ALIVE-AVATAR: STREAMING AUDIO-VIDEO AVATAR GENERATION WITH PERSISTENT IDENTITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-video avatar generation is moving toward real-time, interactive streaming, where the core challenge is to sustain fast response, interactive control, and a persistent identity together over long sessions. We present **ALIVE-AVATAR**, a streaming audio-video avatar generation model that meets these demands. We first adapt the pretrained ALIVE base model through prompt-to-shot teacher forcing to enable promptable shot streaming for interactive control. Then, we design a distillation curriculum, initialized with consistency distillation and refined by our PCM-regularized DMD2, that attains high-quality generation in four sampling steps. At inference, our streaming pipeline uses Context Multiplicity Modulation to regulate attention to recent generated context, mitigating identity drift while retaining bounded historical memory. Together, these designs yield a four-step, 4-NFE model that, on our long-sequence benchmark, outperforms strong avatar baselines in identity consistency, face quality, and speech correctness, renders 480p audio-video at 24 fps faster than playback, and preserves identity even in one-hour rollouts. Building on ALIVE-AVATAR, we further construct a live-streaming agent that delivers planned content and answers viewer comments, demonstrating ALIVE-AVATAR's practicality for real-world avatar live streaming.

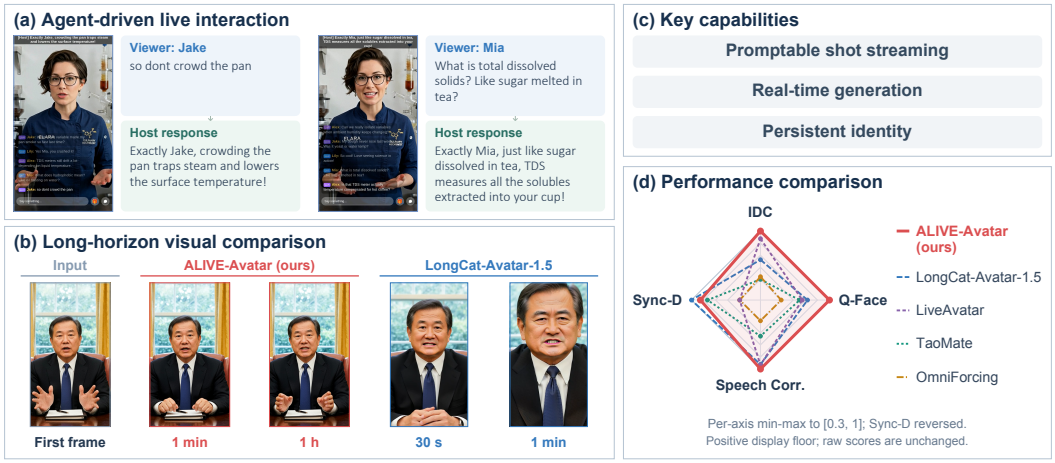

Figure 1: **Interactive avatar streaming with persistent identity.** **(a)** Agent-driven live interaction with enlarged, verbatim viewer comments and host responses. **(b)** Input frame, ALIVE-AVATAR at 1 min and 1 h, and LongCat-Video-Avatar-1.5 at 30 s and 1 min. **(c)** Promptable shot streaming, real-time generation, and persistent identity. **(d)** First-frame-conditioned benchmark scores from Table 1, min–max normalized with Sync-D reversed. All depicted identities are AI-generated.

1 INTRODUCTION

Avatar generation animates digital characters with natural speech, expression, and motion, supporting applications such as virtual assistants and live streaming (Gan et al., 2025; Zeng et al., 2026). While earlier work focused on audio-driven avatar animation (Gao et al., 2025b; Tian et al., 2024), recent

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.

studies have explored joint audio-video generation (Guo et al., 2026; Low et al., 2025; HaCohen et al., 2026). Meanwhile, interactive applications such as live-streaming agents require real-time, streaming, and long-horizon generation. Existing methods extend sequences via sliding windows (Ruhe et al., 2024), anchors (Yang et al., 2026), and generated-history training (Huang et al., 2026), and accelerate sampling through few-step distillation (Yin et al., 2024b;a). Yet jointly achieving real-time speed, interactive control, and persistent identity remains challenging for audio-video avatar generation.

We present **ALIVE-AVATAR**, a streaming audio-video avatar generation model that meets these demands together, responding promptly while sustaining a persistent identity over long interactive sessions (up to one hour). Its capability rests on three pillars: **promptable shot streaming**, where each prompt drives one complete shot for natural interaction; **few-step generation** for real-time streaming; and **persistent-identity modeling** that maintains identity across long horizons.

Starting from pretrained ALIVE (Guo et al., 2026), we post-train it into ALIVE-AVATAR with autoregressive learning for promptable shot streaming and a distillation curriculum for real-time generation. Prior chunk-wise autoregressive video generation methods use fixed-length temporal blocks (Huang et al., 2026; Yin et al., 2025), whose boundaries are agnostic to complete utterances or actions. We instead make the unit of interaction the unit of generation, adopting prompt-to-shot teacher forcing (Hu et al., 2026; Gao et al., 2025a) to produce one complete, variable-length shot per prompt, while allowing prompt updates between shots to control subsequent speech and actions. When distilling this autoregressive model to four steps, direct DMD2 (Yin et al., 2024a) can alter reference identity, whereas consistency distillation (Lu & Song, 2025) can blur rapid motion. We therefore adopt a two-stage curriculum: consistency initialization followed by our proposed PCM-regularized DMD2 refinement. DMD2 improves visual sharpness, while PCM regularization (Wang et al., 2024) suppresses oversaturation. Together, the final model can generate high-quality, identity-preserving audio-video shots in four steps (4 NFE), enabling real-time streaming.

At inference, prior streaming methods can exhibit progressive identity drift during long-horizon generation (Huang et al., 2025; Gan et al., 2026). We therefore propose a streaming inference pipeline centered on Context Multiplicity Modulation (CMM) for long-term identity preservation. CMM softly downweights attention to recent generated context through virtual token multiplicities to limit appearance-error feedback while retaining context for motion continuity. With CMM, our pipeline can generate high-quality audio-video avatars with consistent identity over long horizons.

On our long-horizon benchmark, ALIVE-AVATAR achieves the best identity consistency, face quality, and speech correctness among the evaluated audio-driven and audio-video methods (Fig. 1(d)). It supports real-time generation of 480p audio-video at 24 fps while preserving identity for up to one hour without noticeable drift, as shown in Fig. 1(b). Building on these capabilities, we construct a live-streaming agent that delivers planned content and answers viewer comments (Fig. 1(a)).

Our contributions are fourfold: (1) We present ALIVE-AVATAR, an audio-video avatar model for long-horizon streaming and agent-driven interaction. (2) We introduce promptable shot streaming via prompt-to-shot teacher forcing, producing one complete, variable-length audio-video shot per prompt. (3) We achieve high-quality, four-step generation through a distillation curriculum combining consistency initialization with PCM-regularized DMD2 refinement. (4) We develop a streaming inference pipeline centered on CMM for long-term identity preservation with bounded memory.

2 PRELIMINARIES

Audio-Video Diffusion Models. Video and audio are first encoded by VAEs (Kingma & Welling, 2014) into clean latents $\boldsymbol{x}_0 = (\boldsymbol{x}_0^v, \boldsymbol{x}_0^a)$. We then jointly generate video and audio in latent space (Rombach et al., 2022) using flow matching (Liu et al., 2023; Lipman et al., 2023) [1]. Let $\mathcal{N}$ denote a Gaussian distribution; noisy latents $\boldsymbol{x}_t$ at time $t \in [0, 1]$ are obtained by shared linear interpolation:

$$\boldsymbol{x}_t = (1-t)\,\boldsymbol{x}_0 + t\,\boldsymbol{\epsilon}, \qquad \boldsymbol{x}_t = (\boldsymbol{x}_t^v, \boldsymbol{x}_t^a), \quad \boldsymbol{\epsilon} = (\boldsymbol{\epsilon}^v, \boldsymbol{\epsilon}^a) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \tag{1}$$

whose induced probability-flow ordinary differential equation (PF-ODE) (Song et al., 2021) is $\mathrm{d}\boldsymbol{x}_t/\mathrm{d}t = \boldsymbol{u}_\theta(\boldsymbol{x}_t, t, c)$, with $\boldsymbol{u}_\theta = (\boldsymbol{u}_\theta^v, \boldsymbol{u}_\theta^a)$ a network trained to predict the velocity per modality:

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{\boldsymbol{x}_0,\,\boldsymbol{\epsilon},\,t}\left[\lambda_v \left\|\boldsymbol{u}_\theta^v(\boldsymbol{x}_t, t, c) - (\boldsymbol{\epsilon}^v - \boldsymbol{x}_0^v)\right\|_2^2 + \lambda_a \left\|\boldsymbol{u}_\theta^a(\boldsymbol{x}_t, t, c) - (\boldsymbol{\epsilon}^a - \boldsymbol{x}_0^a)\right\|_2^2\right], \tag{2}$$

[1] Given their equivalence (Albergo et al., 2025), we do not distinguish flow matching from diffusion models.

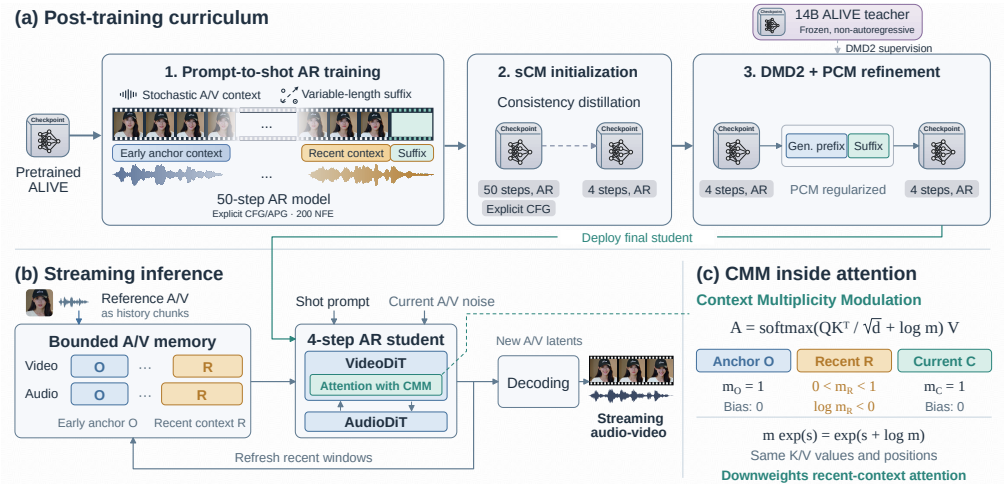

Figure 2: **Overview of ALIVE-AVATAR.** (a) Prompt-to-shot AR training, sCM initialization, and PCM-regularized DMD2 refinement. (b) Four-step streaming inference for identity-preserving audio-video generation. (c) CMM downweights recent-context attention to mitigate identity drift.

where $\lambda_v, \lambda_a > 0$ weight the modalities and c is the conditioning (*i.e.*, reference image, text prompts).

Few-Step Diffusion Distillation. To reduce inference costs for real-time interaction, multi-step diffusion teachers are commonly distilled into few-step students. In this work, we build on two representative distillation techniques: sCM (Lu & Song, 2025) and DMD (Yin et al., 2024b;a).

sCM is the continuous-time formulation of consistency models (Song et al., 2023) that regresses $\boldsymbol{F}_\theta$ toward the instantaneous PF-ODE tangent of a frozen teacher copy $\boldsymbol{f}_{\theta^-}$, where the clean-data predictor is parameterized as $\boldsymbol{f}_\theta(\boldsymbol{x}_t, t) = \cos(t)\boldsymbol{x}_t - \sin(t)\sigma_d \boldsymbol{F}_\theta\left(\boldsymbol{x}_t/\sigma_d,\, c_{\text{noise}}(t)\right)$, with $\boldsymbol{F}_\theta$ the underlying network, σ_d the data standard deviation, and $c_{\text{noise}}(\cdot)$ a fixed noise-level conditioning:

$$\mathcal{L}_{\text{sCM}}(\theta) = \mathbb{E}_{\boldsymbol{x}_0 \sim p_{\text{data}},\, \boldsymbol{\epsilon},\, t}\left[\left\| \boldsymbol{F}_\theta(\boldsymbol{x}_t, t) - \boldsymbol{F}_{\theta^-}(\boldsymbol{x}_t, t) - \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2 + c} \right\|_2^2\right], \boldsymbol{g} = w(t)\frac{\mathrm{d}\boldsymbol{f}_{\theta^-}(\boldsymbol{x}_t, t)}{\mathrm{d}t}. \quad (3)$$

Here, $\boldsymbol{x}_t$ is the noisy sample from clean data $\boldsymbol{x}_0$ and Gaussian noise $\boldsymbol{\epsilon}$ at time t, θ^- is a stop-gradient copy of θ, $w(t)$ is a time-dependent weight, $c > 0$ stabilizes the normalizer, and $\mathrm{d}\boldsymbol{f}_{\theta^-}/\mathrm{d}t$ is obtained via a single JVP. Minimising (3) aligns $\boldsymbol{F}_\theta$ with the teacher's PF-ODE tangent for few-step inference.

DMD optimizes the generator $\boldsymbol{F}_\theta$ by minimizing a time-averaged Kullback–Leibler (KL) divergence between the diffused generated and target distributions, p_{fake} and p_{real}, respectively. The resulting generator gradient is expressed in terms of the difference between their respective score functions:

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = -\mathbb{E}_{t,\boldsymbol{\epsilon}}\left[\left(\boldsymbol{s}_{\text{real}}(\Psi(\boldsymbol{x}_0', t), t) - \boldsymbol{s}_{\text{fake}}(\Psi(\boldsymbol{x}_0', t), t)\right)\frac{\mathrm{d}\boldsymbol{F}_\theta(\boldsymbol{\epsilon})}{\mathrm{d}\theta}\right], \quad (4)$$

where $\boldsymbol{x}_0' := \text{stopgrad}(\boldsymbol{F}_\theta(\boldsymbol{x}_t))$, Ψ applies forward diffusion at time t, and $\boldsymbol{s}_{\text{real}}$, $\boldsymbol{s}_{\text{fake}}$ are scores estimated by diffusion models $\boldsymbol{\mu}_{\text{real}}$ and $\boldsymbol{\mu}^\phi_{\text{fake}}$. During training, $\boldsymbol{\mu}_{\text{real}}$ is a frozen pre-trained teacher, while $\boldsymbol{\mu}^\phi_{\text{fake}}$ is updated jointly with G_θ via denoising matching on generated samples x_0': $\mathcal{L}^\phi_{\text{denoise}} = \|\mathcal{S}(\boldsymbol{\mu}^\phi_{\text{fake}}(\Psi(\boldsymbol{x}_0', t), t)) - \boldsymbol{x}_0'\|_2^2$, where $\mathcal{S}$ is the solver that maps the diffusion outputs to x_0'.

3 METHOD

Our framework combines three-stage post-training with streaming inference (see Fig. 2). Inference-aligned AR training (Sec. 3.1) adapts ALIVE for autoregressive generation. Consistency initialization (Lu & Song, 2025) (Sec. 3.2) distills it into a four-step student, and asymmetric PCM-regularized DMD2 refinement (Wang et al., 2024; Yin et al., 2024a) (Sec. 3.3) further improves generation quality. Streaming inference (Sec. 3.4) enables real-time, identity-preserving avatar generation.

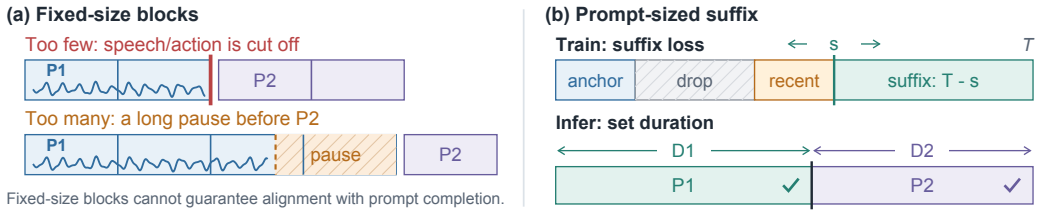

Figure 3: **From fixed-size blocks to promptable shots. (a)** Fixed-size blocks can truncate speech or action, or introduce unnatural pauses. **(b)** We sample s and supervise the full suffix $T - s$, retaining early anchors for identity and recent context for continuity, with stochastic middle dropout. At inference, each prompt is assigned duration D_i, targeting completion at the shot boundary.

3.1 Autoregressive Learning for Promptable Shot-Streaming

Interactive live streaming requires an avatar to respond to viewer feedback while sustaining a presentation. Autoregressive video generation typically advances through fixed-size temporal blocks (Huang et al., 2026; Yang et al., 2025b; Zhu et al., 2026), whose boundaries are agnostic to the units a viewer interacts with (Fig. 3(a)). Instead, our core insight is to make the unit of interaction the unit of generation: we introduce *promptable shot streaming*, where each prompt controls one complete audio-video shot sized to its utterance and action. Successive prompts let the avatar interleave replies with presentations without restarting the stream, powering an interactive live-streaming agent (Sec. E).

Variable-Length Shots with Dual-Timescale Context. As shown in Fig. 3(b), we sample a stride-aligned split s_i in a clip of T video-latent frames and supervise the complete suffix $\boldsymbol{x}_0^i := \boldsymbol{x}_0[s_i : T]$ of length $T_i = T - s_i$, where i indexes shots and audio follows the same physical-time interval. We obtain $\boldsymbol{x}_t^i$ using Eq. (1) with Gaussian noise $\boldsymbol{\epsilon}^i$ and a shared RF time t. To match streaming memory, context $\boldsymbol{x}_0^{<i,m}$ retains the first n_o^m and last n_r^m prefix frames with probability p_{drop}, or the full prefix otherwise, for $m \in \{v, a\}$. We independently sample window lengths from modality-specific discrete uniform ranges, clip windows to available history, and merge overlaps. Cross-modal attention is restricted to shared physical-time intervals; history receives no additive noise.

Prompt-to-Shot Teacher Forcing. To support promptable shot streaming, we adopt a prompt-to-shot formulation of teacher forcing (Hu et al., 2026; Gao et al., 2025a; Zhang et al., 2026b), pairing each complete audio-video shot with its own text prompt. For shot i, its prompt and a single reference image shared across the stream form the control condition c_i. We train the model to denoise the current shot conditioned on both c_i and the ground-truth temporal context $\boldsymbol{x}_0^{<i}$:

$$\mathcal{L}_{\text{AR}}(\theta) = \mathbb{E}_{\boldsymbol{x}_0,\, c_i,\, t,\, \boldsymbol{\epsilon}^i}\left[\sum_{m\in\{v,a\}} \lambda_m \left\| \boldsymbol{u}_\theta^m\left(\boldsymbol{x}_t^i, t, c_i, \alpha_t \boldsymbol{x}_0^{<i}\right) - \left(\boldsymbol{\epsilon}^{i,m} - \boldsymbol{x}_0^{i,m}\right)\right\|_2^2\right]. \tag{5}$$

Only current-shot tokens receive supervision. We use $\alpha_t = \max(1 - t, \alpha_{\min})$ in Eq. (5) to attenuate history according to the target's clean-signal scale while preserving minimum conditioning strength at high noise levels. Training enables inter-shot prompt switching for promptable shot streaming.

3.2 Identity-Preserving Consistency Initialization

DMD's mode-seeking reverse-KL objective makes refinement sensitive to the initial student distribution (Zheng et al., 2026b;a). We therefore initialize the student with consistency distillation, following prior autoregressive pipelines (Zhao et al., 2026; Zhang et al., 2026a).

Consistency Function. We use sCM to distill the 50-step AR model into a four-step student. To apply the TrigFlow-based sCM objective (Eq. (3)), we use an analytic RF-to-TrigFlow wrapper (Zheng et al., 2026b; Chen et al., 2025), preserving ALIVE's native RF inputs and time encoding:

$$\boldsymbol{F}_\theta(\boldsymbol{y}_\tau^i, \tau, c, \alpha_t \boldsymbol{x}_0^{<i}) = (\cos\tau - \sin\tau)\frac{\boldsymbol{x}_t^i}{\sigma_d} + \frac{\boldsymbol{u}_\theta(\boldsymbol{x}_t^i, t, c, \alpha_t \boldsymbol{x}_0^{<i})}{\sigma_d(\cos\tau + \sin\tau)}, \quad \boldsymbol{y}_\tau^i = (\cos\tau + \sin\tau)\frac{\boldsymbol{x}_t^i}{\sigma_d}, \tag{6}$$

where $\tau = \arctan\left(t/(1-t)\right)$ is the TrigFlow time, with $\tau = \pi/2$ for $t = 1$, and α_t is the scale factor in Eq. (5). The wrapper preserves the clean prediction, $\boldsymbol{f}_\theta = \sigma_d\left(\cos\tau\, \boldsymbol{y}_\tau^i - \sin\tau\, \boldsymbol{F}_\theta\right) = \boldsymbol{x}_t^i - t\boldsymbol{u}_\theta$.

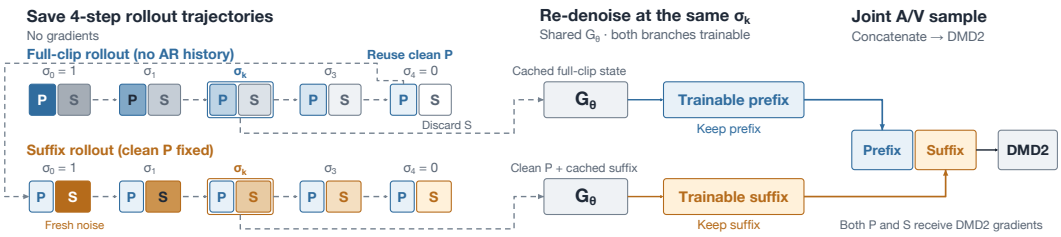

Figure 4: **Shot-aware rollout for autoregressive distillation.** A generated prefix conditions a suffix rollout from fresh noise. Clean prefix and suffix predictions at a shared σ_k are concatenated for DMD2 supervision. Both predictions receive gradients; rollouts and history remain detached.

We apply the sCM objective in Eq. (3) in TrigFlow coordinates. The frozen AR teacher ψ supplies CFG-guided RF velocity $\boldsymbol{u}^g_\psi$, whose wrapped field $\boldsymbol{F}^g_\psi$ defines the target-token trajectory $\mathrm{d}\boldsymbol{y}^i_\tau/\mathrm{d}\tau = \boldsymbol{F}^g_\psi$. Let $\boldsymbol{J}_\tau = \mathrm{d}\boldsymbol{F}_{\theta^-}/\mathrm{d}\tau$ denote the total derivative of the stop-gradient student's wrapped velocity along this trajectory. With predictions evaluated at $(\boldsymbol{y}^i_\tau, \tau, c, \alpha_t \boldsymbol{x}_0^{<i})$, the sCM tangent target is

$$\boldsymbol{g} = \cos\tau \frac{\mathrm{d}\boldsymbol{f}_{\theta^-}}{\mathrm{d}\tau} = \sigma_d \left[\cos^2\tau\left(\boldsymbol{F}^g_\psi - \boldsymbol{F}_{\theta^-}\right) - \sin\tau\cos\tau\left(\boldsymbol{y}^i_\tau + \boldsymbol{J}_\tau\right)\right]. \tag{7}$$

where θ^- and σ_d are defined in Sec. 2. $\boldsymbol{J}_\tau$ is calculated via finite-difference estimation (see Sec. A.1).

3.3 DMD2 REFINEMENT WITH PCM REGULARIZATION.

Although sCM provides a four-step initialization, its outputs may exhibit blur and deformation under rapid motion. We refine the student with DMD2 (Yin et al., 2024a) to improve generation quality, complemented by PCM regularization (Wang et al., 2024) to suppress oversaturation, without increasing the sampling budget. The generator objective is

$$\mathcal{L}_{\text{refine}}(\theta) = \mathcal{L}_{\text{DMD}}(\theta) + \lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}}(\theta) + \lambda_{\text{PCM}}\mathcal{L}_{\text{PCM}}(\theta), \tag{8}$$

with $\lambda_{\text{GAN}} = 0.01$ and $\lambda_{\text{PCM}} = 0.001$. For the DMD objectives in Eq. (4), we use the frozen 14B ALIVE model as the teacher. An auxiliary discriminator provides additional adversarial supervision (Sec. A.2). PCM uses the frozen multi-step AR teacher ψ from Sec. 3.1 to regularize the trajectory.

PCM Regularization. We apply the continuous-time sCM objective independently within each PCM phase. For the four-step RF schedule $\sigma_0 > \cdots > \sigma_4$, we apply PCM regularization only to the first two high-noise phases, $t \in (\sigma_2, \sigma_0]$. Each phase learns a consistent prediction of its lower-noise boundary. Defining $\delta_k = \tau(t) - \tau(\sigma_{k+1})$, we parameterize this prediction as $\boldsymbol{f}_{\theta,k} = \sigma_d\big(\cos\delta_k\, \boldsymbol{y}^i_\tau - \sin\delta_k\, \boldsymbol{F}_\theta\big)$. We retain the RF-to-TrigFlow wrapper in Eq. (6), evaluating all fields at $(\boldsymbol{y}^i_\tau, \tau, c, \alpha_t \tilde{\boldsymbol{x}}_0^{<i})$. Here, $\boldsymbol{y}^i_\tau$ is the wrapped noisy ground-truth suffix, $\tilde{\boldsymbol{x}}_0^{<i}$ is detached generated history, and $c = c_i$ is the current-shot condition. The wrapper uses absolute time τ, while the phase coefficients use relative time δ_k. Following Eq. (7), the phase-specific tangent target is

$$\boldsymbol{g}_k = \cos\delta_k \frac{\mathrm{d}\boldsymbol{f}_{\theta^-,k}}{\mathrm{d}\tau} = \sigma_d \left[\cos^2\delta_k \left(\boldsymbol{F}^g_\psi - \boldsymbol{F}_{\theta^-}\right) - \sin\delta_k\cos\delta_k \left(\boldsymbol{y}^i_\tau + \boldsymbol{J}_\tau\right)\right]. \tag{9}$$

The teacher field $\boldsymbol{F}^g_\psi$ and total derivative $\boldsymbol{J}_\tau$ are defined in Eq. (7). The PCM regularizer follows Eq. (3) with $\boldsymbol{g}$ replaced by $\boldsymbol{g}_k$, retaining the same normalization and stop-gradient convention. Its expectation is restricted to the two selected phases.

Shot-Aware Rollout. To support prompt-controlled shots of varying duration and reduce exposure bias (Schmidt, 2019), we train the student to continue its own completed outputs. As shown in Fig. 4, we first generate a full clip in four steps without AR history. At a randomly sampled temporal split, we retain the clean prefix $\tilde{\boldsymbol{x}}_0^{<i}$ as detached history and regenerate the suffix from fresh Gaussian noise, following the same four-step schedule and conditioning on $\alpha_t \tilde{\boldsymbol{x}}_0^{<i}$ at each RF time t. Both rollouts cache intermediate states without gradients. We then re-evaluate states at a shared noise level σ_k and apply DMD2 supervision to the concatenated prefix and suffix predictions. Using completed four-step outputs as history exposes the student to inference-time errors, while shot-aware rollout aligns the first autoregressive chunk of each shot with its prompt.

Table 1: Comparison on 23 cases with distinct identities. A/V: joint audio-video generation. Params: generation backbones only, excluding text/conditioning encoders, VAEs, and separate LoRA adapters.

Model	Params	Real-time	A/V	Resolution	IDC↑	Sync-D↓	Q-Face↑	Speech Corr.↑
TaoMate	22.1B	✔	✔	768×512	0.240	10.035	0.563	0.476
OmniForcing	19B	✔	✔	768×512	0.270	13.357	0.433	0.241
LongCat-Video-Avatar-1.5†	15.9B	✗	✗	832×480	0.463	**8.389**	0.619	0.909
LiveAvatar†	14B	✔	✗	720×400	0.702	13.347	0.582	0.916
ALIVE-AVATAR	3.6B	✔	✔	864×480	**0.789**	9.308	**0.774**	**0.979**

† denotes audio-driven models supplied with `edge-tts` speech.

3.4 STREAMING INFERENCE

Our inference pipeline is shown in Fig. 2(b). Given a shot prompt, the final autoregressive student jointly generates audio and video latents from noise in four steps. Each modality conditions on a bounded history of early attention-sink anchors and recent latents for continuation, with CMM downweighting the recent context to limit error feedback. After each shot, we update the history and select bounded early-anchor and recent-latent context for the next shot.

Context Multiplicity Modulation. Recent generated context maintains continuity but also feeds rollout errors back into the denoiser. CMM adjusts its attention mass without changing token values, order, or rotary positions. Replicating a key-value token (k_j, v_j) exactly m times is equivalent to adding $\log m$ to its attention logit:

$$\sum_{c=1}^{m} e^{s_{ij}} v_j = m e^{s_{ij}} v_j = e^{s_{ij}+\log m} v_j. \tag{10}$$

This equivalence extends to continuous multiplicities $m > 0$ (Fig. 2(c)). Reweighting group g by m_g, with other groups unchanged, transforms its attention mass P_g as

$$P'_g = \frac{m_g P_g}{1 + (m_g - 1)P_g}, \qquad \frac{P'_g}{1 - P'_g} = m_g \frac{P_g}{1 - P_g}. \tag{11}$$

Thus, m_g directly scales the group's attention odds. We typically leave clean anchors unchanged and set $0 < m_{\mathcal{R}} < 1$ to softly suppress recent generated context.

Remark 1 (Local error attenuation) *Let $\mathcal{R}$ denote the recent-context positions. For $j \in \mathcal{R}$, write $v_j = \bar{v}_j + \delta v_j$, where δv_j is the value error. Let p_{ij} be the attention weight from query i to key j, and define $P_{\mathcal{R}} = \sum_{j \in \mathcal{R}} p_{ij}$, with the dependence on i suppressed. With queries and keys fixed, the recent-context contribution to the output error satisfies*

$$\epsilon_i^{\mathrm{rec}} = \sum_{j \in \mathcal{R}} p_{ij} \delta v_j, \qquad \|\epsilon_i^{\mathrm{rec}}\| \le P_{\mathcal{R}} \max_{j \in \mathcal{R}} \|\delta v_j\|. \tag{12}$$

By (11)*, CMM reduces $P_{\mathcal{R}}$ and thereby tightens this local error bound. This does not guarantee global rollout stability, since errors may also affect queries, keys, and future denoising trajectories.*

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models. We build on ALIVE (Guo et al., 2026) and train on clips up to approximately ten seconds long. Unless otherwise specified, we evaluate the final four-step, 4-NFE student at 480p and 24 fps. Implementation details are provided in Sec. B.

Long-Horizon Benchmark. We evaluate 23 newly constructed cases with distinct identities, each spanning 1.5 to 5 minutes across diverse interaction scenarios. Speech scripts include Chinese and English, with a specified target duration for each prompt. Construction details are in Sec. B.4.

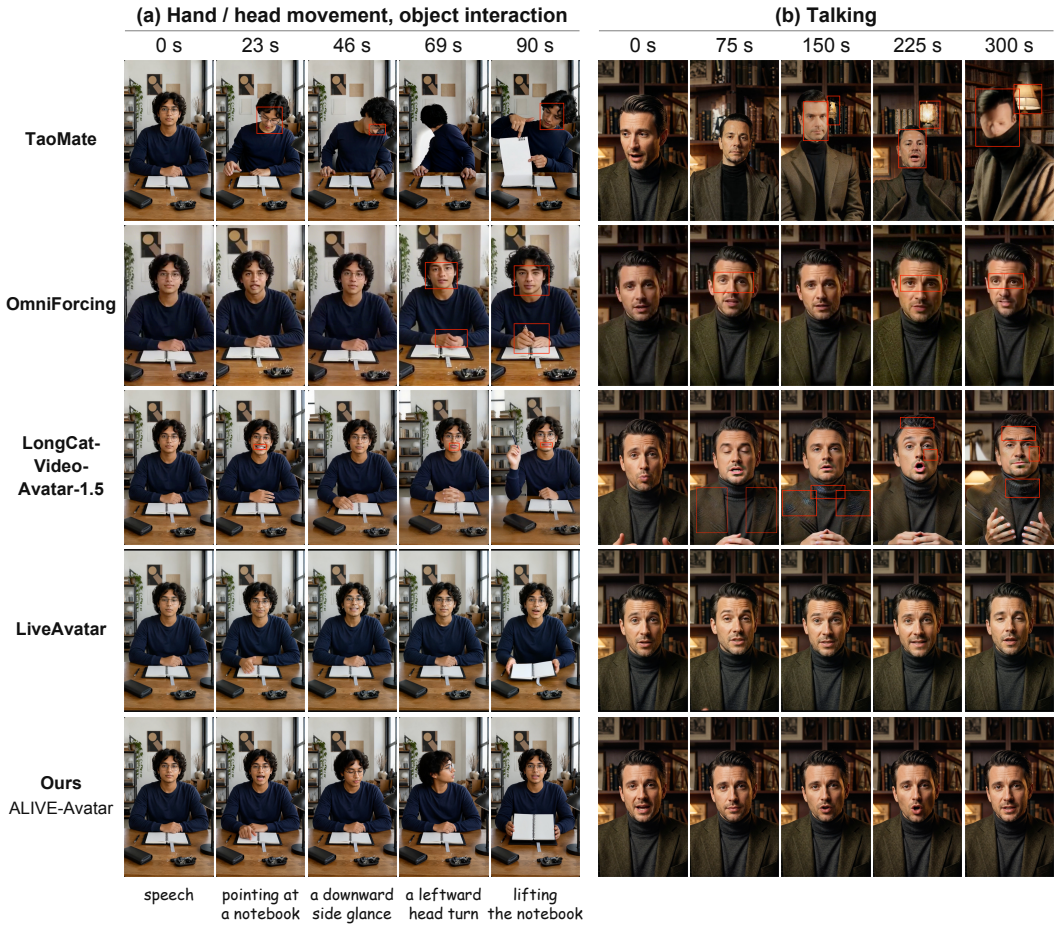

Figure 5: **Comparison with avatar generation baselines across different timestamps.** (a) Speech with object interaction. (b) Front-facing speech. Red boxes highlight noticeable drift.

Baselines and Protocol. We compare ALIVE-AVATAR with joint audio-video generators TaoMate (Gan et al., 2026) and OmniForcing (Su et al., 2026), as well as audio-driven avatar models LongCat-Video-Avatar-1.5 (Team et al., 2026) and LiveAvatar (Huang et al., 2025). Audio-driven baselines use `edge-tts` speech synthesized from the target scripts; joint models generate their own speech. Fixed-chunk methods retain each prompt for a chunk count determined by its target duration and the method's chunk duration, whereas ALIVE-AVATAR generates one variable-length shot matched to that prompt's target duration. Per-model inputs are detailed in Tab. 3.

Metrics. We report IDC for identity consistency, Q-Face for facial quality, Sync-D for audiovisual synchronization, and Speech Correctness (Speech Corr.) for ASR-based script adherence. Speech Corr. evaluates the end-to-end pipeline, not the video generator's speech-generation ability alone. All metrics are averaged over videos; computation details are in Sec. B.5.

4.2 MAIN RESULTS

Table 1 shows that ALIVE-AVATAR achieves the highest mean identity consistency and facial visual quality among the evaluated methods, as well as the highest end-to-end Speech Corr. under the stated input protocol. Figure 1(d) summarizes these results using the normalization defined in its caption. The temporal comparisons in Fig. 5 highlight differences in appearance preservation and action execution. TaoMate exhibits noticeable appearance drift, while OmniForcing shows facial and hand artifacts. LongCat-Video-Avatar-1.5 exhibits mouth-region artifacts and changes in clothing and facial details. LiveAvatar preserves facial appearance more consistently but shows limited responses to the requested head turns. ALIVE-AVATAR maintains consistent appearance while following the

Table 2: Ablations of ALIVE-AVATAR's core modules on the long-horizon benchmark.

Ablation	Variant	IDC↑	Sync-D↓	Q-Face↑	Speech Corr.↑
Distillation Components	w/o sCM initialization	0.710	10.102	0.759	0.943
	w/o DMD2 refinement	0.777	9.414	0.722	0.974
	w/o PCM regularization	0.749	10.248	0.763	0.977
AR Rollout Training	Teacher Forcing Style	0.641	10.555	0.701	0.973
	Self-Forcing Style	0.725	10.174	0.749	0.975
Identity Preservation	w/o CMM	0.611	9.330	0.768	**0.979**
Ours		**0.789**	**9.308**	**0.774**	**0.979**

specified gestures, head turns, and notebook interaction. It also preserves facial features, clothing, and scene composition across the sampled frames of the five-minute talking sequence. These results are achieved with a 3.6B-parameter generation backbone, compared with 14B–22.1B for the baselines, demonstrating strong long-horizon performance with a substantially smaller model.

4.3 Ablation Studies

All ablations vary only the target module, keeping other settings fixed; results are shown in Tab. 2.

Analysis of Few-Step Distillation. Figure 6 (top) and Tab. 2 show complementary benefits of the distillation components. sCM initialization preserves reference facial identity and improves IDC and Q-Face. DMD2 refinement alleviates hand blur and improves Q-Face. PCM regularization mitigates color shifts, further increases IDC, and lowers Sync-D. Together, these three distillation components improve identity preservation and visual quality at four NFE.

Analysis of Rollout Training. Teacher forcing leads to inconsistent facial identity, while self-forcing produces lower face quality than our shot-aware rollout (Fig. 6, bottom). Shot-aware rollout improves IDC and Q-Face and lowers Sync-D over both baselines (Tab. 2), better preserving reference appearance during expressive gestures and head motion. These results support training variable-length continuations on self-generated prefixes for identity-preserving shot streaming.

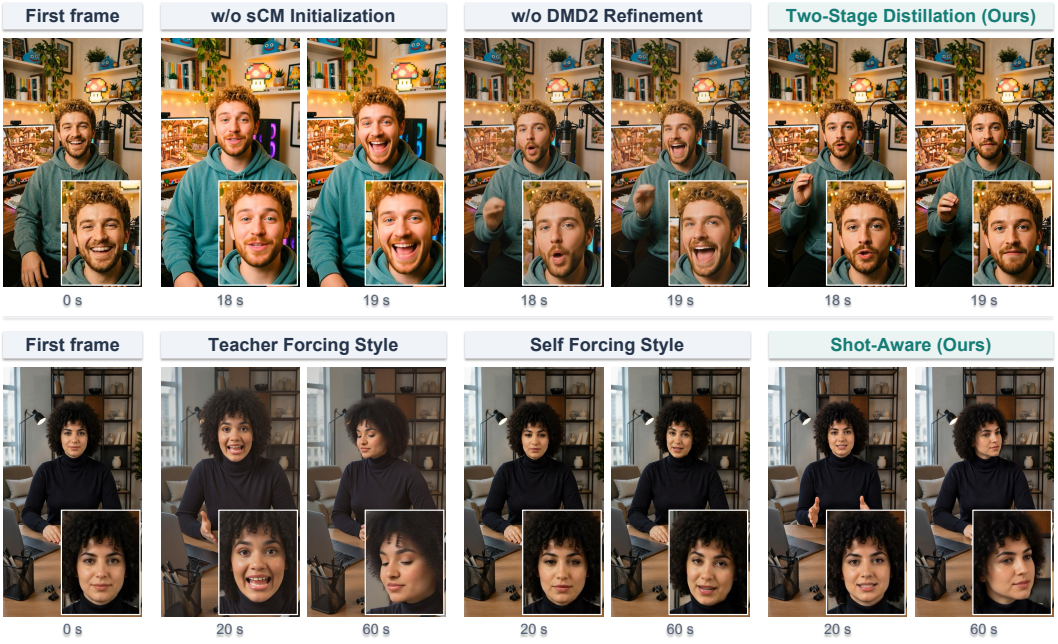

Figure 6: **Qualitative comparisons of different distillation components (top) and rollout strategies (bottom)**. Our method better preserves visual details and maintains temporal consistency.

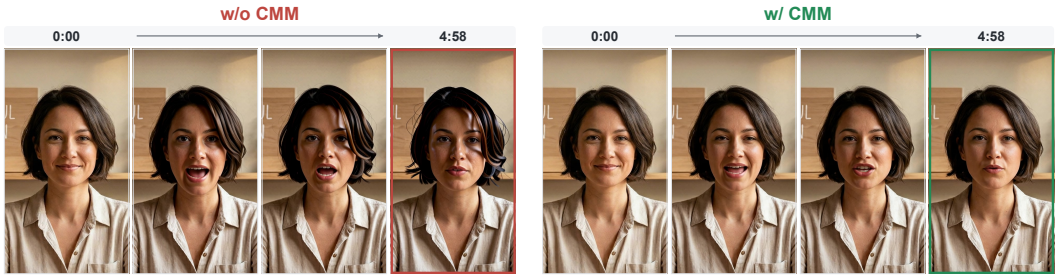

Figure 7: **Effect of CMM during long-horizon autoregressive generation.** At matched timestamps from the same input, CMM ($m_{\mathcal{R}} = 0.6$) better preserves identity and appearance.

Analysis of CMM. As shown in Tab. 2, CMM substantially improves IDC while maintaining comparable Q-Face, Sync-D, and Speech Corr. The five-minute rollout in Fig. 7 further shows that CMM markedly reduces accumulating color shifts and identity drift, better preserving the opening appearance throughout long-horizon autoregressive generation.

4.4 Agent Application

We build a live-streaming agent powered by ALIVE-Avatar that delivers planned content and responds to viewer comments. Given a reference image, voice, and persona, the agent plans speech and actions and updates prompts without generator retraining. Stateful distributed inference enables faster-than-playback generation (Sec. D), while playback-aware scheduling support coherent, continuous interaction. Section E details its design and demonstrations with simulated viewers.

5 Related Work

Avatar Generation. Audio-driven avatars have evolved from facial animation (Prajwal et al., 2020; Zhang et al., 2023) to expressive portrait and upper-body synthesis (Tian et al., 2024; Xu et al., 2024; Cui et al., 2025; Gan et al., 2025; Gao et al., 2025b). Long-form (Tu et al., 2025; Yang et al., 2025a; Team et al., 2026) and streaming variants (Li et al., 2025; Huang et al., 2025) address duration and interactivity. Joint models such as ALIVE (Guo et al., 2026), Ovi (Low et al., 2025), and LTX-2 (HaCohen et al., 2026) instead generate both speech and video, while TaoMate (Gan et al., 2026) and OmniForcing (Su et al., 2026) support real-time joint generation.

Long-Form and Streaming Models. Long-form video models address generation efficiency and temporal memory (Kim et al., 2024; Henschel et al., 2025; Yin et al., 2025). Diffusion Forcing (Chen et al., 2024) uses token-wise noise levels for flexible sequence generation. Self Forcing (Huang et al., 2026) trains on self-generated histories, while Causal Forcing (Zhu et al., 2026) aligns teacher and student causal structures during distillation initialization. LongLive (Yang et al., 2025b) combines long-sequence tuning, windowed attention, and persistent frame sinks. Our work focuses on promptable audio-video streaming with variable-length shots and persistent identity.

6 Conclusion

We presented ALIVE-Avatar, a real-time, identity-preserving audio-video avatar generation model that enables interactive control through promptable shot streaming. Starting from ALIVE, we enable streaming generation with flexible prompt switching through prompt-to-shot autoregressive training via teacher forcing. For real-time generation, we adopt two-stage distillation: consistency initialization followed by PCM-regularized DMD2 refinement, enabling high-quality, identity-preserving four-step generation. At inference, we introduce Context Multiplicity Modulation to mitigate identity drift, maintaining reference identity in qualitative audio-video rollouts of up to one hour. Experiments on our long-horizon benchmark show improved identity consistency and face quality over evaluated baselines while maintaining speech correctness and audiovisual synchronization.

AI USE STATEMENT

We used generative AI tools to polish the manuscript's language, including improving grammar, clarity, and readability. Generative AI was also used to construct synthetic reference portraits and benchmark speech scripts and action descriptions (Sec. B.4). All reference portraits and avatar examples shown in the paper depict AI-generated fictional identities, not real individuals, as clarified in the Ethics Statement. Model-based data annotation and the use of language models in the live-streaming agent are described in Sec. B.3 and Sec. E, respectively. The authors are responsible for reviewing AI-assisted revisions for factual accuracy, consistency with the reported methods and results, and appropriate attribution. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

ALIVE-AVATAR can support communication, entertainment, and interactive virtual assistants. However, real-time audio-video synthesis with persistent identity over long sessions may facilitate impersonation, non-consensual portrayals, fraudulent live streams, and misleading endorsements. Responsible deployment should require appropriate permissions for reference media and consent for depicting identifiable individuals, clearly disclose that the avatar and its speech are synthetic, and incorporate provenance mechanisms and controls against deceptive or abusive use. All reference portraits and avatar examples shown in the paper, including the one-hour rollouts in Fig. 1, depict AI-generated fictional identities rather than real individuals. The quantitative benchmark additionally includes reference images of real individuals from livestream screenshots and publicly available celebrity photographs (Sec. B.4). Neither these real-person reference images nor videos generated from them are displayed in the paper; their evaluation contributes to the reported aggregate metrics. Biases in the pretrained model, training data, or automated evaluation models may affect generation quality and its measurement across demographic groups, motivating further evaluation of potential disparities. Speech Correctness measures adherence to the prompt script and does not establish the factual accuracy or safety of generated statements. For the live-streaming agent, responsible operation should additionally include moderation of viewer comments and generated responses, minimization of personal information retained in interaction histories or passed to language-model services, and mechanisms for human intervention.

REPRODUCIBILITY STATEMENT

To support reproducibility, Sec. 3 describes the autoregressive training formulation, the two-stage distillation curriculum, and the streaming inference procedure. Additional details for consistency distillation are provided in Sec. A.1, while Alg. 1 specifies the PCM-regularized DMD2 training procedure and shot-aware rollout. The derivation of Context Multiplicity Modulation and its numerical implementation requirements are presented in Sec. A.3. Section B.1 and Sec. B.2 summarize training and inference settings, and Sec. B.3 describes the data filtering, annotation, and preprocessing procedures. The experimental setup, case construction, per-model input and audio protocols, and video-level metric aggregation are described in Secs. B.4, B.5 and 4.1, with component-wise ablations reported in Sec. 4.3 and Sec. C.1. Section D documents the distributed inference system, implementation optimizations, and efficiency measurement settings. Finally, Sec. E describes the live-streaming agent's generator interface, module organization, and scheduling strategy.

REFERENCES

Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

Junhao Chen, Mingjin Chen, Jianjin Xu, Xiang Li, Junting Dong, Mingze Sun, Puhua Jiang, Hongxiang Li, Yuhang Yang, Hao Zhao, Xiao-Xiao Long, and Ruqi Huang. Dancetogether: Generating

interactive multi-person video without identity drifting. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7VEECFBzmm.

Junsong Chen, Shuchen Xue, Yuyang Zhao, Jincheng Yu, Sayak Paul, Junyu Chen, Han Cai, Song Han, and Enze Xie. Sana-sprint: One-step diffusion with continuous-time consistency distillation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16185–16195. IEEE, 2025.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21086–21095. IEEE, 2025.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35: 16344–16359, 2022.

Linwei Dong, Ruoyu Guo, Ge Bai, Zehuan Yuan, Yawei Luo, and Changqing Zou. Dual-end consistency model. In *European Conference on Computer Vision*, pp. 423–441. Springer, 2026.

Jiarui Fang and Shangchun Zhao. Usp: A unified sequence parallelism approach for long context generative ai. *arXiv preprint arXiv:2405.07719*, 2024.

Qijun Gan, Ruizi Yang, Jianke Zhu, Shaofei Xue, and Steven Hoi. Omniavatar: Efficient audio-driven avatar video generation with adaptive body animation. *arXiv preprint arXiv:2506.18866*, 2025.

Qijun Gan, Chenwei Zhang, Meiguang Jin, Junfeng Ma, and Qiu Shen. Taomate: Anchor-guided memory bridging evolving and reference states for real-time audio-video digital human generation. *arXiv preprint arXiv:2607.24359*, 2026.

Kaifeng Gao, Jiaxin Shi, Hanwang Zhang, Chunping Wang, Jun Xiao, and Long Chen. Ca2-VDM: Efficient autoregressive video diffusion model with causal generation and cache sharing. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 18550–18565. PMLR, 13–19 Jul 2025a. URL https://proceedings.mlr.press/v267/gao25m.html.

Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, et al. Wan-s2v: Audio-driven cinematic video generation. *arXiv preprint arXiv:2508.18621*, 2025b.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *Advances in Neural Information Processing Systems*, 38:75460–75482, 2026.

Ying Guo, Qijun Gan, Yifu Zhang, Jinlai Liu, Yifei Hu, Pan Xie, Dongjun Qian, Yu Zhang, Ruiqi Li, Yuqi Zhang, et al. Alive: Animate your world with lifelike audio-video generation. *arXiv preprint arXiv:2602.08682*, 2026.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2568–2577. IEEE, 2025.

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

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jinyi Hu, Shengding Hu, Yuxuan Song, Yufei Huang, Mingxuan Wang, Hao Zhou, Zhiyuan Liu, Wei-Ying Ma, and Maosong Sun. ACDit: Interpolating autoregressive conditional modeling and diffusion transformer. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=OuFNXESoCO.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026.

Yubo Huang, Hailong Guo, Fangtai Wu, Weiqiang Wang, Shifeng Zhang, Shijie Huang, Qijun Gan, Lin Liu, Sirui Zhao, Enhong Chen, et al. Live avatar: Streaming real-time audio-driven avatar generation with infinite length. *arXiv preprint arXiv:2512.04677*, 2025.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202. IEEE, 2025a.

Ziyue Jiang, Yi Ren, Ruiqi Li, Shengpeng Ji, Boyang Zhang, Zhenhui Ye, Chen Zhang, Bai Jionghao, Xiaoda Yang, Jialong Zuo, Yu Zhang, Rui Liu, Xiang Yin, and Zhou Zhao. Megatts 3: Sparse alignment enhanced latent diffusion transformer for zero-shot speech synthesis, 2025b. URL https://arxiv.org/abs/2502.18924.

Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. Fifo-diffusion: Generating infinite videos from text without training. *Advances in Neural Information Processing Systems*, 37: 89834–89868, 2024.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

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.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation. *arXiv preprint arXiv:2510.01284*, 2025.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *International Conference on Learning Representations*, volume 2025, pp. 50611–50649, 2025.

KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 484–492, 2020.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling diffusion models, 2024. URL https://arxiv.org/abs/2402.09470.

Seyedmorteza Sadat, Otmar Hilliges, and Romann M Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2024.

Florian Schmidt. Generalization in generation: A closer look at exposure bias. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pp. 157–167, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

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

Yaofeng Su, Yuming Li, Zeyue Xue, Jie Huang, Siming Fu, Haoran Li, Haoyang Huang, and Nan Duan. Omniforcing: Unleashing real-time joint audio-visual generation. In *European Conference on Computer Vision*, pp. 566–584. Springer, 2026.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Meituan LongCat Team, Xunliang Cai, Meng Cheng, Feng Gao, Zhe Kong, Jiamu Li, Le Li, Weiheng Li, Hongyu Liu, Shuai Tan, et al. Longcat-video-avatar 1.5 technical report. *arXiv preprint arXiv:2605.26486*, 2026.

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.

Shuyuan Tu, Yueming Pan, Yinming Huang, Xintong Han, Zhen Xing, Qi Dai, Chong Luo, Zuxuan Wu, and Yu-Gang Jiang. Stableavatar: Infinite-length audio-driven avatar video generation. *arXiv preprint arXiv:2508.08248*, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Fu-Yun Wang, Zhaoyang Huang, Alexander W Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, et al. Phased consistency models. *Advances in neural information processing systems*, 37:83951–84009, 2024.

Zhenzhi Wang, Jiaqi Yang, Jianwen Jiang, Chao Liang, Gaojie Lin, Zerong Zheng, Ceyuan Yang, Yuan Zhang, Mingyuan Gao, and Dahua Lin. Interacthuman: Multi-concept human animation with layout-aligned audio conditions. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=rJilRU8D3c.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching LMMs for visual scoring via discrete text-defined levels. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54015–54029. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wu24ah.html.

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*, 2024.

Shaoshu Yang, Zhe Kong, Feng Gao, Meng Cheng, Xiangyu Liu, Yong Zhang, Zhuoliang Kang, Wenhan Luo, Xunliang Cai, Ran He, et al. Infinitetalk: Audio-driven video generation for sparse-frame video dubbing. *arXiv preprint arXiv:2508.14033*, 2025a.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025b.

Yang Yang, Tianyi Zhang, Wei Huang, Jinwei Chen, Boxi Wu, Xiaofei He, Deng Cai, Bo Li, and Peng-Tao Jiang. Anchor forcing: Anchor memory and tri-region rope for interactive streaming video diffusion. *arXiv preprint arXiv:2603.13405*, 2026.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. IEEE, 2025.

Ailing Zeng, Casper Yang, Chauncey Ge, Eddie Zhang, Garvey Xu, Gavin Lin, Gilbert Gu, Jeremy Pi, Leo Li, Mingyi Shi, et al. Lpm 1.0: Video-based character performance model. *arXiv preprint arXiv:2604.07823*, 2026.

Jintao Zhang, Pengle Zhang, Jun Zhu, Jianfei Chen, et al. Sageattention: Accurate 8-bit attention for plug-and-play inference acceleration. In *International Conference on Learning Representations*, volume 2025, pp. 71566–71585, 2025a.

Jintao Zhang, Kai Jiang, Jintao Chen, Xu Wang, Yang Luo, Yuji Wang, Dechuang Chen, Jungang Li, Chengyang Ye, Marco Chen, et al. Vidu s1: A real-time interactive video generation model. *arXiv preprint arXiv:2607.03118*, 2026a.

Kaiwen Zhang, Liming Jiang, Angtian Wang, Jacob Zhiyuan Fang, Tiancheng Zhi, Qing Yan, Hao Kang, Xin Lu, and Xingang Pan. Storymem: Multi-shot long video storytelling with memory, 2025b. URL https://arxiv.org/abs/2512.19539.

Tianyuan Zhang, Sai Bi, Yicong Hong, Kai Zhang, Fujun Luan, Songlin Yang, Kalyan Sunkavalli, William T. Freeman, and Hao Tan. Test-time training done right. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=Tb9qAxT3xv.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8652–8661. IEEE, 2023.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation, 2026. URL https://arxiv.org/abs/2605.15141.

Kaiwen Zheng, Guande He, Min Zhao, Jintao Zhang, Huayu Chen, Jianfei Chen, Chen-Hsuan Lin, Ming-Yu Liu, Jun Zhu, and Qianli Ma. Causal-rcm: A unified teacher-forcing and self-forcing open recipe for autoregressive diffusion distillation in streaming video generation and interactive world models, 2026a. URL https://arxiv.org/abs/2606.25473.

Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. In *International Conference on Learning Representations*, volume 2026, pp. 2582–2603, 2026b.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=BYInOck3gr.

A Method Details and Algorithms

A.1 Additional Details for Consistency Initialization

We initialize the student through continuous-time consistency distillation (Song et al., 2023) along the CFG/APG-guided (Ho & Salimans, 2022; Sadat et al., 2024) teacher trajectory described in Sec. 3.2. Recent methods (Geng et al., 2026; Lu & Song, 2025; Dong et al., 2026) compute trajectory derivatives using Jacobian-vector products (JVPs). However, integrating PyTorch's native JVP operator with FSDP and FlashAttention (Dao et al., 2022) requires custom implementation support (Zheng et al., 2026b). We therefore use finite differences to approximate the directional derivative through standard forward evaluations, preserving compatibility with our training infrastructure without custom JVP implementations. Meanwhile, continuous-time consistency distillation remains prone to training instability, as large variations in trajectory tangents can induce high-variance gradients. The time-conditioning pathway is a key source of such instability, since timestep embeddings and adaptive modulation directly affect the time derivatives used to construct the training targets (Lu & Song, 2025). To stabilize student training, we combine gradient clipping and adaptive loss weighting with selective freezing of timestep-conditioning parameters, as detailed below.

Finite-Difference Approximation. We approximate the total derivative $\boldsymbol{J}_\tau$ in Eq. (7) by TrigFlow finite differences. For a small angular step $h > 0$, let $\boldsymbol{y}^{i,-} = \cos h\, \boldsymbol{y}^i_\tau - \sin h\, \boldsymbol{F}^g_\psi$. Then

$$\widehat{\boldsymbol{J}}_\tau = \frac{\cos h\, \boldsymbol{F}_{\theta^-}(\boldsymbol{y}^i_\tau, \tau, c, \alpha_t \boldsymbol{x}_0^{<i}) - \boldsymbol{F}_{\theta^-}(\boldsymbol{y}^{i,-}, \tau - h, c, \alpha_{t'} \boldsymbol{x}_0^{<i})}{\sin h}, \tag{13}$$

where $t' = \sin(\tau - h)/[\cos(\tau - h) + \sin(\tau - h)]$. We substitute $\widehat{\boldsymbol{J}}_\tau$ into Eq. (7) to obtain $\boldsymbol{g}$, which is then used to compute the sCM loss in Eq. (3). The loss is evaluated only on target tokens $\boldsymbol{y}^i_\tau$. Only target tokens are perturbed; the same history is rescaled at each evaluation.

Training Stabilization. To stabilize training, we apply gradient clipping and adopt the adaptive loss weighting of sCM (Lu & Song, 2025) to balance loss contributions across timesteps. Motivated by its analysis of instability in time conditioning, we additionally freeze parameters in the timestep-conditioning pathway, including the associated linear projection and adaLN modulation layers, retaining their pretrained values during distillation. For target construction, we apply tangent warmup, clip the resulting tangent target elementwise to $[-1, 1]$, and normalize it by its joint audio-video L_2 norm over the selected loss tokens plus a stabilizing constant c. The resulting target is detached from the computation graph and used in Eq. (3), with λ_v and λ_a controlling the relative contributions of the video and audio losses, respectively. We find that setting the step size to $h = 0.01$ in Eq. (13) further stabilizes training.

A.2 Additional Details for Distillation

Adversarial Supervision. Following DMD2 (Yin et al., 2024a), we use adversarial supervision from real training data in Eq. (8). Our discriminator uses the 3.6B teacher DiT as a feature extractor, augmented with trainable LoRA adapters (Hu et al., 2022). The pretrained backbone weights remain frozen, while the LoRA adapters and modality-specific discriminator heads are optimized. The video head uses 3D convolutions to aggregate spatiotemporal features, whereas the audio head uses an MLP. Each head produces a real/fake logit from the corresponding modality features. The 3.6B feature extractor is separate from the 14B DMD teacher in Sec. 3.3.

Let $\widehat{\mathbf{X}}_\theta$ denote the generated sample from the shot-aware rollout in Sec. 3.3. Before feature extraction, we independently corrupt real latents $\boldsymbol{x}_0$ and generated latents $\widehat{\mathbf{X}}_\theta$ using the RF forward process in Eq. (1), with the same noise-time distribution for both branches. Denote the resulting noisy latents by $\mathbf{X}_r^{\text{real}}$ and $\mathbf{X}_r^{\text{fake}}$, respectively. Writing $\mathcal{E}^m_\rho$ for the modality-m features of the LoRA-adapted DiT and H^m_ρ for its discriminator head, the logits are

$$\ell^m_\rho(\mathbf{X}_r, r, c) = H^m_\rho\big(\mathcal{E}^m_\rho(\mathbf{X}_r, r, c)\big), \qquad m \in \{v, a\}, \tag{14}$$

where ρ collects the trainable discriminator parameters and c denotes the conditioning. Larger logits indicate greater predicted realness. For brevity, let $\ell^m_{\text{real}} = \ell^m_\rho(\mathbf{X}_r^{\text{real}}, r, c)$ and $\ell^m_{\text{fake}} = \ell^m_\rho(\mathbf{X}_r^{\text{fake}}, r, c)$,

with r denoting the noise time sampled for the respective branch. Using $\mathrm{sp}(z) = \log(1 + e^z)$, we minimize the logistic discriminator loss and the non-saturating generator loss:

$$\mathcal{L}^D_{\mathrm{GAN}}(\rho) = \mathbb{E}\left[\sum_{m\in\{v,a\}} \lambda_m^{\mathrm{adv},D}\left(\mathrm{sp}(-\ell^m_{\mathrm{real}}) + \mathrm{sp}(\ell^m_{\mathrm{fake}})\right)\right], \tag{15}$$

$$\mathcal{L}^G_{\mathrm{GAN}}(\theta) = \mathbb{E}\left[\sum_{m\in\{v,a\}} \lambda_m^{\mathrm{adv},G}\,\mathrm{sp}(-\ell^m_{\mathrm{fake}})\right]. \tag{16}$$

The nonnegative weights balance video and audio supervision. Expectations include training data or student rollouts, conditioning, noise times, and Gaussian noise; losses are averaged over batch and logit entries. During discriminator updates, generated samples are detached. During student updates, discriminator parameters are held fixed, but gradients pass through the heads, DiT feature extractor, and noising operation to $\widehat{\mathbf{X}}_\theta$, following the rollout gradient convention in Sec. 3.3. The generator loss in Eq. (16) supplies the adversarial term in Eq. (8).

PCM in the High-Noise Regime. The restriction to the first two denoising phases in Sec. 3.3 is motivated by our empirical observations: applying PCM in the high-noise regime is sufficient to suppress oversaturation, whereas extending the same regularizer to the low-noise phases introduces mild local blurring. We therefore apply the consistency loss in Eq. (3), with the phase-specific target in Eq. (9), only over the selected high-noise phases. This preserves local detail while retaining the saturation control provided by PCM, with the full four-step sampling schedule unchanged.

Exponential Moving Average. We maintain an exponential moving average (EMA) (Tarvainen & Valpola, 2017) of the student parameters during distillation. After each student optimizer update, we set

$$\bar{\theta}_n = \beta\bar{\theta}_{n-1} + (1-\beta)\theta_n, \qquad 0 \le \beta < 1, \tag{17}$$

where n indexes student updates, β is the EMA decay, and $\bar{\theta}_0 = \theta_0$. We use the EMA parameters for inference. In our experiments, the EMA student exhibits fewer distortions and less oversaturation than the corresponding unsmoothed student. Parameter averaging reduces the influence of individual optimization updates without changing the architecture or four-step sampling budget. This EMA model is distinct from θ^- in Eq. (3), which denotes a stop-gradient copy of the current student for target construction.

Overall Training Algorithm. Algorithm 1 summarizes PCM-regularized DMD2 refinement of the sCM-initialized student. At each iteration, we sample a shot split and a student denoising step. A complete four-step rollout without AR history provides a clean generated prefix, from which we select detached history and restart the suffix from fresh Gaussian noise. We then re-evaluate the prefix and suffix predictions at the same noise level σ_k and concatenate them for DMD and adversarial supervision. During student updates, gradients reach both re-evaluated predictions, while the preceding rollouts and conditioning history remain detached.

The fake-score model and discriminator are updated every iteration, whereas student updates begin after $W = 50$ warmup iterations and occur every $R = 10$ iterations thereafter. On student-update iterations, DMD uses re-noised generated samples, while PCM uses independently noised ground-truth suffixes conditioned on the same detached generated history. PCM is restricted to the first two high-noise phases, with targets constructed from the frozen AR teacher and a stop-gradient copy of the current student. The student minimizes the combined objective in Eq. (8), with the discriminator held fixed. We then update the student EMA according to Eq. (17). Auxiliary updates use detached generated samples, with real training data additionally supervising the discriminator; the final EMA student is used for inference.

A.3 Additional Derivation for CMM

We further elaborate on CMM introduced in Sec. 3.4 and derive its exact implementation using attention outputs and log-sum-exp (LSE) statistics.

Algorithm 1 PCM-Regularized DMD2 with Shot-Aware Rollout

Require: sCM-initialized student θ, fake-score model ϕ, discriminator parameters ρ, frozen DMD teacher μ_{real} and AR teacher ψ
Require: Schedule $\mathcal{S} = \{\sigma_0 = 1 > \cdots > \sigma_K = 0\}$, DMD time sampler p_{match}, high-noise PCM sampler p_{PCM}, EMA decay $\beta \in [0, 1)$
Settings: $K = 4$, $\lambda_{\text{PCM}} = 10^{-3}$, $\lambda_{\text{GAN}} = 10^{-2}$, warmup $W = 50$, student update interval $R = 10$
Notation: $f_\theta(x, s; c, h) = x - s\, u_\theta(x, s, c, h)$; sg stops gradients; i indexes the current shot

Initialize EMA parameters $\bar{\theta} \leftarrow \theta$
for $n = 0, \ldots, N-1$ **do**
  Sample (x_0, c_i), a stride-aligned shot split s_i, and $k \in \{0, \ldots, K-1\}$; $x_0^i \leftarrow x_0[s_i : T]$
  $U \leftarrow [n \geq W] \wedge [(n - W) \bmod R = 0]$
  with no gradient
    $(\widetilde{X}_0, X_{\sigma_k}) \leftarrow \text{FullRollout}(\theta, \xi^{\text{full}}, c_i, \mathcal{S}; \text{cache} = k)$ ▷ No AR history
    $h_i \leftarrow \text{SelectHistory}(\widetilde{X}_0^{<i})$
    $z^i_{\sigma_k} \leftarrow \text{SuffixRollout}(\theta, \xi^i, c_i, h_i, \mathcal{S}; 0 \to k)$ ▷ Use $\alpha_{\sigma_\ell} h_i$ at step ℓ
  with student gradients enabled iff U
    $\widehat{x}^{<i}_{0,\theta} \leftarrow [f_\theta(X_{\sigma_k}, \sigma_k; c_i, \varnothing)]^{<i}$ ▷ Re-predict full prefix
    $\widehat{x}^i_{0,\theta} \leftarrow f_\theta(z^i_{\sigma_k}, \sigma_k; c_i, \alpha_{\sigma_k} h_i)$
    $\widehat{X}_\theta \leftarrow \text{Concat}(\widehat{x}^{<i}_{0,\theta}, \widehat{x}^i_{0,\theta})$
  $\overline{X} \leftarrow \text{sg}(\widehat{X}_\theta)$
  if U **then**
    Sample DMD time $r \sim p_{\text{match}}$
    $X_r \leftarrow (1 - r)\overline{X} + r\epsilon_{\text{D}}$
    $\mathcal{L}_{\text{DMD}} \leftarrow \text{DMDLoss}(\widehat{X}_\theta, X_r, r, c_i; \mu_{\text{real}}, \phi)$
    Sample $(j, t) \sim p_{\text{PCM}}$ with $j \in \{0, 1\}$ and $\sigma_{j+1} < t \leq \sigma_j$; $b \leftarrow \sigma_{j+1}$
    $x^i_t \leftarrow (1 - t)x^i_0 + t\epsilon^i_{\text{P}}$ ▷ Re-noise the ground-truth suffix
    $\mathcal{L}_{\text{PCM}} \leftarrow \text{PCMLoss}(\theta, \text{sg}(\theta), \psi; x^i_t, t, b, c_i, h_i)$
    Compute $\mathcal{L}^G_{\text{GAN}}(\theta; \rho)$ from $\widehat{X}_\theta$ as in Eq. (16)
    Update θ using $\mathcal{L}_{\text{DMD}} + \lambda_{\text{GAN}}\mathcal{L}^G_{\text{GAN}} + \lambda_{\text{PCM}}\mathcal{L}_{\text{PCM}}$
    $\bar{\theta} \leftarrow \beta\bar{\theta} + (1 - \beta)\theta$
  Sample fake-training time r_f and fresh noise ϵ_f
  $X_{r_f} \leftarrow (1 - r_f)\overline{X} + r_f\epsilon_f$
  $\mathcal{L}_{\text{fake}} \leftarrow \text{FMLoss}(u_\phi(X_{r_f}, r_f, c_i), \epsilon_f - \overline{X})$
  Update ϕ using $\mathcal{L}_{\text{fake}}$
  Update ρ using $\mathcal{L}^D_{\text{GAN}}(\rho)$ from real x_0 and detached $\overline{X}$ as in Eq. (15)
Output: four-step EMA student $\bar{\theta}$

Attention Reweighting. The replication equivalence in Eq. (10) assumes identical projected key-value pairs with the same positional encoding and visibility, as illustrated in Fig. 8(a). Each copy contributes equally to both the softmax numerator and denominator, so matching these contributions preserves the normalized attention output. CMM applies a shared factor m_g to each group, preserving its conditional attention distribution while changing its total mass according to Eq. (11), thereby adjusting its influence relative to the other groups.

Consider anchor, recent-context, and current-token groups denoted by $\mathcal{O}$, $\mathcal{R}$, and $\mathcal{C}$, respectively. For illustrative baseline masses $(0.25, 0.50, 0.25)$ and multiplicities $(1, 0.6, 1)$, Eq. (11) gives the normalization factor $D = 1 + (0.6 - 1) \times 0.50 = 0.80$. The resulting masses are

$$(P'_{\mathcal{O}}, P'_{\mathcal{R}}, P'_{\mathcal{C}}) = \frac{(0.25,\ 0.6 \times 0.50,\ 0.25)}{0.80} = (0.3125,\ 0.375,\ 0.3125). \tag{18}$$

As illustrated in Fig. 8(b), the recent-context mass decreases from 50% to 37.5%. The anchor and current-token logits remain unchanged, yet their normalized masses increase through the shared denominator, with their relative weighting preserved. This example clarifies that the multiplicity scales attention odds, whereas the resulting attention mass also depends on renormalization. The numerical values are illustrations rather than empirical attention measurements. The implication for the direct value-error contribution is analyzed in the local error attenuation remark in Sec. 3.4.

Algorithm 2 CMM with exact FlashAttention LSE merging

Require: Queries Q, full keys/values (K, V), disjoint selected groups $\{(K_g, V_g, m_g)\}_{g\in\mathcal{G}}$ with $m_g > 0$

$(O, L) \leftarrow \mathrm{FA}(Q, K, V)$
$O \leftarrow \mathrm{FP32}(O)$; $L \leftarrow \mathrm{FP32}(L)$
$D \leftarrow \mathbf{1}$; $U \leftarrow O$
for each selected group g with $m_g \neq 1$ **do**
  $(O_g, L_g) \leftarrow \mathrm{FA}(Q, K_g, V_g)$
  $O_g \leftarrow \mathrm{FP32}(O_g)$; $L_g \leftarrow \mathrm{FP32}(L_g)$
  $\rho_g \leftarrow \exp(\min(L_g - L, 0))$
  $a_g \leftarrow (m_g - 1)\rho_g$
  $U \leftarrow U + a_g O_g$; $D \leftarrow D + a_g$
return $O' = U/D$, $L' = L + \log D$

Exact LSE Merge. FlashAttention need not expose the attention matrix to apply this reweighting. For a fixed query i in one attention head, write the full unnormalized denominator and numerator as

$$Z = \sum_j e^{s_{ij}}, \qquad N = \sum_j e^{s_{ij}} v_j, \qquad O = N/Z, \qquad L = \log Z. \tag{19}$$

All sums are over keys visible to query i. For a group g with at least one visible key, define Z_g and N_g by restricting these sums to $j \in g$, and let $O_g = N_g/Z_g$ and $L_g = \log Z_g$. The group's original mass is $\rho_g = Z_g/Z = e^{L_g - L} = P_g$. Decomposing the full sums into group and complement gives, after multiplying the group contribution by m_g,

$$Z' = Z_{\bar{g}} + m_g Z_g = Z\big[1 + (m_g - 1)\rho_g\big], \tag{20}$$

$$N' = N_{\bar{g}} + m_g N_g = Z\big[O + (m_g - 1)\rho_g O_g\big]. \tag{21}$$

Taking their ratio yields the exact reweighted output. For a collection $\mathcal{G}$ of pairwise disjoint selected groups, let $a_g = (m_g - 1)\rho_g$. Additivity gives

$$O' = \frac{O + \sum_g a_g O_g}{1 + \sum_g a_g}, \tag{22}$$

$$L' = L + \log\left(1 + \sum_g a_g\right). \tag{23}$$

To see why suppression remains a valid attention operation, let $\rho_0 = 1 - \sum_{g\in\mathcal{G}} \rho_g$ be the mass of all unselected tokens. Then

$$D = 1 + \sum_{g\in\mathcal{G}} a_g = \rho_0 + \sum_{g\in\mathcal{G}} m_g \rho_g = Z'/Z > 0. \tag{24}$$

Although $a_g < 0$ when $m_g < 1$, this term subtracts only the excess contribution already contained in the full attention output. The merge is equivalent to softmax attention with the group logit biases specified in Sec. 3.4, so the resulting probabilities remain nonnegative and sum to one. Disjointness is required for these additive corrections; overlapping selections must first be resolved into disjoint groups with their intended effective multiplicities.

Algorithm and Implementation Conditions. In Alg. 2, $\mathrm{FlashAttention}$ returns the attention output and its natural-log LSE without exposing the attention matrix. The full call provides the baseline (O, L), while each group-only call provides its conditional output O_g and normalizer L_g. The LSE difference recovers the group's original mass ρ_g. The accumulators satisfy $U = N'/Z$ and $D = Z'/Z$ after all selected groups have been processed, so the final division and logarithm recover O' and L'. Every group uses the original full-call L when computing ρ_g; it is not updated inside the loop. All scalar operations are applied independently to each query and head, with a_g and D broadcast over the value dimension.

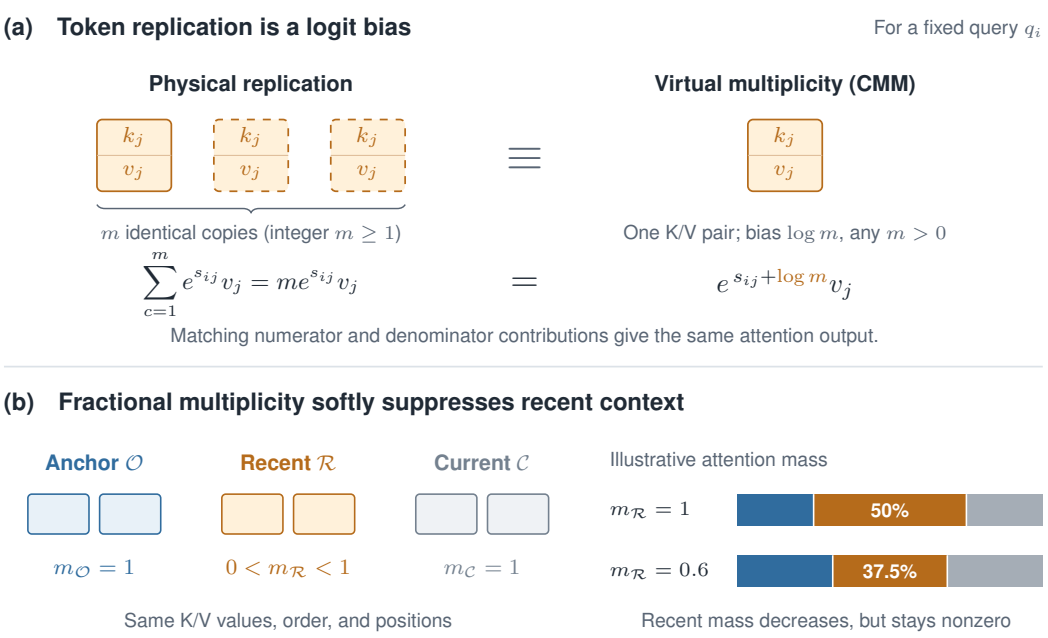

Figure 8: **Context Multiplicity Modulation (CMM). (a)** For a fixed query, m identical copies of a projected key-value pair are equivalent to adding $\log m$ to its logit, with the same normalized attention output and no physical K/V duplication. **(b)** Extending to $0 < m_{\mathcal{R}} < 1$ softly suppresses recent generated context while leaving anchor and current-token logits unchanged. K/V values, order, and positions are preserved. In the illustrative example, baseline group masses $(0.25, 0.50, 0.25)$ become $(0.3125, 0.375, 0.3125)$ at $m_{\mathcal{R}} = 0.6$; these are analytical values, not empirical measurements.

Exactness requires that a shared query-key pair have the same original logit in the full and group-only calls. The calls therefore reuse the same projected Q/K/V tensors, QK normalization, rotary positions, softmax scale, and any other logit transformations. Each group call restricts only the key-value domain and inherits the original mask on those tokens. In particular, extracting a group must not reset its positional indices or change its causal or local-window visibility. We assume each full-attention row has at least one visible key. For a group with no visible keys in a row, its contribution is skipped; equivalently, the group-call convention is $(O_g, L_g) = (0, -\infty)$ for that row.

The LSE differences, exponentials, and accumulation are evaluated in FP32. Since $Z_g \leq Z$, we have $L_g - L \leq 0$ in exact arithmetic; the clamp in Alg. 2 only guards against small positive roundoff errors. With G active groups, the procedure requires one full attention call and G group-only calls, reusing the full call when it is already available. Exactness here refers to the real-arithmetic identity; finite-precision results can differ from a direct biased-softmax evaluation. Under extreme suppression, subtractive cancellation can be reduced by separately evaluating the groups and their complement and merging their positively weighted normalizers and numerators.

B Additional Implementation Details

B.1 Autoregressive Training Configuration

The ALIVE backbone (Guo et al., 2026) comprises a roughly 1.6B VideoDiT and 2B AudioDiT. Video and audio are encoded using Wan2.1 VAE (Wan et al., 2025) and MegaTTS3 WavVAE (Jiang et al., 2025b), respectively. Training uses clips of up to approximately ten seconds.

We train the autoregressive model for 10,000 optimization steps at a learning rate of 10^{-5}, initializing the backbone from a pretrained checkpoint. We use a nine-frame latent chunk stride with no overlap, anchor and recent context ranges of $[1, 30]$ and $[1, 6]$ latent frames, respectively, and independent audio context. The video and audio loss weights are set to 1 and 0.05.

For sCM initialization, we initialize the student from the autoregressive model and train it for $2{,}000$ optimization steps, keeping the teacher frozen. We use LoRA (Hu et al., 2022), excluding the adaptive-normalization modulation projections within the transformer blocks. The video and audio learning rates are set to 5×10^{-6} and 2×10^{-5}, respectively, with a $1{,}000$-step warmup followed by cosine decay. The sCM loss is evaluated only on the current suffix tokens, with video and audio loss weights of 1 and 0.3, respectively. We approximate the tangent derivative using finite differences with an angular step size of $h = 0.01$. Training uses bfloat16 precision, gradient checkpointing, and gradient clipping with a maximum norm of 1.0.

For refinement, we train the student using PCM-regularized DMD2 for $2{,}000$ training iterations. The student learning rate is 10^{-6} for both modalities, with a five-update warmup followed by cosine decay, while the fake-score model and discriminator use a learning rate of 3×10^{-5}. The generator and discriminator adversarial loss coefficients are 0.01 and 0.1, respectively. We jointly optimize the student objective with a PCM regularization weight of 10^{-3}, restricting PCM to the first two high-noise denoising phases. The video and audio loss weights are set to 1 and 0.3.

B.2 Inference configuration

We use the student to generate 480×864 videos at 24 fps with four denoising steps (4 NFE), guidance scales of 1.0, a fixed time-shift factor of 7.0, and four-way sequence parallelism. Video context retains 18 anchor latent frames and 1 recent frames. Independent audio anchor/recent budgets are $18/1$ on the video-latent temporal grid. Cross-modal attention is restricted to temporally overlapping regions. CMM uses anchor/recent multiplicities of $1.0/0.6$.

B.3 Dataset Documentation

The training corpus is the same as that used for ALIVE (Guo et al., 2026); no additional raw training data are collected for this work. Data preprocessing follows the pipeline described by Guo et al. (2026). Starting from videos containing valid audio tracks, the pipeline performs OCR-based text and watermark detection, aesthetics assessment, RAFT-based motion analysis (Teed & Deng, 2020), and model-based quality classification. It further applies clarity filtering to construct the high-resolution training subsets and category balancing across speaking, non-speaking, and general-scene categories.

Annotations follow the two-stage captioning procedure described by Guo et al. (2026): a visual captioning model first describes the visual content, and a multimodal model subsequently incorporates relevant audio information. The multimodal model is also used to assess audio quality and audio–visual coherence. For human-centric scenes, the same subject–dialogue correspondence correction pipeline is applied to reduce erroneous speaker assignments. Unless otherwise stated, the preprocessing settings, annotation procedures, and filtering criteria are unchanged from those described for ALIVE (Guo et al., 2026).

B.4 Benchmark Construction and Evaluation Protocol

Cases and Provenance. The benchmark contains 23 cases with distinct identities, each specifying a reference image and a sequence of prompts for 1.5 to 5 minutes of generation. Each prompt pairs speech and action descriptions with a target duration; speech scripts include both Chinese and English. The cases span diverse age groups and skin tones across product demonstrations, eating broadcasts, and makeup application, including head and hand movements and hand-object interactions. Reference images are drawn from livestream screenshots, publicly available photographs of celebrities, and synthetic portraits generated with `gpt-image-2`. Speech scripts and action descriptions are constructed using GPT-5.5 through Codex. These tools are used to construct inputs, not to assign quantitative evaluation scores.

Scope of displayed examples. All reference portraits and avatar examples shown in the paper, including the one-hour rollouts in Fig. 1, depict AI-generated fictional identities, not real individuals. Benchmark cases using real-person references contribute to the quantitative results, but neither their reference images nor generated videos are displayed in the paper.

Table 3: Per-model core inputs and speech sources for the 23-case benchmark. Text conditions contain the case script and action descriptions for joint generators; audio-driven inputs use speech synthesized from the target script with `edge-tts`. All output audio is evaluated with Whisper large. Resolutions follow Tab. 1.

Model	Core inputs	Speech source	Resolution
TaoMate	Reference image + text	Jointly generated	768×512
OmniForcing	Reference image + text	Jointly generated	768×512
LongCat-Video-Avatar-1.5	Reference image + driving audio	`edge-tts`	832×480
LiveAvatar	Reference image + driving audio	`edge-tts`	720×400
ALIVE-AVATAR	Reference image + text	Jointly generated	864×480

Separation from Training and Model Selection. The test cases are newly constructed independently of the training dataset, rather than extracted from training examples, and are not used for training. Our hyperparameters are selected on separate validation cases constructed during development, not on these 23 test cases. This separates test-case construction from training and model selection, but does not by itself certify identity-level disjointness from all pretraining data, particularly for publicly known individuals. We therefore do not claim an audited identity-disjoint split across the compared pretrained models.

Input and Audio Protocol. The comparison uses the same case definitions while retaining the models' different conditioning modalities. Joint audio-video generators receive the reference image and text conditions and synthesize both speech and video. For LongCat-Video-Avatar-1.5 and LiveAvatar, the target script is converted to driving speech using the same TTS engine, `edge-tts`; these baselines are not supplied with real recordings or speech generated by a competing joint model. Table 3 lists the input and speech-source protocol for each model. Output resolutions are model-specific rather than resolution-matched, so Tab. 1 compares the evaluated systems at the reported settings, not under a common rendering budget.

Per-Prompt Duration Control. For fixed-chunk-size methods, we determine each prompt's chunk count from its target duration and the method's fixed chunk duration. The prompt conditions all allocated chunks, after which generation proceeds to the next prompt. For ALIVE-AVATAR, each prompt instead controls one complete, variable-length audio-video shot matched to its target duration.

Sampling and scope. Each model is evaluated with one random seed per case, so the reported scores do not estimate variability across generation seeds. Aggregate results are descriptive and do not by themselves quantify uncertainty across cases or establish statistical significance.

B.5 METRIC COMPUTATION

Video-Level Aggregation. All metrics are computed separately for each generated video before averaging across cases. Let $s_{m,i}$ denote the score for metric m on the video generated for case i. The main table reports the equal-weight case mean

$$\bar{s}_m = \frac{1}{23} \sum_{i=1}^{23} s_{m,i}. \tag{25}$$

We do not pool valid frames or speech segments across different videos. In particular, the IDC entry is a mean of per-video medians, not a median over all detected faces in the benchmark.

ID Consistency. For each video, we extract face embeddings from the reference image and sampled video frames, measure the cosine similarity between the reference identity and the detected face in each valid frame, and take the median similarity over that video's valid frames as its IDC score.

Synchronization Distance. For each video, we measure audio-visual synchronization using SyncNet (Chung & Zisserman, 2016) by computing the pairwise distance between visual mouth-motion

Table 4: **Ablations of PCM weight and phase coverage.** (a) We vary λ_{PCM} while applying PCM to the first two high-noise phases. (b) We vary phase coverage with $\lambda_{\mathrm{PCM}} = 10^{-3}$. All configurations use four NFE. Bold rows indicate the default configuration.

Configuration	IDC↑	Sync-D↓	Q-Face↑	Speech Corr.↑
(a) PCM weights				
$\lambda_{\mathrm{PCM}} = 10^{-2}$	0.783	9.411	0.737	0.964
$\boldsymbol{\lambda_{\mathrm{PCM}} = 10^{-3}}$	**0.789**	**9.308**	**0.774**	**0.979**
$\lambda_{\mathrm{PCM}} = 10^{-4}$	0.751	9.478	0.771	0.976
(b) PCM phase coverage				
All four phases	0.772	9.243	0.747	0.979
First two phases	**0.789**	**9.308**	**0.774**	**0.979**

features and audio features across temporal offsets. The minimum averaged distance is the video's Sync-D score, where a lower value indicates better synchronization.

Q-Face Score. We uniformly sample frames from each generated video, detect and crop the largest face in each valid sampled frame, score each face crop with the Q-Align (Wu et al., 2024) image-quality scorer, and average the crop scores within that video to obtain its Q-Face score.

Speech Correctness. We use Whisper (Radford et al., 2023) large to transcribe every evaluated output: jointly generated speech for audio-video methods and the TTS-derived audio track for audio-driven methods. Within each video, we align the transcript with the prompt-level speech segments and average the text similarities between recognized segments and their full target texts to obtain the per-video Speech Corr. score. Thus, speech truncated at the end of a prompt's allocated chunks is still compared against the full target text, penalizing incomplete utterances.

The score measures script adherence of the generated-speech or TTS-based pipeline, not the video generator's speech-generation ability in isolation. Using a common ASR model standardizes the evaluator but does not eliminate recognition errors; differences may reflect speech generation or TTS, audio processing, and ASR. The metric does not assess factual correctness of the script.

C Additional Results and Ablations

C.1 Additional Ablations

PCM Weights. We evaluate the effect of PCM regularization strength in Tab. 4. Our default weight of 10^{-3} achieves the best results across all four metrics. However, increasing the weight to 10^{-2} degrades all four metrics, with Q-Face and speech correctness dropping to 0.737 and 0.964, respectively. These results favor an intermediate regularization strength and support our choice of $\lambda_{\mathrm{PCM}} = 10^{-3}$.

PCM Phase Coverage. We compare applying PCM to the first two high-noise phases with applying it to all four phases. Both configurations retain four denoising steps at inference. As shown in Tab. 4, restricting PCM to the first two phases improves IDC from 0.772 to 0.789 and Q-Face from 0.747 to 0.774, while maintaining the same speech correctness of 0.979. Applying PCM to all four phases yields a slightly lower Sync-D (9.243 versus 9.308), but reduces identity consistency and face quality. This trade-off is consistent with our qualitative observation that high-noise PCM suppresses oversaturation, whereas extending it to low-noise phases introduces mild local blurring. We therefore restrict PCM to the first two phases to favor identity consistency and face quality.

C.2 Additional Visualizations

We provide additional long-horizon qualitative comparisons on the moonlit hanfu, orbital botanist, and steampunk workshop examples in Figs. 9 to 11. These scenarios involve distinct character appearances and interactions with a handheld fan, a potted plant, and a rigid mechanical device.

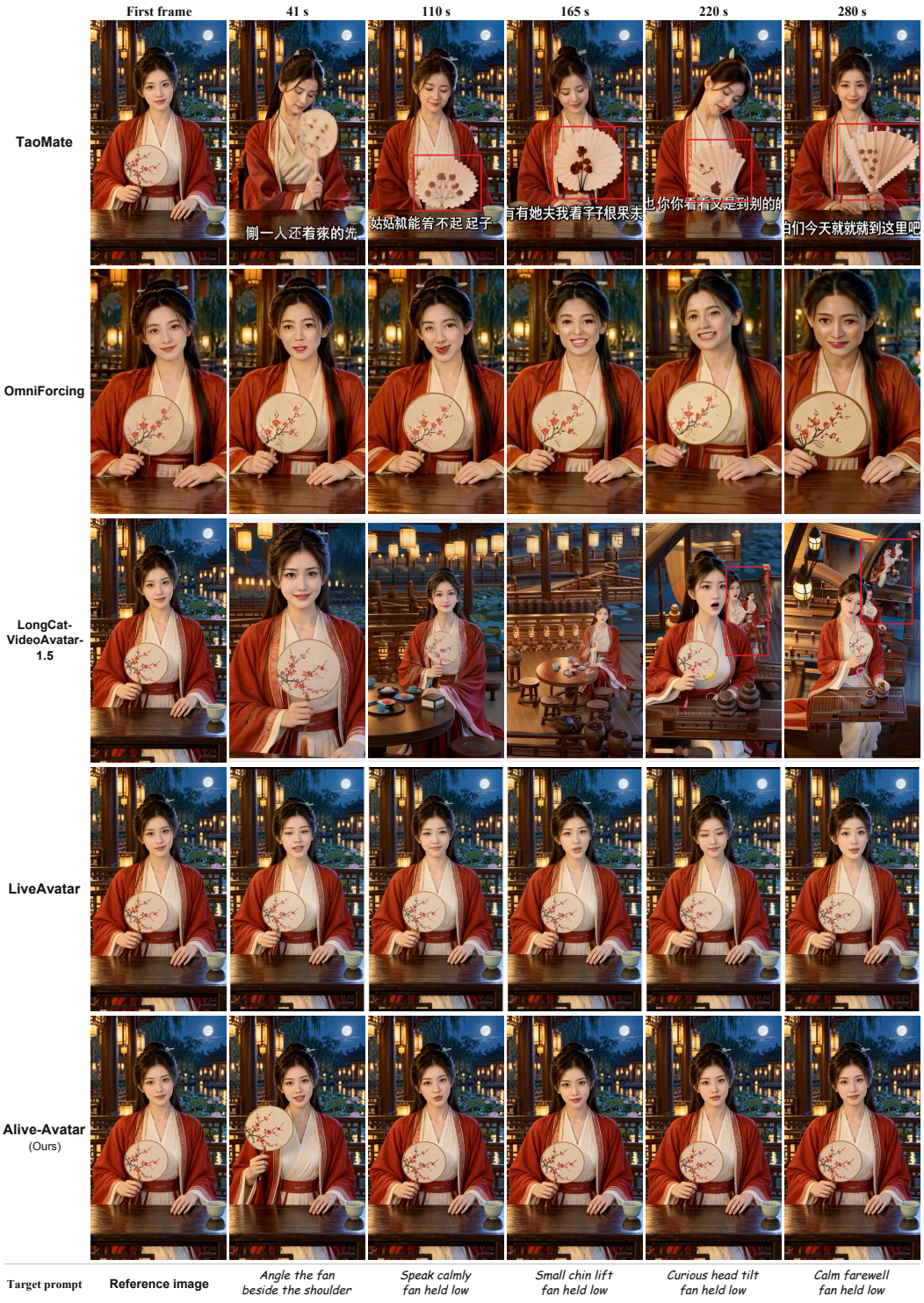

Figure 9: **Long-horizon qualitative comparison on the moonlit hanfu example.** Red boxes highlight selected fan-shape changes and duplicated characters. Columns show common playback times, with each method's first frame (*i.e.*, reference image) on the left. Italic labels below the remaining columns summarize target actions.



Figure 10: **Long-horizon qualitative comparison on the orbital botanist example.** Red boxes highlight selected ghosting artifacts, changes in plant and pot appearance, and duplicated pots. Columns show common playback times, with each method's first frame (*i.e.*, reference image) on the left. Italic labels below the remaining columns summarize target actions.



Figure 11: **Long-horizon qualitative comparison on the steampunk workshop example.** Red boxes highlight selected duplicated goggles, extra hands, changes in device structure and appearance, and duplicated characters. Columns show common playback times, with each method's first frame (*i.e.*, reference image) on the left. Italic labels below the remaining columns summarize target actions.

Across the sampled frames, ALIVE-AVATAR maintains relatively consistent character appearance, scene composition, and object structure, while some baseline outputs exhibit changes in object shape, ghosting artifacts, extra hands, or duplicated characters and objects, as highlighted by the red boxes. These visualizations complement the quantitative evaluation (Tab. 1) by illustrating appearance preservation and visible structural artifacts over extended durations.

D INFERENCE SYSTEM AND EFFICIENCY

The model is served as a stateful distributed inference service while preserving its native autoregressive generation semantics. The system combines cross-shot state reuse, USP sequence parallelism, bounded context caching, CMM-aware attention, compilation stabilization, and parallel VAE decoding. A *shot* here is one generated segment in the sense of Sec. 3.1.

D.1 STATEFUL AND LONG-CONTEXT INFERENCE

We maintain persistent DiT/VAE states, reference latents, text conditions, and video/audio histories across shots, avoiding redundant encoding and state reconstruction.

For long-context generation, we decouple logical temporal positions from physical latent storage and retain only the historical context required by subsequent shots. The original temporal indices are preserved for CA-RoPE (Su et al., 2024) and audio-video alignment. This keeps the physical cache bounded while limiting the number of historical tokens involved in attention.

D.2 USP AND CMM ATTENTION

We parallelize DiT inference with **USP sequence parallelism** (Fang & Zhao, 2024) by partitioning video tokens across devices, reducing per-device memory and accelerating long-sequence attention.

The model uses CMM to reweight historical context without explicit token duplication. Following the continuous-virtual-copy interpretation in Section 3.4, we make CMM compatible with USP by preserving the original sequence partition and synchronizing compact group/span weights instead of replicated KV tensors. We write the resulting attention as

$$\mathrm{Attn}(Q, K, V; B_{\mathrm{CMM}}), \tag{26}$$

where B_{CMM} encodes the CMM group biases while the physical K/V sequence remains unchanged.

For video attention, we further fuse the CMM span bias into **SageAttention** (Zhang et al., 2025a). This avoids KV expansion and preserves the original USP communication pattern and sequence shape.

D.3 COMPILATION AND VAE DECODING

We use BF16 inference, cached `torch.compile`, text-length bucketing, and workload-aware warmup to reduce compilation overhead and stabilize inference shapes.

Wan VAE decoding uses context parallelism and temporal chunking. Its average RTF decreases from 0.560 without context parallelism to 0.080 at CP=8. Video encoding and writing are performed asynchronously and excluded from the compute measurements.

D.4 INFERENCE EFFICIENCY

Table 5 summarizes the scaling of DiT inference with sequence parallelism. USP consistently reduces RTF, with stronger scaling at longer contexts.

The results also highlight the cost of autoregressive context: increasing the retained history consistently increases DiT RTF. Bounded context caching and CMM without physical KV expansion are therefore important for maintaining efficient long-horizon generation.

Overall, our serving optimizations target the dominant long-context DiT workload: USP distributes sequence computation, bounded caching controls history growth, and USP-aware CMM reweights

Table 5: **DiT RTF with sequence parallelism.** Workload: the DiT at the landscape 864×480 configuration, with audio conditioning, 4 denoising steps, CMM, and a single recent latent (N_r=1). Rows vary the number of retained anchor latents N_o. We report the average over 3–8 s shots.

N_o	SP=1	SP=2	SP=4	SP=8
3	1.077	0.670	0.432	0.337
9	1.372	0.840	0.527	0.381
18	1.884	1.120	0.664	0.458

context without increasing the physical attention sequence. Parallel VAE decoding further reduces the decoding overhead.

E AGENTIC LIVE-STREAMING APPLICATION

Building on promptable shot streaming (Sec. 3.1) and the bounded-memory, faster-than-playback inference of Sec. 3.4 and Sec. D, we construct a live-streaming agent that delivers planned content over a scheduled session and answers viewer comments. It drives the same four-NFE student evaluated in Sec. 4; no part of the generator is retrained. This section describes its generator interface (Sec. E.1), modules (Sec. E.2), and time-aware scheduling (Sec. E.3).

E.1 GENERATOR INTERFACE

A session is initialized once with a reference avatar image, a reference voice, and a persona description, for example:

> *Warm but deeply analytical, treating cooking as a rigorous science rather than just an art. I have a soothing, methodical voice, often geeking out over the exact chemistry of caramelization or the precise extraction rate and TDS (total dissolved solids) of my morning coffee. I am fiercely detail-oriented but remain patient and encouraging with my audience. I have a playful, slightly nerdy sense of humor, especially when a culinary experiment goes unexpectedly wrong. I am a perfectionist who streams to share the joy of technical mastery and the hidden logic behind delicious food.*

The agent then drives the generator through a single operation, SUBMIT(p, d), which appends a text prompt p with target duration d to the generation queue and returns decoded audio-video shots in physical-time order. The prompt carries the subject, the scene, the action, and the utterance in one string, so speech, lip motion, and gesture are specified together:

```
Subjects: Subject1: {subject} Visual: {scene}
Narration: Subject1 {action}, saying <W>{speech}</W>
```

The target duration is estimated from the utterance at two English words per second. For instance, midway through a session with the persona above, the agent submits the following prompt, whose 15-word utterance gives $d = 7.5$ s:

> `Subjects:` Subject1: A woman with short, wavy dark brown hair styled in a side part, wearing tortoise-shell glasses and small teardrop earrings. She has a warm, friendly smile showing her teeth and is wearing a navy blue chef coat. The coat features the name "ELARA" embroidered on the right chest and a gold molecular structure above the text "CULINARY ALCHEMIST" on the left chest. `Visual:` A clean, brightly lit head-and-shoulders portrait shot against a plain light gray wall. On the right background, out-of-focus laboratory glassware is subtly visible. `Narration:` Subject1 gestures with hands and looks thoughtfully into the camera, saying <W>Today, we're taking two classic breakfast staples—scrambled eggs and coffee—and turning them into pure chemistry.</W>

Since each shot is conditioned on its own prompt, moving to the next prompt, or revising one not yet submitted, is an ordinary prompt transition (Secs. 3.1 and 3.4). The avatar's appearance and voice are therefore kept consistent by the generator alone, and revisions never require re-encoding

history, because the generator's conditioning state is bounded and carried forward regardless of what is submitted next. The agent itself is responsible only for semantic consistency, such as persona and speech topics, which it maintains through its own memory (Sec. E.2).

E.2 MODULES

The agent consists of five modules that form a closed loop around the generator (Tab. 6). A meta-planner maintains a running outline of the session, and a planner turns what comes next into timed steps. An executor submits these steps as prompts in short batches. Viewer comments reach the plan by two paths: an observer reads each incoming comment and decides whether it warrants an immediate reaction, in which case a patcher splices the reaction into the steps not yet submitted; otherwise the comment is left to the planner, which takes it up in later steps.

The agent's memory holds the persona description, the running outline with the session's progress through it, and an append-only textual record of what has been said, of which planning reads a bounded tail. It tells later planning where the session currently stands, keeps it consistent with earlier claims and free of repeated material, and gives viewer comments the context they are interpreted against.

Table 6: Modules of the live-streaming agent. Every module is driven by the same language model, Gemini 3.6 Flash, and differs only in its prompt and in the reasoning effort selected at call time (Sec. E.3).

Module	Role	Invocation
Meta-planner	Maintains a running outline of the session and the notes later planning will draw on	Session start, and whenever the outline is advanced or revised
Planner	Turns the next stretch of the outline into timed steps, taking up viewer comments raised earlier	When the current steps are close to exhausted
Executor	Submits steps as prompts in short batches	Continuously, subject to the lead cap
Observer	Reads incoming comments and decides whether one warrants an immediate reaction	Continuously, subject to the lead floor
Patcher	Inserts or replaces steps not yet submitted with a reaction	On an immediate-reaction decision

To demonstrate the agent's interactive capability, we additionally run several simulated viewers, each with its own persona. Running independently of the agent, they never see the plan, only the speech of prompts the avatar has already delivered, i.e., the text inside <W>…</W> (Sec. E.1). At random times, one of them posts a short, in-character comment on what has just been said. These comments are written to the comment feed that the observer reads, so the agent reacts to them as it would to any viewer.

E.3 TIME-AWARE SCHEDULING

Let $\ell(t)$ be the lead at wall-clock time t: the playback duration of media that has been rendered but not yet played. Because the four-NFE model generates faster than playback, ℓ grows while the executor keeps submitting and shrinks whenever submission pauses, for instance while a planning call is in flight. The executor stops submitting at a lead cap $\ell_{\max}$, which bounds how much already-committed media a reaction must wait behind. The agent measures ℓ before each language-model call and treats it as a thinking budget, selecting a reasoning effort from it (Table 7): the planners deliberate while the lead is comfortable and fall back to shallow, low-latency decisions as it shrinks.

The floor row is the one property the schedule must guarantee: when the lead approaches exhaustion, only the minimum planning needed to keep the queue fed is started, so planning cannot stall the generator and interrupt playback. The lead cap matters for a second reason. A generator slower than playback would need a large $\ell_{\max}$ simply to survive scheduling jitter, and every reaction would then

Table 7: Effort selection from the measured lead. The thresholds separating the three regimes are deployment-dependent and are set from the observed call latencies.

Lead ℓ	Planning	Comment handling
Comfortable	Full deliberation	Reaction decided on its merits
Tight	Shallow, low-latency decisions	Reaction decided on its merits
Below the floor	Shortest available path	Immediate reactions suppressed; the comment is left to later planning

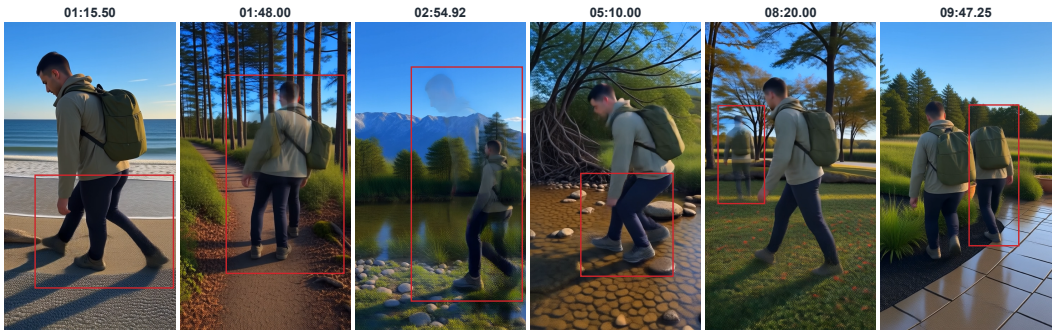

Figure 12: **Failure cases in dynamic scenes.** Selected frames at different timestamps from a single rollout. Red boxes highlight character ghosting and body distortions.

queue behind minutes of committed media; faster-than-playback generation lets $\ell_{\max}$ be set by jitter alone.

F Failure Cases, Limitations, and Future Work

F.1 Failure Cases

Figure 12 illustrates a challenging dynamic rollout in which body pose, viewpoint, and apparent scale change rapidly and substantially as the character moves across different environments. These rapid transitions require the model to preserve character appearance while adapting to markedly different spatial configurations. The selected frames exhibit duplicated limbs, overlapping silhouettes, semi-transparent character ghosts, and incomplete body structures. A plausible contributing factor is that retained historical context can become spatially inconsistent with the current generation during abrupt visual changes. Although this context still provides useful appearance cues, its outdated body configurations may interfere with the synthesis of the new pose, producing overlapping contours or structural distortions. CMM downweights recent generated context to limit error feedback, but does not explicitly align historical and current body geometry. This case highlights a remaining challenge in maintaining structural and temporal coherence under rapid, large changes in visual configuration.

F.2 Limitations and Future Work

Multi-Character Interaction. Our study focuses on streaming generation of a single avatar. More complex scenarios involving multiple characters require distinct identity preservation, correct associations between speech and speakers, and coordinated actions during interactions, which have not been systematically evaluated in our framework. Building on person-specific conditioning (Wang et al., 2026) and explicit binding between identity and motion (Chen et al., 2026), future work could extend our framework to support coordinated dialogue and actions among multiple characters.

Scene Transitions. Our evaluation does not systematically cover controlled transitions between distinct environments. Beyond changes in speech and actions, such transitions require adapting backgrounds, lighting, and spatial layouts while preserving avatar identity and avoiding unintended carryover from the previous scene. Inspired by memory-conditioned multi-shot generation (Zhang

et al., 2025b), future work could explore scene-aware history updates that retain identity-relevant information while refreshing context associated with the environment.

Multiple Scene-Specific References. Our framework supports multiple reference images from a single temporal sequence, treating them as historical context. However, it does not support independent references from different scenes as separate conditioning inputs, limiting scene-specific visual control. Building on multi-reference video conditioning (Jiang et al., 2025a), future work could explore representing scene-specific references separately from temporal history, enabling control over different environments while maintaining shared avatar identity.