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

Anonymous authors
Paper under double-blind review

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 generates each frame within 34 ms, keeping the model-side latency, from the current audio timestamp to the delivery of the rendered frame including the lookahead, under 100 ms. Besides, it achieves state-of-the-art lip-sync quality, with offline and online LipSync Confidence scores of 8.13 and 7.67 on HDTF, respectively. The model, weights and codes are available.

1 Introduction

Recent breakthroughs in full-duplex speech generation systems (Wang et al., 2024a; Défossez et al., 2024) 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.

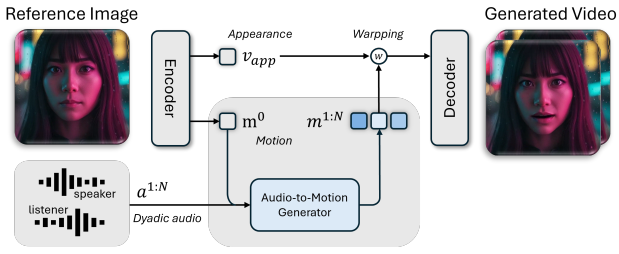

Figure 1: **System pipeline.** DyStream generates talking-head videos from a single reference image and dyadic stream. First, an Autoencoder disentangles the reference image into a static appearance feature $\mathbf{v}_{app}$ and an initial, identity-agnostic motion feature $\mathbf{m}^0$. Next, the 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 $\mathbf{m}^{1:N}$.

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.

Current state-of-the-art approaches (Zhu et al., 2025) have primarily focused on exploring offline solutions. These systems, typically based on two-stage, chunk-by-chunk models, achieve good generation 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. Rather than a new dyadic formulation, our focus is making dyadic generation streamable. 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. To solve this, we design a causal audio encoder with a minimal audio lookahead module, which conditions the current frame on an extremely short segment of future audio (*e.g.*, 60 ms). This lookahead is shorter than the typical packet size of streaming audio APIs (*e.g.*, $\sim$100 ms for the OpenAI Realtime API), so a frame can be generated as soon as a packet arrives. We also analyze alternative causal encoder designs, such as distillation and generative encoders, and find 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 7.67 on HDTF, respectively. Furthermore, each frame is generated within 34 ms, shorter than the 40 ms frame interval at 25 FPS, with a model-side latency below 100 ms including the 60 ms lookahead. 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, reducing the generation latency of previous chunk-based methods.
- We introduce a causal audio encoder with a minimal lookahead (*e.g.*, 60 ms) for the speaker track and a purely causal listener track, which is simple yet effective for high-quality lip-sync with 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 (Tian et al., 2024; Xu et al., 2024a; Jiang et al., 2024; Lin et al., 2025; Chen et al., 2025b; Wei et al., 2025; Zhou et al., 2025b; Bigata et al., 2025a; Li et al., 2025; Hong et al., 2025; Zhong et al., 2025; Guan et al., 2025; Yi et al., 2025; Yang et al., 2025; Liu et al., 2024; 2025a; Qin et al., 2025; Bigata et al., 2025b; Mu et al., 2025; Sun et al., 2024; Wang et al., 2025a; Chen et al., 2025c). Early works in this domain were often based on U-Net architectures from Stable Diffusion (Rombach et al., 2022). For instance, EMO (Tian et al., 2024) integrates audio conditions via an additional cross-attention layer, while Loopy (Jiang et al., 2024) designs a past-frame compression scheme to better support longer video generation. More recent works have shifted towards Transformer-based backbones (DiT), such as Hallo (Xu et al., 2024a), HunyuanVideo-Avatar (Chen et al., 2025b), and MOCHA (Wei et al., 2025), 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 (Drobyshev et al., 2022; 2024; Xu et al., 2024b; Zhu et al., 2025; Lu et al., 2021; Zhou et al., 2020; Zhang et al., 2023; Wang et al., 2025b; Qi et al., 2025; Huang et al., 2025; Chen et al., 2025a; Gong et al., 2025; Fu et al., 2025; Shen et al., 2025; Liu et al., 2025b; Zhen et al., 2025; Liu et al., 2025c; Guo et al., 2025). Representative works such as MegaPortraits (Drobyshev et al., 2022), VASA (Xu et al., 2024b), and INFP (Zhu et al., 2025) have explored various designs for a robust, identity-agnostic facial motion space. Recent variants further improve flexibility and style controllability, including OmniTalker (Wang et al., 2025b), and ChatAnyone (Qi et al., 2025). 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.

Streaming and Real-time Talking Heads Autoregressive models such as FaceFormer (Fan et al., 2022) and CodeTalker (Xing et al., 2023) have long been used for speech-driven facial animation, but they run offline with non-causal audio encoders over the whole utterance. Recent real-time systems, *e.g.*, Live Avatar (Huang et al., 2026) and SoulX-FlashHead (Yu et al., 2026), distill large video diffusion models into causal few-step generators that output a chunk of frames per step, and thus must wait for each audio chunk (0.96 s for SoulX-FlashHead). PersonaLive (Li et al., 2026) adopts micro-chunk streaming, but for video-driven portrait animation. In contrast, DyStream generates motion frame by frame with a bounded audio lookahead, responding after a single short audio packet.

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 (Park et al., 2024) generated speaker behaviors conditioned on the user's previous utterance, but failed to produce synchronized listener reactions. To jointly model both roles, ViCo-X (Zhou et al., 2025a) proposes a Role Switcher bridging separate generators for speaker and listener, but the explicit switching often caused unnatural transitions. DIM (Tran et al., 2024) pretrained a unified model to capture dyadic context but still required manual role assignment during adaptation. AgentAvatar (Wang et al., 2023a) synthesized photo-realistic avatars driven by textual descriptions, but its motions lacked precise audio alignment. Recent person-specific studies (Sun et al., 2025; Ng et al., 2024; Huang et al., 2024a; Yan et al., 2024) achieve high realism but perform limited on unseen identities. INFP (Zhu et al., 2025) employs a unified generator to learn dynamic conversational states from dual-track audio without explicit role labeling, yet its clip-based paradigm limits real-time applications. Like INFP, we use a single, unified model to generate both speaker and listener behaviors from dual-track audio, without role-switching rules or multiple expert models; our main difference from INFP lies in streaming generation rather than in the dyadic formulation itself.

3 METHODOLOGY

Given a single image and a streaming dyadic audio signal, our method generates streaming talking-head videos with a two-stage framework: an off-the-shelf model based on LIA (Wang et al., 2024b) disentangles appearance and motion from the source image (Section 3.1), and a flow matching-based autoregressive model generates motion from audio (Section 3.2). For real-time responsiveness, we design a causal audio encoder with a lookahead module to minimize its reliance on future audio (Section 3.3).

3.1 MOTION-AWARE AUTOENCODER

The first stage contains an Autoencoder (AE) to disentangle a source image $\mathbf{I}_s$ into 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. The AE includes an appearance encoder ($\mathcal{E}_{app}$), a motion encoder ($\mathcal{E}_m$), a flow estimator ($\mathcal{F}$), and a decoder ($\mathcal{D}_{vae}$), trained jointly in a self-supervised manner by reconstructing a driving video from the appearance of a source image and the motion of the driving video itself. We adopt the architecture and training objective from LivePortrait (Guo et al., 2024) and LIA (Wang et al., 2024b).

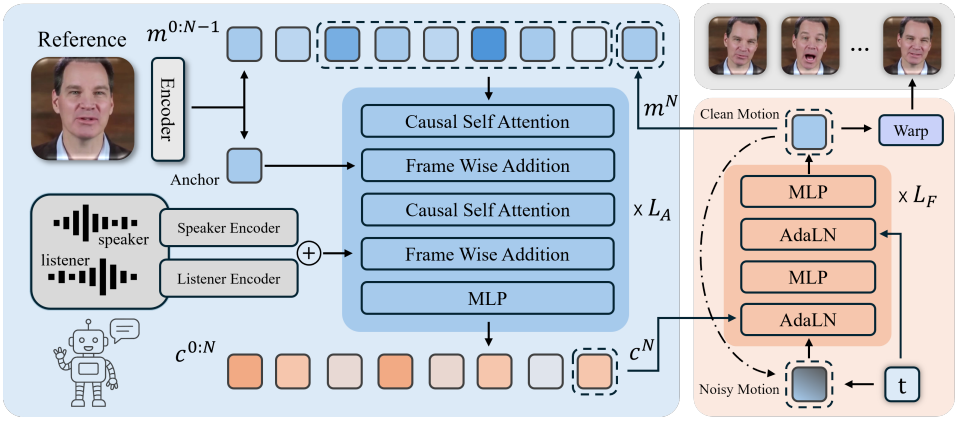

Figure 2: **The architecture of our audio-to-motion generator.** An autoregressive network (blue) of causal self-attention and MLP blocks processes the audio, anchor, and previous motion into a condition c^N, which is injected via AdaLN into a flow matching head (orange) of stacked MLPs to produce the motion m^N with multi-step flow matching. m^N warps the reference image into the output frame and is fed back into the autoregressive network.

The motion encoder $\mathcal{E}_m$ extracts a source motion feature $\mathbf{m}_s$ from $\mathbf{I}_s$ and driving motion features $\mathbf{m}_{dri}^{1:N}$ from a driving video $\mathcal{V}_{dri}$. The flow estimator $\mathcal{F}$ then predicts a dense flow field $\mathbf{F}_{s\to d}$, with which the decoder $\mathcal{D}_{vae}$ warps the appearance feature $\mathbf{v}_{app}$ to synthesize the frames $\hat{\mathbf{I}}_{pred}^{1:N}$:

$$\begin{aligned} \mathbf{m}_s = \mathcal{E}_m(\mathbf{I}_s), \quad \mathbf{v}_{app} = \mathcal{E}_{app}(\mathbf{I}_s), \quad \mathbf{m}_{dri}^{1:N} = \mathcal{E}_m(\mathcal{V}_{dri}), \\ \mathbf{F}_{s\to d} = \mathcal{F}(\mathbf{m}_s, \mathbf{m}_{dri}^{1:N}), \quad \hat{\mathbf{I}}_{pred}^{1:N} = \mathcal{D}_{vae}(\text{Warp}(\mathbf{v}_{app}, \mathbf{F}_{s\to d})). \end{aligned} \tag{1}$$

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 C.

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 Li et al. (2024b), the AR model regresses the continuous, deterministic feature, and a single frame-level flow matching head is for stochastic modeling. This targets the one-to-many nature of dyadic motion, especially for the listener: the same conversational audio admits many plausible expressions and head movements, so a deterministic regressor trained with an ℓ_2 loss tends to collapse to their conditional mean, whereas the flow matching head samples the next-frame motion from its conditional distribution. 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 (Wan et al., 2025). 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 (Baevski et al., 2020) 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 a frame-by-frame autoregressive model, our approach inevitably faces error accumulation in long sequences: the head gradually drifts from its initial position in the reference image, causing severe artifacts. Since our inference generates only one new frame per forward pass in the sliding window, with previously generated frames as historical context, we randomly select one of the last ten frames of each training sequence as the appearance anchor, while always using the initial reference frame as the anchor during inference. This randomized anchor training effectively mitigates 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:

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

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 ($\mathbf{S}$), listener ($\mathbf{L}$), and anchor ($\mathbf{R}$) 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 rearranging the standard CFG formulation:

$$\begin{aligned}\mathbf{m}_{\text{cond}} = (1 - w_\Sigma)\hat{\mathbf{m}}_{\text{uncond}} + w_s\hat{\mathbf{m}}(\mathbf{S}, \mathbf{0}, \mathbf{0}) + w_l\hat{\mathbf{m}}(\mathbf{0}, \mathbf{L}, \mathbf{0}) \\ + w_r\hat{\mathbf{m}}(\mathbf{0}, \mathbf{0}, \mathbf{R}) + w_{\text{all}}\hat{\mathbf{m}}(\mathbf{S}, \mathbf{L}, \mathbf{R}),\end{aligned} \tag{3}$$

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 our model could generate video in a causal frame-by-frame manner in theory, its lip-sync performance is limited with a straightforward implementation, because standard audio encoders such as Wav2Vec2 (Baevski et al., 2020) require non-causal context to extract current features, and enforcing causality during inference severely degrades these features. As shown in Table 3, audio-lip synchronization improves as the current audio frame can access more future audio frames. We therefore customize a causal Wav2Vec2 with a minimal lookahead module, restricting its receptive field to at most 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 GroupNorm 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) (Su et al., 2024) into the Transformer attention layers.

These modifications remove implicit, hard-coded lookaheads from the architecture, so the temporal receptive field is governed solely by the attention mask, giving explicit control over how much future audio 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 (Hurst et al., 2024) 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, adding no latency beyond receiving the packet itself.

In contrast, the listener track uses a purely causal attention mask: the average human reaction time in a conversation is approximately 300 milliseconds (Meyer, 2023), and unlike a speaker, a listener does not require future audio context to determine precise lip movements, so its responses are based only on past and current audio.

Training Details. The Rope-Wav2Vec2 model with a full attention mask is trained from scratch with the same pre-training task as the original Wav2Vec2. The restricted-attention variants are not trained from scratch but distilled from this full-attention model as the teacher, and the distilled models are then used for our generation task.

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS

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 with a hidden dimension of 768 and 8 attention heads, and the Flow matching Head is composed of 6 blocks with a hidden dimension of 768. For inference, we employ a 5-step Euler ODE sampler with CFG (Equation 3). Training details and hyperparameters are provided in Appendix E.

4.2 DATASET

We use two types of data for our experiments: public datasets and an additional collected dataset. Our data preprocessing pipeline is similar to that of LatentSync (Li et al., 2024a). The core steps include detecting and cropping face regions from raw videos, followed by filtering out clips with low SyncNet (Chung & Zisserman, 2016) 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 (Zhang et al., 2021) and Realtalk (Geng et al., 2023). After processing, our open-source training set contains approximately 50k video clips, each with a duration of about 8 seconds. For RealTalk, we prepare an audio track and a motion sequence for each participant, and randomly choose one participant as the target in each training sample: the target's own audio serves as the speaker track and the partner's audio as the listener track, so no role labels are required. We do not annotate silence, turn-taking, overlap, or backchannels; the model learns these behaviors from data. For single-speaker data such as HDTF, the listener track is set to zero (unconditional). Our collected 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

Unless stated otherwise, Sec. 4.3–4.5 use the offline version, *i.e.*, the audio encoder can see future audio; the online version and its latency are discussed in Sec. 4.6 and 4.7. Following VBench (Huang et al., 2024b), we use automated, no-reference metrics for lip-sync quality (Chung & Zisserman, 2016), content similarity (Radford et al., 2021), image quality, and motion dynamics (Huang et al., 2024b), detailed in Appendix D.

We compare DyStream with several representative or state-of-the-art talking-head generation models: SadTalker (Zhang et al., 2023), Real3DPortrait (Ye et al., 2024), Hallo3 (Cui et al., 2025), Sonic (Ji et al., 2025), and our re-implementation of INFP (Zhu et al., 2025) (INFP*). As the code, weights, and DyConv data of INFP are not public, we re-implement it strictly following its paper; INFP* may not match the original model, but its Sync-C is higher than that of SadTalker, a basic validation. 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.

For a fair comparison, we establish two additional baselines designed for frame-by-frame generation. The first, which we denote as INFP-I (INFP-Inpainting), is an adaptation of our reproduced INFP

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 re-implementation. INFP_I* generates n frames per step, *i.e.*, a latency of $n \times 40$ ms; for our online variants, we report the speaker-audio lookahead.

	Sync-C↑	CS↑	Quality↑	Dynamic↑
GroundTruth	7.614	0.928	0.642	0.660
Hallo3 (Cui et al., 2025)	6.814	0.915	0.638	**0.870**
Sonic (Ji et al., 2025)	**8.495**	0.935	0.626	0.600
SadTalker (Zhang et al., 2023)	6.704	**0.965**	**0.697**	0.080
Real3DPortrait (Ye et al., 2024)	6.811	0.943	0.637	0.030
INFP* (Zhu et al., 2025)	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	<u>8.136</u>	<u>0.947</u>	<u>0.642</u>	<u>0.580</u>
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 (lookahead 40 ms)	6.896	0.937	0.647	0.570
Ours (lookahead 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 (Zhu et al., 2025)	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

baseline, chosen because INFP is the closest method to ours in supporting both speaker and listener generation and no open-source streaming dyadic method is available. 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, in our 25 FPS setting. This strategy must be applied in both training and inference, as applying it only at inference to a model trained on full chunks causes severe jitter. The second baseline removes the flow matching head from our method, yielding a deterministic model that autoregressively predicts the motion frame by frame.

Results. In Table 1, our method achieves the best Sync-C among warp-based methods, second only to the diffusion-based Sonic, while INFP_I* degrades sharply as its latency decreases. No method is best on all metrics: CS measures similarity to the first frame and favors nearly static outputs (*e.g.*, SadTalker has the highest CS but a Dynamic score of 0.080), while INFP* and Hallo3 exceed the Dynamic score of real videos. Our CS, Quality, and Dynamic stay close to those of real videos, and limiting the lookahead mainly affects lip-sync rather than visual quality or dynamics.

4.4 Evaluation of Listener

To quantitatively evaluate the listener performance, we conduct evaluations on the RealTalk test set. Following the protocol of prior listener generation works (Ng et al., 2022; Zhu et al., 2025), we only keep clips in which the target participant listens while only the partner speaks; out of the 407 test clips (1,896 clips in total), 38 satisfy this condition, ranging from 3 to 14 seconds in length. We could not additionally evaluate on ViCo (Zhou et al., 2022), as its public download link is no longer available. We compare our method against our INFP re-implementation. On facial expression coefficients (Exp) and head poses (Pose) extracted with MediaPipe (Lugaresi et al., 2019), we report Fréchet Distance (FD) and Mean Squared Error (MSE) for realism and accuracy, and Variance (Var) and SI Diversity (SID) (Ng et al., 2022) for diversity (Appendix D).

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	∞	∞	8.114
Wav2vec2 (random init)	∞	∞	4.568
Wav2vec2 (freeze)	∞	∞	6.037
Rope-Wav2vec2	∞	∞	8.136
Hubert	∞	∞	8.007
Wavlm	∞	∞	7.719
Wav2vec	0	0	1.761
Mimi	0	0	5.628
Encodec(24KHz)	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
Rope-Wav2vec (lookahead)	0	0	3.017
	20	20	5.976
	40	40	6.896
	60	60	7.387
	80	80	7.671
Rope-Wav2vec2-VAE	0	0	3.495

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 cross-attention. As shown in Tables 1 and 4, this yields better lip-sync, suggesting that audio features within a local window suffice for lip-syncing, and increases inference speed by 23.1%.

Anchor Frame Selection. Sampling the training anchor from the last 10 frames (ours) prevents long-term pose drift more effectively than sampling it from the entire sequence or using no anchor (Appendix B).

Flow Matching Head. Removing the flow matching head converts our model into a deterministic autoregressive model. As detailed in Tables 1 and 2, the listener mode then shows a notable decrease in motion diversity, while the speaker mode degrades only marginally. We hypothesize that listening 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, where the choice of audio encoder is critical. We evaluate non-causal encoders, Wav2vec2 (Baevski et al., 2020), Hubert (Hsu et al., 2021), and WavLM (Chen et al., 2022), and causal encoders, Wav2vec (Schneider et al., 2019), Mimi (Défossez et al., 2024), Encodec (Défossez et al., 2022), and our Rope-Wav2vec2.

As shown in Table 3 (group 1), wav2vec2, Hubert, and WavLM yield comparable performance, which we attribute to their similar architectures and pre-training objectives, while a randomly initialized wav2vec2 leads to a substantial drop, confirming that a well-pre-trained audio encoder is critical. Furthermore, unlike INFP, which freezes the audio encoder, we finetune it during training; freezing it significantly degrades performance, which likely explains why our method outperforms INFP even without more complex operations on audio features, such as a memory bank.

We then compare four design choices for making the audio encoder causal: (1) directly using pretrained causal encoders (Mimi, EnCodec-24kHz, and Wav2Vec); (2) distilling a causal Rope-Wav2Vec2 from its full-attention counterpart; (3) allowing limited lookahead by modifying the attention mask; and (4) training Rope-Wav2Vec2 with generative objectives, such as a VAE-style loss during distillation. The results in Table 3 (groups 2–4) strongly suggest that for high-fidelity

Table 4: **Inference performance** on a 4090 GPU. APD: delay from audio packet arrival to motion output (incl. chunk buffering for INFP, excl. rendering). Step: flow matching steps. Ours: 60 ms lookahead.

Method	Step	FPS	APD(s)	Sync-C↑
INFP	5	31.15	0.9921	6.104
INFP	10	15.17	1.0259	6.381
Ours	1	84.75	0.0118	7.165
Ours	5	68.49	0.0146	7.371
Ours	10	57.80	0.0173	7.387
Ours (w/o Add)	5	52.08	0.0192	7.119

Table 5: **Comparison with SoulX-FlashHead** on HDTF-100. Resp.: required audio packet duration plus the time to generate the first output block.

Method	Sync-C↑	FPS↑	Packet(s)↓	Resp.(s)↓
FlashHead-Lite	6.394	**96**	0.96	1.21
FlashHead-Pro	**7.438**	10	0.96	3.27
Ours	7.371	68	**0.10**	**0.1146**

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.

4.7 Latency Analysis and Comparison with Streaming Methods

Latency definitions. (1) *Per-frame generation time*: audio feature extraction (≈4 ms), motion generation (Table 4), and warping-based rendering (≈10 ms) for one frame once its audio is available, with the reference appearance feature extracted once and cached; it stays below 34 ms, shorter than the 40 ms frame interval at 25 FPS. (2) *Model-side latency*: the delay from the current audio timestamp to the delivery of its rendered frame, *i.e.*, the 60 ms speaker-audio lookahead plus the per-frame generation time, below 100 ms. (3) *Response time*: the required audio packet duration plus the time to generate the first output block, measuring how fast a model first responds rather than its throughput. These exclude audio capture, upstream speech generation, dialogue management, network transmission, video encoding, and display buffering, which depend on the deployment.

Comparison with streaming methods. Our model runs at 68.49 FPS with an APD of 14.6 ms (5 steps), whereas the chunk-based INFP must buffer a 0.96 s audio chunk (Table 4). As no open-source warping-based streaming method is available, we also compare with the recent real-time SoulX-FlashHead (Yu et al., 2026) (Table 5). Generating 24 frames (0.96 s) per step gives its Lite variant a higher throughput, but each step must wait for 0.96 s of audio, leading to response times of 1.21 s (Lite) and 3.27 s (Pro), versus 0.11 s for DyStream with a Sync-C comparable to Pro. Our VBench Image Quality on HDTF-100 (0.642) is also on par with Lite (0.636) and Pro (0.640). Although SoulX-FlashHead builds on a much larger 1.3B-parameter DiT, its response time cannot fall below the 0.96 s chunk regardless of computation speed.

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 accessories like jewelry or glasses. Second, it is challenging to deal with hand occlusion, and the generated videos exhibit static hand motion (Appendix Figure 5). Additionally, the input face must be cropped and aligned to a fixed position and scale, similar to the training data. Moreover, while the flow matching head improves short-term motion diversity (Table 2), the model lacks long-term diversity of speaking style, *e.g.*, semantic control via text or emotion labels, largely due to the scarcity of continuous, long-form training data. Finally, our listener evaluation relies on automatic motion statistics, which do not directly measure whether reactions are timely, natural, and appropriate; a user study and stratified evaluations under overlapping speech, crosstalk, or leakage between the two audio tracks remain future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript translation, language polishing, and writing visualization code. Our use of generative AI tools was limited to these tasks. We have reviewed all AI-assisted text and checked the visualization code and its outputs. We take full responsibility for the final content of this work, including all AI-assisted text, code, and visualizations.

ETHICS STATEMENT

This work generates realistic talking-head videos of a person from a single reference image and audio, which could be misused to create deceptive content or to impersonate individuals without consent. Our experiments use the public HDTF and RealTalk datasets together with an additional collected dataset that is used only for training and not released. We encourage responsible use of the released model, including obtaining consent from depicted individuals and clearly labeling generated content, and we support ongoing research on the detection of synthetic media.

REPRODUCIBILITY STATEMENT

We describe the model architecture and training objectives in Section 3, and the causal audio encoder with the lookahead module in Section 3.3. Model sizes, hyperparameters, optimizers, and training budgets are listed in Section 4.1 and Appendix E, the latency measurement is defined in Section 4.7, the data processing pipeline is described in Section 4.2, and the details of the autoencoder and all evaluation metrics are given in Appendix C and Appendix D. The model, weights, and code will be released.

REFERENCES

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.

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*, 2025a.

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*, pp. 5477–5488, 2025b.

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*, 2025a.

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.

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*, 2025b.

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*, volume 39, pp. 2403–2410, 2025c.

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, 2016.

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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21086–21095, 2025.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

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.

Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 2663–2671, 2022.

Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8498–8507, 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.

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

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

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*, pp. 5523–5534, 2025.

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*, pp. 10678–10689, 2025.

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.

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.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Fa-Ting Hong, Zunnan Xu, Zixiang Zhou, Jun Zhou, Xiu Li, Qin Lin, Qinglin Lu, and Dan Xu. Audio-visual controlled video diffusion with masked selective state spaces modeling for natural talking head generation. *arXiv preprint arXiv:2504.02542*, 2025.

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.

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.

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*, 2024a.

Yubo Huang, Hailong Guo, Fangtai Wu, Shifeng Zhang, Shijie Huang, Qijun Gan, Lin Liu, Sirui Zhao, Enhong Chen, Jiaming Liu, and Steven Hoi. Live avatar: Streaming real-time audio-driven avatar generation with infinite length. In *European Conference on Computer Vision*, 2026.

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*, pp. 21807–21818, 2024b.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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*, pp. 193–203, 2025.

Jianwen Jiang, Chao Liang, Jiaqi Yang, Gaojie Lin, Tianyun Zhong, and Yanbo Zheng. Loopy: Taming audio-driven portrait avatar with long-term motion dependency. *arXiv preprint arXiv:2409.02634*, 2024.

Chunyu Li, Chao Zhang, Weikai Xu, Jingyu Lin, Jinghui Xie, Weiguo Feng, Bingyue Peng, Cunjian Chen, and Weiwei Xing. Latentsync: Taming audio-conditioned latent diffusion models for lip sync with syncnet supervision. *arXiv preprint arXiv:2412.09262*, 2024a.

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, 2024b.

Tianqi Li, Ruobing Zheng, Minghui Yang, Jingdong Chen, and Ming Yang. Ditto: Motion-space diffusion for controllable realtime talking head synthesis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9704–9713, 2025.

Zhiyuan Li, Chi-Man Pun, Chen Fang, Jue Wang, and Xiaodong Cun. Personalive! expressive portrait image animation for live streaming. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18118–18128, 2026.

Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, Chao Liang, Yuan Zhang, and Jingtuo Liu. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13847–13858, 2025.

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.

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

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*, 2025b.

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*, 2025c.

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

Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (ToG)*, 40(6):1–17, 2021.

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.

Antje S. Meyer. Timing in conversation. *Journal of Cognition*, 6(1), 2023. doi: 10.5334/joc.268. URL https://doi.org/10.5334/joc.268.

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. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10437–10446, 2025.

Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20395–20405, 2022.

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*, pp. 1001–1010, 2024.

Se Jin Park, Chae Won Kim, Hyeongseop Rha, Minsu Kim, Joanna Hong, Jeong Hun Yeo, and Yong Man Ro. Let's go real talk: Spoken dialogue model for face-to-face conversation. *arXiv preprint arXiv:2406.07867*, 2024.

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.

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. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 8253–8262, 2025.

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*, pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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

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.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

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*, 133(5):2910–2926, 2025.

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

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pp. 244–260. Springer, 2024.

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

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

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

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*, pp. 17979–17989, 2023b.

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*, pp. 26212–26221, 2025a.

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, 2024a.

Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Lia: Latent image animator. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024b.

Zhongjian Wang, Peng Zhang, Jinwei Qi, Guangyuan Wang, Chaonan Ji, Sheng Xu, Bang Zhang, and Liefeng Bo. Omnitalker: One-shot real-time text-driven talking audio-video generation with multimodal style mimicking. *arXiv preprint arXiv:2504.02433*, 2025b.

Cong Wei, Bo Sun, Haoyu Ma, Ji Hou, Felix Juefei-Xu, Zecheng He, Xiaoliang Dai, Luxin Zhang, Kunpeng Li, Tingbo Hou, et al. Mocha: Towards movie-grade talking character synthesis. *arXiv preprint arXiv:2503.23307*, 2025.

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.

Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024a.

Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *Advances in Neural Information Processing Systems*, 37:660–684, 2024b.

Yichao Yan, Zanwei Zhou, Zi Wang, Jingnan Gao, and Xiaokang Yang. Dialoguenerf: Towards realistic avatar face-to-face conversation video generation. *Visual Intelligence*, 2(1):24, 2024.

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*, volume 39, pp. 9256–9264, 2025.

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.

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.

Tan Yu, Qian Qiao, Le Shen, Ke Zhou, Jincheng Hu, Dian Sheng, Bo Hu, Haoming Qin, Jun Gao, Changhai Zhou, Shunshun Yin, and Siyuan Liu. Soulx-flashhead: Oracle-guided generation of infinite real-time streaming talking heads. *arXiv preprint arXiv:2602.07449*, 2026.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. 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*, pp. 8652–8661, 2023.

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.

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*, pp. 21075–21085, 2025.

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*, pp. 3101–3110, 2025.

Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, Tiejun Zhao, and Tao Mei. Responsive listening head generation: A benchmark dataset and baseline. In *European Conference on Computer Vision*, 2022.

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

Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, 39(6): 1–15, 2020.

Ziqi Zhou, Weize Quan, Hailin Shi, Wei Li, Lili Wang, and Dong-Ming Yan. Gohd: Gaze-oriented and highly disentangled portrait animation with rhythmic poses and realistic expressions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10914–10922, 2025b.

Yongming Zhu, Longhao Zhang, Zhengkun Rong, Tianshu Hu, Shuang Liang, and Zhipeng Ge. Infp: Audio-driven interactive head generation in dyadic conversations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10667–10677, 2025.

APPENDIX

This supplemental document contains six sections:

- Video and Codes (Section A).
- Head Pose Stability in Long Video Generation (Section B).
- Disentangled AE for Motion Representation (Section C).
- Metric Definitions (Section D).
- Training Details (Section E).
- Guidance Weights (Section F).

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, including out-of-domain portraits such as AI-generated faces, cartoons, sculptures, oil paintings, and sketches.

B HEAD POSE STABILITY (ANCHOR)

To quantitatively evaluate the anti-drifting efficacy. We measure the 6-DoF head pose drift relative to the initial frame using MediaPipe, and report Mean/Max Drift (average/maximum deviation) and Std Drift (fluctuation stability). We test on 10 audio clips (3 minutes each) driven by 10 different reference images. As shon in Figure 3 and Table 6, our method demonstrates exceptional stability across all durations (10s, 60s, 180s). Specifically, our rotation drift remains under 8° even after 3 minutes, significantly outperforming the "Random Anchor" and "No Anchor" baselines, which suffer from severe cumulative error (drifting $> 34^\circ$).

Table 6: Head Pose Drift Evaluation. We evaluate head pose stability across three different anchor strategies. The metrics reported for Rotation are in degrees, and for Position are in relative units from MediaPipe. Lower values indicate better stability.

Duration	Method	Rotation (degrees)			Position		
		Mean Drift ↓	Max Drift ↓	Std Drift ↓	Mean Drift ↓	Max Drift ↓	Std Drift ↓
10s	Ours	**6.61**	**12.48**	**2.25**	**1.04**	**2.45**	**0.53**
	Random Anchor	11.65	20.76	3.83	1.62	3.32	0.74
	Without Anchor	30.12	41.08	8.47	2.19	4.17	0.84
60s	Ours	**7.73**	**17.61**	**2.45**	**1.16**	**3.26**	**0.57**
	Random Anchor	12.69	24.27	3.49	1.62	3.87	0.76
	Without Anchor	33.96	43.25	4.59	2.38	4.56	0.73
180s	Ours	**7.62**	**19.31**	**2.72**	**1.20**	**3.69**	**0.59**
	Random Anchor	12.86	26.90	4.62	1.62	4.09	0.75
	Without Anchor	34.52	44.45	3.31	2.42	4.68	0.71

C 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})



Figure 3: **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 without anchor.

with the same network architecture as LivePortrait (Guo et al., 2024). 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 (Wang et al., 2023b; 2024b).

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 Wang et al. (2024b) for more details of this architecture.

Training Objective. The VAE is trained with a composite loss function. We adopt the base losses from LIA (Wang et al., 2024b), which include reconstruction ($\mathcal{L}_{rec}$), perceptual ($\mathcal{L}_{feat}$), and adversarial ($\mathcal{L}_{adv}$) losses. We adopt two additional losses following Drobyshev et al. (2022): 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 (Drobyshev et al., 2022) 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:

$$\mathcal{L}_{VAE} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{feat}\mathcal{L}_{feat} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{gaze}\mathcal{L}_{gaze} \quad (4)$$

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

D Metric Definitions

Lip-sync Consistency (Sync-C): We employ the confidence score derived from SyncNet (Chung & Zisserman, 2016), referred to as SyncScore, 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 (Teed & Deng, 2020) 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 (Ng et al., 2022), we assess the diversity of generated facial and head motions by applying k-means clustering to the ground-truth motions ($k = 40$ for expression and $k = 20$ for rotation), and then computing the entropy of the histogram of cluster IDs within each generated sequence. Higher values indicate more diverse motions.

Variance (Var): Following L2L (Ng et al., 2022), 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 the motion generation delay 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 therefore excludes the warping-based rendering, which is included in the latency breakdown of Section 4.7. 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.

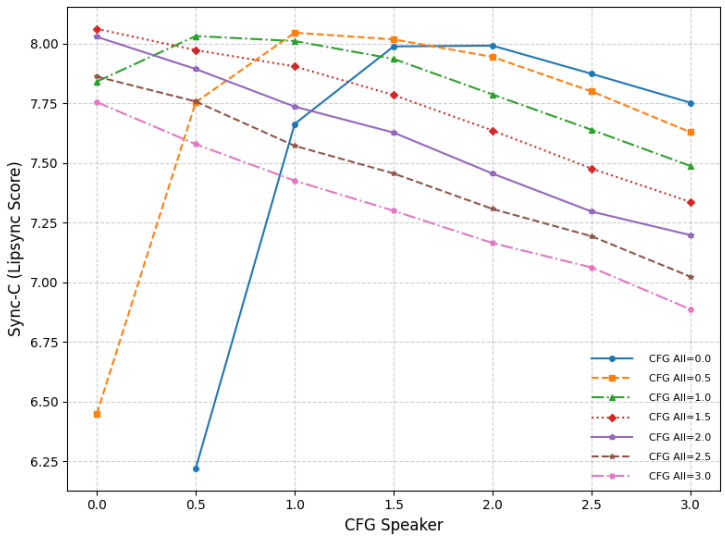

Figure 4: Effect of classifier-free guidance (CFG) on lipsync performance. We evaluate Sync-C across a grid of CFG All and CFG Speaker.

E Training Details

Autoencoder Training. We train the AE using the AdamW optimizer (Loshchilov & Hutter, 2017) 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) (Ho & Salimans, 2022), as described in Equation 3, with guidance weights set to $w_{all} = 1.0$, $w_s = 0.5$, $w_l = 0.5$, and $w_r = 0.5$ (see Appendix F). The model is trained for 120,000 iterations using 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.

F Guidance Weights

The lip-sync performance is mainly determined by the guidance weights w_{all} and w_s. As shown in Figure 4, 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.



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.