

Avatar-Forever: Decoupled Parallel Training for High-Quality Real-Time Infinite Avatars

Ruibin Li^{1*} Tao Yang² Zhiyuan Ma¹ Fangzhou Ai³ Shilei Wen² Lei Zhang^{†1}

¹The Hong Kong Polytechnic University ²ByteDance ³AMD

* Work done during internship in ByteDance. [†] Corresponding author (cslzhang@comp.polyu.edu.hk).

Project Page

Code

Demo

BibTeX

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 digital humans.

KEYWORDS: Interactive Avatars, Streaming Video Generation, Real-Time Video Generation

1 Introduction

Generating realistic and expressive digital humans is highly demanded in applications such as virtual assistants, content creation, education, and entertainment [1–3]. 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 [4–8] have significantly advanced audio-driven avatar generation, producing photorealistic talking avatars with improved visual fidelity, motion realism, and lip synchronization [9–12]. 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 challenging [13–15]. To generate longer outputs, streaming systems typically reuse previously generated chunks as conditions for future chunks [16–20], which introduces distribution drift and error accumulation: early artifacts, identity drift, or lip-sync errors become part of the historical context and gradually degrade subsequent generation [21, 22].

Recent works [15, 23, 24] 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 [13, 14, 19, 25–

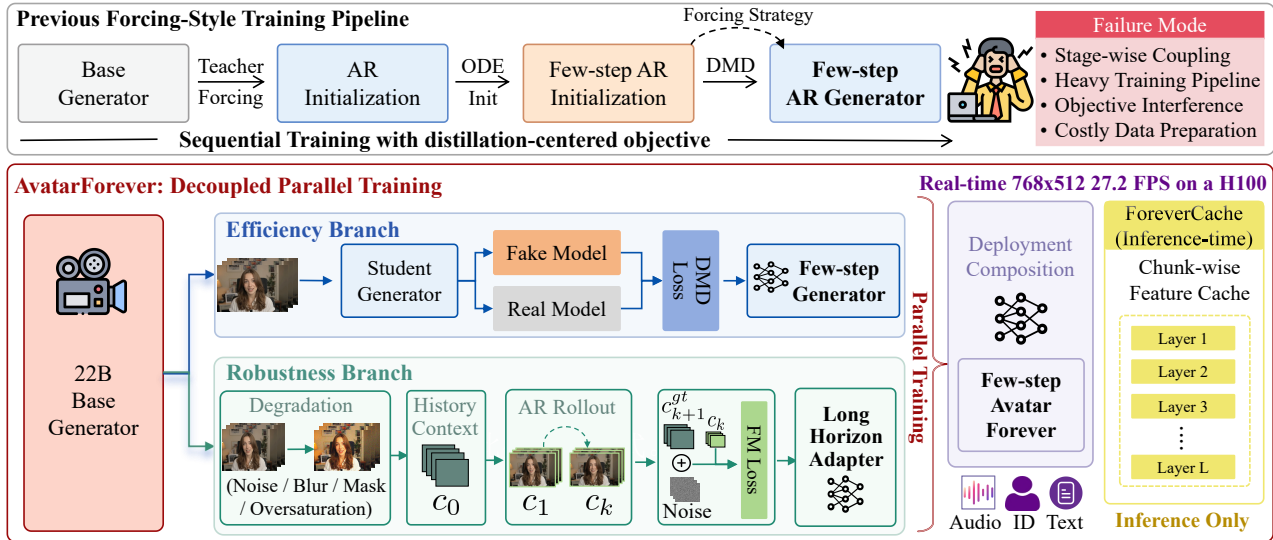


Figure 1. Overview of Avatar-Forever. *Top:* Existing streaming avatar systems employ sequential distillation pipelines, leading to stage-wise dependence and objective interference. *Bottom:* Avatar-Forever learns few-step efficiency and long-horizon robustness in parallel. A lightweight adapter is trained 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.

27] 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 [6] 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** 768×512 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 768×512 resolution inference at **27.2 FPS** on a single H100 GPU.

Remarks. Few-step efficiency and long-horizon stability operate on different temporal scales. Coupling them as one distillation objective entangles speed with robustness, complicating the training pipeline. Avatar-Forever learns to achieve them in a parallel manner, significantly improving the training process.

2 Related Work

2.1 Audio-Driven Avatar Video Generation

Traditional and 3D-based Methods. Audio-driven avatar generation has been widely studied through 2D talking-head synthesis [28], explicit facial representations [29], and neural rendering [30]. Early systems primarily focus on accurate lip synchronization and identity preservation from limited visual context [31]. To improve structural stability, later methods introduce 3D priors [32], facial motion coefficients [1], or neural radiance fields [33], mapping speech to geometrically meaningful facial dynamics before rendering. These methods provide strong priors for lip motion, head pose, and identity consistency, but they are basically task-specific generators or subject-specific renders, offering limited flexibility to diverse identities, expressive body motion, complex scenes, and open-ended behaviors [27]. Their reliance on predefined structures or identity-specific rendering also limits the generalization to unconstrained avatar scenarios, especially for long-duration generation.

Diffusion-based Avatars. Diffusion models [34–36] have recently become increasingly important backbones for avatar synthesis, as they can capture subtle facial dynamics, photorealistic appearance, and temporal variations. Task-specific avatar systems such as DiffTalk [37], EMO [3], Hallo [38], AniPortrait [39] and EchoMimic [40] adapt diffusion models to audio-driven portrait animation through audio-conditioned denoising, hierarchical control, or editable motion guidance. Meanwhile, large video foundation models, exemplified by Wan [5], Seedance 2.0 [4], Kling-Omni [41], and Sora [42], have demonstrated that scaled diffusion-transformer video generators can serve as strong spatiotemporal priors, although they are not specifically designed for audio-driven avatar generation. However, these models introduce substantially high inference cost, making fast sampling a critical bottleneck. While diffusion acceleration methods have been developed to reduce the denoising steps [43–48], they mainly address the problem of efficient short-horizon generation, whereas long-duration avatar synthesis also requires robustness under recursive autoregressive rollout.

2.2 Long-Horizon and Streaming Video Diffusion Models

Streaming video generation [19, 27, 49, 50] reuse previously generated frames as conditions for future predictions to conduct long-horizon video generation. However, this creates a train–test mismatch: models

are trained on clean ground-truth contexts but must condition on their own imperfect outputs at inference, causing artifacts, motion drift, and temporal inconsistency to accumulate over time [21, 23, 24]. To reduce the train–test mismatch in autoregressive long-video generation, recent methods expose models to self-generated contexts during training or distillation. Self-Forcing++ [25], Causal Forcing [19], Rolling Forcing [51], and Hybrid Forcing [50] introduce mechanisms such as teacher-guided rollout, auto-regressive teacher initialization, progressive denoising, and hybrid attention design to improve long-horizon stability. These forcing strategies substantially improve the stability and quality of streaming generation. However, they often rely on tightly coupled multi-stage pipelines, making component-wise diagnosis difficult and scaling to large video foundation models cumbersome.

Long-duration avatar generation faces the same autoregressive streaming challenge, but with stricter requirements on identity preservation, appearance stability, lip synchronization, facial dynamics, and human motion coherence [8, 23, 52, 53]. Recent systems such as LiveAvatar [27], StreamAvatar [15], SoulX-FlashTalk [24], and LPM 1.0 [23] adapt large video priors to streaming or long-horizon human synthesis, showing the potential of video foundation models for interactive digital humans. However, these methods largely follow the sequential training paradigm of generic long-video generation, where multiple stages are coupled through autoregressive rollout, distillation, and complicated forcing strategies. As a result, errors and distribution shifts [21, 25] introduced in early stages can propagate through the pipeline, making the final behavior hard to diagnose and costly to scale to large video foundation models.

3 Method

As illustrated in the bottom of Figure 1, Avatar-Forever consists of three core designs. An *efficiency branch* and a *robustness branch* are trained separately via decoupled adaptation. The full-parameter efficiency branch is distilled to preserve high-quality few-step generation, while the robust branch is adapted to improve long-horizon stability. The resulting long-horizon LoRA adapter is merged into the distilled generator at inference. 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. In addition, 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) [44] to compress the original multi-step LTX base model [6] 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:

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

where $\tilde{\mathbf{x}}_t$ is obtained by applying the forward diffusion process to the student sample $G_{\theta}(\epsilon, \mathbf{c})$, 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

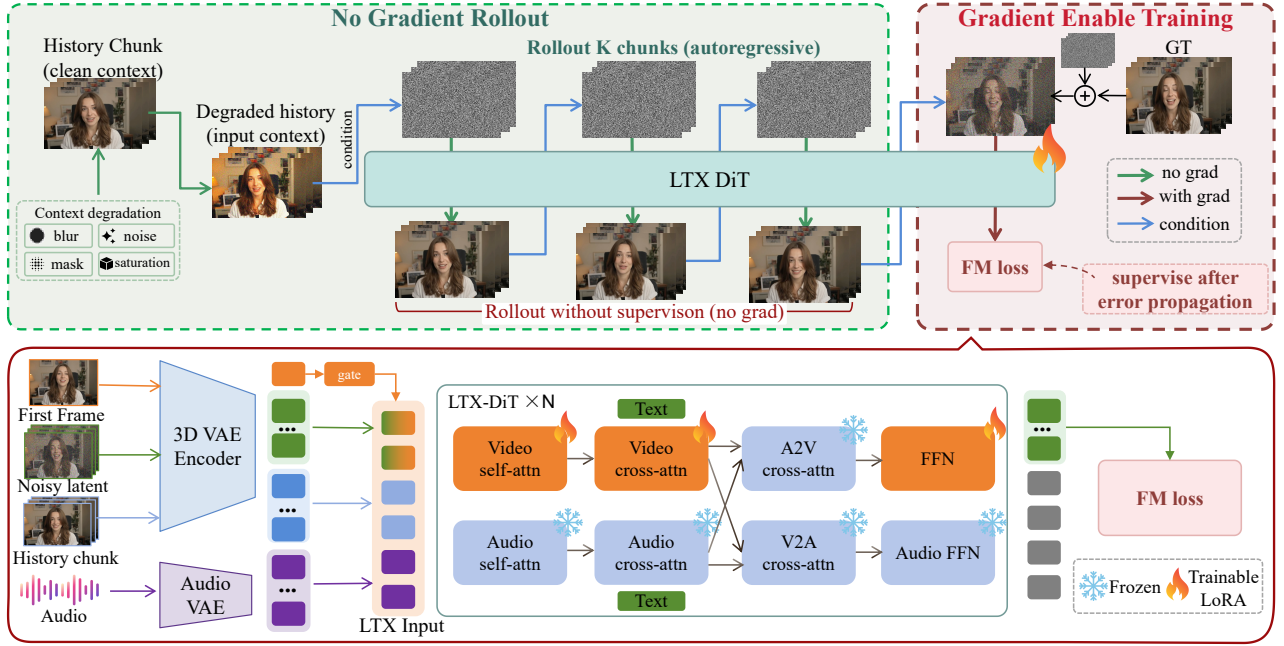


Figure 2. Recovery-oriented Rollout Training (RRT). *Top:* RRT degrades an early history chunk, propagates the resulting error through K autoregressive rollout steps without gradients, and applies standard FM supervision only to the subsequent recovery step. *Bottom:* In LTX DiT, video and audio conditions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

reused and amplified, causing identity drift, motion incoherence, and appearance artifacts [21, 24]. Training only on clean histories may lead to failures when exposed to deployment-time error distribution. A natural remedy is corrupted-history training, as explored in Helios [21]. 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 after errors have accumulated. RRT thus trains recovery under model-induced context drift rather than one-step synthetic corruption. The details of RRT are illustrated in Figure 2.

Global Reference Conditioning. We use the first frame as a persistent visual anchor. It is encoded into a latent representation $\mathbf{r}$ and injected into the denoising tokens through a lightweight gated channel-conditioning module. Unlike autoregressive history, which may drift during rollout, $\mathbf{r}$ remains fixed and provides stable guidance for identity, appearance, and scene layout.

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:

$$\hat{\mathbf{c}}_0 = \mathcal{D}(\mathbf{c}_0), \quad (2)$$

where $\mathcal{D}(\cdot)$ is a stochastic degradation operator. Inspired by Helios [21], $\mathcal{D}$ 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 $k = 1, \dots, K$, 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:

$$\hat{\mathbf{c}}_{k,t} = \text{sg} \left(G_{\theta} \left(\hat{\mathbf{c}}_{k,t+1}; \hat{\mathbf{c}}_{k-1,0}, \mathbf{r}, \mathbf{a}_k, y \right) \right), \quad t = T-1, \dots, 0, \quad (3)$$

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 $\text{sg}(\cdot)$ stops gradients through the rollout trajectory. As illustrated in the top left of Figure 2, rollout is executed without gradient tracking. The intermediate chunks $\hat{\mathbf{c}}_1, \dots, \hat{\mathbf{c}}_K$ are not directly supervised; their role is to

expose the final prediction to model-induced context drift, so that optimization targets recovery under the same autoregressive 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 $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and a flow-matching noise level $\sigma \sim p(\sigma)$, we corrupt only the target tokens:

$$\hat{\mathbf{c}}_{K+1,\sigma} = (1 - \sigma)\mathbf{c}_{K+1} + \sigma\epsilon, \quad (4)$$

while keeping the rolled-out historical context $\hat{\mathbf{c}}_{K,0}$ clean. We then optimize a simple flow-matching [35] objective on the final target chunk:

$$\mathcal{L}_{\text{RRT}} = \mathbb{E}_{\mathbf{c}, \mathbf{a}, \mathbf{y}, \sigma, \epsilon} \left[\left\| v_{\theta}(\mathbf{c}_{K+1,\sigma}, \sigma; \hat{\mathbf{c}}_{K,0}, \mathbf{r}, \mathbf{a}_{K:K+1}, \mathbf{y})_i - (\epsilon - \mathbf{c}_{K+1})_i \right\|_2^2 \right], \quad (5)$$

where v_{θ} is the velocity network, $\mathbf{a}_{K:K+1}$ denotes the audio condition aligned with the visible context-target window. We do not impose 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.

Insight. We do not supervise immediate local reconstruction from a degraded context; we supervise recovery after degradation has propagated through autoregressive rollout.

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

$$\theta^* = \theta_0 + \Delta\theta_{\text{DMD}} + \Delta\theta_{\text{RRT}}. \quad (6)$$

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 additional training and 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 $\mathcal{H}_k = \{\mathbf{c}_0, \mathbf{c}_{k-1}\}$, consisting of the first and 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, as shown in Figure 3(a). However, this 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 access to long-range context while restricting repeated computation to the evolving current chunk, as illustrated in Figure 3(b).

Formally, let σ_t denote the noise level at denoising step t . A standard autoregressive implementation evaluates the velocity network on the complete visible window at every step:

$$\mathbf{v}_{k,t} = v_{\theta}([\mathcal{H}_k, \mathbf{c}_{k,t}], \sigma_t; \mathbf{r}, \mathbf{a}_{\mathcal{H}_k:k}, \mathbf{y}), \quad (7)$$

where $[\mathcal{H}_k, \mathbf{c}_{k,t}]$ is the history-current sequence input to LTX [6]. Only the velocity prediction for the current chunk is retained for updating $\mathbf{c}_{k,t}$; the historical chunks act solely as conditioning context.

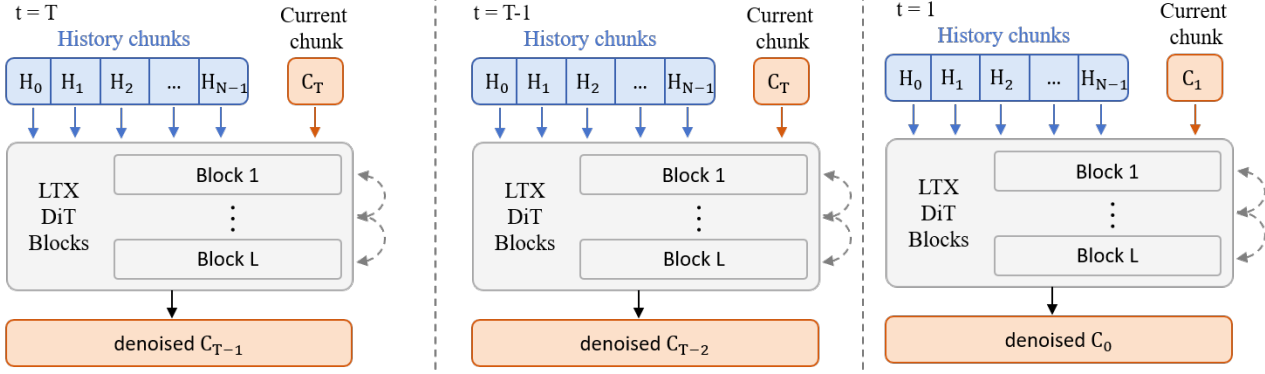
ForeverCache performs full computation only once at $t = T$, while collecting historical features from all the L transformer blocks:

$$\mathbf{v}_{k,T}, C_k = v_{\theta}^{\text{populate}}([\mathcal{H}_k, \mathbf{c}_{k,T}], \sigma_T; \mathbf{r}, \mathbf{a}_{\mathcal{H}_k:k}, \mathbf{y}), \quad C_k = \{C_k^{\ell}\}_{\ell=1}^L. \quad (8)$$

For all subsequent steps, the model forwards only the current-chunk tokens and retrieves C_k as historical conditioning memory:

$$\mathbf{v}_{k,t} = v_{\theta}^{\text{reuse}}(\mathbf{c}_{k,t}, \sigma_t; C_k, \mathbf{r}, \mathbf{a}_k, \mathbf{y}), \quad t = T - 1, \dots, 0. \quad (9)$$

(a) Standard AR denoising (no cache)



(b) ForeverCache / chunk-wise history feature cache

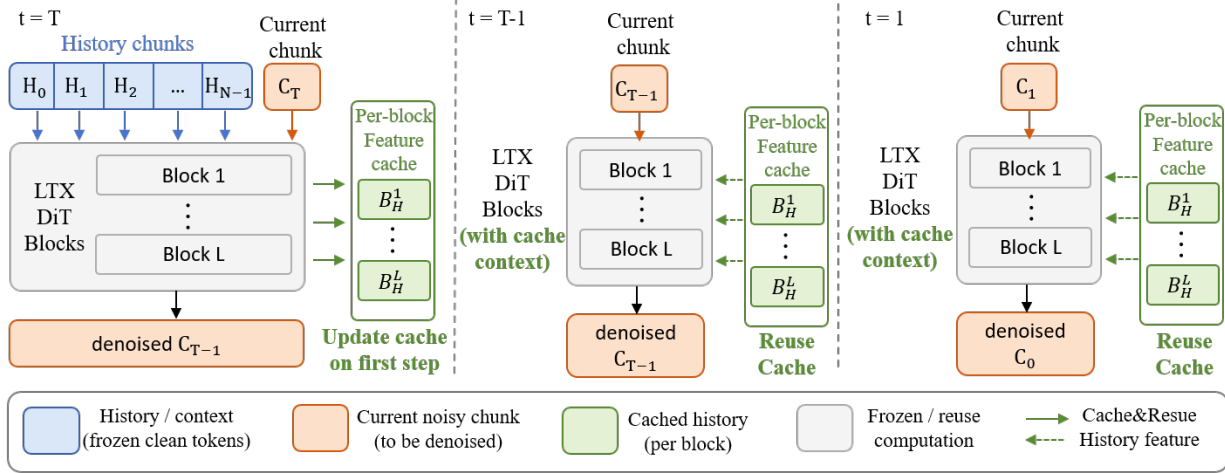


Figure 3. ForeverCache for autoregressive streaming inference. *Top:* Standard autoregressive denoising recomputes fixed history features at every step. *Bottom:* ForeverCache populates per-block history features once at $t = T$ and reuses them while forwarding only the evolving current chunk thereafter. The cache is reset for each new autoregressive chunk, providing bounded-memory streaming inference without redundant history computation.

Here, $v_{\theta}^{\text{populate}}$ and $v_{\theta}^{\text{reuse}}$ 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 Synthetic Data Construction Pipeline

Long-horizon avatar adaptation requires 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 fully synthetic pipeline that generates and filters avatar videos using the same video foundation model adapted by our framework. This construction pipeline provides controllable training data that are naturally 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 [54] and use GPT [55] to filter and refine samples that are incoherent, low-quality, or visually uninformative. The

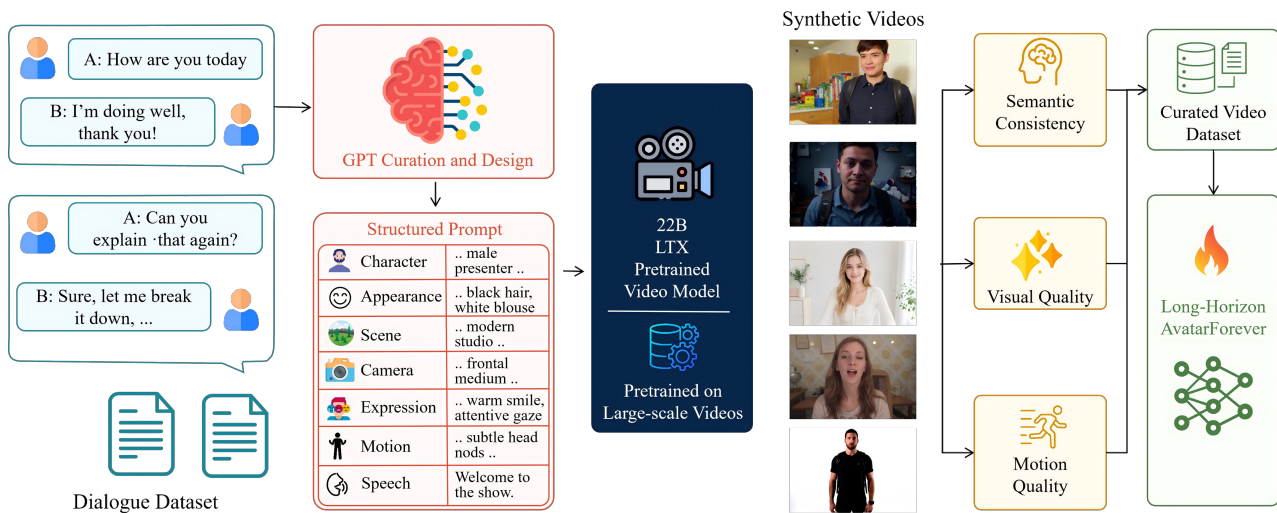


Figure 4. Synthetic data construction pipeline for long-horizon avatar adaptation. Public dialogues are filtered and converted into structured video prompts, from which the pretrained LTX model synthesizes controllable avatar videos. Quality-aware filtering removes semantically mismatched, visually degraded,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

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 [56], CLAP [57], and Unified Reward Model [58], and use Gemini [59] 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 produced by global affine camera motion, such as uniform vertical camera movement, translation, or zooming, rather than local facial and body dynamics. For dialogue-driven samples, we retain only those videos with sufficient local non-rigid motion, including visible facial articulation or body movement. This procedure yields a curated synthetic corpus with semantic consistency, visual fidelity, and motion diversity for long-horizon adaptation.

5 Experiments

5.1 Experimental Setup

Training Data. We train the long-horizon adaptation branch using the fully synthetic pipeline described in Section 4. Each sample contains an avatar video, paired audio, and a text prompt. Videos are partitioned into context-target pairs: the model conditions on a context window and predicts a video-only target chunk under the full audio condition. Unless otherwise stated, we use four latent frames as context and supervise the subsequent four latent frames. To improve robustness to temporal misalignment, the context start is sampled either from the first latent frame or from a randomly selected later position.

Implementation Details. Avatar-Forever is built on the 22B LTX-2.3 [6]. The efficiency branch distills the base model into a four-step generator using DMD [43], while the robustness branch independently trains video-side LoRA [60] adapters with rank and alpha both set to 128. For RRT, we perturb the historical context, roll

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/ <u>7.95</u>	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	<u>775.11/858.06</u>	<u>5.24/38.85</u>
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	<u>7.06/7.60</u>	26.71/19.48	390.69/793.74	51.50/309.20
	LiveAvatar	3.90/3.80	<u>3.95/3.80</u>	3.08/3.00	<u>3.67/3.56</u>	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	<u>8.56/7.56</u>	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	<u>5.24/38.85</u>
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/ <u>51.41</u>	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	<u>554.00/994.42</u>	<u>5.24/38.85</u>

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 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 Protocol. We evaluate short-horizon quality and long-horizon stability on TalkVid [61], EMTD [62], and HDTF [63]. 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 [52], LiveAvatar [27], SoulX-FlashTalk [24], and InfiniteTalk [26]. For each baseline, we follow the official inference configuration whenever available and use identical audio and reference inputs for comparison.

Evaluation Metrics. We evaluate visual quality, motion realism, audio-visual synchronization, and video-distribution similarity. For perceptual evaluation, we use Gemini-Flash-3.5 [59] as a strict multimodal judge, which scores audio-visual consistency, visual quality, and motion naturalness on a 1–5 scale. The final score is the weighted average of them with weights 0.35, 0.35, and 0.30, respectively. The complete prompt is provided in **Appendix A**. We additionally 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 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.

5.2 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. Figure 5 provides qualitative comparison for these methods. Even in the short-horizon regime, the baselines already exhibit visual degradation, 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 natural, diverse head movements and expressions with accurate lip synchronization. Beyond foreground animation, Avatar-Forever also retains coherent scene dynamics: for example, the large

Table 2. Double-blind user study results on the EMTD long-video evaluation split. Each score is computed from user ratings on a 1–5 scale and normalized to a 0–100 scale. The best and second-best results are highlighted in bold and underlined, respectively.

Model	User Study			
	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

globe continues to rotate naturally while preserving its geographic structure and map textures, rather than becoming static or visually distorted. Please view the videos in the [project page](#) for better comparison.

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 6 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.

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× 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× 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 6 show that cache-enabled inference retains the key visual behavior of the full model without visible collapse or systematic temporal degradation.

Extended-Duration Generation. Figure 7 presents an audio-driven avatar video generated continuously for more than 11 minutes. Avatar-Forever 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.

Videos of the above examples and more visual examples can be found in [Appendix B](#) and [project page](#).

Human Perceptual Validation. Table 2 reports the double-blind user study results on the EMTD long-video split. Twenty participants rate anonymized videos on the same four perceptual dimensions as the LLM judge using integer scores from 1 to 5; scores are then 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 the ForeverCache variant consistently ranks second with only a small gap from the full model.

The advantage in human validation is particularly pronounced for Visual and Motion quality, highlighting aspects of temporal naturalness that are not fully captured by automatic evaluation. While several baselines preserve identity and produce locally plausible frames, their long videos often exhibit limited upper-body

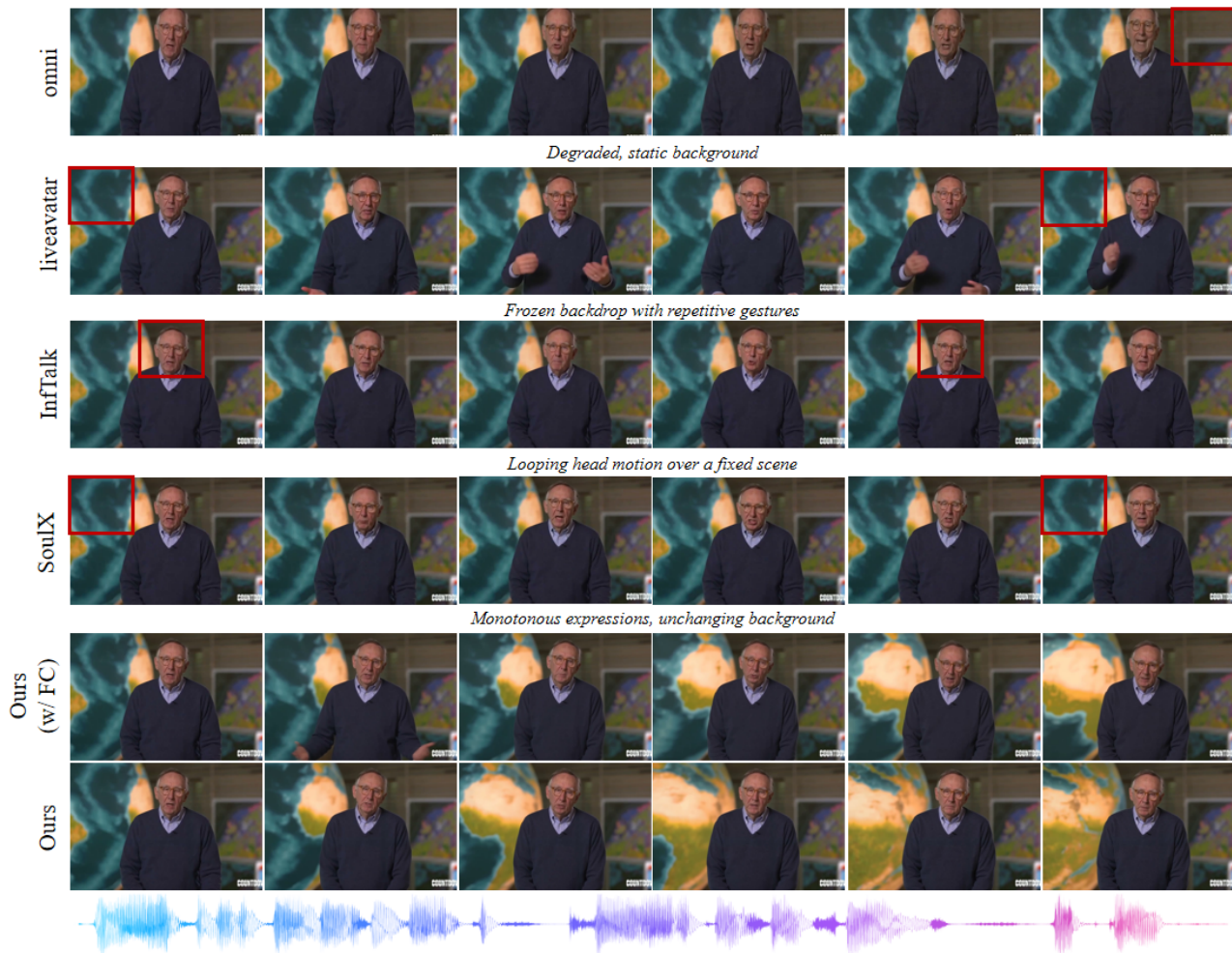


Figure 5. Visual comparison on 5-second generation. Avatar-Forever produces more natural speech-driven facial motion, hand gestures, and background dynamics than prior methods, which often restrict motion to the face and mouth. The ForeverCache variant preserves the visual performance of the standard model while reducing inference cost.

motion, static backgrounds, repetitive gestures, rigid posture changes, or exaggerated mouth shapes. In contrast, Avatar-Forever produces more expressive facial dynamics, smoother gesture transitions, and motion that better follows the rhythm of speech. The resulting improvement in temporal naturalness also strengthens perceived visual realism, explaining the highest Overall score assigned by human raters. ForeverCache preserves most of these perceptual gains while providing substantially faster inference.

5.3 Analysis of Decoupled Training

Table 3 and Figure 8 test our central hypothesis that few-step efficiency and long-horizon robustness are distinct capabilities. The full Decoupled DMD + RRT model achieves the best result on every metric, spanning perceptual quality, motion, synchronization, and distributional realism. Relative to DMD only, adding RRT improves the LLM Overall score by 3.6%, increases Sync-C by 4.2%, and reduces FID and FVD by 11.5% and 16.5%, respectively. Although DMD provides a strong few-step generator, it is not exposed to the accumulated self-conditioning errors encountered during autoregressive inference. Correspondingly, Figure 8 shows weaker long-horizon appearance stability.

The FM only variant also shows limited performance. Optimizing long-horizon flow matching without the distilled generator substantially degrades visual fidelity and synchronization, producing blurred facial details and poor perceptual quality. Compared with this variant, the full model improves LLM Overall by 77.9%, increases Sync-C by 229.6%, and reduces FID by 54.3%. These results show that robustness cannot be obtained by optimizing long-horizon recovery alone. DMD preserves the few-step generation prior, whereas RRT adapts it to the rollout-time error distribution; independently learning and composing these capabilities yields both high local fidelity and stable long-horizon generation.



Figure 6. 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 degraded hand details, repetitive gestures, or instability over time. ForeverCache retains comparable long-horizon visual stability with substantially faster inference.

5.4 Effect of Recovery-oriented Rollout Training

Figure 9 isolates the effect of the rollout horizon K in RRT. The key observation is that corrupted-history training alone is insufficient: when $K = 0$, supervision is applied immediately after context perturbation, reducing the task to local reconstruction. The resulting videos remain susceptible to accumulated autoregressive errors, exhibiting spot-like artifacts and global color drift. Increasing the horizon to $K = 1$ partially exposes the model to self-generated errors, but visible appearance fluctuations remain.

Larger rollout horizons match more closely the error accumulation process encountered at inference time. Both $K = 2$ and $K = 4$ improve appearance, background, and color consistency, with $K = 4$ producing the most stable long-video results. This progression supports the design of RRT: its benefit comes from supervising recovery after model-induced errors have propagated through rollout, rather than from merely injecting synthetic corruption into historical context.

6 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

Long-Horizon Generation

11:44 | 768 × 512 | 25 FPS

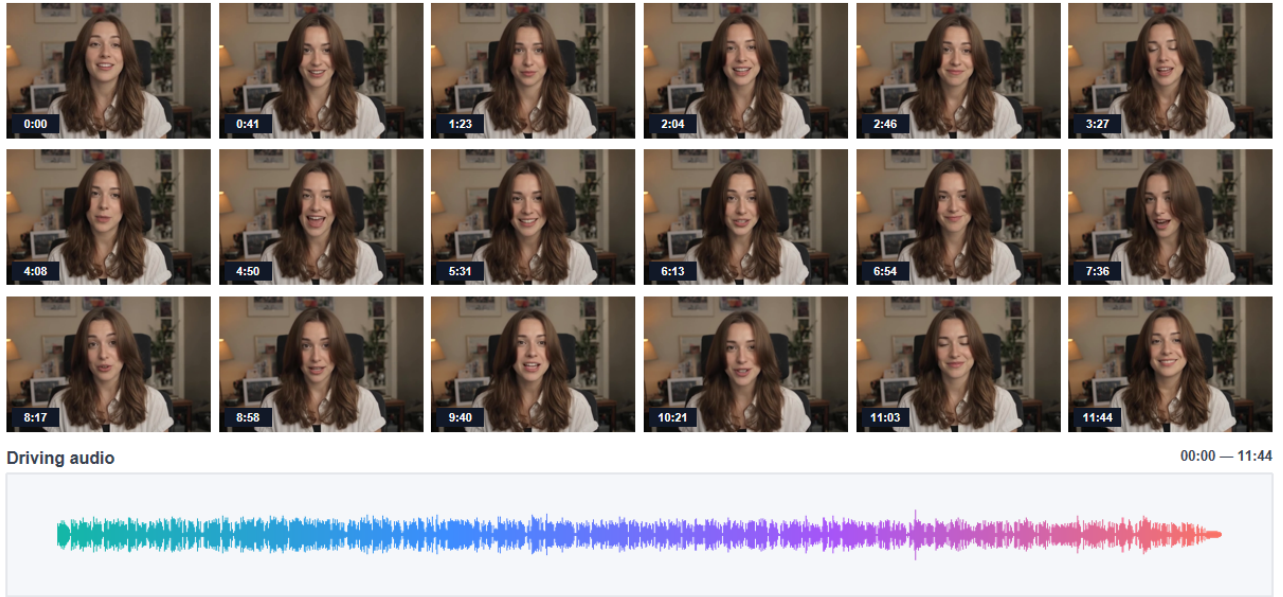


Figure 7. Visual results of extended-duration generation. We uniformly sample frames from a continuous video generated for more than 11 minutes. Avatar-Forever maintains stable identity, facial structure, and scene content while producing natural expressions, coherent head and body motion, and accurate audio–visual synchronization throughout the extended autoregressive rollout. The bottom row shows the corresponding driving-audio waveform.

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

Setting	Audio-Visual↑	LLM Judge			Automatic Metrics					
		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
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>



Figure 8. Effect of decoupled training. DMD only preserves short-horizon appearance but is susceptible to long-horizon drift, whereas FM-only adaptation degrades visual fidelity. Combining DMD with RRT preserves detailed appearance while improving stability under autoregressive generation.

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 it is not yet optimized 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



Figure 9. Effect of the RRT rollout horizon. Immediate supervision after context degradation ($K = 0$) 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.

scenarios, as illustrated in our [project page](#). Future work will explore domain-specific data construction and training strategies to further unlock Avatar-Forever’s potential for general long-horizon video generation.

Ethics Statement

This work aims to advance long-duration audio-driven avatar generation for constructive applications such as virtual communication, education, accessibility, and digital content creation. The evaluation data used in this paper are drawn from publicly available and openly accessible research datasets. Our synthetic training data are generated from the underlying video foundation model and filtered for quality, motion, and semantic consistency, rather than collected from private or restricted personal sources. We recognize that high-fidelity avatar synthesis is a dual-use technology. While it can support beneficial interactive media applications, it may also be misused for impersonation, deceptive content generation, or misinformation. We do not endorse any use of this technology to create unauthorized representations of real individuals, manipulate public opinion, or violate privacy, consent, or intellectual property rights.

To reduce potential misuse, we encourage deployment together with responsible safeguards, including identity-consent verification, synthetic-content disclosure, invisible watermarking, provenance tracking, and robust forgery detection. We also recommend that generated videos be clearly labeled as synthetic when used in public-facing scenarios. Our goal is to contribute to safer and more transparent digital human generation, and we will continue to align future development with responsible AI principles.

References

- [1] 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*, pages 8652–8661, 2023.
- [2] 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*, pages arXiv–2412, 2024.
- [3] 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.

- [4] Team Seedance. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.
- [5] 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.
- [6] 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.
- [7] 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*, 2025.
- [8] 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.
- [9] 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*, pages 12956–12965, 2025.
- [10] 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*, pages 9704–9713, 2025.
- [11] Chetwin Low and Weimin Wang. Talkingmachines: Real-time audio-driven facetime-style video via autoregressive diffusion models. *arXiv preprint arXiv:2506.03099*, 2025.
- [12] 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.
- [13] 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*, pages 22963–22974, 2025.
- [14] 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.
- [15] 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.
- [16] 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*, 2025.
- [17] 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*, pages 34385–34397, 2026.
- [18] 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*, 2025.
- [19] 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.
- [20] 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*, pages 40256–40265, 2026.

- [21] 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.
- [22] 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*, 2026.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] 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*, 2025.
- [27] 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*, 2025.
- [28] Trong Thang Pham, Tuong Do, Nhat Le, Ngan Le, Hung Nguyen, Erman Tjiputra, Quang Tran, and Anh Nguyen. Style transfer for 2d talking head generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7500–7509, 2024.
- [29] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5154–5163, 2020.
- [30] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10101–10111, 2019.
- [31] 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*, pages 484–492, 2020.
- [32] 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*, pages 958–968, 2024.
- [33] 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*, pages 1–10, 2023.
- [34] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [35] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.
- [36] 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*, pages 10684–10695, 2022.
- [37] 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*, pages 1982–1991, 2023.
- [38] 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.

- [39] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024.
- [40] 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.
- [41] Kling Team. Kling-omni technical report. *arXiv preprint arXiv:2512.16776*, 2025.
- [42] OpenAI. Video generation models as world simulators. <https://openai.com/index/video-generation-models-as-world-simulators/>, 2024. Technical report.
- [43] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6613–6623, 2024.
- [44] Tianwei Yin, Michaël 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.
- [45] 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*, pages 87–103, 2024.
- [46] 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.
- [47] 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.
- [48] Tianhe Wu, Ruibin Li, Lei Zhang, and Kede Ma. Diversity-preserved distribution matching distillation for fast visual synthesis. *arXiv preprint arXiv:2602.03139*, 2026.
- [49] 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*, 2025.
- [50] 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.
- [51] Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.
- [52] 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.
- [53] 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.
- [54] 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.
- [55] OpenAI. Chatgpt. <https://chatgpt.com/>, 2026. Accessed: 2026-07-17.
- [56] 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.
- [57] 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.
- [58] 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.
- [59] Google DeepMind. Gemini 3.5 flash. <https://deepmind.google/models/gemini/flash/>, 2026. Accessed: 2026-07-08.
- [60] 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.

- [61] 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*, pages 3492–3500, 2026.
- [62] Rang Meng, Xingyu Zhang, Yuming Li, and Chenguang Ma. Echomimicv2: Towards striking, simplified, and semi-body human animation. *arXiv preprint arXiv:2411.10061*, 2024.
- [63] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021.

Appendix

In this appendix, we provide the LLM-based perceptual evaluation prompt, and more qualitative results to complement the main paper. Specifically, the supplement includes:

- A. The LLM-based perceptual evaluation prompt (referring to Sec. 5.1 in the main paper);
- B. Additional visual comparisons for 5-second and 30-second audio-driven avatar generation, including results with and without ForeverCache (referring to Sec. 5.2 in the main paper).

For better viewing experience, we uploaded all the videos to our project page [project page](#), where the videos can be played directly in the browser. Note that, due to the significant number of high-quality video files included in our demonstrations, initial page loading may require several minutes to complete. We appreciate your patience during this process, as the complete visual experience is essential to understand the capabilities and performance of our approach.

A LLM-based Perceptual Evaluation Prompt

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 expressions, and unnatural motion transitions can be difficult to characterize using individual automatic metrics. We therefore employ Gemini-Flash-3.5 [59] as a multimodal perceptual judge. For each generated sample, the evaluator receives only the corresponding audio and video to assess 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:

$$S_{\text{overall}} = 0.35S_{\text{A-V}} + 0.35S_{\text{visual}} + 0.30S_{\text{motion}}, \quad (\text{A.1})$$

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 A.1.

B Visual Results and Comparisons

30s Long-Video Generation. Figures B.2–B.3 compare 30-second results on two long-horizon samples. In each figure, from top to bottom, the rows show OmniAvatar, LiveAvatar, InfiniteTalk, SoulX-FlashTalk, and our method with and without ForeverCache (*Ours (w/ FC)* and *Ours*); the columns show uniformly sampled frames from start to end, and the bottom strip visualizes the driving audio. 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,

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 A.1. 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.

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. At the same time, AvatarForever with ForeverCache closely matches the base AvatarForever, confirming that ForeverCache preserves long-horizon quality while substantially reducing redundant history computation.

5s Short-Video Generation. Figures B.4–B.5 compare 5-second results on two EMTD samples. 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. AvatarForever with ForeverCache is visually on par with base AvatarForever, showing that ForeverCache accelerates streaming inference without sacrificing quality.



Figure B.2. Visual comparison on 30-second generation (sample 1). From top to bottom, the rows show OmniAvatar, LiveAvatar, InfiniteTalk, SoulX-FlashTalk, *Ours* (w/ FC), and *Ours*; the columns show 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.

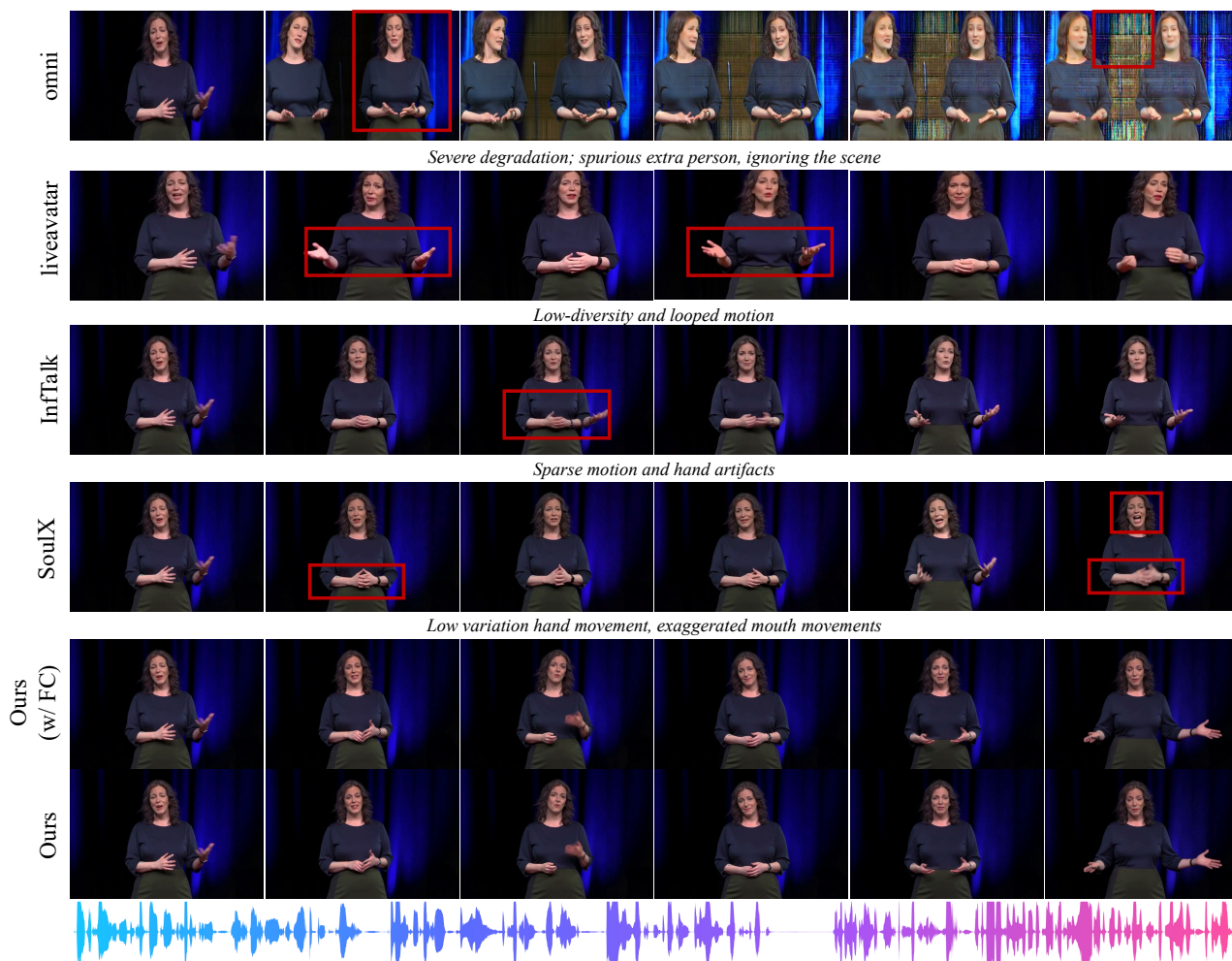


Figure B.3. Visual comparison on 30-second generation (sample 2). Same row and column layout as Figure B.2. The baselines show severe degradation, including a spurious extra person that ignores the scene, together with low motion diversity. Avatar-Forever (last two rows) remains stable over the full 30 seconds with larger, more diverse hand and arm motion. ForeverCache retains comparable visual stability with substantially faster inference.



Figure B.4. Visual comparison on 5-second generation (sample 1). From top to bottom, the rows show OmniAvatar, LiveAvatar, InfiniteTalk, SoulX-FlashTalk, *Ours* (w/ FC), and *Ours*; the columns show 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.

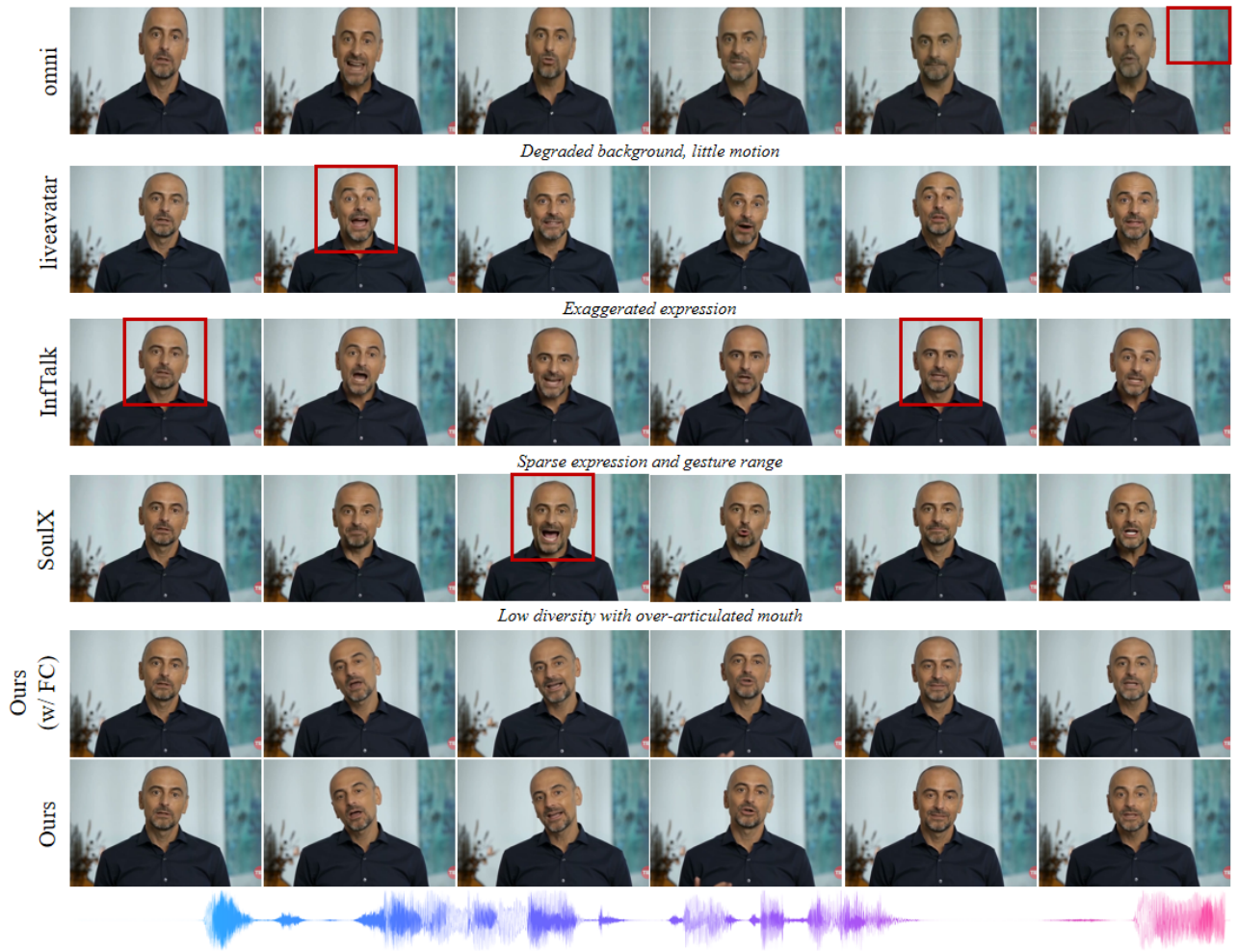


Figure B.5. Visual comparison on 5-second generation (sample 2). Same row and column layout as Figure B.4. The baselines exhibit background degradation, over-articulated expression and mouth motion. Avatar-Forever (last two rows) achieves richer motion diversity, such as the head leaning and turning, while keeping a stable appearance. ForeverCache retains comparable visual quality with substantially faster inference.