DyStream: Streaming Dyadic Talking Heads Generation via Flow Matching-based Autoregressive Model

Bohong Chen[1]   Haiyang Liu[2]
[1]Zhejiang University   [2]The University of Tokyo

https://robinwitch.github.io/DyStream-Page

Abstract

Generating realistic, dyadic talking head video requires ultra-low latency. Existing chunk-based methods require full non-causal context windows, introducing significant delays. This high latency critically prevents the immediate, non-verbal feedback required for a realistic listener. To address this, we present DyStream, a flow matching-based autoregressive model that could generate video in real-time from both speaker and listener audio. Our method contains two key designs: (1) we adopt a stream-friendly autoregressive framework with flow-matching heads for probabilistic modeling, and (2) We propose a causal encoder enhanced by a lookahead module to incorporate short future context (e.g., 60 ms) to improve quality while maintaining low latency. Our analysis shows this simple-and-effective method significantly surpass alternative causal strategies, including distillation and generative encoder. Extensive experiments show that DyStream could generate video within 34 ms per frame, guaranteeing the entire system latency remains under 100 ms. Besides, it achieves state-of-the-art lip-sync quality, with offline and online LipSync Confidence scores of 8.13 and 7.61 on HDTF, respectively. The model, weights and codes are available.

1. Introduction

Recent breakthroughs in full-duplex speech generation systems [11, 61] have paved the way for highly interactive, streaming AI agents. However, these speech-only agents lack a visual presence, limiting their application in areas like education, sales, and virtual companionship. Bridging this gap requires a model that can generate real-time talking-head video from a single image and the streaming dual-track audio source. Current state-of-the-art approaches [75, 79] have primarily focused on exploring offline solutions. These systems, typically based on two-stage, chunk-by-chunk models, achieve good generation
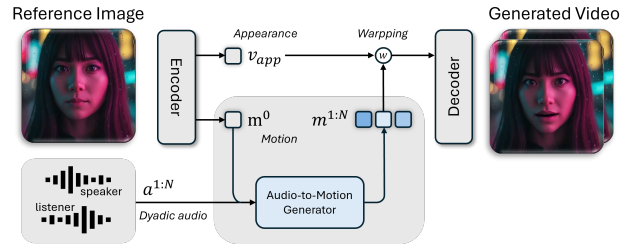

Figure 1. **System pipeline.** DyStream generates talking-head videos from a single reference image and dyadic stream. First, an Autoencoder disentangles the reference image into a static appearance feature $\mathbf{v}_{app}$ and an initial, identity-agnostic motion feature $\mathbf{m}^0$. Next, the Audio-to-Motion Generator takes the initial motion $\mathbf{m}^0$ and the audio stream as input to generate a new sequence of audio-aligned motion features $\mathbf{m}^{1:N}$. Finally, the Autoencoder's decoder synthesizes the output video by warping the appearance feature $\mathbf{v}_{app}$ according to the generated motion sequence $\mathbf{m}^{1:N}$.

quality. However, their design requires the entire audio chunk's content before generation begin, introduces a fundamental latency problem. This accumulated latency, often hundreds of milliseconds, is acceptable when the agent is speaking but becomes a critical bottleneck when it needs to act as an active listener. A listener must react instantaneously to the user's speech with subtle non-verbal cues like nods and facial expressions to create a sense of engagement. The inherent delay in chunk-based systems limits their performance on immediate, naturalistic responses.

To address the latency issue, we propose a new flow-matching based autoregressive model, DyStream. It accepts a streaming, dual-track audio input—capturing both the user's speech and the agent's own generated speech—and generates the corresponding talking head video frame-by-frame. This autoregressive (AR) architecture fundamentally reduces the latency bottleneck and could generate a listener response with one-token latency in theory.

However, achieving high-fidelity lip synchronization in this causal, frame-by-frame manner is challenging. As

human speech involves anticipatory coarticulation, where mouth movements prepare for upcoming phonemes slightly before the sound is produced. This makes it difficult for a purely causal model, without future information, to generate accurate lip shapes. To solve this, we design a causal audio encoder with a minimal audio lookahead module. Specifically, the model generates the current video frame by conditioning on an extremely short segment of future audio (*e.g.*, 60ms). Crucially, this lookahead duration is designed to be shorter than the typical packet size of streaming audio APIs (*e.g.*, 100ms from the OpenAI Realtime API). This ensures that our system can process an entire incoming audio packet and generate the corresponding video before the next packet arrives, thus achieving system-level real-time performance and operating seamlessly within a streaming infrastructure. In addition, we analyzed alternative causal encoder designs, such as distillation and generative encoders, and found the lookahead module to be simple yet effective.

With the above designs, our system achieves state-of-the-art lip-sync quality, with offline and online LipSync Confidence scores of 8.13 and 7.61 on HDTF, respectively. Furthermore, it ensures an immediate and fluid interactive experience by processing any incoming audio packet within 40ms. This level of responsiveness makes our model suitable for creating interactive visual agents that can not only speak but also listen. Our contributions can be summarized as follows:

- We propose a new flow-matching based autoregressive model that generates talking head videos frame-by-frame, theoretically reducing the generation latency compared to previous chunk-based methods.
- We introduce a causal audio encoder with a minimal audio lookahead mechanism (*e.g.*, 60ms) and demonstrate it is simple yet effective for enabling high-quality lip-sync while maintaining system-level real-time responsiveness.
- Our method achieves state-of-the-art performance in lip-sync accuracy and guarantees a generation latency under 34ms.

2. Related Work

Video Diffusion-based Talking Heads Generation A prominent approach involves finetuning a pre-trained image-to-video (I2V) model to align the generated video content with an audio condition, mainly for achieving accurate lip synchronization [2, 3, 6, 7, 16, 20, 26, 29–32, 39, 45, 52, 54, 60, 64, 65, 67, 69, 74, 77]. Early works in this domain were often based on U-Net architectures from Stable Diffusion [47]. For instance, EMO [54] integrates audio conditions via an additional cross-attention layer, while Loopy [26] designs a past-frame compression scheme to better support longer video generation. More recent works have shifted towards Transformer-based back-

bones (DiT), such as Hallo [65], HunyuanVideo-Avatar [6], and MOCHA [64], which largely follow the cross-attention mechanism to inject audio features. While all these methods demonstrate high-quality results with few artifacts and strong generalization, they suffer from high computational costs and slow inference speeds.

Two-Stage Warp-based Talking Heads Generation Two-stage methods aim to first predict disentangled appearance and motion representations from audio, and then use these representations to warp a source image [4, 12, 13, 15, 18, 23, 33, 34, 36, 44, 49, 63, 70, 71, 73, 75, 78, 79]. Representative works such as MegaPortraits [70], VASA [75], and INFP [79] have explored various designs for a robust, identity-agnostic facial motion space. Recent variants further improve flexibility and style controllability, including OmniTalker [63], and ChatAnyone [44]. For the audio-to-motion stage, they typically adopt a chunk by chunk diffusion strategy. However, the requirement of the entire audio chunk's content leads to a latency problem. Different with them, we introduce a flow-matching based autoregressive model to guarantee one token latency in theory.

Dyadic Talking Heads Generation The early approaches in dyadic talking heads generation typically treated conversations as multi-turn exchanges, where the agent acted as either a speaker or listener in each turn. MultiDialog [43] generated speaker behaviors conditioned on the user's previous utterance, but failed to produce synchronized listener reactions. To jointly model both roles, ViCo-X [76] proposes a Role Switcher bridging separate generators for speaker and listener, but the explicit switching often caused unnatural transitions. DIM [56] pretrained a unified model to capture dyadic context but still required manual role assignment during adaptation. AgentAvatar [58] synthesized photo-realistic avatars driven by textual descriptions, but its motions lacked precise audio alignment. Recent person-specific studies [22, 41, 51, 66] achieve high realism but perform limited on unseen identities. In contrast to these approaches, which often rely on explicit role-switching rules, multiple expert models, or offline processing of dual audio tracks, our method is designed for online, streaming applications. We utilize a single, unified model to generate both speaker and listener behaviors concurrently from dual-track audio inputs.

3. Methodology

Given a single image and a streaming dyadic audio signal, our method generates streaming talking-head videos. It is based on a two-stage framework. The first stage disentangles appearance and motion from the source image using an off-the-shelf model based on LIA [62], which we briefly
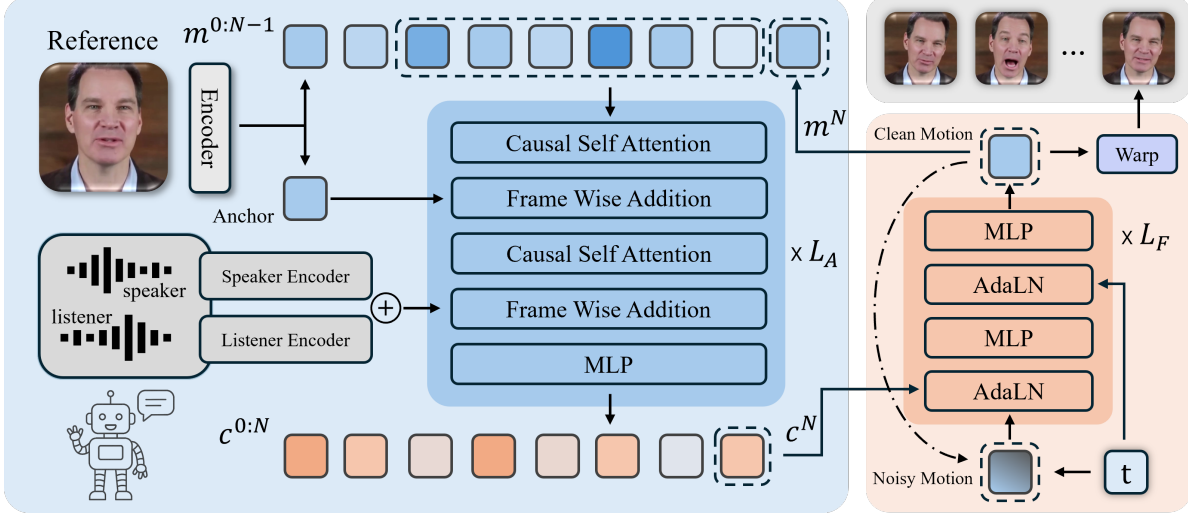
Figure 2. **The architecture of our audio-to-motion generator.** Our model comprises two core modules: an autoregressive network (blue) and a flow matching head (orange). The autoregressive network, built from causal self-attention and MLP blocks, processes the audio, anchor, and previous motion inputs to generate a conditioning signal c^N. This signal is fed into the flow matching head, a stack of MLPs and AdaLN layers. Here, it is injected via AdaLN to guide a multi-step flow matching process to produce the final motion m^N. Finally, the newly generated motion m^N is used to warp the reference image into the output frame, while simultaneously being fed back into the autoregressive network as input for the subsequent generation.

review in Section 3.1. The second stage implements audio-to-motion generation using a flow matching-based autoregressive model, as detailed in Section 3.2. Furthermore, to facilitate real-time responsiveness, we design a causal audio encoder with a lookahead module (Section 3.3) to minimize its reliance on future audio frames.

3.1. Motion-aware Autoencoder

The first stage contains an Autoencoder (AE) to disentangle a source image $\mathbf{I}_s$ into two distinct representations: a static appearance feature $\mathbf{v}_{app}$ that encodes the subject's identity, and a dynamic motion feature $\mathbf{m}$ that describes the pose and expression. Conceptually, this AE includes: an appearance encoder ($\mathcal{E}_{app}$), a motion encoder ($\mathcal{E}_m$), a flow estimator ($\mathcal{F}$), and a final decoder ($\mathcal{D}_{vae}$). These components are trained jointly in a self-supervised manner, using a video reconstruction task that reconstruct a driving video by combining the appearance from a source image with the motion from the driving video itself. In our implementation, we adopt the architecture and training objective from LivePortrait [17] and LIA [62].

First, the motion encoder $\mathcal{E}_m$ extracts a source motion feature $\mathbf{m}_s$ from the source image $\mathbf{I}_s$, and a sequence of driving motion features $\mathbf{m}_{dri}^{1:N}$ from a driving video $\mathcal{V}_{dri}$.

$$\mathbf{m}_s = \mathcal{E}_m(\mathbf{I}_s), \quad \mathbf{v}_{app} = \mathcal{E}_{app}(\mathbf{I}_s), \quad \mathbf{m}_{dri}^{1:N} = \mathcal{E}_m(\mathcal{V}_{dri}) \quad (1)$$

Subsequently, a motion flow estimator $\mathcal{F}$ predicts a dense flow field $\mathbf{F}_{s \to d}$ from the motion features. The decoder

$\mathcal{D}_{vae}$ then synthesizes the final video frames $\hat{\mathbf{I}}_{pred}^{1:N}$ by warping the source appearance feature $\mathbf{v}_{app}$ with this flow field.

$$\mathbf{F}_{s \to d} = \mathcal{F}(\mathbf{m}_s, \mathbf{m}_{dri}^{1:N}); \quad \hat{\mathbf{I}}_{pred}^{1:N} = \mathcal{D}_{vae}(\text{Warp}(\mathbf{v}_{app}, \mathbf{F}_{s \to d})) \quad (2)$$

The module is trained with a composite loss function $\mathcal{L}_{AE}$ combining reconstruction, adversarial, perceptual, cycle consistency, and gaze direction losses. More details are provided in Appendix A.

3.2. Flow matching-based Autoregressive Model

The second stage employs an Audio-Driven Motion Generation Transformer, which we denote as $\mathcal{D}_\theta$. This model generates a sequence of motion latents $\mathbf{m}^{1:N}$ conditioned on a corresponding dyadic audio feature sequence $\mathbf{a}^{1:N}$ and the initial motion latent $\mathbf{m}_s$ extracted from the source image.

Model Architecture. As shown in Figure 2, the core architecture of our network is a combination of a flow matching head and an autoregressive (AR) model. Motivated by [28], the AR model regresses the continuous, deterministic feature, and a single frame-level flow matching head is for stochastic modeling. Specifically, the feature from the AR model, along with the denoising timestep and noise, are fed as conditions for the flow matching head. The denoised clean motion latent is then fed to the AR model for next-frame generation.

Implementation Details. For the AR model, we adapt the 1D Transformer block design from Wan2.1 [57]. A key change is our use of block-independent projections for timestamp embeddings instead of the original shared projections. The flow matching head is composed of stacked MLPs. For audio conditioning, the dyadic audio feature is from the concatenation of encoded speaker and listener audio features. We adopt a customized Wav2Vec2 [1] as an audio encoder; the details are in Section 3.3. As shown in Figure 2, we leverage linear interpolation to align with the temporal dimension of the motion features, and fuse the audio feature via element-wise addition. We found that this is similar to a cross-attention with a diagonal attention mask, which shows better performance in our case.

Anchor Conditioning for Anti-Drifting. As an inherently frame-by-frame autoregressive model, our approach inevitably faces error accumulation during long-sequence generation. The subject's head position gradually drifts from its initial location in the reference image, causing severe artifacts and a decline in visual quality. Considering that our inference procedure generates only one new frame per forward pass in the sliding window—with previously generated frames serving as historical context—we introduce a corresponding modification to the training strategy. Specifically, during training, we randomly select one frame from the last ten frames of the sequence to serve as the appearance anchor during training. During inference, however, we consistently use only the initial reference frame as the anchor. This randomized anchor training strategy effectively mitigating pose drift during autoregressive generation.

Training and Inference. For training, we adopt the m_0-prediction objective, where the model learns to predict the clean motion latent $\mathbf{m}_0$ from its noised version $\mathbf{m}_t$. The loss function is formulated as:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{t,\boldsymbol{\epsilon}} \left[\|\mathcal{D}_\theta(\mathbf{m}_t, t, \mathbf{c}) - \mathbf{m}_0\|^2 \right] \tag{3}$$

where $\mathbf{m}_t = (1 - \sigma_t)\mathbf{m}_0 + \sigma_t\boldsymbol{\epsilon}$ with $t \sim \mathcal{U}[0, 1]$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The conditioning signal $\mathbf{c}$ is the output from AR model.

During inference, we compute the velocity field from the predicted $\hat{\mathbf{m}}_0$ and integrate using an Euler-based ODE solver. To enable fine-grained control, we employ a multi-conditional guidance strategy. Let $\hat{\mathbf{m}}(\mathbf{c}_S, \mathbf{c}_L, \mathbf{c}_R) = \mathbf{m}_\theta(\mathbf{m}_t, t, \mathbf{c}_S, \mathbf{c}_L, \mathbf{c}_R)$ denote the model's prediction given the speaker (**S**), listener (**L**), and anchor (**R**) conditions. The unconditional prediction is $\hat{\mathbf{m}}_{\text{uncond}} = \hat{\mathbf{m}}(\mathbf{0}, \mathbf{0}, \mathbf{0})$.

The final guided prediction $\mathbf{m}_{\text{cond}}$ is computed by rear-

ranging the standard CFG formulation:

$$\begin{aligned} \mathbf{m}_{\text{cond}} = {}&(1 - w_\Sigma)\hat{\mathbf{m}}_{\text{uncond}} + w_s\hat{\mathbf{m}}(\mathbf{S}, \mathbf{0}, \mathbf{0}) \\ &+ w_l\hat{\mathbf{m}}(\mathbf{0}, \mathbf{L}, \mathbf{0}) + w_r\hat{\mathbf{m}}(\mathbf{0}, \mathbf{0}, \mathbf{R}) \\ &+ w_{\text{all}}\hat{\mathbf{m}}(\mathbf{S}, \mathbf{L}, \mathbf{R}), \end{aligned} \tag{4}$$

where w_s, w_l, w_r, and w_{all} are the respective guidance scales, and $w_\Sigma = w_s + w_l + w_r + w_{\text{all}}$.

Besides, we maintain a sliding window of N frames matching the training length. Initially, the first N-1 frames are set to static motion with the corresponding dyadic audio input to predict the next frame. We then shift the window by removing the oldest frame and appending the newly generated frame, repeating this process autoregressively to generate the entire sequence frame by frame.

3.3. Make It Realtime with Customized Wav2Vec2

While current models could generate video in a causal frame-by-frame manner in theory, their lipsync performance is limited with a straightforward implementation. A primary reason is that standard audio encoders, such as Wav2Vec2 [1], typically require non-causal context to extract current features. This leads to severe feature degradation when causality is enforced during inference. As shown in Table 3, the model's audio-lip synchronization improves as the current audio frame can access more future audio frames. To solve this, we customize a causal Wav2Vec2 with a minimal lookahead module, restricting its receptive field to a maximum of n future audio frames.

To implement this causal Wav2Vec2, we introduce two key modifications to eliminate its inherent non-causal operations. First, the standard architecture utilizes Group-Norm in its initial convolutional layer. Since GroupNorm computes statistics across the entire input sequence, it introduces a dependency on future frames. We replace it with LayerNorm, which operates independently at each timestep, thus ensuring causality. Second, a more significant source of non-causality lies in its positional encoding, implemented as a convolution with a large kernel size of 128. This layer inherently accesses up to 64 future frames (a 1.28-second lookahead). We therefore remove this convolutional positional encoding entirely and instead integrate Rotary Positional Embeddings (RoPE) [50] into the Transformer attention layers.

These modifications are important to remove implicit, hard-coded lookaheads from the architecture. As a result, the model's temporal receptive field is governed solely by the attention mask. This grants us precise and explicit control over the amount of future audio context the model can access.

We then decide the number of lookahead frames based on the demands of real-time processing. For instance, systems like the OpenAI Realtime API [42] deliver audio in

100ms chunks. Our target video generation rate is 25 fps, which corresponds to one video frame per 40ms of audio. By configuring our modified audio encoder to access an additional 60ms or less of future audio, we can generate a video frame immediately upon receiving the first 100ms audio segment, thereby incurring no additional latency. This lookahead is implemented by adjusting the attention mask inside our customized Wav2Vec2.

In contrast, for the listener part, we adopt a different approach. The average human reaction time in a conversation is approximately 300 milliseconds [38]. Unlike a speaker, a listener may not require future audio context to determine precise lip movements. Therefore, for the listener's audio encoder, we implement a purely causal attention mask. This ensures that the model's responses are generated based only on past and current audio information.

Training Details. Our training strategy for the Rope-Wav2Vec2 models differs based on their attention mechanism. The Rope-Wav2Vec2 model with a full attention mask is trained from scratch, using the same pre-training task as the original Wav2Vec2. Subsequently, for the variant with restricted attention, we do not train it from scratch. Instead, we employ knowledge distillation, using the pre-trained full-attention model as the teacher to train the restricted-attention model. These final distilled models are then used for our generation task.

4. Experiments

4.1. Implementation Details

Architecture. Our model has three components: an Autoencoder (AE), an Autoregressive model, and a Flow matching Head, with 110M, 182M, and 80M parameters, respectively. The Autoregressive model consists of 12 blocks, each with a hidden dimension of 768 and 8 attention heads. The Flowmatching Head is composed of 6 blocks, also featuring a hidden dimension of 768.

Autoencoder Training. We train the AE using the AdamW optimizer [35] with a learning rate of 1e-5 and a batch size of 48. The dimension of the latent motion codes is set to 512. The AE is trained for a total of 150,000 iterations on H200.

Autoregressive Flow Matching Training. We apply a dropout of 0.5 to the speaker and listener audio, and 0.1 to the reference anchor. For inference, we employ a 5-step Euler ODE sampler and utilize Classifier-Free Guidance (CFG) [19], as described in Equation 4, with guidance weights set to $w_{all} = 1.0$, $w_s = 0.5$, $w_l = 0.5$, and $w_r = 0.5$. The model is trained for 120,000 iterations us-ing the AdamW optimizer with a learning rate of 2e-5 and a global batch size of 64, and costs 12 hours on a single H200.

4.2. Dataset

We use two types of data for our experiments: public datasets and an additional colloected dataset. Our data pre-processing pipeline is similar to that of LatentSync [27]. The core steps include detecting and cropping face regions from raw videos, followed by filtering out clips with low SyncNet [8] scores to ensure data quality. All data are finally processed into talking-head-only videos at a resolution of 512×512, matching the input size of our model. We use two public datasets, HDTF [72] and Realtalk [14]. After processing, our open-source training set contains approximately 50k video clips, each with a duration of about 8 seconds. Our colloected dataset contains approximately 100k clips of human faces, processed using the same pipeline. This dataset is used exclusively as an additional training source to improve model robustness and is not used for evaluation. The AE is trained on both datasets and the Audio-to-Motion Transformer is trained on Open-Source Datasets only. All videos are resampled to 25FPS.

4.3. Evaluation of Speaker

We first compare all methods in offline version in Sec. 4.3, 4.4 and 4.5, *i.e.*, the audio feature are encoded in once via audio encoder (the encoder could see the future audio). Then discuss the online version in Sec 4.6.

We focus on automated, no-reference metrics for evaluation in this subsection. Metrics that require a ground-truth video are discussed in our ablation studies. Following the evaluation protocol of recent video generation benchmarks [24], we evaluate our method across four key dimensions: lip-sync quality [8], content similarity [46], image quality [24], motion dynamics [24]. The specific metrics used to quantify each dimension are detailed in Appendix.

We compare DyStream with several representative or state-of-the-art talking-head generation models: SadTalker [71], Real3DPortrait [68], Hallo3 [9], Sonic [25], and our reproduced INFP [79]. For evaluation, we construct a unified test set by randomly sampling 100 clips each from the HDTF and our internal datasets, these clips were filtered out from training. All input faces are cropped and resized to 512×512.

To ensure a fair and comprehensive comparison, we establish two additional baselines specifically designed for frame-by-frame generation. The first, which we denote as INFP-I (INFP-Inpainting), is an adaptation of our reproduced INFP baseline. We modify its chunk-by-chunk framework: for a fixed chunk size $k = 32$, we decrease the number of generated frames to n and use $k - n$ frames as past context. We tested $n \in \{1, 2, 4, 8\}$, which corresponds to latencies of $\{40, 80, 160, 320\}$ ms, respectively,
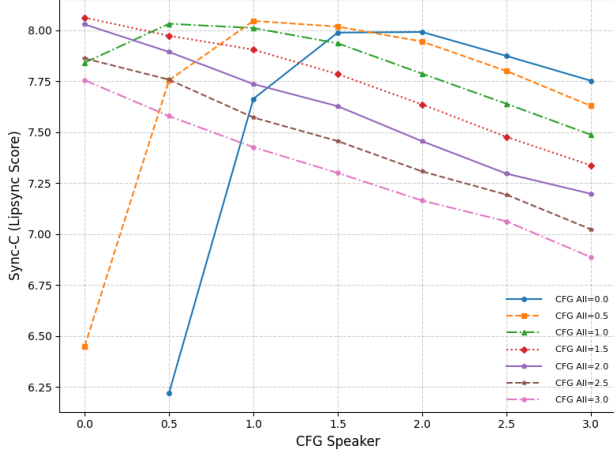
Figure 3. Effect of classifier-free guidance (CFG) on lipsync performance. We evaluate Sync-C across a grid of CFG All and CFG Speaker.

Table 1. **Comparison with existing methods.** We compare our method with state-of-the-art video diffusion models (*slower*) and other warp-based methods (*faster*) on the HDTF-100 test set. ∗ denotes our reproduced version. Latency means the reaction time to the most recent incoming audio.

	Sync-C↑	CS↑	Quality↑	Dynamic↑
GroundTruth	7.614	0.928	0.642	0.660
Hallo3 [9]	6.814	0.915	0.638	**0.870**
Sonic [25]	**8.495**	0.935	0.626	0.600
SadTalker [71]	6.704	**0.965**	**0.697**	0.080
Real3DPortrait [68]	6.811	0.943	0.637	0.030
INFP* [79]	6.894	0.909	0.643	0.950
INFP_I* (latency 40 ms)	1.718	0.877	0.630	0.580
INFP_I* (latency 80 ms)	3.112	0.897	0.655	0.270
INFP_I* (latency 160 ms)	5.335	0.939	0.644	0.040
INFP_I* (latency 320 ms)	6.637	0.945	0.638	0.210
Ours	8.136	0.947	0.642	0.580
Ours (w/o Flow Matching Head)	7.867	0.940	0.643	0.680
Ours (w/o Frame Wise Addition)	7.660	0.925	0.648	0.470
Ours (latency 40 ms)	6.896	0.937	0.647	0.570
Ours (latency 80 ms)	7.671	0.959	0.645	0.520

Table 2. Quantitative comparison on FD, MSE, SID and Var for expression (Exp) and pose (Pose). Lower FD/MSE and higher SID/Var are better.

Method	FD↓		MSE↓		SID↑		Var↑	
	Exp	Pose	Exp	Pose	Exp	Pose	Exp	Pose
INFP [79]	0.141	**3.158**	0.019	**1.286**	4.263	2.562	0.185	0.200
Ours	0.074	3.192	**0.018**	1.636	4.586	**3.070**	**0.275**	**0.596**
Ours (w/o fm)	0.079	3.236	0.020	1.472	4.570	2.896	0.246	0.556
Ours (80ms)	**0.069**	3.489	0.019	1.503	**4.602**	2.831	0.263	0.338

in our 25 FPS setting. It is important to note that this strategy must be consistently applied during both training and inference; applying it only at inference time to a model trained on full chunks results in severe temporal instability and jitter. The second baseline is a deterministic ablation of our own proposed method. We achieve this by removing the flow matching head from our architecture. This transforms our probabilistic model into a deterministic one that autoregressively predicts the motion frame by frame.

The lip-sync performance is mainly determined by the guidance weights w_{all} and w_s. As shown in Figure 3, we empirically found that optimal performance is typically achieved when their sum, $w_{all} + w_s$, is approximately 1.5. Therefore, we set $w_{all} = 1.0$ and $w_s = 0.5$ as the chosen hyperparameters for our model.

4.4. Evaluation of Listener

To quantitatively evaluate the listener performance, we conduct evaluations on the RealTalk test set, which comprises 38 videos ranging from 3 to 14 seconds in length. We compare our method against our own reproduced of INFP. The evaluation relies on facial expression coefficients (Exp) and head poses (Pose) extracted using MediaPipe [37]. We employ four key metrics: Fréchet Distance (FD) and Mean Squared Error (MSE) to measure the realism and accuracy of the generated motions, and two metrics to evaluate diversity: Variance (Var) and SI Diversity (SID). For SID, we follow the methodology of [40], which assesses the diversity of facial and head motions by applying k-means clustering (with $k = 40$ for expression and k=20 for rotation) on ground truth and then calculating the entropy of the resulting histogram of cluster IDs within each sequence.

4.5. Ablation Study

Feature Frame Wise Addition. We inject dyadic audio and anchor information via direct, frame-wise addition along the temporal dimension, instead of the cross-attention mechanism. As shown in Tables 1 and 4, this approach yields better lip-sync performance. The suggests for this task, the audio feature within a local window is enough for lip-syncing. Furthermore, by replacing computationally expensive cross-attention with simple addition, we achieve a 23.1% increase in inference speed.

Anchor Frame Selection. The anchor frame is crucial for mitigating pose drift in long-sequence autoregressive generation. We designed an ablation study with three settings: (1) our method, where the anchor is randomly sampled from the last 10 frames of the training sequence, (2) an anchor frame randomly sampled from the entire sequence, and (3) no anchor. As shown in Figure 4, our approach proves most effective at preventing pose drift over long generation periods.

Figure 4. **Ablation Study on Anchor Frame Selection Strategy.** Visual comparison of long-sequence generation results using the same input audio and initial reference image. The rows from top to bottom correspond to three different anchor selection strategies used during training: our method (anchor sampled from the last 10 frames), a random anchor, and no anchor.

Flow Matching Head. By removing the flow matching head, our model is converted into a deterministic autoregressive model. As detailed in Tables 1 and 2, the listener mode shows a notable decrease in motion diversity, as the model loses its ability to generate varied, stochastic reactions. In contrast, the speaker mode's performance degrades only marginally. We hypothesize this is because the listener mode is a more pronounced one-to-many mapping, where stochastic modeling provides a clear benefit.

4.6. Wav Encoder for Online Generation

We then discuss the online version, the selection of the audio encoder will influence the performance on downstream tasks. We conducted a comprehensive evaluation of several prominent audio encoders to determine their suitability for our task. The candidates included non-causal encoders: Wav2vec2 [1], Hubert [21], WavLM [5], and causal encoders: Wav2vec [48], Mimi [11], Encodec [10], and our Rope-Wav2vec2.

As shown in the Table 3, group 1. our experiments provide several insights into the role of the audio encoder. First, we observe that wav2vec2, Hubert, and WavLM yield comparable performance, which we attribute to their similar network architectures and pre-training objectives. To emphasize the importance of pre-training, we also trained a model with a randomly initialized wav2vec2, which led to a substantial drop in performance. This confirms that a well-pre-trained audio encoder is critical for this task. Furthermore, unlike INFP, which freezes the audio encoder during training, our method finetunes its parameters throughout the process. An ablation study where we froze the wav2vec2 encoder resulted in a significant performance degradation. This finding likely explains why our method outperforms INFP, even without resorting to more complex operations on the audio features, such as memory bank. The continu-

Table 3. **Comparison of different Wav encoders.** ∞ and 0 denotes full-attention and pure-causal architecture, respectively. Top to bottom: (1) different pretrained full-attention wav encoder performs similar; (2) all pure-causal wav encoder show limited performance; (3) training with distillation from full-attention version could slightly improve the results; (4) lookahead improves the performance of Rope-Wav2vec2 (w. distill) (5) training Rope-Wav2vec2 with VAE style generation loss have a limited improvement, suggest that lookahead is simple yet effective.

Wav Encoder	Lookahead Time (ms)		Sync-C ↑
	Train	Inference	
Wav2vec2 [1]	∞	∞	8.114
Wav2vec2 (random init)	∞	∞	4.568
Wav2vec2 (freeze)	∞	∞	6.037
Rope-Wav2vec2	∞	∞	8.136
Hubert [21]	∞	∞	8.007
Wavlm [5]	∞	∞	7.719
Wav2vec [48]	0	0	1.761
Mimi [11]	0	0	5.628
Encodec(24KHz) [10]	0	0	2.034
Rope-Wav2vec2	∞	∞	8.136
Rope-Wav2vec2 (w/o distill)	0	0	2.861
Rope-Wav2vec2 (w. distill)	0	0	3.017
	0	0	3.017
	20	20	5.976
Rope-Wav2vec (lookahead)	40	40	6.896
	60	60	7.387
	80	80	7.671
Rope-Wav2vec2-VAE	0	0	3.495

ous fintuning of the encoder during training is the key factor in our model's better performance.

Figure 5. **Qualitative results and limitations.** *Top:* Given reference images from different views, our approach can generate realistic videos that maintain consistent head pose over time. *Bottom:* Our method have the limitation when hands are overlapped with face.

Table 4. **Inference performance.** Inference speeds measured on an H200 GPU, where APD (Audio Packet Delay) denotes the latency between the arrival of a new audio packet and the generation of the corresponding facial motion, and Step denotes the flow matching denoising step. Result of our model is with lookahead 60 ms version.

Method	Step	FPS	APD(s)	Sync-C↑
INFP	5	21.27	1.0070	6.104
INFP	10	11.37	1.0479	6.381
Ours	1	38.46	0.0262	7.165
Ours	5	29.24	0.0342	7.371
Ours	10	21.88	0.0457	7.387
Ours (w/o Add)	5	22.47	0.0445	7.119

Then, we discuss the different design choices to make the audio encoder causal.

1. **Apply pretrained causal encoders directly.** Instead of using Wav2Vec2 [1], we test wav encoder origianlly designed with causal architectures, such as Mimi [11], Encodec(24KHz) [10] and Wav2vec [48].

2. **Train a causal encoder with distillation.** We modified our Wav2Vec2 encoder to make it causal, noted as Rope-Wav2vec2, and first train a the full-attention version, then distill a causal version.

3. **Train a causal encoder with lookahead.** We operate on the attention mask to allow Rope-Wav2vec2 could use a few future audio features.

4. **Train a generative causal encoder.** We apply generative model style training for the Rope-Wav2vec2, *e.g.*, during the distillation phase, apply a VAE style loss.

As shown in Table 3, group $2-4$, The empirical evidence strongly suggests that for high-fidelity audio-driven talking head generation, a small window of future audio is not just beneficial but essential for correctly modeling the co-articulation and anticipatory dynamics of human speech.

5. Limitations and Future Works

Our method inherits some structural limitations from its warp-based foundation. First, inaccuracies in the warping process can occasionally lead to rigid deformations on ac-

cessories like jewelry or glasses. Second, it is challenging for our method to deal with hand occlusion. As shown in Figure 5, the generated videos exhibit static hand motion. Additionally, our inference pipeline requires the input face to be cropped and aligned to a fixed position and scale, similar to the training data. Finally, while our autoregressive model can access historical context, it lacks a dedicated design to ensure long-term diversity. This is largely due to the scarcity of continuous, long-form training data, presenting a key challenge we will explore in the future.

References

[1] 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. 4, 7, 8

[2] Antoni Bigata, Rodrigo Mira, Stella Bounareli, Michał Stypułkowski, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Keysync: A robust approach for leakage-free lip synchronization in high resolution. arXiv preprint arXiv:2505.00497, 2025. 2

[3] Antoni Bigata, Michał Stypułkowski, Rodrigo Mira, Stella Bounareli, Konstantinos Vougioukas, Zoe Landgraf, Nikita Drobyshev, Maciej Zieba, Stavros Petridis, and Maja Pantic. Keyface: Expressive audio-driven facial animation for long sequences via keyframe interpolation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 5477–5488, 2025. 2

[4] Jiahui Chen, Yang Huan, Runhua Shi, Chanfan Ding, Xiaoqi Mo, Siyu Xiong, and Yinong He. Audio-driven gesture generation via deviation feature in the latent space. arXiv preprint arXiv:2503.21616, 2025. 2

[5] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 16(6): 1505–1518, 2022. 7

[6] Yi Chen, Sen Liang, Zixiang Zhou, Ziyao Huang, Yifeng Ma, Junshu Tang, Qin Lin, Yuan Zhou, and Qinglin Lu. Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters. arXiv preprint arXiv:2505.20156, 2025. 2

[7] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 2403–2410, 2025. 2

[8] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In Workshop on Multi-view Lip-reading, ACCV, 2016. 5, 2

[9] Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. arXiv preprint arXiv:2412.00733, 2024. 5, 6

[10] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. arXiv preprint arXiv:2210.13438, 2022. 7, 8

[11] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. arXiv preprint arXiv:2410.00037, 2024. 1, 7, 8

[12] Aleksandr Drobyshev, Nikita Drobyshev, Egor Zakharov, and Victor Lempitsky. Emoportraits: Emotionally controllable talking head generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. 2

[13] Ao Fu, Ziqi Ni, and Yi Zhou. Dual audio-centric modality coupling for talking head generation. arXiv preprint arXiv:2503.22728, 2025. 2

[14] Scott Geng, Revant Teotia, Purva Tendulkar, Sachit Menon, and Carl Vondrick. Affective faces for goal-driven dyadic communication. arXiv preprint arXiv:2301.10939, 2023. 5

[15] Shengjie Gong, Haojie Li, Jiapeng Tang, Dongming Hu, Shuangping Huang, Hao Chen, Tianshui Chen, and Zhuoman Liu. Monocular and generalizable gaussian talking head animation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 5523–5534, 2025. 2

[16] Jiazhi Guan, Kaisiyuan Wang, Zhiliang Xu, Quanwei Yang, Yasheng Sun, Shengyi He, Borong Liang, Yukang Cao, Yingying Li, Haocheng Feng, et al. Audcast: Audio-driven human video generation by cascaded diffusion transformers. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 10678–10689, 2025. 2

[17] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. arXiv preprint arXiv:2407.03168, 2024. 3, 1

[18] Ying Guo, Xi Liu, Cheng Zhen, Pengfei Yan, and Xiaoming Wei. Arig: Autoregressive interactive head generation for real-time conversations. arXiv preprint arXiv:2507.00472, 2025. 2

[19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598, 2022. 5

[20] Fa-Ting Hong, Zunnan Xu, Zixiang Zhou, Jun Zhou, Qin Lin, and Dan Xu. Actalker: Audio-visual controlled video diffusion with masked selective state spaces modeling for natural talking head generation. arXiv preprint arXiv:2504.02542, 2025. 2

[21] 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. 7

[22] Yinghao Huang, Leo Ho, Dafei Qin, Mingyi Shi, and Taku Komura. Interact: Capture and modelling of realistic, expressive and interactive activities between two persons in daily scenarios. arXiv preprint arXiv:2405.11690, 2024. 2

[23] Yihuan Huang, Jiajun Liu, Yanzhen Ren, Wuyang Liu, and Juhua Tang. Se4lip: Speech-lip encoder for talking head synthesis to solve phoneme-viseme alignment ambiguity. arXiv preprint arXiv:2504.05803, 2025. 2

[24] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 21807–21818, 2024. 5

[25] Xiaozhong Ji, Xiaobin Hu, Zhihong Xu, Junwei Zhu, Chuming Lin, Qingdong He, Jiangning Zhang, Donghao Luo, Yi Chen, Qin Lin, et al. Sonic: Shifting focus to global audio perception in portrait animation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 193–203, 2025. 5, 6

[26] Fangzhou Jiang, Yuqian Liu, Yongan Zhang, Fen Lin, Jinxiao Lian, Zeyi Yang, Xuan Shan, and Jie Chen. Loopy: Long-term talking head generation via past-frame compression. arXiv preprint arXiv:2409.02634, 2024. 2

[27] Chunyu Li, Chao Zhang, Weikai Xu, Jinghui Xie, Weiguo Feng, Bingyue Peng, and Weiwei Xing. Latentsync: Audio conditioned latent diffusion models for lip sync. arXiv preprint arXiv:2412.09262, 2024. 5

[28] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. Advances in Neural Information Processing Systems, 37:56424–56445, 2024. 3

[29] Tianqi Li, Ruobing Zheng, Minghui Yang, Jingdong Chen, and Ming Yang. Ditto: Motion-space diffusion for controllable realtime talking head synthesis. arXiv preprint arXiv:2411.19509, 2024. 2

[30] Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, and Chao Liang. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. arXiv preprint arXiv:2502.01061, 2025.

[31] Haiyang Liu, Xingchao Yang, Tomoya Akiyama, Yuantian Huang, Qiaoge Li, Shigeru Kuriyama, and Takafumi Taketomi. Tango: Co-speech gesture video reenactment with hierarchical audio motion embedding and diffusion interpolation. arXiv preprint arXiv:2410.04221, 2024.

[32] Haiyang Liu, Zhan Xu, Fa-Ting Hong, Hsin-Ping Huang, Yi Zhou, and Yang Zhou. Video motion graphs. arXiv preprint arXiv:2503.20218, 2025. 2

[33] Kangwei Liu, Junwu Liu, Yun Cao, Jinlin Guo, and Xiaowei Yi. Disenttalk: Cross-lingual talking face generation via semantic disentangled diffusion model. arXiv preprint arXiv:2503.19001, 2025. 2

[34] Yujian Liu, Shidang Xu, Jing Guo, Dingbin Wang, Zairan Wang, Xianfeng Tan, and Xiaoli Liu. Syncanimation: A real-time end-to-end framework for audio-driven human pose and talking head animation. arXiv preprint arXiv:2501.14646, 2025. 2

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017. 5

[36] Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live speech portraits: Real-time photorealistic talking-head animation. ACM Transactions on Graphics (Proc. SIGGRAPH Asia), 40(6), 2021. 2

[37] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172, 2019. 6

[38] Antje S. Meyer. Timing in conversation. Journal of Cognition, 6(1), 2023. 5

[39] Lingzhou Mu, Baiji Liu, Ruonan Zhang, Guiming Mo, Jiawei Jin, Kai Zhang, and Haozhi Huang. Flap: Fully-controllable audio-driven portrait video generation through 3d head conditioned diffusion model. arXiv preprint arXiv:2502.19455, 2025. 2

[40] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, , Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 6, 2

[41] Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. From audio to photoreal embodiment: Synthesizing humans in conversations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1001–1010, 2024. 2

[42] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric

Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunningham, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. 4

[43] Se Park, Chae Kim, Hyeongseop Rha, Minsu Kim, Joanna Hong, Jeonghun Yeo, and Yong Ro. Let's go real talk: Spoken dialogue model for face-to-face conversation. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16334–16348, Bangkok, Thailand, 2024. Association for Computational Linguistics. 2

[44] Jinwei Qi, Chaonan Ji, Sheng Xu, Peng Zhang, Bang Zhang, and Liefeng Bo. Chatanyone: Stylized real-time portrait video generation with hierarchical motion diffusion model. arXiv preprint arXiv:2503.21144, 2025. 2

[45] Zheng Qin, Ruobing Zheng, Yabing Wang, Tianqi Li, Zixin Zhu, Sanping Zhou, Ming Yang, and Le Wang. Versatile multimodal controls for expressive talking human animation. arXiv preprint arXiv:2503.08714, 2025. 2

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021. 5

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 2

[48] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. arXiv preprint arXiv:1904.05862, 2019. 7, 8

[49] Shuai Shen, Wanhua Li, Yunpeng Zhang, Weipeng Hu, and Yap-Peng Tan. Audio-plane: Audio factorization plane gaussian splatting for real-time talking head synthesis. arXiv preprint arXiv:2503.22605, 2025. 2

[50] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with

rotary position embedding. Neurocomputing, 568:127063, 2024. 4

[51] Mingze Sun, Chao Xu, Xinyu Jiang, Yang Liu, Baigui Sun, and Ruqi Huang. Beyond talking–generating holistic 3d human dyadic motion for communication. International Journal of Computer Vision, pages 1–17, 2024. 2

[52] Wenzhang Sun, Xiang Li, Donglin Di, Zhuding Liang, Qiyuan Zhang, Hao Li, Wei Chen, and Jianxun Cui. Uni-avatar: Taming lifelike audio-driven talking head generation with comprehensive motion and lighting control. arXiv preprint arXiv:2412.19860, 2024. 2

[53] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In ECCV, 2020. 2

[54] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive – generating expressive portrait videos with audio. arXiv preprint arXiv:2402.17485, 2024. 2

[55] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. arXiv preprint physics/0004057, 2000. 1

[56] Minh Tran, Di Chang, Maksim Siniukov, and Mohammad Soleymani. Dim: Dyadic interaction modeling for social behavior generation. In European Conference on Computer Vision, pages 484–503. Springer, 2024. 2

[57] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianx-iao Yang, et al. Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314, 2025. 4

[58] Duomin Wang, Bin Dai, Yu Deng, and Baoyuan Wang. Agentavatar: Disentangling planning, driving and rendering for photorealistic avatar agents. arXiv preprint arXiv:2311.17465, 2023. 2

[59] Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 17979–17989, 2023. 1

[60] Haotian Wang, Yuzhe Weng, Yueyan Li, Zilu Guo, Jun Du, Shutong Niu, Jiefeng Ma, Shan He, Xiaoyan Wu, Qiming Hu, et al. Emotivetalk: Expressive talking head generation through audio information decoupling and emotional video diffusion. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 26212–26221, 2025. 2

[61] Peng Wang, Songshuo Lu, Yaohua Tang, Sijie Yan, Wei Xia, and Yuanjun Xiong. A full-duplex speech dialogue scheme based on large language model. Advances in Neural Information Processing Systems, 37:13372–13403, 2024. 1

[62] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Lia: Latent image animator. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024. 2, 3, 1

[63] Zhongjian Wang, Peng Zhang, Jinwei Qi, Guangyuan Wang, Sheng Xu, Bang Zhang, and Liefeng Bo. Om-nitalker: Real-time text-driven talking head generation with in-context audio-visual style replication. arXiv preprint arXiv:2504.02433, 2025. 2

[64] Cong Wei, Bo Sun, Haoyu Ma, Ji Hou, Xintao Wang, Bin Zhou, Jingtao Hu, and Jing Liao. Mocha: Towards movie-grade talking character synthesis. arXiv preprint arXiv:2503.23307, 2025. 2

[65] Jialiang Xu, Feng Cheng, Qianyi Wu, Ming Ding, Youqiang Li, Hangbo Qi, Tao Kong, Hongxia Yang, and Jingren Zhou. Hallo: High-fidelity audio-lips learned optimization for talking head generation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024. 2

[66] Yichao Yan, Zanwei Zhou, Zi Wang, Jingnan Gao, and Xiaokang Yang. Dialoguenerf: Towards realistic avatar face-to-face conversation video generation, 2023. 2

[67] Shurong Yang, Huadong Li, Juhao Wu, Minhao Jing, Linze Li, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Jin Wang. Megactor-sigma: Unlocking flexible mixed-modal control in portrait animation with diffusion transformer. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 9256–9264, 2025. 2

[68] Zhenhui Ye, Tianyun Zhong, Yi Ren, Jiaqi Yang, Weichuang Li, Jiawei Huang, Ziyue Jiang, Jinzheng He, Rongjie Huang, Jinglin Liu, et al. Real3d-portrait: One-shot realistic 3d talking portrait synthesis. arXiv preprint arXiv:2401.08503, 2024. 5, 6

[69] Hongwei Yi, Tian Ye, Shitong Shao, Xuancheng Yang, Jiantong Zhao, Hanzhong Guo, Terrance Wang, Qingyu Yin, Zeke Xie, Lei Zhu, et al. Magicinfinite: Generating infinite talking videos with your words and voice. arXiv preprint arXiv:2503.05978, 2025. 2

[70] Egor Zakharov, Nikita Drobyshev, Jenya Chelshev, Taras Khakhulin, Aleksei Ivakhnenko, and Victor Lempitsky. Megaportraits: One-shot megapixel neural head avatars. In Proceedings of the ACM International Conference on Multimedia (ACM MM), 2022. 2, 1

[71] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Wenping Wang, and Qifeng Chen. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 2, 5, 6

[72] 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, pages 3661–3670, 2021. 5

[73] Dingcheng Zhen, Shunshun Yin, Shiyang Qin, Hou Yi, Ziwei Zhang, Siyuan Liu, Gan Qi, and Ming Tao. Teller: Real-time streaming audio-driven portrait animation with autoregressive motion generation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 21075–21085, 2025. 2

[74] Tianyun Zhong, Chao Liang, Jianwen Jiang, Gaojie Lin, Jiaqi Yang, and Zhou Zhao. Fada: Fast diffusion avatar synthesis with mixed-supervised multi-cfg distillation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 3101–3110, 2025. 2

[75] Hang Zhou, Shiyang Yan, Junting Dong, Xihui Liu, Ziwei Liu, and Bolei Zhou. Vasa-1: Visual affective skills agent. In Advances in Neural Information Processing Systems (NeurIPS), 2024. 1, 2

[76] Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, and Tiejun Zhao. Interactive conversational head generation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025. 2

[77] SZ Zhou, YB Wang, JF Wu, T Hu, JN Zhang, ZJ Li, and Y Liu. Paha: Parts-aware audio-driven human animation with diffusion model. arXiv e-prints, pages arXiv–2505, 2025. 2

[78] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makeittalk: Speaker-aware talking-head animation. In ACM Transactions on Graphics (Proc. SIGGRAPH Asia), 2020. 2

[79] Yucheng Zhu, Anpei Chen, Lingjie Liu, Zhixin Piao, Duygu Ceylan, Nicolas Chappuis, Tuanfeng Wang, and Christian Theobalt. Infp: Identity-neutral facial performance. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025. 1, 2, 5, 6

DyStream: Streaming Dyadic Talking Heads Generation via Flow Matching-based Autoregressive Model

Supplementary Material

This supplemental document contains four sections:
- Video and Codes (Section A).
- Disentangled AE for Motion Representation (Section B).
- Inference Performance (Section C).
- Metric Definitions (Section D).

A. Video and Codes

This supplemental provides a comprehensive HTML that contains separate videos, including:
- Results for dyadic talking heads generation.
- Comparison between our method and INFP baseline.
- Ablation study for the lookahead module and anchor.
- More results for in-the-wild image and audios.

B. Disentangled AE for Motion Representation

Architecture. The VAE is an image-level autoencoder that disentangles a source image I_s into an appearance feature $\mathbf{v}_{app}$ and a motion feature $\mathbf{m}$ in the latent space. To enhance the robustness of facial representations, we leverage a 3D-aided appearance encoder (E_{app}) and a face decoder (D_{vae}) with the same network architecture as Live-Portrait [17]. The 3D appearance feature volume provides a more accurate characterization of facial appearance in 3D compared to traditional 2D feature maps. We also employ a 2D-based motion encoder (E_m) to extract motion codes and compress them into low-dimensional 1D representations. This design effectively captures the core semantics of facial motion while avoiding entanglement with appearance information [55, 59, 62].

Following LIA, we build a motion flow-based generation pipeline. Concretely, E_m first extracts the source motion code $\mathbf{m_s}$ from the input portrait image I_s, and driving motion codes $\mathbf{m_{dri}^{1:N}}$ from the driving video V_{dri}. A motion flow estimation model F then takes these motion codes as input to predict the *self→driving* flow $Flow_{s \to d}$. We utilize the E_{app} to extract a 3D appearance feature volume ($\mathbf{v}_{app}$) from the portrait image I_s. $\mathbf{v}_{app}$ is then warped using the estimated motion flow and passed to the decoder D_{vae} to synthesize the final predicted video $I_{pred}^{1:N}$. The overall pipeline is formulated as follows:

$$\mathbf{m_s} = E_m(I_s); \mathbf{m_{dri}^{1:N}} = E_m(V_{dri})$$

$$\mathbf{v}_{app} = E_{app}(I_s)$$

$$Flow_{s \to d} = F(\mathbf{m_s}, \mathbf{m_{dri}^{1:N}})$$

$$I_{pred}^{1:N} = D_{vae}(Warp(\mathbf{v}_{app}, Flow_{s \to d}))$$

Readers are referred to [62] for more details of this architecture.

Training Objective. The VAE is trained with a composite loss function. We adopt the base losses from LIA [62], which include reconstruction ($\mathcal{L}_{rec}$), perceptual ($\mathcal{L}_{feat}$), and adversarial ($\mathcal{L}_{adv}$) losses. We adopt two additional losses following [70]: a cycle consistency loss ($\mathcal{L}_{cyc}$) to prevent appearance leakage and enhance the disentanglement between motion and appearance features; and a gaze direction loss ($\mathcal{L}_{gaze}$) to encourage diversity in eye movement. To further enhance the disentanglement between motion and appearance features, we follow MegaPortrait [70] and incorporate severl additional losses: a cross-identity consistency loss ($\mathcal{L}_{cyc}$), an identity consistency loss ($\mathcal{L}_{id}$) a gaze direction loss ($\mathcal{L}_{gaze}$). We additionally incorporate a gaze direction loss ($\mathcal{L}_{gaze}$) to encourage diversity in eye movement. The final objective is a weighted sum of these terms:

$$\mathcal{L}_{VAE} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{feat}\mathcal{L}_{feat} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{gaze}\mathcal{L}_{gaze} \quad (5)$$

The weights λ for each term are determined empirically ($\lambda_{rec} = 1$, $\lambda_{feat} = 2$, $\lambda_{adv} = 0.2$, $\lambda_{gaze} = 5$, $\lambda_{cyc} = 1$).

C. Inference Performance

In Table 4, we report that our method achieves a FPS of 29.24 and an APD (Audio Packet Delay) of 0.0342 with a flow matching step of 5. By reducing the data offload between the CPU and GPU, we could archive an FPS of 38.61 and an APD of 0.0259 on the same server.

Furthermore, we conduct tests on a consumer-grade host equipped with a consumer-grade GPU (4090) and an i9-13900KF CPU. On this setup, our method achieves a FPS of 58.48 and an APD of 0.0171 with a flow matching step of 5. The entire process consumed only 2.37G of VRAM. This is mainly because our inference pipeline is latency-bound and involves frequent CPU–GPU synchronization, where high CPU single-thread performance plays a crucial role. As our current implementation already meets real-time performance requirements, we focus the main paper on the modeling and algorithmic rather than extensive system-level optimization. We note that additional engineering efforts, such as reducing CPU–GPU offloading and synchronization during inference, could further improve runtime performance, and we leave these optimizations to future work.

D. Metric Definitions

Lip-sync Consistency (Sync-C): We employ the confidence score derived from SyncNet [8], referred to as Sync-Score, to evaluate the temporal synchronization quality between the generated lip movements and the corresponding audio signals. Specifically, we extract audio features and visual features from consecutive video frames, and compute the cross-correlation between audio and visual embeddings across multiple temporal shifts. The confidence score represents the maximum correlation value, indicating the degree of audio-visual synchronization. The final metric is computed as a weighted average across all frames, where higher values indicate better lip-sync quality.

Content Similarity (CS): To evaluate the temporal consistency of the generated content, we employ Content Similarity metric. Specifically, we extract deep visual features from all video frames using a pre-trained encoder, resizing each frame to 160×160 pixels. The first frame serves as the reference, and we compute the cosine similarity between its feature embedding and those of all subsequent frames. The final score is the mean cosine similarity across all frame pairs, with values closer to 1 indicating higher temporal consistency.

Imaging Quality (Quality): We assess the visual quality of generated videos using a pre-trained image quality assessment model. Each frame in the video is independently evaluated by the model, producing a quality score. The Imaging Quality metric is computed as the average score across all frames, normalized to the range [0, 1], where higher values indicate superior visual fidelity and perceptual quality.

Dynamic Degree (Dynamic): To quantify the presence of motion in generated videos, we utilize the Dynamic Degree metric based on optical flow analysis. We employ the RAFT model [53] to estimate inter-frame optical flow, sampling frames at intervals of fps/8. For each consecutive frame pair, we compute the flow magnitude and take the mean of the top 5% largest magnitudes as the motion score. A video is classified as dynamic if the number of frames exceeding an adaptive threshold (scaled by resolution and video length) surpasses a minimum count threshold. The final metric represents the proportion of videos exhibiting significant motion.

Fréchet Distance (FD): Motion realism is quantified by measuring the distributional discrepancy between the generated motion sequences and the ground-truth motions. Specifically, we calculate the Fréchet Distance (FD) across the feature domains of facial expressions and head poses over the entire sequence.

Mean Squared Error (MSE): We calculate the Mean Squared Error (MSE) between the generated motion parameters and the ground-truth motion sequences to evaluate the deterministic reconstruction accuracy.

Shannon Index for Diversity (SID): Following L2L [40], this metric quantifies the magnitude and dynamism of generated facial motions by calculating the variance along the temporal dimension. We extract facial motion representations using MediaPipe Face Landmarker, obtaining 6-dimensional head motion parameters (3D translation and 3D rotation in Rodrigues form) and 52-dimensional expression blendshapes for each frame. The variance is computed separately for head motion (Var_head) and facial expressions (Var_exp) as $\text{Var} = \sum_{d=1}^{D} \text{Var}(\mathbf{x}^{(d)})$, where $\mathbf{x}^{(d)}$ denotes the motion sequence along dimension d. Higher variance indicates more dynamic and expressive motions.

Variance (Var): Following L2L [40], this metric quantifies the magnitude and dynamism of generated facial motions by calculating the variance along the temporal dimension. We extract facial motion representations using MediaPipe Face Landmarker, obtaining 6-dimensional head motion parameters (3D translation and 3D rotation in Rodrigues form) and 52-dimensional expression blendshapes for each frame. The variance is computed separately for head motion (Var_head) and facial expressions (Var_exp) as $\text{Var} = \sum_{d=1}^{D} \text{Var}(\mathbf{x}^{(d)})$, where $\mathbf{x}^{(d)}$ denotes the motion sequence along dimension d. Higher variance indicates more dynamic and expressive motions.

Audio Packet Delay (APD): This metric evaluates system latency in real-time streaming generation scenarios, defined as the time interval between the reception of an audio packet and the output of the corresponding motion. It is important to note that chunk-based methods, such as INFP, inherently require the accumulation of a full audio chunk before motion synthesis can commence. Consequently, the calculation of APD must encompass not only the model's computational inference time but also the buffering latency equivalent to the duration of the audio chunk. In our implementation of INFP, this chunk duration is set to 0.96 seconds.