-dimensional shape, 50-dimensional expression, and 6-dimensional pose parameters. Resampling preserves the original audio-motion alignment. Predictions and ground truth are reconstructed on the same FLAME topology for vertex-wise comparison.

C.2 Fairness of Baseline Evaluation

Dataset-specific checkpoints. Checkpoint selection is handled separately for the two datasets. On TFHP, we retrain or fine-tune every baseline on the official training split using its released implementation code and evaluate it on the official test split. On VOCASET, FaceFormer (Fan et al., 2022), CodeTalker (Xing et al., 2023), UniTalker (Fan et al., 2024), and StreamingTalker (Yang et al., 2025) use released models trained with VOCASET. DiffPoseTalk (Sun et al., 2024), ARTalk (Chu et al., 2025), and FlowTalker are evaluated zero-shot, without training or fine-tuning on VOCASET. The asterisks in Tab. 1 mark these zero-shot results.

Unified evaluation. All predictions are evaluated at 25 fps using the same FLAME geometry, lip and upper-face vertex masks, and metric implementations. We report LVE and MOD in mm and FDD in $\times 10^{-5}$ m. Reported values may differ from the original papers because of different tracking, geometry, frame rate, and metric implementations.

Speaker conditioning. We retain each method's original speaker-conditioning mechanism. On TFHP, FlowTalker, ARTalk, and DiffPoseTalk use reference-motion conditioning, while the remaining baselines use their native speaker or identity conditioning. Reference motion is taken from a separate utterance of the same speaker and never overlaps with the evaluated audio or motion. Each method follows its required reference length; FlowTalker uses 100 frames. Following DiffPoseTalk and ARTalk, FlowTalker encodes the reference once before streaming into a temporally pooled style embedding, which is fixed throughout inference without test-time optimization. For the style-conditioning ablation in Fig. 8, we train and evaluate a FlowTalker variant without speaker-style conditioning while keeping all other settings unchanged.

On VOCASET, FlowTalker, ARTalk, and DiffPoseTalk use the same style encoder on reference motion from the same separate utterance of the target subject, which does not overlap with the evaluated audio or motion. For methods that rely on learned speaker or identity codes rather than reference-motion conditioning, we follow their original VOCASET conditioning setup, as using a motion clip to infer style would require modifying or retraining the model. No method uses motion from the evaluated utterance or performs test-time optimization. Here, zero-shot indicates that the model is neither trained nor fine-tuned on VOCASET.

C.3 User Study Protocol

We compare FlowTalker with UniTalker (Fan et al., 2024), DiffPoseTalk (Sun et al., 2024), StreamingTalker (Yang et al., 2025), and ARTalk (Chu et al., 2025) in a user study with 30 participants. The study includes 20 randomly sampled comparison cases, comprising 10 mesh-animation cases and 10 photorealistic-avatar cases. Each case contains outputs from multiple methods generated for the same subject and input audio. All videos are presented with audio in a blind and randomized order. Participants rate each result from 1 (worst) to 5 (best) on four criteria:

- **Audio-Visual Synchronization** – how well the mouth movements align with the audio;
- **Expression Naturalness** – whether the facial expressions appear natural and expressive;
- **Temporal Coherence** – whether facial and head motion remain smooth and stable over time;
- **Emotional Alignment** – whether the facial affect matches the emotional cues in the audio.

Tab. 2 reports the mean scores across participants and samples. FlowTalker ranks highest on all four criteria.

D Model and Training Details

FLAME Region Masks. We derive the region masks from the official FLAME annotations (Li et al., 2017) and use them for both the masked vertex reconstruction loss and region-based evaluation. The lower-face mask contains the same 254 lip vertices used for LVE, while the upper-face mask contains 884 vertices from the eye and forehead regions.

Loss Weights. In our stage-wise training framework, the loss weights are set as follows: For Stage 1, we set $\lambda_{global} = 1.0$, $\lambda_{up} = 1.0$, $\lambda_{low} = 2.0$, $\lambda_{param} = 1.0$, $\lambda_{vq} = 1.0$, and $\beta_{\mathrm{vq}} = 0.25$. For Stage 3, the rectified flow refinement uses $\lambda_{mesh} = 0.1$.

E Representation and Training Analyses

E.1 Audio Coupling of Motion Components

The factorized motion representation in Sec. 3.2 is motivated by the observation that head pose, expression, and mouth motion evolve at different temporal scales and couple to audio differently. Fig. 10 provides empirical evidence for this observation using ground-truth full-head motion. As illustrated in Fig. 10(a), mouth motion varies rapidly with syllabic activity, expression follows slower prosodic changes, and head pose evolves most slowly and is only weakly tied to speech activity.

To quantify this coupling, we derive multi-scale features from the audio loudness envelope: temporal offsets within $\pm$240 ms capture local audio-motion dependencies, while smoothing windows of 0.5-3.2 s capture longer-term envelope dynamics associated with speech rhythm and prosody. A ridge regressor predicts the frame-wise speed of each motion component from these features. We report five-fold cross-validated R^2 over 62 TFHP test clips, where higher values indicate stronger predictability from the audio envelope. As shown in Fig. 10(b), R^2 increases consistently from head pose (0.008) to expression (0.035) and mouth motion (0.059). This ordering reveals different degrees of audio-envelope coupling across motion components and motivates dedicated token streams for modeling their distinct temporal dynamics and audio dependencies.

E.2 What the Factorized Tokens Encode

We next examine whether the learned streams acquire their intended functional specialization. The pose branch is structurally separated from facial deformation through an independent decoder. For facial motion, the expression branch encodes the FLAME expression code ψ, while the mouth branch jointly encodes ψ and jaw pose to capture articulation-related dynamics. Although the two branches use separate encoders and codebooks, they share ψ as input and are jointly decoded by the face decoder. We therefore apply region-specific masked vertex losses to the upper-face and lip regions to encourage specialization (Sec. D). The ablation in Sec. E.6 verifies the contribution of this supervision, while the token interventions in Fig. 11 show that the two branches learn distinct yet complementary roles.

For each motion token stream, we replace the entire sequence with a single codebook entry repeated across all frames, while keeping the other streams unchanged. We repeat this intervention over all codebook entries and 15 test clips, and measure the induced vertex displacement relative to the original reconstruction after removing global head rotation. To account for the different motion scales across facial regions, particularly between the lips and upper face, we normalize each region's

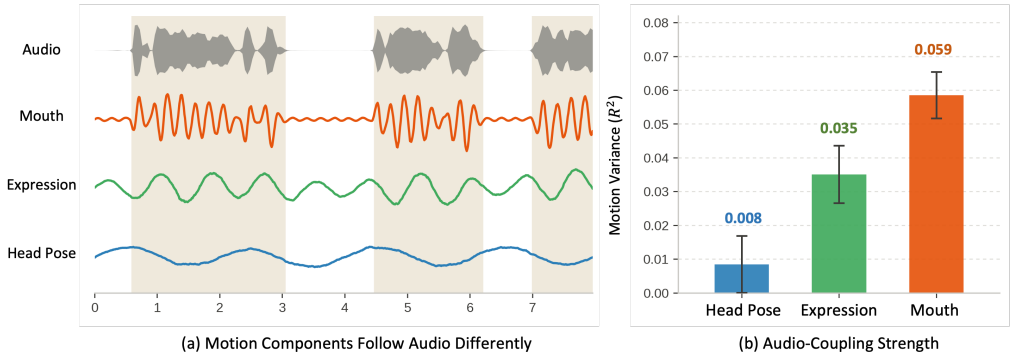
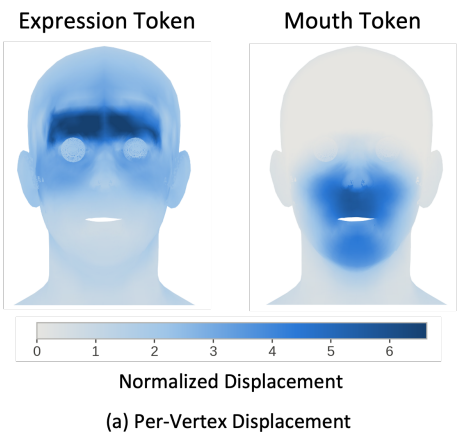


Figure 10: **Distinct temporal scales and audio coupling across motion components.** (a) Schematic traces reflecting measured temporal statistics over a real 8 s audio track; shaded regions indicate speech activity, and amplitudes are normalized for display. (b) Five-fold cross-validated R^2 for predicting component-wise motion speed from loudness-envelope features over 62 TFHP test clips (mean $\pm$ SEM; higher indicates stronger audio coupling).



	Upper Face	Lip Region
Expression Token	3.41	1.03
Mouth Token	0.65	2.19

(b) Region-Wise Response

Figure 11: **Functional roles of factorized motion tokens.** (a) Normalized per-vertex deformation when one token stream is replaced and the others are fixed. (b) Average response in the upper-face and lip regions. Expression tokens mainly affect the upper face, whereas mouth tokens primarily control the lips and jaw. Pose is omitted because it uses a separate decoder.

displacement by its motion magnitude in the original sequence and average the resulting scores. The resulting dimensionless score measures the intervention strength relative to natural speech-driven motion: 1 indicates a comparable change, values above or below 1 indicate stronger or weaker perturbations, and values near 0 indicate little effect.

The interventions reveal distinct roles. Expression tokens produce a stronger response in the upper face than the lips (3.41 vs. 1.03), while mouth tokens show the opposite pattern (0.65 vs. 2.19). Some cross-region response is expected from the shared decoder, but the maps clearly localize expression control to the brow and eye regions and articulation control to the lips and jaw.

E.3 Directed Generation across Factorized Token Streams

The spatially factorized VQ-VAE is pretrained to learn separate pose, expression, and mouth latent spaces and is then frozen. The self-attention mask determines how the autoregressive drafter composes these factorized motion token streams during generation, both across history segments and within the current segment. Let $\mathcal{H}_k$ denote the audio, emotion, and style conditions for segment k,

Table 6: **Directed cross-stream attention ablation on TFHP Short** under causal sliding-window inference. Both variants use the same temporal causal mask and differ only in their cross-stream dependencies.

Attention structure	Cross-stream dependency	LVE↓	FDD↓	MOD↓
Unrestricted self-attention	Symmetric interaction among all streams	9.77	18.46	1.91
Directed mask (ours)	Pose → Expression; Pose, Expression → Mouth	**9.03**	**17.52**	**1.79**

and let $\mathbf{I}_b^k$ and $\mathbf{I}_b^{<k}$ denote the current and historical token sequences of branch b, respectively. Our directed mask induces the following conditional factorization:

$$\begin{aligned} & p\left(\mathbf{I}_{\text{pose}}^k, \mathbf{I}_{\text{expr}}^k, \mathbf{I}_{\text{mouth}}^k \mid \mathcal{H}_k, \mathbf{I}^{<k}\right) \\ & = p\left(\mathbf{I}_{\text{pose}}^k \mid \mathbf{I}_{\text{pose}}^{<k}, \mathcal{H}_k\right) \\ & \quad p\left(\mathbf{I}_{\text{expr}}^k \mid \mathbf{I}_{\text{pose}}^{\leq k}, \mathbf{I}_{\text{expr}}^{<k}, \mathcal{H}_k\right) \\ & \quad p\left(\mathbf{I}_{\text{mouth}}^k \mid \mathbf{I}_{\text{pose}}^{\leq k}, \mathbf{I}_{\text{expr}}^{\leq k}, \mathbf{I}_{\text{mouth}}^{<k}, \mathcal{H}_k\right), \end{aligned} \tag{13}$$

where $\mathbf{I}_b^{\leq k} = [\mathbf{I}_b^{<k}; \mathbf{I}_b^k]$. Thus, pose attends only to pose history, expression additionally uses pose as structural context, and mouth additionally uses both pose and expression. We compare this directed structure with unrestricted cross-stream self-attention, keeping the temporal mask, pretrained factorized VQ-VAE, model architecture, conditioning inputs, and training protocol unchanged.

As shown in Tab. 6, the directed mask reduces LVE from 9.77 to 9.03, FDD from 18.46 to 17.52, and MOD from 1.91 to 1.79. These consistent improvements show that structured coarse-to-fine dependencies are more effective than unrestricted cross-stream interaction for composing the factorized motion streams. Together with the token-intervention analysis in Fig. 11, these results clarify the complementary roles of the two stages: the VQ-VAE learns component-specific motion representations, while the drafter models how these components interact during generation.

E.4 Why Refinement Leaves Pose Fixed

Building on the factorized motion representation, we refine only the expression and mouth branches while keeping head pose fixed. Head pose exhibits lower-frequency temporal variations and the weakest coupling to the audio loudness envelope ($R^2 = 0.008$ in Fig. 10), whereas expression and mouth motion contain fine-grained dynamics that can be attenuated during discrete drafting. We therefore focus continuous refinement on these facial components, preserving the drafted global trajectory and using pose as a condition to maintain coordination between head and facial motion.

E.5 Training-Inference Gap of the Drafter

Tab. 7 compares three training strategies, all evaluated with predicted history. During training, teacher forcing uses ground-truth history, scheduled sampling gradually replaces it with predictions, and free-running training uses predicted history throughout. We also evaluate the teacher-forced model with GT history as a non-deployable upper bound and remove the refiner as a separate control.

The LVE gap between GT and predicted history grows from 0.14 on Short to 0.78 on Long, indicating greater sensitivity to history errors in the Long setting. Among the training strategies, teacher forcing and scheduled sampling each outperform fully free-running training across all three metrics in both settings. Their relative performance varies with sequence length: teacher forcing yields slightly better results on Short sequences, whereas scheduled sampling performs slightly better on Long sequences.

We retain teacher forcing as the default: it achieves the best Short results among the evaluated deployable configurations and remains competitive on Long sequences without an additional history-sampling schedule. Scheduled sampling is a useful alternative when long-sequence accuracy is prioritized.

Table 7: **Training-inference gap and training strategies on TFHP.** GT history is a non-deployable upper bound; all other rows use free-running inference. The last row removes the refiner from our full model. [Key: **Best**, Second Best, excluding the upper bound].

Configuration	Short			Long		
	LVE↓	FDD↓	MOD↓	LVE↓	FDD↓	MOD↓
Teacher forcing, GT history (upper bound)	8.89	17.13	1.75	9.07	17.21	1.78
Teacher forcing, free-running (ours)	**9.03**	**17.52**	**1.79**	9.85	18.67	1.88
Scheduled sampling, free-running	9.14	17.68	1.82	**9.72**	**18.32**	**1.86**
Free-running training and inference	9.28	17.87	1.85	9.95	18.81	1.93
Ours w/o refiner	9.77	19.34	1.98	11.02	20.82	2.21

Table 8: **Loss-component ablation on TFHP Short** under causal sliding-window inference. Each row independently removes one term from the full objective. [Key: **Best**].

Configuration	LVE↓	FDD↓	MOD↓
Full objective	**9.03**	**17.52**	**1.79**
w/o upper-face/lip masked vertex losses	9.72	18.74	1.90
w/o mesh regularizer $\mathcal{L}_{\text{mesh}}$	9.48	18.37	1.85
w/o global vertex reconstruction loss	9.45	18.26	1.87
w/o parameter regularizer $\mathcal{L}_{\text{param}}$	9.27	17.69	1.84

Refined motion is reused as history for subsequent updates. Removing the refiner increases LVE by 0.74 on Short and 1.17 on Long and worsens FDD and MOD, demonstrating its contribution to motion quality in both settings.

E.6 Loss Components

We independently remove each loss term while keeping the architecture, training schedule, and remaining objective fixed. For the spatially factorized VQ-VAE, we evaluate the global and region-masked vertex losses together with the parameter-space regularizer; for refinement, we evaluate the mesh regularizer added to flow matching. We do not remove the quantization objective because it defines the discrete representation and is already contrasted with continuous regression in Fig. 8.

Removing the upper-face and lip losses causes the largest degradation, increasing LVE, FDD, and MOD by 0.69, 1.22, and 0.11, respectively. This highlights the importance of region-specific supervision for facial motion quality and, together with the interventions in Fig. 11, supports its role in promoting functional specialization between the expression and mouth branches. Removing either the mesh regularizer or the global vertex loss also consistently degrades all metrics, showing that mesh-level regularization and overall facial reconstruction provide complementary supervision. Removing the parameter regularizer degrades performance across all metrics, indicating that direct parameter-space supervision provides additional benefits beyond vertex-space losses. These results support the complete training objective.

F Photorealistic Avatar Results

For photorealistic evaluation, we fix the CAP4D (Taubner et al., 2025) avatar, camera, and renderer and vary only the predicted FLAME motion. We render ten TFHP Short test audios on two avatars, giving twenty 200-frame sequences per method. Global head pose is removed. The reference is the same avatar driven by ground-truth FLAME parameters rather than the original video, so the metrics primarily reflect differences in the predicted motion.

In Tab. 9, FlowTalker achieves the best performance across all five metrics. Its LSE-C and LSE-D (Chung & Zisserman, 2016a;b) scores deviate from the reference values by only 0.27 and 0.38, indicating more accurate audio-motion synchronization. For the reconstruction metrics, FlowTalker improves PSNR from 29.72 to 30.71 and SSIM (Wang et al., 2004) from 0.9594 to 0.9734, while

Table 9: **Photorealistic driving results on TFHP.** Scores are averaged over twenty sequences generated from ten audios and two CAP4D avatars under fixed cameras and rendering settings. LSE-C/D evaluate audio-visual synchronization; PSNR, SSIM, and LPIPS compare each result with the same avatar driven by ground-truth FLAME parameters. [Key: **Best**, Second Best, excluding the reference].

Method	LSE-C↑	LSE-D↓	PSNR↑	SSIM↑	LPIPS↓
StreamingTalker (Yang et al., 2025)	4.39	10.12	29.55	0.9479	0.0176
DiffPoseTalk (Sun et al., 2024)	5.21	9.82	29.72	0.9594	0.0151
ARTalk (Chu et al., 2025)	5.43	9.64	29.63	0.9581	0.0123
FlowTalker (ours)	**5.86**	**8.93**	**30.71**	**0.9734**	**0.0085**
Reference (GT parameters)	6.13	8.55	–	–	–

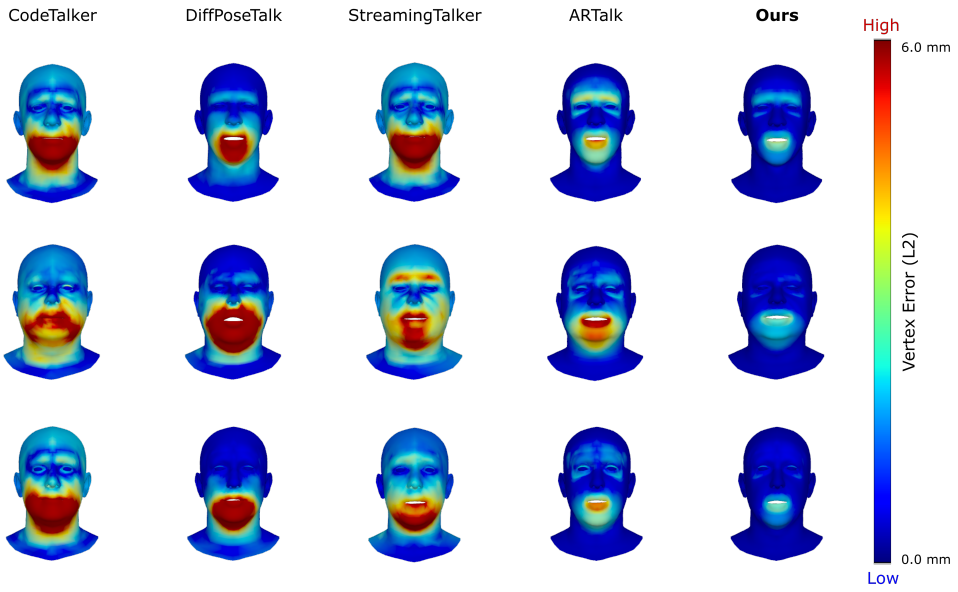


Figure 12: **Mean per-vertex error on TFHP Short.** Each row corresponds to a test clip, and each column corresponds to a method. Colors indicate the frame-averaged ℓ_2 error after removing global head pose, using a shared 0-6.0 mm scale (blue: lower; red: higher).

reducing LPIPS (Zhang et al., 2018) from 0.0123 to 0.0085. Because the avatar and renderer are fixed, these gains reflect more accurate motion driving rather than differences in avatar reconstruction. This experiment is a controlled comparison over twenty rendered sequences; the mesh-domain results in Tab. 1 provide evaluation over the full test split.

G Additional Visualizations and Supplementary Videos

Static figures cannot fully convey dynamic performance. The supplementary video therefore presents the following results in order, with FlowTalker outputs generated under causal sliding-window inference and baseline outputs under their official inference protocols:

1. TFHP Short mesh comparisons without global head pose, followed by the corresponding per-vertex error heatmaps;
2. VOCASET Short mesh comparisons without global head pose, temporally resampled back to the dataset's native 60 fps for visualization;
3. CAP4D photorealistic results without global head pose, comprising two cases with two avatar identities per case;

4. TFHP Short full-head results, showing both the generated meshes and their corresponding photorealistic renderings with global head pose;
5. Long-sequence results. Owing to the submission size limit, the video shows a continuous 30 s segment of the full evaluation sequence.
6. In-the-wild audio results covering emotional speech, languages unseen during training, and spontaneous conversations.

Within each comparison, all methods use the same input audio, camera, avatar, and playback frame rate. Global head pose is removed from the facial-motion comparisons and retained only in the dedicated full-head results. VOCASET metrics are computed on the 25 fps sequences described in Sec. C.1; only the supplementary visualizations are resampled to 60 fps for visualization.

Fig. 12 shows the mean mesh per-vertex error heatmaps. FlowTalker produces consistently lowest errors across all methods.

H Limitations and Future Work

While FlowTalker advances causal streaming audio-driven 3D head animation, several limitations remain that suggest promising directions for future research:

Sensitivity to Unreliable AR Drafting: Our Draft-then-Refine framework relies on the autoregressive (AR) drafter to provide a reasonable coarse motion prior. In rare cases where the drafted motion tokens become severely inaccurate, the subsequent rectified-flow refinement may not fully correct the errors, which can lead to transient visual artifacts. Future work could explore draft-quality estimation or confidence-aware refinement mechanisms to further improve robustness when the initial AR draft is unreliable.

Trade-off Between Emotional Granularity and Temporal Stability: To maintain temporal stability under the causal streaming setting, FlowTalker uses window-level temporal pooling to extract an emotion embedding. This design is effective for preserving a consistent affective tone, but it may smooth out short-term emotional variations and reduce sensitivity to brief, fine-grained affective changes. Future work could explore multi-scale emotion representations that better capture both long-term emotion trends and short-term affective fluctuations. Incorporating explicit emotion control may also further improve the practicality and flexibility of the system.

Rigid Head-Torso Articulation: FLAME represents head pose as a rigid rotation about a fixed neck pivot. Natural head motion is more complex: the effective center of rotation shifts along the cervical spine during nodding and turning, the head translates slightly relative to the torso, and the neck deforms non-rigidly rather than rotating as part of a rigid block. This limitation is inherited from the FLAME parameterization and its tracking protocol, and is therefore shared by all compared methods as well as the tracked ground truth. Future work could augment the pose representation with head translation, exploit FLAME's articulated neck joint, or adopt body-aware parameterizations such as SMPL-X that distribute head motion along the neck while anchoring it to the torso, bringing more natural full-head movement into the streaming setting.