AVATAR-FOREVER: DECOUPLED PARALLEL TRAINING FOR HIGH-QUALITY REAL-TIME INFINITE AVATARS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing streaming video systems often rely on sequential, distillation-centered training pipelines to enable few-step long-video generation. However, this paradigm suffers from two limitations. First, failures or distribution shifts introduced in earlier stages affect later optimization, complicating the training process to converge. Second, the distillation-centric objective favours short-term generation but is prone to quality degradation when autoregressive errors accumulate over long rollouts. We propose *Avatar-Forever*, a decoupled parallel training framework for high-quality real-time infinite interactive avatars. Instead of coupling generation efficiency and long-horizon robustness under a sequential distillation pipeline, we treat them as two independent capabilities that can be trained in parallel. One branch performs full-parameter distillation to train an efficient generator with high visual quality, while another trains a lightweight long-horizon adapter via *Recovery-oriented Rollout Training* (RRT), which improves generation robustness under long-horizon inference conditions. Our decoupled parallel training design simplifies the overall training process and avoids unnecessary objective conflicts between few-step generation and long-horizon adaptation. We further introduce *ForeverCache*, a chunk-wise feature caching mechanism to substantially reduce redundant history computation during streaming inference. Built upon a 22B video foundation model, Avatar-Forever supports unbounded audio-driven avatar generation while maintaining identity consistency, motion coherence, and visual fidelity, enabling an end-to-end throughput of high-resolution 768×512 videos at 27.2 FPS on a single H100 GPU and providing a practical path toward stable avatar.

1 INTRODUCTION

Generating realistic and expressive digital humans is highly demanded in applications such as virtual assistants, content creation, education, and entertainment (Zhang et al., 2023; Li et al., 2024; Tian et al., 2024). A practical avatar system should continuously synthesize coherent video over extended durations while maintaining identity consistency, audio-driven motion, visual fidelity, and efficient inference. Recent diffusion-based video foundation models (Team Seedance, 2026; Wan et al., 2025; HaCohen et al., 2026; Yang et al., 2025c; SII-GAIR et al., 2026) have significantly advanced audio-driven avatar generation, producing photorealistic talking avatars with improved visual fidelity, motion realism, and lip synchronization (Guo et al., 2025; Li et al., 2025b; Low & Wang, 2025; Cui et al., 2026). However, most models are trained to generate a short clip within a fixed temporal window, and extending these advances from short clips to real-time long-duration avatar generation remains a significant practical challenge (Yin et al., 2025; Huang et al., 2026; Sun et al., 2026). To generate longer outputs, streaming systems typically reuse previously generated chunks as conditions for future chunks (Yang et al., 2025b; Lu et al., 2026; Li et al., 2025c; Zhu et al., 2026; Yesiltepe et al., 2026), which inevitably introduces distribution drift and progressive error accumulation: early artifacts, identity drift, or lip-sync errors become part of the historical context and gradually degrade subsequent generation quality. (Yuan et al., 2026; Chen et al., 2026b).

Recent works (Zeng et al., 2026; Shen et al., 2026; Sun et al., 2026) usually address this challenge with distillation-centered sequential training pipelines, as illustrated in the top of Figure 1. Starting from a multi-step bidirectional video model, these methods combine teacher forcing, ODE initialization, DMD distillation, and additional forcing strategies (Yin et al., 2025; Huang et al., 2026; Zhu et al.,

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

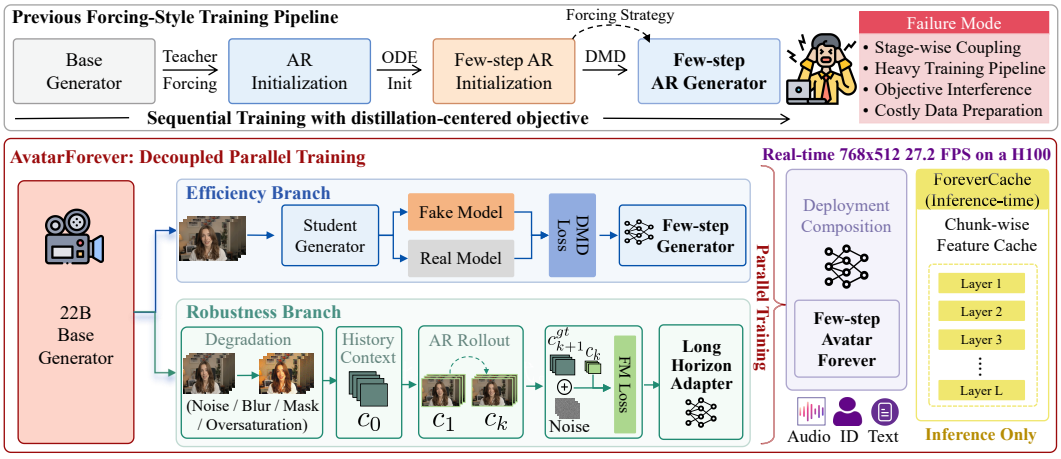

Figure 1: **Overview of Avatar-Forever.** *Top:* Existing streaming avatar systems employ sequential distillation pipelines, leading to stage-wise dependence and objective interference. *Bottom:* Avatar-Forever learns few-step efficiency and long-horizon robustness in parallel. A lightweight adapter is trained under accumulated rollout errors, and it is merged with the distilled generator in inference. In addition, ForeverCache is proposed to reuse historical features for efficient streaming inference.

2026; Cui et al., 2025; Yang et al., 2025a; Huang et al., 2025b) with supervision imposed in fixed denoising steps to obtain a few-step model. This sequential design entangles two distinct objectives in the optimization pipeline: efficient few-step generation, which preserves short-term visual quality under reduced denoising steps, and long-horizon robustness, which stabilizes generation against accumulated autoregressive errors. Such an entanglement makes the training fragile and difficult to scale: distribution shifts introduced at earlier stages affect later optimization, making it hard to diagnose and obscuring the contribution of each stage.

We argue that the two objectives of efficient few-step generation and long-horizon robustness should be decoupled and learned in parallel, and propose *Avatar-Forever*, a decoupled parallel training framework for high-quality real-time long-duration audio-driven avatars, as illustrated in the bottom of Figure 1. Instead of intertwining distillation, forcing schedules and long-context adaptation in a sequential pipeline, we train for efficiency and robustness in parallel. One branch performs full-parameter distillation to obtain a high-quality few-step generator, while the other trains a lightweight long-horizon adapter for stability under extended rollout conditions. This design simplifies the training procedure, making each objective easier to optimize and diagnosis. For the long-horizon adaptation branch, we introduce *Recovery-oriented Rollout Training* (RRT), which targets the main failure mode of autoregressive generation, *i.e.*, errors accumulate after recursive rollout. Starting from a degraded historical context that simulates generation artifacts, the model rolls out multiple future chunks using its own predictions as conditions. Supervision is applied when the degradation has propagated through the rollout, rather than at every local reconstruction step. Importantly, RRT uses a standard flow matching objective and does not require specialized long-video distillation losses or complicated regularization schemes.

During streaming inference, each new chunk needs to attend to historical chunks. One simple way is to recompute the entire visible history window at every denoising step, although only the current chunk is updated. We introduce *ForeverCache*, a chunk-wise autoregressive history feature caching mechanism to remove redundant computation. For each generated chunk, we perform a full forward pass at the first denoising step to populate per-block historical context features, and then reuse these stable features in subsequent denoising steps while forwarding only the current chunk tokens. This inference-time mechanism preserves the historical context needed for identity consistency, motion continuity, and audio-driven synchronization, while improving throughput by **23%** during long-horizon generation.

To support scalable adaptation, we construct a fully synthetic data pipeline by using video foundation models to synthesize training videos from high-quality conversational prompts, followed by automatic filtering to remove low-quality, static, or semantically inconsistent samples. Built upon a 22B video

foundation model (HaCohen et al., 2026) and our synthetic dataset, Avatar-Forever scales audio-driven avatar generation to extended durations while maintaining visual fidelity, identity consistency, and motion coherence. Specifically, Avatar-Forever achieves an **end-to-end** (including DiT inference and VAE decoding) generation throughput of **27.2 FPS** for **high-quality** $\mathbf{768 \times 512}$ avatar videos on a single H100 GPU, achieving state-of-the-art performance on long-horizon audio-driven avatar generation. Our main contributions are summarized as follows:

- We propose **Avatar-Forever**, a decoupled parallel training paradigm for real-time long-duration avatar generation, showing that few-step efficiency and autoregressive robustness can be effectively learned in parallel, rather than in a complicated sequential manner.
- We introduce **Recovery-oriented Rollout Training** (RRT) in order to recover from errors after propagating through multi-chunk autoregressive rollout. Unlike local corrupted-context reconstruction or long-video distillation objectives, RRT uses standard flow matching supervision under the rollout distribution encountered at inference time.
- We introduce **ForeverCache**, a chunk-wise history feature cache that reuses stable context features across denoising steps, removing redundant recomputation of historical chunks during streaming inference and improving long-horizon generation throughput by **23%**.
- We implement the framework with a fully synthetic data pipeline and a 22B video foundation model, achieving high-quality long-duration audio-driven avatar generation with persistent identity, coherent motion, accurate lip synchronization, and real-time $\mathbf{768 \times 512}$ resolution inference at **27.2 FPS** on a single H100 GPU.

2 Related Work

Audio-driven Avatar Generation. Audio-driven avatar generation has evolved from 2D talking-head synthesis to 3D-aware neural rendering. Early methods focus on lip synchronization and identity preservation, while later approaches introduce facial motion coefficients, 3D priors, or neural radiance fields to improve facial dynamics and pose stability (Prajwal et al., 2020; Zhang et al., 2023; Xu et al., 2023; Cao et al., 2024). However, these designs are often task- or subject-specific, limiting their flexibility for diverse identities, expressive body motion, complex scenes, and long-duration generation (Huang et al., 2025b). Diffusion-based (Ho et al., 2020; Lipman et al., 2022; Rombach et al., 2022) avatar systems improve photorealism and motion controllability through audio-conditioned denoising and motion guidance (Shen et al., 2023; Tian et al., 2024; Xu et al., 2024; Chen et al., 2024). Meanwhile, large video foundation models provide strong general spatiotemporal priors for avatar synthesis (Wan et al., 2025; Team Seedance, 2026; OpenAI, 2024), but their high inference cost makes few-step generation essential. Existing acceleration methods reduce denoising steps effectively (Yin et al., 2024; Sauer et al., 2024; Luo et al., 2023; Wang et al., 2024; Wu et al., 2026; Li et al., 2025a), yet mainly target short-horizon sampling rather than robustness to accumulated autoregressive errors.

Long-Horizon Streaming Video Diffusion Models. Streaming video generation (Huang et al., 2025a; Zhu et al., 2026; Huang et al., 2025b; Li et al., 2026) reuses previously generated frames as conditions for future predictions. This creates a train–test mismatch: models are trained on clean ground-truth contexts but must condition on imperfect self-generated outputs at inference, causing artifacts, motion drift, and temporal inconsistency to accumulate over time (Shen et al., 2026; Yuan et al., 2026; Zeng et al., 2026). Recent methods therefore expose models to self-generated contexts during training or distillation. Self-Forcing++ (Cui et al., 2025), Causal Forcing (Zhu et al., 2026), and Hybrid Forcing (Li et al., 2026) use teacher-guided rollout, autoregressive initialization, progressive denoising, or hybrid attention to improve long-horizon stability. However, these methods typically rely on tightly coupled multi-stage pipelines, making component-wise diagnosis difficult and large-scale adaptation cumbersome. Long-duration avatar generation inherits this challenge while imposing stricter requirements on identity, appearance, lip synchronization, facial dynamics, and human motion coherence (Zeng et al., 2026; Gan et al., 2025; Chen et al., 2025). Systems such as LiveAvatar (Huang et al., 2025b), StreamAvatar (Sun et al., 2026), SoulX-FlashTalk (Shen et al., 2026), and LPM 1.0 (Zeng et al., 2026) adapt large video priors to streaming human synthesis. Yet they largely follow the sequential paradigm of generic long-video generation, coupling autoregressive rollout, distillation, and forcing strategies. Consequently, distribution shifts introduced at early stages can propagate through later optimization, making the final system difficult to diagnose and scale.

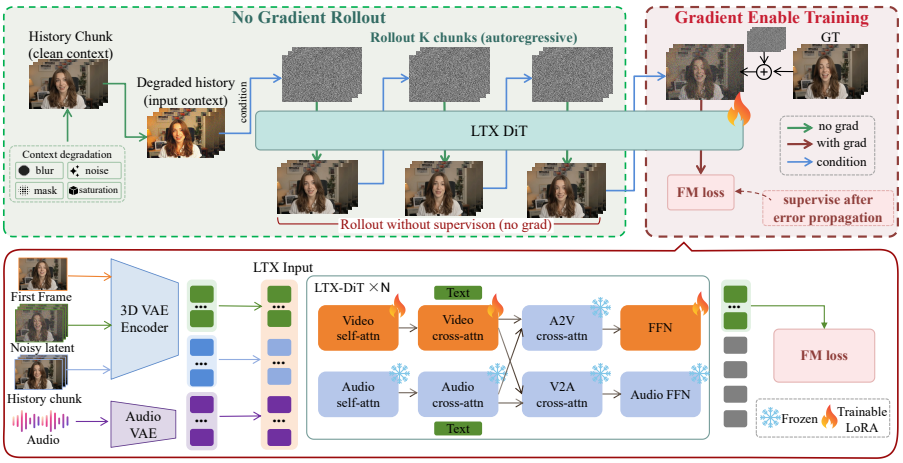

Figure 2: **Recovery-oriented Rollout Training (RRT).** *Top:* RRT degrades an early history chunk, propagates the resulting error through K autoregressive rollout steps without gradients, and applies standard FM supervision only to the subsequent recovery step. *Bottom:* In LTX DiT, video and audio conditions are encoded into multimodal tokens; video-side LoRA modules are trained, while the remaining components are frozen.

3 METHOD

As illustrated in the bottom of Figure 1, Avatar-Forever consists of three core designs. First, the efficiency branch and robustness branch are trained separately with *Decoupled Adaptation*: a full-parameter distillation branch is used to preserve high-quality few-step generation, while a lightweight rollout-adaptation branch is used to improve long-horizon stability; the resulting long-horizon LoRA adapter is merged into the distilled generator at deployment. Second, within the robustness branch, we propose *Recovery-oriented Rollout Training* (RRT), which perturbs early historical context, propagates the degradation through multi-step autoregressive rollout, and applies standard flow-matching supervision only after errors have accumulated. Third, a *ForeverCache* strategy is proposed to accelerate streaming inference by caching stable historical context features across denoising steps, avoiding redundant recomputation of clean history chunks while preserving long-range conditioning.

3.1 EFFICIENCY BRANCH: FEW-STEP DISTILLATION

The efficiency branch aims to obtain a high-quality few-step generator from the pretrained video foundation model. We adopt Distribution Matching Distillation (DMD) (Yin et al., 2024) to compress the original multi-step LTX base model (HaCohen et al., 2026) into a four-step generator while preserving conditional generation quality. Let θ_0 denote the pretrained model parameters, G_θ denote the student generator, and $\mathbf{c}$ denote the conditioning inputs, including text and optional visual conditions. DMD optimizes the student by matching the generated distribution to the teacher distribution through a reverse-KL objective, formulated as follows:

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = -\mathbb{E}_{t,\boldsymbol{\epsilon},\mathbf{c}} \left[\left(s_{\text{real}}(\tilde{\mathbf{x}}_t) - s_{\text{fake}}(\tilde{\mathbf{x}}_t) \right) \frac{\partial G_\theta(\boldsymbol{\epsilon}, \mathbf{c})}{\partial \theta} \right], \tag{1}$$

where $\tilde{\mathbf{x}}_t$ is obtained by applying the forward diffusion process to the student sample $G_\theta(\boldsymbol{\epsilon}, \mathbf{c})$, and s_{real} and s_{fake} are score functions. This branch is responsible only for few-step efficiency. We do not introduce autoregressive rollout, corrupted history, or long-horizon objectives during distillation. To preserve compatibility with downstream avatar generation, we train the distilled model under mixed conditioning: each sample is randomly used either as a text-to-video instance or as a first-frame-conditioned instance, where the first frame serves as the visual condition. This preserves both T2V and I2V capabilities of the base model while keeping the distillation objective focused on short-horizon quality, motion realism, and efficient sampling.

3.2 Robustness Branch: Recovery-oriented Rollout Training

The robustness branch addresses the train–test mismatch introduced by autoregressive self-conditioning. At inference time, generated chunks become historical context for future chunks, so early errors can be recursively reused and amplified, causing identity drift, motion incoherence, and appearance artifacts (Yuan et al., 2026; Shen et al., 2026). Training only on clean histories therefore fails to expose the model to its deployment-time error distribution. A natural remedy is corrupted-history training, as explored in Helios (Yuan et al., 2026). However, local corrupted-context reconstruction only approximates an isolated perturbation, while long-horizon failure arises from accumulated self-generated drift. We therefore introduce *Recovery-oriented Rollout Training* (RRT): we perturb an early history chunk, let the model propagate this perturbation through multi-step autoregressive rollout, and apply supervision only after errors have sufficiently accumulated over time. RRT thus explicitly trains recovery under realistic model-induced context drift rather than one-step synthetic corruption. The detailed training procedure is illustrated in Figure 2.

Early-history perturbation. We partition each training video latent $\mathbf{x}_0$ into chunks $\{\mathbf{c}_k\}_{k=0}^{K+1}$, where $\mathbf{c}_0$ denotes the initial context, $\mathbf{c}_{1:K}$ are intermediate rollout chunks, and $\mathbf{c}_{K+1}$ is the final supervised target chunk. We construct the initial perturbed history as $\hat{\mathbf{c}}_0 = \mathcal{D}(\mathbf{c}_0)$, where $\mathcal{D}(\cdot)$ is a stochastic degradation operator. Inspired by Helios (Yuan et al., 2026), $\mathcal{D}$ randomly samples from a lightweight family of perturbations, including photometric distortion, additive noise, resolution degradation, partial latent masking, and identity mapping. Importantly, the perturbation is applied only to the earliest conditioning history. All subsequent histories are generated by the model itself and recursively reused as conditions. The initial perturbation therefore serves as a trigger for distribution shift, while later errors emerge from the model's own rollout dynamics.

Autoregressive rollout. Let $\mathbf{a}_k$ denote the audio condition aligned with chunk $\mathbf{c}_k$, and let y denote the text condition. For each intermediate chunk $k = 1, \ldots, K$, we initialize from Gaussian noise $\hat{\mathbf{c}}_{k,T} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and perform T denoising steps conditioned on the previous generated chunk:

$$\hat{\mathbf{c}}_{k,t} = \mathrm{sg}\left(G_\theta\left(\hat{\mathbf{c}}_{k,t+1}; \hat{\mathbf{c}}_{k-1,0}, \mathbf{r}, \mathbf{a}_k, y\right)\right), \tag{2}$$

where $\hat{\mathbf{c}}_{k-1,0}$ is the generated result of the previous chunk, G_θ denotes one denoising step of the sampler, and $\mathrm{sg}(\cdot)$ stops gradients through the rollout trajectory. As shown in Figure 2(a), rollout is executed without gradient tracking. The intermediate chunks $\hat{\mathbf{c}}_1, \ldots, \hat{\mathbf{c}}_K$ are not directly supervised, but to expose the final prediction to model-induced context drift, so that optimization targets recovery under the same error-propagation pattern encountered at inference time.

Masked flow matching objective. After K rollout steps, we supervise only the next ground-truth chunk $\mathbf{c}_{K+1}$, conditioned on the generated history $\hat{\mathbf{c}}_{K,0}$. Given noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and a flow-matching noise level $\sigma_t \sim p(\sigma)$, we corrupt only the target tokens:

$$\hat{\mathbf{c}}_{K+1,\sigma} = (1 - \sigma_t)\mathbf{c}_{K+1} + \sigma_t \boldsymbol{\epsilon}, \tag{3}$$

while keeping the rolled-out historical context $\hat{\mathbf{c}}_{K,0}$ clean. We then optimize a simple flow-matching (Lipman et al., 2022) objective on the final target chunk:

$$\mathcal{L}_{\mathrm{RRT}} = \mathbb{E}\big[\|v_\theta\big(\mathbf{c}_{K+1,\sigma}, \sigma_t; \hat{\mathbf{c}}_{K,0}, \mathbf{r}, \mathbf{a}_{K:K+1}, y\big) - \big(\boldsymbol{\epsilon} - \mathbf{c}_{K+1}\big)\|_2^2\big], \tag{4}$$

where v_θ is the velocity network, $\mathbf{a}_{K:K+1}$ denotes the audio condition aligned with the visible context-target window. We impose no auxiliary losses on the intermediate rollout chunks. Once realistic autoregressive degradation is induced through model-in-the-loop rollout, a standard flow-matching objective on the final target chunk is sufficient to train recovery from accumulated long-horizon errors.

Deployment. Both the efficiency and robustness branches are initialized from the same pretrained base model θ_0 so that they can be composed directly at deployment. Let $\Delta\theta_{\mathrm{DMD}}$ denote the update learned by the efficiency branch, and let $\Delta\theta_{\mathrm{RRT}}$ denote the LoRA update learned by the robustness branch. The final Avatar-Forever generator is obtained as

$$\theta^\star = \theta_0 + \Delta\theta_{\mathrm{DMD}} + \Delta\theta_{\mathrm{RRT}}. \tag{5}$$

This composition keeps the two objectives decoupled during training: full-parameter distillation provides a strong few-step generator, while the LoRA-based RRT branch adds long-horizon robustness with minimal deployment cost.

3.3 FOREVERCACHE: CHUNK-WISE AUTOREGRESSIVE HISTORY FEATURE CACHING

When producing chunk k, the autoregressive model denoises the current latent $\mathbf{c}_{k,t}$ while attending to a compact history window $\mathcal{H}_k = \{\mathbf{c}_0, \mathbf{c}_{k-1}\}$ that consists of the first and the most recent generated chunks. A naive implementation forwards the entire visible window $\{\mathcal{H}_k, \mathbf{c}_{k,t}\}$ through every transformer block at each denoising step. However, such a computation is structurally redundant: $\mathbf{c}_{k,t}$ changes throughout denoising, whereas the historical chunks remain fixed clean context. We introduce *ForeverCache*, an inference-time feature reuse mechanism that eliminates repeated computation on fixed historical context. For each autoregressive chunk, ForeverCache executes a full-window forward pass only at the initial denoising step, then reuses the resulting historical features throughout the remaining denoising trajectory. This design retains reliable access to long-range temporal context while restricting repeated computation to the dynamically evolving current chunk. The detailed illustration of ForeverCache can be found in the **Appendix**.

Formally, a standard autoregressive implementation evaluates the velocity network on the complete visible window at every step:

$$\mathbf{v}_{k,t} = v_\theta\left([\mathcal{H}_k, \mathbf{c}_{k,t}], \sigma_t; \mathbf{r}, \mathbf{a}_{\mathcal{H}_k:k}, y\right), \tag{6}$$

where $[\mathcal{H}_k, \mathbf{c}_{k,t}]$ is the history-current sequence input to LTX. Only the velocity prediction for the current chunk is retained for updating $\mathbf{c}_{k,t}$; the historical chunks act solely as conditioning context.

ForeverCache performs this full computation once, at $t = T$, while collecting the historical features from all L transformer blocks:

$$\mathbf{v}_{k,T}, \mathcal{C}_k = v_\theta^{\text{populate}}\left([\mathcal{H}_k, \mathbf{c}_{k,T}], \sigma_T; \mathbf{r}, \mathbf{a}_{\mathcal{H}_k:k}, y\right), \tag{7}$$

where $\mathcal{C}_k = \{\mathcal{C}_k^\ell\}_{\ell=1}^L$. For all subsequent steps, the model forwards only the current-chunk tokens and retrieves $\mathcal{C}_k$ as historical conditioning memory:

$$\mathbf{v}_{k,t} = v_\theta^{\text{reuse}}\left(\mathbf{c}_{k,t}, \sigma_t; \mathcal{C}_k, \mathbf{r}, \mathbf{a}_k, y\right), \quad t = T-1, \ldots, 0. \tag{8}$$

Here, $v_\theta^{\text{populate}}$ and v_θ^{reuse} share the same model parameters but differ in their execution mode. At each transformer block, only current tokens are processed as active tokens, while cached features provide historical context for video self-attention, audio self-attention, and cross-modal audio-video attention. The resulting prediction is scattered back to the original autoregressive layout, preserving the external denoising interface of the non-cached implementation. The cache is reset for each autoregressive chunk, ensuring bounded memory and preventing stale reuse as the history window changes. After the initial cache-population step, ForeverCache reduces the dominant transformer cost by forwarding only the evolving current-chunk tokens, while reusing cached video and audio history features to preserve multimodal conditioning for identity consistency, motion continuity, and lip synchronization. Note that ForeverCache is used only at inference time and applies directly to the distilled Avatar-Forever generator without modifying any learned weights.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Training data. We train our model using fully synthetic data. The details about our data synthetic pipeline can be found in the **Appendix**. Each sample contains an avatar video, paired audio, and a text prompt. Videos are partitioned into context-target pairs: a context chunk is used as the condition to predict a target chunk under the audio condition. Unless otherwise stated, we use four latent frames as the context to predict the subsequent four latent frames.

Implementation details. Avatar-Forever is built on the 22B LTX-2.3 (HaCohen et al., 2026). The efficiency branch distills the base model into a four-step generator using DMD (Yin et al., 2024), while the robustness branch independently trains video-side LoRA (Hu et al., 2022) adapters with both rank and alpha set to 128. For RRT, we perturb the historical context, roll out $K = 4$ autoregressive chunks, and apply the flow-matching loss to the following target chunk. Rollout uses the default 30-step denoising schedule without CFG. History degradation is applied with probability 0.5, including additive noise, blur, saturation, and latent masking. We always condition on the first frame through a zero-initialized gated module applied only to target denoising tokens. Both branches are optimized

Table 1: Quantitative comparison on the short- and long-video evaluation splits. Each entry reports the 5-second / 30-second results. Latency is measured in seconds. 'w/ FC' means Avatar-Forever with ForeverCache. The best and second-best results are marked in bold and underline, respectively.

Dataset	Model	LLM Judge				Automatic Metrics						Latency (s)↓
		A-V↑	Visual↑	Motion↑	Overall↑	IQA↑	ASE↑	Sync-C↑	Sync-D↓	FID↓	FVD↓	
EMTD	OmniAvatar	3.77/2.92	3.81/1.83	3.06/2.00	3.57/2.26	4.15/2.98	2.84/2.42	5.57/6.83	9.47/8.59	61.62/115.84	979.07/1834.79	850.00/> 1h
	InfiniteTalk	3.58/4.17	3.71/4.00	2.81/3.67	3.39/3.96	4.40/4.79	3.11/3.64	6.73/6.81	8.00/7.95	39.25/37.63	786.25/1125.80	51.50/309.20
	LiveAvatar	3.74/4.25	4.06/4.08	3.06/4.00	3.65/4.12	4.38/4.59	3.07/3.57	6.67/6.62	8.45/8.08	39.23/61.17	795.87/1373.52	53.01/320.95
	SoulX	3.87/4.08	3.97/4.08	3.13/3.75	3.68/3.98	4.44/4.55	3.14/3.45	6.83/6.83	7.97/7.96	38.91/33.46	785.20/868.38	25.43/125.26
	Ours w/ FC	3.81/4.33	4.00/4.25	3.13/4.00	3.67/4.23	4.77/4.84	3.28/3.70	7.11/6.68	8.39/8.01	48.76/34.52	**759.78**/905.97	**4.24/26.71**
	Ours	**3.90/4.42**	**4.16/4.50**	**3.32/4.08**	**3.82/4.32**	**4.80/4.88**	**3.33/3.73**	**7.56/6.85**	**7.95/7.94**	**38.37/33.33**	775.11/**858.06**	5.24/38.85
HDTF	OmniAvatar	3.73/3.80	3.75/3.70	2.88/3.00	3.48/3.53	4.01/4.03	2.70/3.06	6.14/7.30	9.14/7.61	21.31/24.01	390.61/956.60	850.00/> 1h
	InfiniteTalk	3.34/4.00	3.58/3.80	2.47/3.00	3.16/3.63	4.03/4.15	2.75/3.16	8.14/7.43	7.06/7.60	26.71/19.48	390.69/793.74	51.50/309.20
	LiveAvatar	3.90/3.80	3.95/3.80	3.08/3.00	3.67/3.56	4.19/4.07	2.74/3.23	7.39/6.23	8.39/8.45	23.42/95.35	426.26/1675.81	53.01/320.95
	SoulX	3.82/4.00	3.80/3.70	2.95/3.20	3.55/3.66	4.05/4.17	2.75/3.20	8.48/7.50	7.10/7.54	21.14/19.41	401.39/1066.36	25.43/125.26
	Ours w/ FC	3.73/**4.10**	3.88/3.90	3.15/3.20	3.61/3.80	**4.42/4.42**	**3.38/3.38**	8.56/7.56	7.06/7.56	**20.19**/17.37	393.92/686.96	**4.24/26.71**
	Ours	**3.90**/4.00	**4.03/4.10**	**3.25/3.40**	**3.75/3.82**	4.30/4.38	2.89/3.33	**8.68/7.56**	**6.94/7.50**	20.91/**17.19**	**378.17/593.92**	5.24/38.85
TalkVid	OmniAvatar	3.66/4.00	3.71/3.80	2.84/3.50	3.43/3.78	3.68/4.12	2.41/3.13	4.45/6.49	9.77/8.94	48.15/64.70	618.67/1064.48	850.00/> 1h
	InfiniteTalk	3.25/4.10	3.13/3.90	2.60/3.40	3.01/3.82	3.51/4.30	2.31/3.26	5.48/6.35	8.87/8.90	53.10/52.09	669.80/1074.68	51.50/309.20
	LiveAvatar	3.65/4.00	3.58/3.90	2.95/3.40	3.41/3.79	3.79/4.35	2.48/3.35	4.91/6.07	9.11/8.64	47.47/70.21	668.96/1308.11	53.01/320.95
	SoulX	3.63/4.20	3.65/4.20	2.90/3.50	3.42/3.99	3.70/4.17	2.45/3.20	5.61/6.60	8.89/8.24	47.58/51.41	720.48/1066.36	25.43/125.26
	Ours w/ FC	3.78/**4.30**	3.83/**4.40**	3.23/3.80	3.63/4.19	3.90/4.47	2.59/3.35	5.68/6.51	8.81/8.29	49.56/53.06	**543.87**/998.75	**4.24/26.71**
	Ours	**3.85/4.30**	**3.85/4.40**	**3.35/3.90**	**3.70/4.22**	**3.96/4.51**	**2.62/3.38**	**6.01/6.62**	**8.61/8.23**	**47.43/49.82**	554.00/**994.42**	5.24/38.85

with AdamW at a learning rate of 1×10^{-5} and a global batch size of 256; DMD distillation and RRT training run for 5,000 and 3,000 steps, respectively.

Evaluation. We evaluate short-horizon quality and long-horizon stability on TalkVid (Chen et al., 2026a), EMTD (Meng et al., 2024), and HDTF (Zhang et al., 2021), using 5-second and 30-second splits with 40 samples each following previous evaluation protocal. Ablations are conducted on EMTD unless otherwise specified. We compare with OmniAvatar (Gan et al., 2025), LiveAvatar (Huang et al., 2025b), SoulX-FlashTalk (Shen et al., 2026), and InfiniteTalk (Yang et al., 2025a) under their official inference settings, using identical audio and reference inputs. We evaluate audio–visual consistency, visual quality, and motion naturalness with `Gemini-Flash-3.5` (Google DeepMind, 2026), which assigns 1–5 scores combined with weights 0.35, 0.35, and 0.30, respectively. We additionally report FID, FVD, Q-Align image quality (IQA), aesthetics (ASE), and Sync-C/Sync-D. Finally, 20 participants rate anonymized and randomly ordered videos on the EMTD long-horizon split using the same perceptual criteria. Further details on the evaluation splits, metric implementation, LLM prompt, and user-study protocol are provided in the **Appendix**.

4.2 Main Results

Short-horizon quality. The 5-second results in Table 1 show that Avatar-Forever preserves the generation quality of the video foundation model under few-step inference. Compared to the strongest competing baseline, Avatar-Forever improves the LLM Overall score by 4.6% on average, together with average gains of 5.1% in IQA, 5.6% in ASE, and 6.7% in Sync-C, while consistently reducing Sync-D. The Motion score improves by up to 13.6%, extending the advantage beyond frame-level fidelity and synchronization to more natural speech-driven dynamics.

Long-horizon stability. The 30-second results in Table 1 further demonstrate Avatar-Forever's robustness under autoregressive rollout. It achieves the best LLM Overall score on all three datasets, outperforming the strongest prior baseline by 5.0% on average. It also improves LLM Visual quality by 7.6% on average and Motion quality by up to 11.4%, while obtaining the best FID and FVD across all datasets. In particular, FID is reduced by 5.0% on average and FVD by 25.2% on HDTF, supporting improved distributional realism under accumulated context errors. Figure 3 provides complementary qualitative evidence for these gains. As highlighted by the red boxes, competing methods exhibit progressive visual degradation and appearance drift, repetitive low-diversity gesture cycles, motion-induced hand blur and structural distortion, or over-articulated mouth motion accompanied by stereotyped upper-body gestures. In contrast, our Avatar-Forever preserves identity, facial and hand details, and diverse natural motion throughout the extended rollout. Its ForeverCache variant retains comparable long-horizon stability despite the substantial inference acceleration. Qualitative comparisons for both short-video generation and extended-duration rollouts are provided in the **Appendix**, where we show that Avatar-Forever generates more diverse and natural human motion, better preserves identity, and even maintains natural background dynamics.



Figure 3: **Visual comparison on 30-second generation.** Avatar-Forever maintains identity, facial detail, hand structure, and motion diversity over extended autoregressive rollout. In contrast, competing methods exhibit progressive appearance drift, repetitive low-diversity gestures, motion-induced hand artifacts, or over-articulated mouth motion with stereotyped upper-body actions. ForeverCache retains comparable visual stability with substantially faster inference.

Table 2: Double-blind user study results on the EMTD long-video evaluation split. Best results are highlighted in bold, and second-best results are underlined.

Model	Audio-Visual↑	Visual↑	Motion↑	Overall↑
OmniAvatar	57.5	32.5	29.0	40.20
InfiniteTalk	68.0	51.5	45.0	55.33
LiveAvatar	70.0	54.5	47.0	57.68
SoulX-FlashTalk	71.0	56.5	52.0	60.23
Ours w/ ForeverCache	<u>73.0</u>	<u>76.5</u>	<u>75.0</u>	<u>74.83</u>
Ours	**74.0**	**78.0**	**76.0**	**76.00**

Efficiency-quality balance. ForeverCache improves streaming efficiency by reusing fixed historical context features across denoising steps. On short videos, it reduces latency by 19.1% and increases throughput by 23.6% compared to standard Avatar-Forever, while remaining approximately $6.0\times$ faster than the fastest prior baseline. On 30-second generation, it reduces runtime from 38.85s to 26.71s, corresponding to a 31.2% latency reduction and a 45.5% throughput increase; it remains approximately $4.7\times$ faster than the fastest prior method. Despite this acceleration, ForeverCache preserves most quality gains of the standard model: it remains competitive across perceptual, synchronization, and distributional metrics on short videos, and achieves the second-best LLM Overall score on all long-video splits while outperforming the strongest prior baseline by 3.8% on average. Figures 3 show that cache-enabled inference retains the key visual behavior of the full model without visible collapse or systematic temporal degradation.

Human perceptual validation. Table 2 reports a double-blind user study on the EMTD long-video split. Twenty participants rate videos on four perceptual dimensions as the LLM judge using integer scores from 1 to 5; scores are averaged and normalized to a 0–100 scale. The human evaluation is consistent with the automatic results: Avatar-Forever ranks first on all dimensions, while its variant with ForeverCache ranks second with a small gap from the full model.

Table 3: Ablation study on decoupled training. We compare the proposed decoupled training strategy with the sequential training baseline (DMD only) and FM-only long-horizon adaptation. Best results are highlighted in bold, and second-best results are underlined.

Setting	LLM Judge				Automatic Metrics					
	Audio-Visual↑	Visual↑	Motion↑	Overall↑	IQA↑	ASE↑	Sync-C↑	Sync-D↓	FID↓	FVD↓
Decoupled DMD + RRT	**4.250**	**4.267**	**3.750**	**4.105**	**4.851**	**3.684**	**6.694**	**8.032**	**34.978**	**901.811**
Sequential (DMD only)	<u>4.083</u>	<u>4.167</u>	<u>3.583</u>	<u>3.962</u>	<u>4.793</u>	<u>3.671</u>	<u>6.427</u>	<u>8.085</u>	<u>39.515</u>	1080.230
FM only	2.750	1.917	2.250	2.308	3.409	2.590	2.031	10.760	76.577	<u>1071.200</u>

Avatar-Forever's advantage is pronounced in Visual and Motion quality, reflecting temporal naturalness not captured by automatic metrics. Although several baselines preserve identity and produce plausible frames, their long videos often exhibit limited upper-body motion, static backgrounds, repetitive gestures, rigid postures, or exaggerated mouth shapes. In contrast, Avatar-Forever produces more expressive facial dynamics, smoother gesture transitions, and speech-aligned motion, improving perceived realism and yielding the highest Overall human score. ForeverCache preserves most of these perceptual gains while providing faster inference.

Analysis of Decoupled Training. Table 3 and Figure 7 in the **Appendix** contrast three strategies for combining few-step efficiency with long-horizon robustness. FM only trains RRT under the original multi-step denoising schedule. Stable autoregressive rollouts show that rollout-oriented flow matching can learn recovery; however, without distillation, four-step sampling degrades visual quality. A natural alternative incorporates the same recovery signal into prior sequential distillation pipelines (Huang et al., 2026; Cui et al., 2025). Sequential (DMD only) therefore performs RRT-like rollout during DMD training. In this setting, a single DMD objective must match the few-step generation distribution and correct long-horizon rollout errors. Moreover, later stages inherit errors introduced by earlier ones. These competing demands make optimization unstable and the multi-stage pipeline difficult to diagnose. Avatar-Forever instead separates the two objectives into parallel branches, optimizes each independently, and composes their parameter updates only at deployment. The decoupled model outperforms Sequential (DMD only), improving LLM Overall and Sync-C by 3.6% and 4.2% while reducing FID and FVD by 11.5% and 16.5%. Beyond these gains, the parallel design simplifies training and eliminates cross-stage error propagation.

Effect of Recovery-oriented Rollout Training. Figure 8 in the **Appendix** analyzes the effect of the rollout horizon K in RRT. When $K = 0$, supervision is applied immediately after context corruption which matches the Helios setting (Yuan et al., 2026), reducing training to local reconstruction and leaving the model vulnerable to accumulated autoregressive errors, such as spot-like artifacts and global color drift. Increasing the horizon progressively exposes the model to self-generated rollout errors: $K = 1$ alleviates these issues but still exhibits noticeable appearance fluctuations, while $K = 2$ and $K = 4$ further improve appearance, background, and color consistency, with $K = 4$ yielding the most stable long-horizon generation. These results confirm that the effectiveness of RRT comes from supervising recovery after model-induced errors have propagated through autoregressive rollout, rather than from simply injecting synthetic corruption into the historical context.

5 Conclusion

We presented Avatar-Forever, a framework for real-time long-duration audio-driven avatar generation. Our key insight was that the quality of few-step generation and the robustness to autoregressive drift are distinct adaptation objectives. We therefore learned them independently: DMD produced an efficient few-step generator, RRT trained a lightweight adapter to recover from accumulated rollout errors, and ForeverCache reduced redundant historical computation during streaming inference. Built on a 22B video foundation model and a fully synthetic data pipeline, Avatar-Forever improved perceptual quality, audio–visual synchronization, long-horizon stability, and inference efficiency. These results supported a simple principle for scalable video adaptation: learn efficiency and long-horizon robustness separately, then compose them at deployment.

Limitations and future work. Avatar-Forever achieves 27.2 FPS at 768×512 resolution on a single H100 GPU, but we have not yet optimized it for consumer-grade hardware. In addition, although our training and optimization are tailored to audio-driven avatars, we observe a good generalization to broader video-generation scenarios. Future work will explore domain-specific data construction and training to further unlock Avatar-Forever's potential for general long-horizon video generation.

AI USE STATEMENT

Generative AI tools were used solely to improve the readability of author-written text, including grammar correction, wording refinement, and language polishing. They were not used to generate research ideas or hypotheses, develop theoretical or conceptual frameworks, formulate mathematical claims or proofs, design or implement methods or experiments, process datasets, or analyze or interpret experimental results; other tasks requiring disclosure are not applicable to this work. All AI-assisted edits were carefully reviewed by the authors to ensure that they preserved the intended technical meaning and introduced no unsupported claims, results, or citations. The authors 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

This work focuses on methodological research in audio-driven avatar video generation and evaluates generated content solely in a research setting. It does not target individuals or demographic groups and is not intended for high-stakes decision-making, surveillance, impersonation, or other harmful applications. The user study is limited to subjective comparisons of generated video quality and does not involve sensitive personal information or consequential decisions. Nevertheless, as with other video generation technologies, the proposed method could potentially be misused to produce misleading or inappropriate content. We therefore encourage its responsible use for research purposes in accordance with applicable safeguards and regulations. Beyond these general risks associated with generative video models, we are not aware of additional ethical concerns specific to this work.

REPRODUCIBILITY STATEMENT

We provide the training budget, implementation settings, optimization hyperparameters, and other experimental details required to reproduce our results in Sec. 4.1 and Appendix. Additional algorithmic descriptions, experimental configurations, and qualitative evaluation details are provided in the main paper and supplementary material to facilitate independent reproduction and verification of our results.

AUTHOR CONTRIBUTIONS

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 958–968, 2024.

Shunian Chen, Hejin Huang, Yexin Liu, Zihan Ye, Pengcheng Chen, Chenghao Zhu, Michael Guan, Rongsheng Wang, Junying Chen, Jianye Hou, et al. Talkvid: A large-scale diversified dataset for audio-driven talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3492–3500, 2026a.

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

Yukang Chen, Luozhou Wang, Wei Huang, Shuai Yang, Bohan Zhang, Yicheng Xiao, Ruihang Chu, Weian Mao, Qixin Hu, Shaoteng Liu, et al. Longlive-2.0: An nvfp4 parallel infrastructure for long video generation. *arXiv preprint arXiv:2605.18739*, 2026b.

Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditioning. *arXiv preprint arXiv:2407.08136*, 2024.

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. *arXiv preprint arXiv:2510.02283*, 2025.

Liyuan Cui, Wentao Hu, Wenyuan Zhang, Zesong Yang, Fan Shi, and Xiaoqiang Liu. Avatarforcing: One-step streaming talking avatars via local-future sliding-window denoising. *arXiv preprint arXiv:2603.14331*, 2026.

Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander Miller, Arthur Szlam, and Jason Weston. Evaluating prerequisite qualities for learning end-to-end dialog systems, 2016.

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.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023.

Google DeepMind. Gemini 3.5 flash. `https://deepmind.google/models/gemini/flash/`, 2026. Accessed: 2026-07-08.

Ying Guo, Xi Liu, Cheng Zhen, Pengfei Yan, and Xiaoming Wei. Arig: Autoregressive interactive head generation for real-time conversations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12956–12965, 2025.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025a.

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, Jiaming Liu, and Steven Hoi. Live avatar: Streaming real-time audio-driven avatar generation with infinite length. *arXiv preprint arXiv:2512.04677*, 2025b.

Chunyu Li, Chao Zhang, Weikai Xu, Jinghui Xie, Weiguo Feng, Bingyue Peng, and Weiwei Xing. Latentsync: Audio conditioned latent diffusion models for lip sync. *arXiv e-prints*, pp. arXiv–2412, 2024.

Ruibin Li, Tao Yang, Song Guo, and Lei Zhang. Rorem: Training a robust object remover with human-in-the-loop. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14024–14035, 2025a.

Ruibin Li, Tao Yang, Fangzhou Ai, Tianhe Wu, Shilei Wen, Bingyue Peng, and Lei Zhang. Long-horizon streaming video generation via hybrid attention with decoupled distillation. *arXiv preprint arXiv:2604.10103*, 2026.

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, 2025b.

Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable video infinity: Infinite-length video generation with error recycling. *arXiv preprint arXiv:2510.09212*, 2025c.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Chetwin Low and Weimin Wang. Talkingmachines: Real-time audio-driven facetime-style video via autoregressive diffusion models. *arXiv preprint arXiv:2506.03099*, 2025.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, et al. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34385–34397, 2026.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Rang Meng, Xingyu Zhang, Yuming Li, and Chenguang Ma. Echomimicv2: Towards striking, simplified, and semi-body human animation. *arXiv preprint arXiv:2411.10061*, 2024.

OpenAI. Video generation models as world simulators. `https://openai.com/index/video-generation-models-as-world-simulators/`, 2024. Technical report.

OpenAI. Chatgpt. `https://chatgpt.com/`, 2026. Accessed: 2026-07-17.

K R Prajwal, Rudrabha Mukhopadhyay, Vinay Namboodiri, and C V 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.

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

Axel Sauer, Dominik Lorenz, Andreas Blattmann, Puneet Dokania, Stefano Ermon, Andreas Geiger, Patrick Esser, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103, 2024.

Le Shen, Qian Qiao, Tan Yu, Ke Zhou, Tianhang Yu, Yu Zhan, Zhenjie Wang, Dingcheng Zhen, Ming Tao, Shunshun Yin, and Siyuan Liu. Soulx-flashtalk: Real-time infinite streaming of audio-driven avatars via self-correcting bidirectional distillation. *arXiv preprint arXiv:2512.23379*, 2026.

Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1982–1991, 2023.

SII-GAIR, Sand. ai, Ethan Chern, Hansi Teng, Hanwen Sun, Hao Wang, Hong Pan, Hongyu Jia, Jiadi Su, Jin Li, Junjie Yu, Lijie Liu, Lingzhi Li, Lyumanshan Ye, Min Hu, Qiangang Wang, Quanwei Qi, Steffi Chern, Tao Bu, Taoran Wang, Teren Xu, Tianning Zhang, Tiantian Mi, Weixian Xu, Wenqiang Zhang, Wentai Zhang, Xianping Yi, Xiaojie Cai, Xiaoyang Kang, Yan Ma, Yixiu Liu, Yunbo Zhang, Yunpeng Huang, Yutong Lin, Zewei Tao, Zhaoliang Liu, Zheng Zhang, Zhiyao Cen, Zhixuan Yu, Zhongshu Wang, Zhulin Hu, Zijin Zhou, Zinan Guo, Yue Cao, and Pengfei Liu. Speed by simplicity: A single-stream architecture for fast audio-video generative foundation model. *arXiv preprint arXiv:2603.21986*, 2026.

Zhiyao Sun, Ziqiao Peng, Yifeng Ma, Yi Chen, Zhengguang Zhou, Zixiang Zhou, Guozhen Zhang, Youliang Zhang, Yuan Zhou, Qinglin Lu, and Yong-Jin Liu. Streamavatar: Streaming diffusion models for real-time interactive human avatars. *arXiv preprint arXiv:2512.22065*, 2026. Accepted by CVPR 2026.

Team Seedance. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive: Generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*, 2024.

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

Fu-Yun Wang, Zhaoyang Huang, Alexander William Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, Xiaogang Wang, and Hongsheng Li. Phased consistency models. In *Advances in Neural Information Processing Systems*, 2024.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025.

Tianhe Wu, Ruibin Li, Lei Zhang, and Kede Ma. Diversity-preserved distribution matching distillation for fast visual synthesis. *arXiv preprint arXiv:2602.03139*, 2026.

Yusong Wu*, Ke Chen*, Tianyu Zhang*, Yuchen Hui*, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.

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

Yuelang Xu, Hongwen Zhang, Lizhen Wang, Xiaochen Zhao, Han Huang, Guojun Qi, and Yebin Liu. Latentavatar: Learning latent expression code for expressive neural head avatar. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–10, 2023.

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.

Tao Yang, Ruibin Li, Yangming Shi, Yuqi Zhang, Qide Dong, Haoran Cheng, Weiguo Feng, Shilei Wen, Bingyue Peng, and Lei Zhang. Many-for-many: Unify the training of multiple video and image generation and manipulation tasks. *arXiv preprint arXiv:2506.01758*, 2025c.

Hidir Yesiltepe, Tuna Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40256–40265, 2026.

Tianwei Yin, Micha"el Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. Improved distribution matching distillation for fast image synthesis. *arXiv preprint arXiv:2405.14867*, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22963–22974, 2025.

Shenghai Yuan, Yuanyang Yin, Zongjian Li, Xinwei Huang, Xiao Yang, and Li Yuan. Helios: Real real-time long video generation model. *arXiv preprint arXiv:2603.04379*, 2026.

Ailing Zeng, Casper Yang, Chauncey Ge, Eddie Zhang, Garvey Xu, Gavin Lin, Gilbert Gu, Jeremy Pi, Leo Li, Mingyi Shi, Shawn Wang, Sheng Bi, Steven Tang, Thorn Hang, Tobey Guo, Vincent Li, Xin Tong, Yikang Li, Yuchen Sun, Yue Zhao, Yuhan Lu, Yuwei Li, Zane Zhang, Zeshi Yang, and Zi Ye. Lpm 1.0: Video-based character performance model. *arXiv preprint arXiv:2604.07823*, 2026.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8652–8661, 2023.

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

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. *arXiv preprint arXiv:2602.02214*, 2026.

APPENDIX

In this appendix, we provide additional implementation details, evaluation protocols, ablation analyses, and qualitative results that complement the main paper. Specifically, the supplement includes:

- **A.** Details of the synthetic data construction pipeline, including dialogue curation, prompt construction, LTX video synthesis, and quality-aware filtering (referring to Sec. 4.1 in the main paper);
- **B.** A detailed illustration of ForeverCache and its layer-wise history feature reuse mechanism (referring to Sec. 3.3 in the main paper);
- **C.** Training infrastructure;
- **D.** The complete evaluation protocol, including evaluation split construction, metric definitions, the LLM-based perceptual evaluation prompt (referring to Sec. 4.1 in the main paper);
- **E.** Visual Comparison of Decoupled Training and Different Rollout Horizons K (referring to Sec. 4.2).
- **F.** Additional visual comparisons for 5-second and 30-second audio-driven avatar generation, including results with and without ForeverCache (referring to Sec. 4.2 in the main paper), further illustrating the long-horizon stability of Avatar-Forever.

Because temporal quality cannot be fully assessed from sparsely sampled frames, we provide the corresponding **video** in the **supplementary material**. For better viewing experience, we uploaded the video demo and comparison videos to a dedicated anonymous website https://anonymous.4open.science/w/avatarforever-ano-page-23AC/, where the videos can be played directly in the browser. We recommend viewing these videos with audio when assessing lip synchronization, motion naturalness, identity preservation, and long-horizon stability.

A SYNTHETIC DATA CONSTRUCTION PIPELINE

Long-horizon avatar adaptation requires training videos with persistent identity, meaningful human motion, and reliable semantic alignment over extended durations. Such data are difficult to collect from the web: real videos often exhibit identity changes, weak text–video correspondence, limited facial or body motion, and inconsistent visual quality. We address this limitation with a synthetic data construction pipeline that generates and filters avatar videos using the same video foundation model adapted by our framework. This construction pipeline provides controllable training data aligned with the target generation distribution, without large-scale manual curation of long-form real videos.

Dialogue and prompt synthesis. We collect conversational text from the publicly available MDD corpus (Dodge et al., 2016) and use GPT (OpenAI, 2026) to filter and refine incoherent, low-quality, or visually uninformative samples. The retained dialogues are converted into structured video prompts following the LTX prompting format. Each prompt specifies the character, appearance, scene, camera view, facial expression, body motion, and speech content, encouraging visible articulation and interaction-relevant motion in the resulting videos, as illustrated in the red part of Figure 4.

LTX video synthesis. We feed the synthesized prompts to the pretrained LTX base model and generate avatar videos with standard multi-step sampling. Unlike heterogeneous web videos, these samples are produced under controllable visual and motion conditions and remain close to the distribution of the foundation model being adapted. The resulting corpus therefore provides a targeted source of training data for long-horizon audio-driven avatar generation.

Quality-aware filtering. Although synthetic generation improves controllability, individual samples may still exhibit semantic mismatch, visual artifacts, static content, or camera-dominated motion. We therefore filter the generated videos along three complementary dimensions. We assess semantic consistency using multimodal similarity and reward models, including ImageBind (Girdhar et al., 2023), CLAP (Wu* et al., 2023), and Unified Reward Model (Wang et al., 2025), and use Gemini (Google DeepMind, 2026) to remove visually degraded videos. To assess motion quality, we compare adjacent frames to identify clips with negligible temporal change or motion dominated by simple global transformations. In particular, we reject nearly static videos and videos whose frame-to-frame variation is primarily explained by global affine camera motion, such as uniform

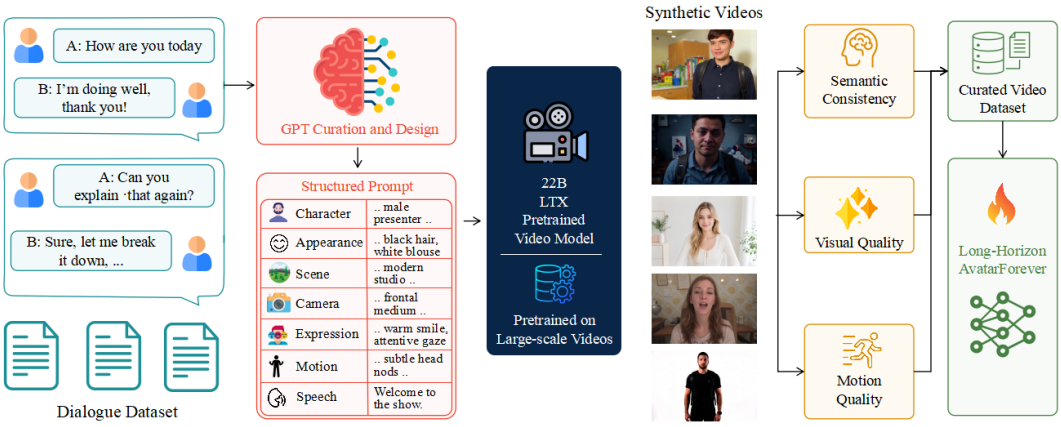

Figure 4: **Synthetic data contruction pipeline for long-horizon avatar adaptation.** Public dialogues are filtered and converted into structured video prompts, from which the pretrained LTX model synthesizes controllable avatar videos. Quality-aware filtering removes semantically mismatched, visually degraded, static, and camera-dominated samples, yielding a curated synthetic corpus for long-horizon avatar adaptation.

vertical camera movement, translation, or zooming, rather than local facial and body dynamics. For dialogue-driven samples, we further retain only videos with sufficient local non-rigid motion, including visible facial articulation or body movement. This procedure yields a curated synthetic corpus with stronger semantic consistency, visual fidelity, and motion diversity for long-horizon adaptation.

B Detailed Illustration of ForeverCache

Figure 5 provides a detailed comparison between standard autoregressive denoising and ForeverCache. When generating a new video chunk, the current latent evolves from C_T to C_0 over T denoising steps, whereas the history chunks $\{H_i\}_{i=0}^{N-1}$ remain fixed throughout the entire denoising trajectory. As shown in the top row, standard inference nevertheless processes the complete sequence $[H_0, \ldots, H_{N-1}, C_t]$ through all L LTX DiT blocks at every step. Since only the current chunk is progressively updated, repeatedly recomputing the representations of the unchanged history introduces substantial redundant computation.

ForeverCache removes this redundancy through layer-wise history feature reuse, as illustrated in the bottom row. At the first denoising step, it performs a full-window forward pass and stores the historical representation B_H^ℓ produced at each transformer block ℓ. During all subsequent steps, only the evolving current-chunk tokens are actively forwarded, while the corresponding cached history features are retrieved at each block as conditioning context. Because the visible history remains unchanged within the denoising trajectory, the cache can be safely reused until the current chunk is fully denoised. It is then refreshed when generation advances to the next autoregressive chunk and the history window changes.

By caching block-specific historical representations, ForeverCache preserves the contextual information required for identity consistency, motion continuity, and audio–visual synchronization while substantially reducing repeated transformer computation. The mechanism is applied only at inference time, requires no modification to the learned model parameters, and maintains bounded memory by resetting the cache for each autoregressive chunk.

C Training Infrastructure

The model training is conducted on a Linux-based high-performance computing cluster equipped with 32 NVIDIA H100 GPUs. Each GPU provides 80 GB of high-bandwidth memory, resulting in a total aggregate GPU memory capacity of 2.56 TB. The large-scale GPU configuration enables

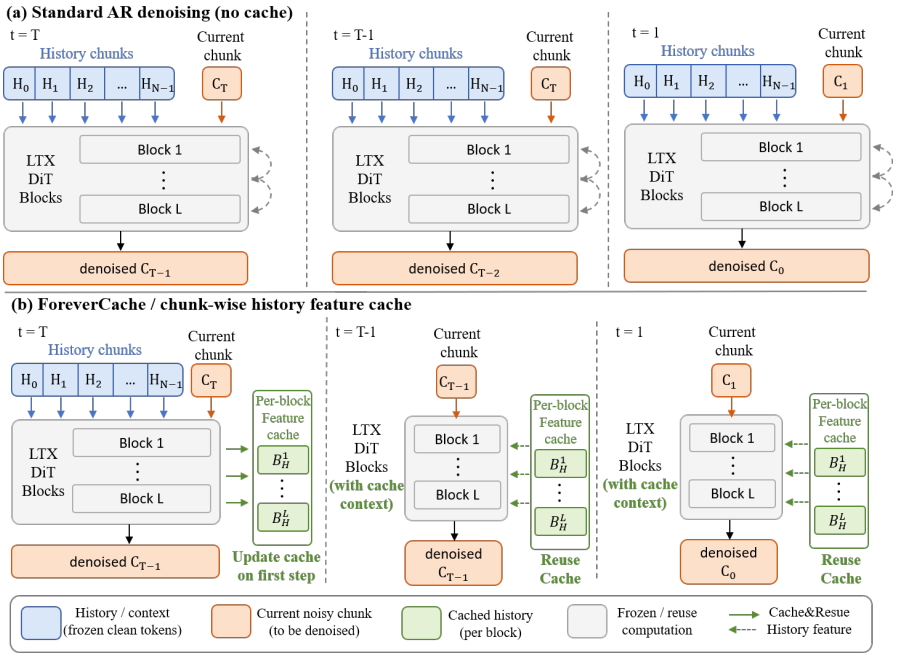

Figure 5: **Detailed illustration of ForeverCache.** *Top:* Standard autoregressive denoising repeatedly forwards the fixed history chunks together with the evolving current chunk through all LTX DiT blocks at every denoising step. *Bottom:* ForeverCache performs a full-window forward pass only at the first denoising step to populate a per-block cache of historical features. Subsequent steps forward only the current-chunk tokens while reusing the cached history features as layer-wise conditioning context. The cache is refreshed once for each newly generated autoregressive chunk.

distributed training of the proposed model while accommodating high-resolution inputs, long temporal sequences, and memory-intensive intermediate representations. The training pipeline is implemented using PyTorch and executed in a distributed multi-GPU environment. Model parameters, optimizer states, input batches, and computational workloads are partitioned across GPUs to improve training throughput and memory efficiency. NVIDIA CUDA and cuDNN are used to accelerate tensor operations, convolutional computations, and other GPU-intensive components. Where applicable, mixed-precision computation is employed to reduce memory consumption and improve computational efficiency on the H100 architecture.

D Evaluation Protocol

We evaluate short-horizon quality and long-horizon stability on TalkVid (Chen et al., 2026a), EMTD (Meng et al., 2024), and HDTF (Zhang et al., 2021). We construct a 5-second split for short-horizon quality and a 30-second split for long-horizon stability, each containing 40 samples with paired audio, reference identity, and text conditions. Unless otherwise specified, ablations are conducted on EMTD. We compare with representative audio-driven and long-horizon avatar methods, including OmniAvatar (Gan et al., 2025), LiveAvatar (Huang et al., 2025b), SoulX-FlashTalk (Shen et al., 2026), and InfiniteTalk (Yang et al., 2025a). For each baseline, we follow the official inference configuration whenever available and use identical audio and reference inputs for comparison. Our choices of benchmark datasets and metrics follow the established protocols in these prior works. We use the same evaluation protocol for the 5-second and 30-second comparisons; within each duration, every method is evaluated on the same 40 samples with identical audio, reference identity, and text conditions, and all outputs are scored using the same metric implementations.

Evaluation metrics. We evaluate visual quality, motion realism, audio–visual synchronization, and video-distribution similarity. We report Fréchet Inception Distance (FID), Fréchet Video Distance (FVD), Q-Align image quality (IQA), Q-Align aesthetic score (ASE), and Sync-C/Sync-D for lip

```
You are an expert evaluator for audio-driven talking avatar videos. Your task is to evaluate the quality of a generated avatar video based only on the provided audio and video. Please act as a strict but fair human judge.
Evaluate the video from the following three aspects:
Audio-Visual Consistency / Lip-Sync Quality
Assess whether the facial movements, mouth shapes, lip motion, head motion, and speaking rhythm are well synchronized with the input audio.5: The mouth movements and facial dynamics are highly synchronized with the audio. Lip shapes, speaking rhythm, pauses, and expressions are natural and convincing.
4: Mostly synchronized, with only minor lip-sync or timing errors.
3: Roughly synchronized, but noticeable mismatches exist in mouth motion, rhythm, or expression.
2: Poor synchronization. The mouth often moves at the wrong time or does not match the speech well.
1: Severe mismatch between audio and video. The avatar appears unrelated to the audio.

Visual Quality
Assess the overall visual fidelity of the video, including face quality, sharpness, identity consistency, lighting, texture realism, background stability, and absence of artifacts.5: High visual quality with realistic face details, stable identity, clean textures, natural lighting, and almost no artifacts.
4: Good visual quality with minor artifacts, slight blur, or small temporal inconsistencies.
3: Acceptable quality, but with visible artifacts, flickering, blur, identity drift, or unnatural textures.
2: Low quality with strong artifacts, unstable face, obvious distortion, or degraded background.
1: Very poor visual quality with severe artifacts, broken face structure, or unusable video.

Motion Naturalness / Content Naturalness
Assess whether the avatar behaves naturally over time. Focus on head movement, blinking, facial expressions, body gestures, breathing, and whether the motion is diverse and contextually appropriate.5: The avatar shows natural, smooth, and diverse motion. Expressions, blinking, head movement, and small gestures are realistic and not repetitive.
4: Mostly natural motion, with only slight stiffness or occasional repetition.
3: Somewhat natural, but the motion is limited, mechanical, repetitive, or lacks expressive variation.
2: Clearly unnatural motion. The avatar repeats the same gestures, has frozen expressions, or shows robotic movement.
1: Extremely unnatural. The avatar is almost static, heavily repetitive, jittery, or physically implausible.

Important evaluation rules:
Do not judge based on the identity, gender, ethnicity, or appearance of the person.
Do not reward a video simply because it has large motion; motion should be natural and appropriate.
Penalize repetitive behaviors, such as repeatedly nodding, blinking, smiling, moving hands, or making the same head motion without semantic variation.
Penalize long-term degradation, including identity drift, accumulated blur, increasing artifacts, or motion becoming frozen/repetitive over time.
If the video contains a speaking avatar, prioritize whether the avatar looks like a natural person speaking the given audio.
Be strict. A score of 5 should only be given when the video is nearly indistinguishable from a high-quality real talking video.
Please output your evaluation in the following JSON format only:
{
"audio_visual_consistency": {
"score": <integer from 1 to 5>,
"reason": "<brief explanation>"
},
"visual_quality": {
"score": <integer from 1 to 5>,
"reason": "<brief explanation>"
},
"motion_naturalness": {
"score": <integer from 1 to 5>,
"reason": "<brief explanation>"
},
"overall_score": <weighted average score from 1 to 5>,
"overall_reason": "<brief final judgment>"
}
Compute the overall_score using the following weights:
Audio-Visual Consistency: 0.35
Visual Quality: 0.35
Motion Naturalness: 0.30
Round the overall_score to two decimal places.
```

Figure 6: **System prompt for LLM-based perceptual evaluation.** The same evaluation prompt is applied to all methods, datasets, and video durations. It defines the scoring criteria for audio–visual consistency, visual quality, and motion naturalness, as well as the weighted aggregation used to compute the overall score.

synchronization, where higher Sync-C and lower Sync-D are better. We further conduct a user study on the EMTD long-horizon split with 20 participants. Participants rate anonymized and randomly ordered videos from Avatar-Forever and competing methods using the same three perceptual criteria as the LLM judge. We report the mean participant score, with randomization and anonymization used to reduce ordering and model-identity bias.

Conventional automatic metrics measure image quality, video realism, and audio–visual synchronization, but they do not fully capture whether an avatar behaves like a natural speaker throughout a continuous video. In particular, temporally accumulated artifacts, repetitive gestures, frozen expres-

sions, and unnatural motion transitions can be difficult to characterize using individual automatic metrics. We therefore employ `Gemini-Flash-3.5` (Google DeepMind, 2026) as a multimodal perceptual judge. For each generated sample, the evaluator receives only the corresponding audio and video and assesses the result using a fixed evaluation rubric shared across all methods, datasets, and video durations.

The evaluation covers three complementary dimensions. *Audio–Visual Consistency* measures whether mouth shapes, facial dynamics, head motion, speaking rhythm, and pauses are synchronized with the input audio. *Visual Quality* evaluates facial fidelity, sharpness, identity consistency, lighting, texture realism, background stability, and temporal artifacts such as flickering, blur, and appearance drift. *Motion Naturalness* focuses on the temporal behavior of the avatar, including facial expressions, blinking, head and upper-body movement, breathing, gesture diversity, contextual appropriateness, and physical plausibility.

Each dimension is assigned an integer score from 1 to 5 according to explicit quality anchors, with a score of 5 reserved for videos that are nearly indistinguishable from high-quality real talking videos. The overall perceptual score is computed as:

$$S_{\text{overall}} = 0.35S_{\text{A-V}} + 0.35S_{\text{visual}} + 0.30S_{\text{motion}}, \tag{9}$$

and rounded to two decimal places. The evaluator additionally returns a brief justification for each score in a structured JSON format, enabling consistent parsing and inspection of the judgments.

The prompt is designed to be particularly sensitive to long-horizon generation failures. It explicitly penalizes accumulated blur, identity or appearance drift, visual artifacts, repetitive gesture cycles, and motion that gradually becomes frozen or mechanical. It also prevents large but inappropriate motion from being rewarded solely for its magnitude. To avoid evaluation based on irrelevant personal attributes, the judge is instructed not to consider the identity, gender, ethnicity, or appearance of the depicted person. The complete system prompt used for evaluation are provided in Figure 6.

E Visual Comparison of Decoupled Training and Different Rollout Horizons K

We provide additional qualitative comparisons to further analyze the two key components of our training framework: the decoupled optimization of DMD and RRT, and the rollout horizon K used in RRT. These visual results complement the quantitative ablations in the main paper by illustrating how different training configurations affect appearance fidelity and long-horizon temporal stability during autoregressive generation.

As shown in Figure 7, the DMD-only model retains relatively strong local appearance quality but gradually suffers from visual drift as autoregressive errors accumulate over time. As in Table 3, DMD only denotes the sequential training baseline. In contrast, the FM-only variant improves adaptation to long-horizon rollout but noticeably weakens fine-grained visual fidelity. The full decoupled DMD + RRT model combines the advantages of both objectives, preserving detailed appearance while maintaining substantially stronger stability over extended generation.

Figure 8 further visualizes the effect of the rollout horizon K. When $K = 0$, the model is supervised immediately after context perturbation and therefore mainly learns to correct local synthetic corruption, leaving noticeable artifacts and color drift during long-horizon inference. This $K = 0$ configuration is the Helios setting (Yuan et al., 2026), which applies immediate supervision after corrupting the history rather than first rolling the error forward through additional autoregressive chunks. Increasing K allows model-induced errors to propagate through autoregressive rollout before supervision, making the training condition more consistent with the error accumulation encountered at inference time. As K increases from 1 to 4, we observe progressively improved appearance, background, and identity consistency, with $K = 4$ producing the most stable qualitative results.

F Visual Results and Comparisons

Extended-Duration Generation. Figure 9 presents an audio-driven avatar video generated continuously for more than 11 minutes. Despite the extended autoregressive horizon, Avatar-Forever



Figure 7: **Effect of decoupled training.** The sequential DMD-only baseline preserves short-horizon appearance but is more susceptible to long-horizon drift, whereas FM-only adaptation degrades visual fidelity. Combining DMD with RRT preserves appearance while improving stability under autoregressive generation.



Figure 8: **Effect of the RRT rollout horizon.** Immediate supervision after context degradation ($K = 0$, the Helios setting (Yuan et al., 2026)) leaves visible artifacts and appearance drift. Increasing the autoregressive rollout horizon exposes the model to accumulated self-generated errors before supervision, improving color, background, and identity consistency; $K = 4$ yields the most stable result.

maintains stable identity, facial structure, appearance, and scene content without progressive drift or visible quality collapse. The avatar exhibits natural speech-driven dynamics, including realistic facial expressions, coherent gaze, subtle head movements, smooth posture changes, and non-repetitive gestures. These motions remain temporally coherent and appropriately follow the rhythm and emphasis of the speech, avoiding the frozen expressions, mechanical motion, and repetitive behavior commonly observed in long-form generation. Moreover, lip movements and facial articulation remain consistently aligned with the driving audio throughout the rollout. This result demonstrates that Avatar-Forever preserves visual fidelity, motion naturalness, and audio–visual consistency well beyond the 30-second evaluation setting, supporting its capability for stable generation without a predefined duration.

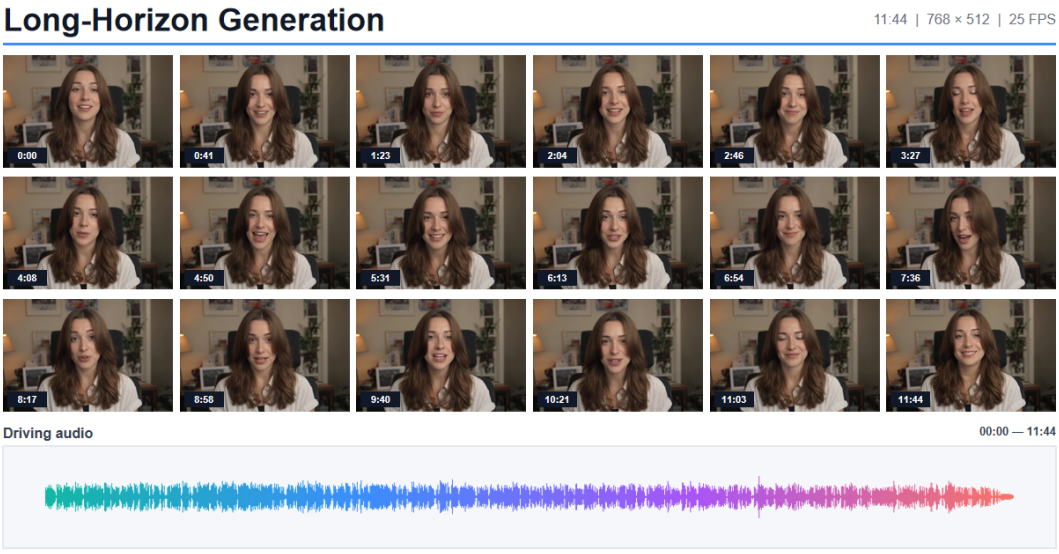

Figure 9: **Visual results of extended-duration generation.** We uniformly sample frames from a continuous video generated for more than 11 minutes. Avatar-Forever maintains stable identity, facial structure, and scene content while producing natural expressions, coherent head and body motion, and accurate audio–visual synchronization throughout the extended autoregressive rollout. The bottom row shows the corresponding driving-audio waveform.

30s Long-Video Generation. Figures 10–12 compare 30-second results on three long-horizon samples. Over the longer horizon, the baselines accumulate errors that are far more pronounced than in the short setting: background and appearance drift, severe visual degradation and spurious content (e.g., an extra person that ignores the scene), skin-color shift, and motion that is either nearly static or dominated by camera movement. Our results (the last two rows) remain stable across the full duration, keeping the background and identity consistent while generating more vivid and diverse head and body motion, such as the head leaning and rotating rather than repeating a fixed pose. As in the short setting, *Ours (w/ FC)* closely matches *Ours*, confirming that ForeverCache preserves long-horizon quality while substantially reducing redundant history computation.

5s Short-Video Generation. Figures 13–15 compare 5-second results on three EMTD samples. In each figure, rows show, from top to bottom, OmniAvatar, LiveAvatar, InfiniteTalk, SoulX-FlashTalk, and our method with and without ForeverCache (*Ours (w/ FC)* and *Ours*); columns are uniformly sampled frames from start to end, and the bottom strip visualizes the driving audio. Even in the short-horizon regime, the baselines already exhibit visual degradation, repetitive and stereotyped facial gestures, limited motion diversity, and over-articulated mouth motion. In contrast, our results (the last two rows) preserve identity and visual fidelity while producing more natural and diverse head motion and expressions with accurate lip synchronization. *Ours (w/ FC)* is visually on par with *Ours*, showing that ForeverCache accelerates streaming inference without sacrificing quality.



Figure 10: **Visual comparison on 30-second generation (sample 1).** Rows are, top to bottom, OmniAvatar, LiveAvatar, InfiniteTalk, SoulX-FlashTalk, *Ours (w/ FC)*, and *Ours*; columns are frames uniformly sampled over time, with the driving audio shown below. The baselines drift with a degraded background and a shifted face, and the body barely moves. Avatar-Forever (last two rows) keeps identity and background stable while the head and body lean and rotate for more vivid, diverse motion. ForeverCache retains comparable visual stability with substantially faster inference.

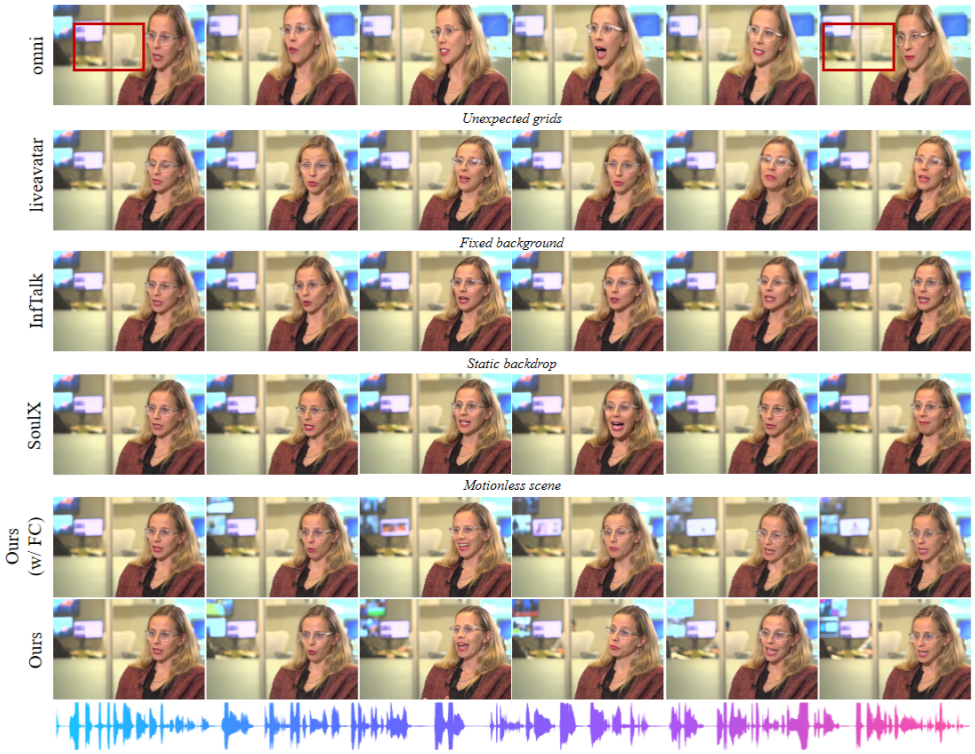

Figure 11: **Visual comparison on 30-second generation (sample 2).** Same row and column layout as Figure 10. The baselines freeze the background over the full rollout. Avatar-Forever (last two rows) renders a coherent, dynamic scene that looks more realistic while preserving identity and appearance. ForeverCache retains comparable visual stability with substantially faster inference.

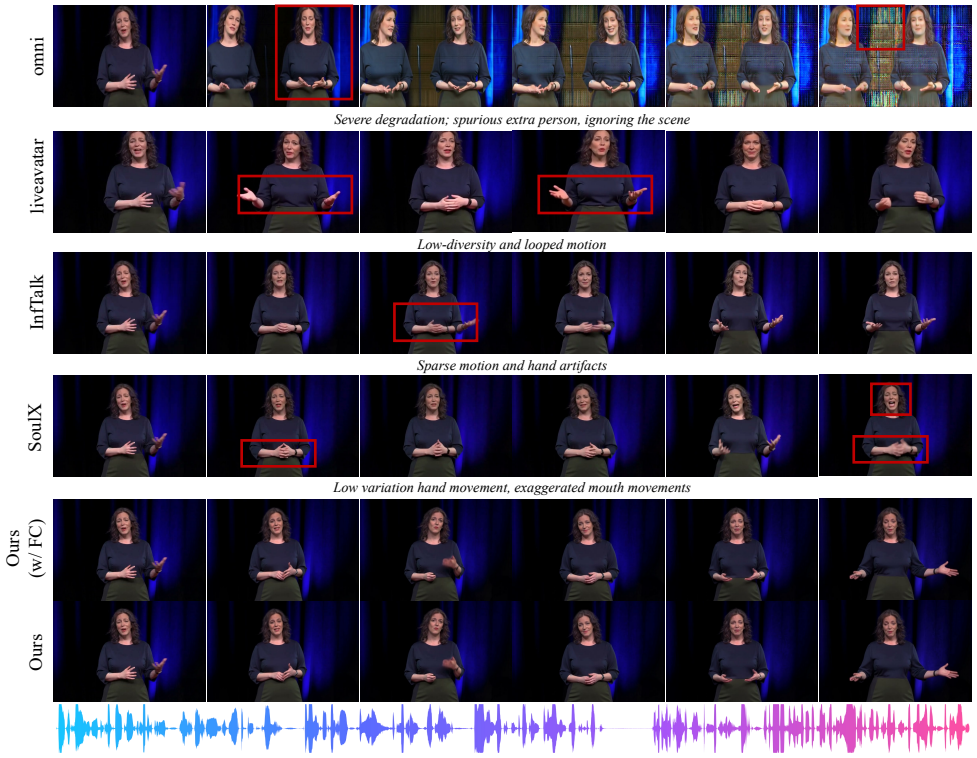

Figure 12: **Visual comparison on 30-second generation (sample 3).** Same row and column layout as Figure 10. The baselines show severe degradation, including a spurious extra person that ignores the scene, together with low motion diversity. Avatar-Forever (last two rows) remains stable over the full 30 seconds with larger, more diverse hand and arm motion. ForeverCache retains comparable visual stability with substantially faster inference.



Figure 13: **Visual comparison on 5-second generation (sample 1).** Rows are, top to bottom, OmniAvatar, LiveAvatar, InfiniteTalk, SoulX-FlashTalk, *Ours (w/ FC)*, and *Ours*; columns are frames uniformly sampled over time, with the driving audio shown below. The baselines show unexpected grids, repetitive and stereotyped facial gestures, and over-articulated mouth motion. Avatar-Forever (last two rows) preserves identity and facial detail with more natural, diverse expressions and accurate lip synchronization. ForeverCache retains comparable visual quality with substantially faster inference.

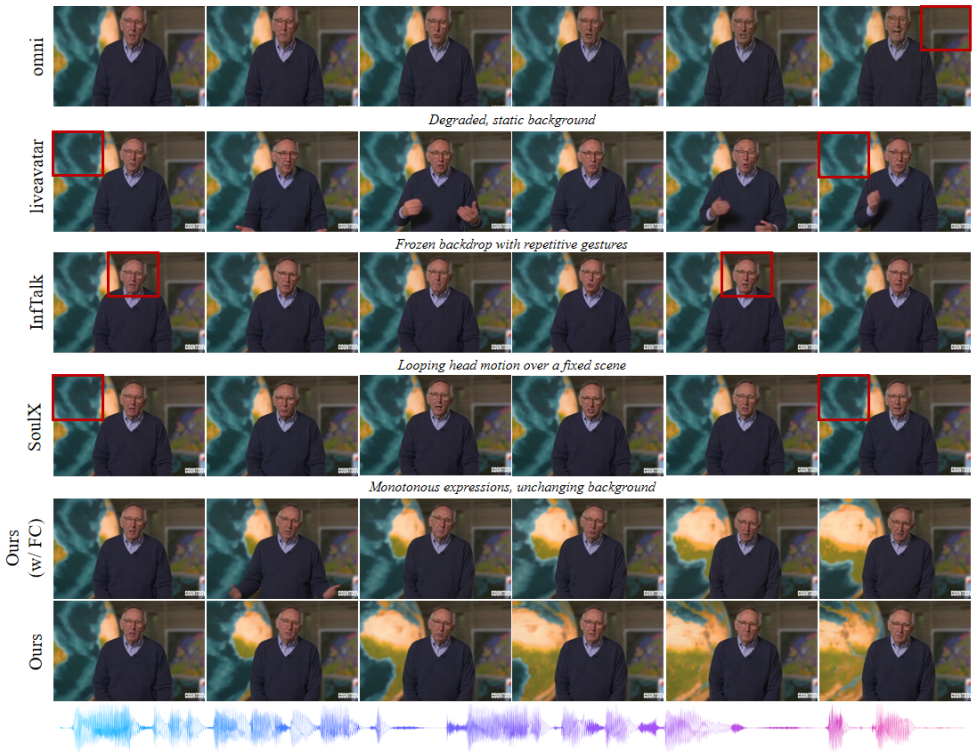

Figure 14: **Visual comparison on 5-second generation (sample 2).** Same row and column layout as Figure 13. The baselines keep a nearly static background with limited body motion. Avatar-Forever (last two rows) produces livelier, more diverse motion over a more dynamic scene while preserving identity and appearance. ForeverCache retains comparable visual quality with substantially faster inference.

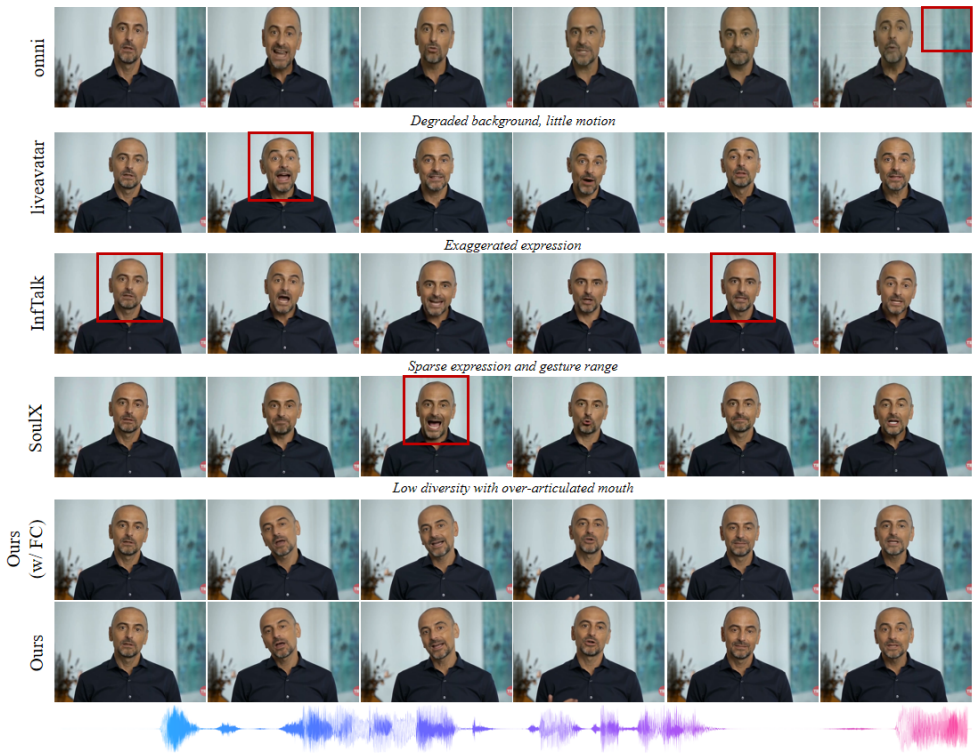

Figure 15: **Visual comparison on 5-second generation (sample 3).** Same row and column layout as Figure 13. The baselines exhibit background degradation, over-articulated expression and mouth motion. Avatar-Forever (last two rows) achieves richer motion diversity, such as the head leaning and turning, while keeping a stable appearance. ForeverCache retains comparable visual quality with substantially faster inference.