DEFORCING: EFFICIENT ROLLING-FREE TRAINING AND DISTILLATION FRAMEWORK FOR STREAMING REAL-TIME AVATAR GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming real-time avatar generation is important for live video conferencing, virtual agents, and interactive digital humans, where visual responses must be synthesized causally and with low latency while preserving identity, motion coherence, and audio–visual synchronization. We present **DeForcing**, a rolling-free training and distillation framework that converts a bidirectional diffusion avatar model into a streaming causal generator. For training, the framework uses optional causal warm-up (Causal SFT or CFG distillation) followed by DMD, without pre-generated ODE trajectories from expensive multi-step teacher inference or sequential self-forcing rollouts. Under noisy-cache inference, noise-aligned history and a fixed block-causal mask make denoising invariant to temporal chunking. The same model therefore supports autoregressive (AR) rollout for low-latency streaming, multi-token prediction (MTP) for parallel denoising of multiple temporal chunks, and single-pass prefill for higher throughput, without retraining. We further combine hybrid local–global temporal attention across layers with attention-sink caching under a fixed KV-cache budget to reduce attention cost and enable long-horizon streaming. Instantiated in our audio-driven avatar system with reference-conditioned generation, the framework empirically supports the equivalence of these decoding schedules and demonstrates stable appearance and scene layout over long-horizon continuation. The resulting model maintains a favorable balance among visual quality, audio–visual synchronization, and identity preservation relative to the bidirectional teacher.

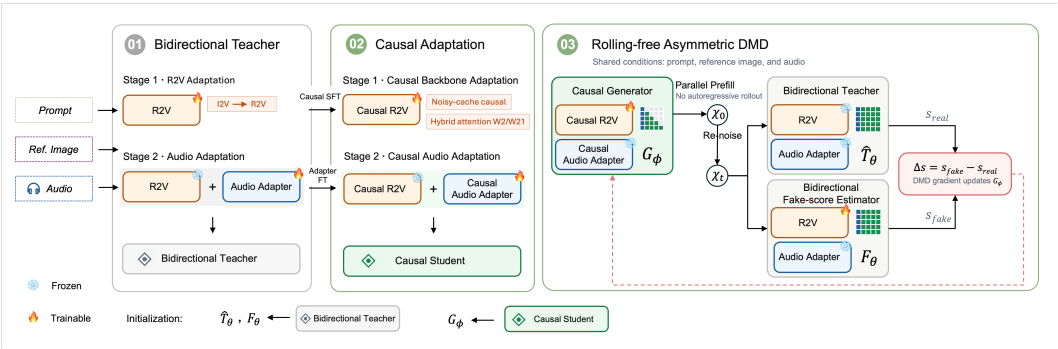


Figure 1: **Overview of DeForcing Pipeline.** (1) We adapt a pretrained image-to-video model into a bidirectional reference-to-video (R2V) backbone, then train a causal audio adapter with the backbone frozen to obtain the bidirectional avatar teacher. (2) We adapt the R2V backbone through Causal SFT under the noisy-cache paradigm, then fine-tune a copy of the causal audio adapter with the backbone frozen to obtain the causal student. (3) We perform asymmetric DMD to obtain the few-step causal generator. During DMD, the student generator produces $\mathbf{x}_0$ with a parallel prefill pass and therefore requires no temporal autoregressive rollout.

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.

1 INTRODUCTION

Avatar generation supports both offline content production and online interactive applications such as live streaming: offline production can refine a complete temporal window before rendering, whereas live avatars must respond causally to incoming audio and user input with low latency while preserving temporal coherence. Recent streaming-avatar methods address this challenge by converting bidirectional diffusion models into causal generators through ODE initialization and self-forcing distillation Chern et al. (2025); Huang et al. (2025; 2026); Cui et al. (2026). However, ODE initialization requires repeated multi-step teacher sampling, while self-forcing relies on sequential temporal rollout, reducing training parallelism and potentially causing instability Shen et al. (2025); Guo et al. (2025).

Temporal causality does not necessarily imply temporal rollout. The key distinction is whether later inputs are recursively constructed from earlier outputs. In autoregressive language models, the next input contains the previously generated token, which is why they are typically trained with teacher forcing and distilled or evaluated with autoregressive rollout.

Motivated by this distinction, we analyze causal diffusion through the way historical keys and values are cached during inference. In *clean-cache inference*, the current unit attends to fully denoised history; in *noisy-cache inference*, it attends to history maintained at the same denoising level. As illustrated in Fig. 2, these choices induce different training and distillation procedures. Our noisy-cache initialization, *Causal SFT*, uses a block-wise causal mask while aligning the history noise level with the current denoising step, unlike Diffusion Forcing Chen et al. (2024), which assigns independent noise levels.

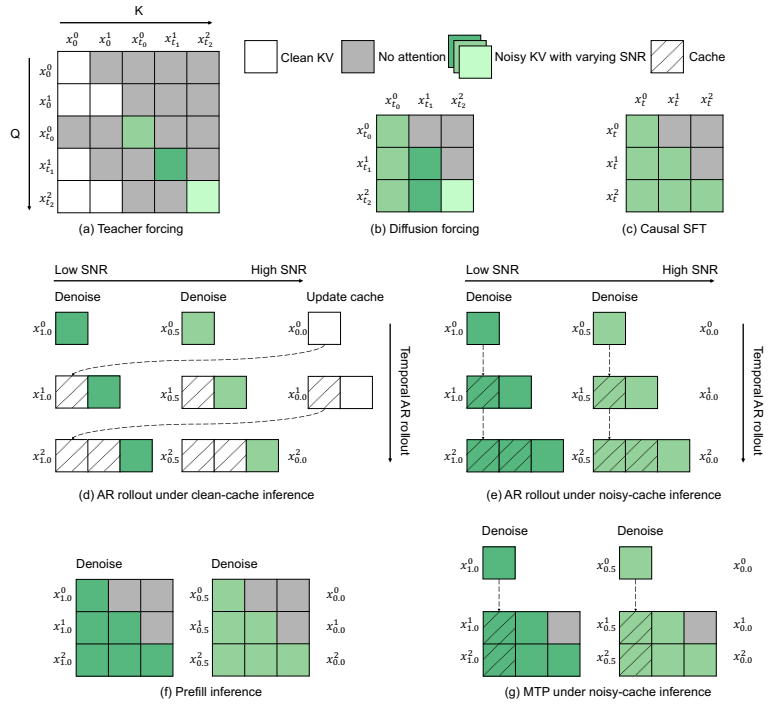


Figure 2: **Core Principle of DeForcing: Noise-Aligned History Enables Rollout-Free Distillation and Flexible Inference.** Clean- and noisy-cache training/inference. $x^i_{t_j}$ is the ith temporal latent at flow time t_j. (a–c) parallel training schemes; (d) clean-cache AR rollout; and (e–g) noisy-cache AR, prefill, and MTP. Noise-aligned history makes the latter executions share the same block-causal computation while exposing different latency–throughput trade-offs.

Distillation must address both diffusion-state and temporal-context mismatch. DMD Yin et al. (2024) uses backward simulation to construct the student's intermediate denoising states, while clean-cache inference additionally requires temporal autoregressive rollout to construct its history.

Under noisy-cache inference, however, a fixed block-wise causal mask and noise-aligned history make the denoising computation invariant to temporal chunking. Therefore, a full-sequence prefill pass implements the same causal computation as frame-by-frame rollout, enabling rolling-free distillation and flexible AR, MTP, or prefill inference with one model.

Fig. 1 summarizes the three stages of our framework: bidirectional avatar teacher construction, noisy-cache causal adaptation, and rolling-free few-step distillation. At inference, the same model supports AR, MTP, and prefill execution without retraining, and uses hybrid local–global attention with attention-sink caching for bounded-memory streaming. We apply the framework to audio-driven, reference-image-conditioned avatar generation and obtain a four-step streaming causal model with stable long-horizon continuation.

Our main contributions are summarized as follows.

Rolling-free training and distillation. We propose a rolling-free training and distillation framework that converts a bidirectional avatar diffusion model into a causal streaming model without ODE trajectory data or temporal autoregressive rollouts.

Flexible chunk-wise inference. We show that, under noisy-cache inference, arbitrary temporal partitions yield equivalent denoising results. This unifies full-sequence denoising and frame-by-frame streaming as two extreme cases, enabling a training-free trade-off between latency and throughput.

End-to-end streaming avatar deployment. We instantiate the framework for audio-driven, reference-conditioned avatars and distill a four-step causal model with hybrid local–global attention and attention-sink streaming. It enables bounded-memory, visually stable long-horizon generation with comparable qualitative quality.

2 Related Work

Audio-Driven Talking Avatar Generation. Recent diffusion-based avatar models typically use bidirectional temporal attention over fixed clips Ma et al. (2024); Yang et al. (2024); Kong et al. (2026); Team et al. (2025); Yang et al. (2025a). We convert this paradigm into an open-ended causal generator while preserving reference identity and audio–visual alignment.

Autoregressive and Causal Video Diffusion. Diffusion Forcing, CausVid, and Self-Forcing study causal video diffusion and autoregressive distillation Chen et al. (2024); Sun et al. (2025); Yin et al. (2025); Huang et al. (2026). Self-Forcing and recent context-matching extensions construct student prefixes sequentially, introducing a training bottleneck Cui et al. (2026); Zheng et al. (2026); Bandyopadhyay et al. (2026). Our noise-aligned history and fixed block-causal mask remove this rollout from DMD and support flexible AR, MTP, and prefill inference.

Streaming Inference and Attention Sinks. LiveAvatar uses noisy historical KV states and timestep forcing for long-horizon streaming Huang et al. (2025). We instead exploit prefill–rollout equivalence for parallel DMD back-simulation and chunk-size changes without retraining, while using attention sinks for bounded-memory continuation. See detailed comparisons in Appendix B.2.

3 Method

3.1 Preliminaries

Flow Matching. We use the standard linear flow-matching path (Liu et al., 2022; Esser et al., 2024): $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\boldsymbol{\epsilon}$, where $\mathbf{x}_0$ is a clean latent, $\boldsymbol{\epsilon}$ is Gaussian noise, and $t \in [0, 1]$. The model predicts the conditional velocity and is integrated from $t = 1$ to $t = 0$. See full objective in Appendix A.2.

Distribution Matching Distillation. DMD (Yin et al., 2024; Fan et al., 2026) distills a multi-step model into a few-step generator by matching intermediate flow-state distributions. It uses a trainable generator $\boldsymbol{G_\phi}$, a frozen real-distribution score estimator $\boldsymbol{T_{\hat{\theta}}}$, and a trainable fake-distribution estimator $\boldsymbol{F_\theta}$. With conditioning omitted, its reverse-KL objective is

$$\mathcal{L}_{\mathrm{DMD}} = \mathbb{E}_t \left[D_{\mathrm{KL}} \left(p_{\mathrm{fake},t} \,\|\, p_{\mathrm{real},t} \right) \right]. \tag{1}$$

The fake score estimator tracks the generator distribution, and the generator is updated with the teacher–fake score discrepancy. Appendix A.2 gives the full gradient derivation.

3.2 Rolling-free Training and Distillation

3.2.1 Clean-Cache and Noisy-Cache Inference Formulations

We index the video by temporal latent units and denote unit i at flow time t by $\mathbf{x}_t^i$. Depending on the VAE temporal compression, each unit may represent one or more consecutive RGB frames. Typically, a Wan (Wan et al., 2025) temporal latent represents either one or four RGB frames. Let I, P, and A^i denote the reference image, global prompt embedding, and audio embedding aligned with unit i, respectively. Let w denote the number of preceding temporal latent units visible at a given layer, which may differ between local and global attention layers.

At flow time t, the predicted velocity of the current unit is conditioned on its latent state, the temporal cache, and the reference, audio, and prompt inputs.

In clean-cache inference, the context contains fully denoised historical latents at flow time 0:

$$\widehat{\mathbf{v}}_t^i = \boldsymbol{\psi}_{\boldsymbol{\theta}}\Big(\mathbf{x}_t^i, \mathbf{x}_0^{(i-w):(i-1)}, I, A^i, P, t\Big)\,. \tag{2}$$

In noisy-cache inference, the context is maintained at the same flow time as the current unit:

$$\widehat{\mathbf{v}}_t^i = \boldsymbol{\psi}_{\boldsymbol{\theta}}\Big(\mathbf{x}_t^i, \mathbf{x}_t^{(i-w):(i-1)}, I, A^i, P, t\Big)\,. \tag{3}$$

Here, $\widehat{\mathbf{v}}_t^i$ denotes the velocity predicted for the current temporal latent unit. The clean-cache formulation in Eq. (2), illustrated in Fig. 2(d), conditions the current unit on fully denoised historical latents. In contrast, the noisy-cache formulation in Eq. (3), illustrated in Fig. 2(e), conditions the current unit on the visible historical latents maintained at the current flow time t.

3.2.2 Flexible Chunk-Wise Inference

Given a latent video sequence of n temporal latent units, the sequence can be partitioned into m consecutive chunks of arbitrary sizes $f_1, \ldots, f_m$ satisfying $\sum_{i=1}^m f_i = n$. The chunks are processed sequentially, while all temporal latent units within each chunk are denoised in parallel. Single-pass full-sequence denoising and unit-by-unit streaming are the two extreme cases corresponding to $m = 1$ and $f_i = 1$ for all i, respectively. By analogy with common terminology in large language model inference, we refer to unit-by-unit streaming as AR rollout, single-pass full-sequence denoising as prefill inference, and intermediate chunk-wise denoising as MTP inference (Gloeckle et al., 2024). We use this terminology only as a conceptual analogy to language-model inference, referring to how temporal latent units are grouped across forward passes. Each temporal latent unit itself comprises multiple DiT tokens rather than corresponding to a single language-model token.

Design-space summary. Clean- and noisy-cache formulations admit different combinations of causal initialization and DMD teacher choice. Our final pipeline uses noise-aligned causal initialization (Causal SFT or Causal CFG distillation), followed by rolling-free DMD with a bidirectional fake score estimator and frozen bidirectional teacher. Appendix A.1 compares clean-cache and causal-teacher alternatives, including their rollout requirements.

3.3 Bidirectional Avatar Teacher Design

We instantiate our framework with a bidirectional teacher that jointly ensures reference-conditioned appearance preservation, cross-segment continuity, and audio-aligned motion.

3.3.1 Reference-Conditioned Long-Horizon R2V Design

We adapt an image-to-video (I2V) foundation model (Wan et al., 2025) into a reference-to-video (R2V) teacher. Unlike I2V, which preserves the input image as the first frame, R2V uses it as an appearance condition: even the first output frame must follow the driving audio rather than copy the reference mouth shape or expression. Fig. 3 summarizes the resulting teacher.

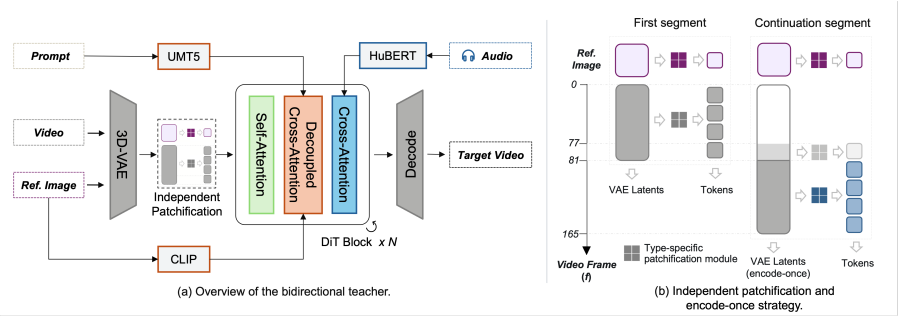


Figure 3: **Overview of the bidirectional avatar teacher.** It combines reference and multimodal conditioning with encode-once temporal processing for segment-wise continuation.

Reference Conditioning and Identity Preservation. We combine low-level reference-latent injection (Lin et al., 2025) with high-level image feature cross-attention. Mixing in-segment and out-of-segment references, together with a reference-state embedding, balances identity preservation against repeated copying (Yang et al., 2025a; Jiang et al., 2025). Appendix A.5.1 describes reference injection and sampling details.

Long-Horizon R2V Continuation. For continuation, each segment conditions on motion latents from the preceding segment's tail. Our *encode-once* strategy encodes context, motion, and target frames in one 3D-VAE pass, so the reused tail latents have matching temporal context and avoid decode–re-encode error. Type-specific patchification and segment sampling are detailed in Appendix A.5.2, with an ablation in Appendix A.4.6.

3.3.2 Lightweight and Temporally Aligned Audio Conditioning

We freeze the R2V visual backbone and train a lightweight audio adapter that projects speech into temporally aligned tokens A^i, injected through residual cross-attention at selected layers. The adapter is causal from the outset, so its conditioning semantics remain unchanged during causal adaptation and distillation. Appendix A.6 details encoder, window, and timestep-modulation.

3.4 Causal Adaptation and Distillation

3.4.1 Causal Model Adaptation

Causal R2V Backbone Adaptation. We adapt the pretrained bidirectional R2V model into a causal student which initializes the generator for the subsequent DMD stage. We consider two alternative objectives for initializing the causal visual backbone: Causal SFT and Causal CFG distillation. Both use the same causal mask and conditioning inputs, and differ only in their regression targets. Let I denote the reference-image condition, P the text prompt condition, and M the causal attention mask. Let $\boldsymbol{\psi}_{\boldsymbol{\phi}}$ and $\boldsymbol{T}_{\hat{\boldsymbol{\theta}}}$ denote the causal student and the frozen bidirectional teacher models, respectively. Causal SFT adapts the backbone to the causal computation pattern using the standard flow-matching loss:

$$\mathcal{L}_{\text{CausalSFT}} = \mathbb{E}_{\mathbf{x}_0,t,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\psi}_{\boldsymbol{\phi}}\left(\mathbf{x}_t, t, I, P; M\right) - \left(\boldsymbol{\epsilon} - \mathbf{x}_0\right)\right\|_2^2\right]. \tag{4}$$

Alternatively, Causal CFG distillation trains the causal student to match the frozen bidirectional teacher's classifier-free-guided prediction under the same causal mask. It shares Causal SFT 's conditioning and optimization setup, with the full target and loss given in Appendix A.3.

Audio Adapter Adaptation for the Causal Backbone. We freeze the causally adapted R2V backbone and fine-tune only the audio adapter initialized from the teacher (Sec. 3.3.2). Keeping its architecture and causal audio window unchanged, we adapt it to the causal visual representation to preserve audio–visual synchronization.

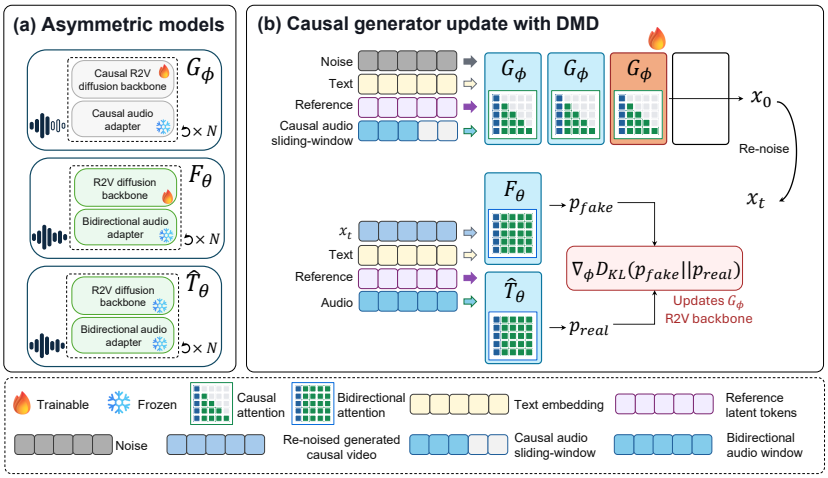


Figure 4: **Asymmetric distribution matching distillation.** (a) The causally adapted student initializes the generator $\boldsymbol{G}_\phi$, while the pretrained bidirectional R2V model initializes the fake score estimator $\boldsymbol{F}_\theta$ and frozen teacher $\boldsymbol{T}_{\hat{\theta}}$. The generator uses the audio adapter adapted to the causal visual backbone, whereas the bidirectional teacher and fake score estimator use the teacher audio adapter; all three use causal sliding-window audio conditioning. Only the generator and fake score estimator backbones are updated, while all audio adapters remain frozen. (b) Back-simulation uses a parallel prefill pass at each denoising step, avoiding temporal autoregressive rollout.

3.4.2 Asymmetric Distribution Matching Distillation

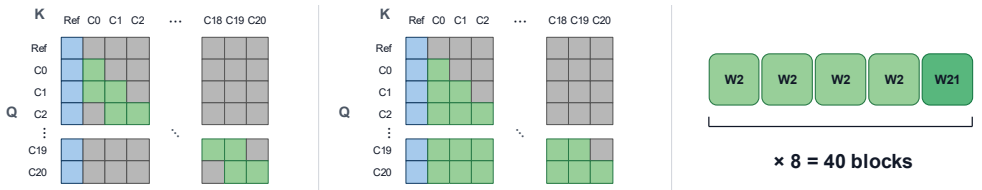
Building on the causally adapted backbone and audio adapter, we distill the causal student into the final few-step causal generator. Following CausVid Yin et al. (2025), we adopt an asymmetric distribution matching distillation setup. As shown in Fig. 4, the generator $\boldsymbol{G}_\phi$ is initialized from the causally adapted model and uses block-wise causal attention, whereas the frozen teacher $\boldsymbol{T}_{\hat{\theta}}$ and trainable fake score estimator $\boldsymbol{F}_{\boldsymbol{\theta}}$ are initialized from the bidirectional R2V model in Sec. 3.3 and retain bidirectional attention. The corresponding audio adapters also participate in this multi-conditional distillation. At each distillation step, the generator, teacher, and fake score estimator receive the same reference-image, text, and audio conditions. The generator uses the audio adapter adapted to the causal visual backbone, whereas the teacher and fake score estimator use the teacher audio adapter from Sec. 3.3.2; all three use the same causal audio-window design. We freeze the audio adapters so that the reduction in sampling steps is achieved entirely by updating the backbone parameters, while the fixed adapters preserve the audio-conditioning pathway and shape the gradients received by the trainable components. By the prefill–rollout equivalence established in Sec. 3.5, the inference-time causal dependencies can be reproduced by a parallel prefill forward pass at each denoising step, thereby avoiding temporal autoregressive rollout during distillation. Appendix A.4.5 further analyzes direct-distillation stability under clean- and noisy-cache inference, together with the effect of causal initialization.

3.4.3 Efficient Hybrid Local–Global Attention

Full-horizon block-wise causal attention preserves long-range context but increases attention and KV-cache costs, while a narrow window limits distant interactions. Inspired by sliding window attention (Jiang et al., 2023), we combine local and global temporal windows across layers.

For target chunks $\mathbf{z}_1, \ldots, \mathbf{z}_T$, W_k lets $\mathbf{z}_i$ attend to the reference latent $\mathbf{z}_{\text{ref}}$ and target chunks $\mathbf{z}_j$ with $\max(1, i-k+1) \le j \le i$. Thus, $W2$ covers the current and previous chunk, whereas $W21$ covers the full 21-chunk horizon; the reference remains visible in every layer, as illustrated in Fig. 5(a,b).

The 40-layer generator repeats $[W2, W2, W2, W2, W21]$ eight times. This 4:1 schedule reduces average temporal context and KV-cache demand while preserving periodic global access. We use the same schedule during causal adaptation and asymmetric distribution matching distillation.



(a) W2 causal window (b) W21 causal window (c) Hybrid block schedule

Figure 5: **Hybrid local–global block-wise causal attention.** Blue, green, and gray cells indicate reference keys, visible target keys, and masked positions, respectively. Both windows include the reference latent, plus (a) the current and immediately preceding target chunks ($W2$) or (b) up to 21 current-or-previous target chunks ($W21$). (c) The generator repeats four $W2$ blocks followed by one $W21$ block eight times, totaling 40 transformer blocks.

3.5 Flexible and Long-Horizon Causal Inference

Flexible Inference. As described in Sec. 3.2.2, noise-aligned history and a fixed block-causal mask supports three equivalent modes: prefill, autoregressive (AR) rollout, and multi-token prediction (MTP). Prefill processes all temporal units in one pass, AR processes one unit at a time, and MTP processes intermediate-sized chunks, allowing the same model to switch modes without retraining.

Long-Horizon Streaming with Attention Sink. Following StreamLLM Xiao et al. (2024), we combine layer-specific attention sinks with a rolling KV cache to maintain bounded memory during long-horizon streaming. We defer the cache and window-relative RoPE update rules to Appendix A.4.4 and evaluate their efficiency and long-horizon behavior in Sec. 4.4.

4 Experiments

4.1 Experimental Setup

We train on 1.1K hours of paired human-centric video, audio, and prompt data and evaluate on held-out SekoBench-Short and SekoBench-Long splits. They contain 100 480p cases with up to 20-second audio and 25 480p cases with four-minute audio, respectively. We compare offline and streaming methods at 480p using identity, audio–visual synchronization, perceptual, and aesthetic metrics. Starting from Wan2.1-I2V, we build a bidirectional R2V teacher, causally adapt its visual backbone and audio adapter, and distill the final generator with asymmetric DMD. Applying the same pipeline to MiniMax H3 Ref2VA produces DeForcing-H3. Detailed data, schedules, and optimization settings are provided in Appendix A.4.

4.2 Quantitative Evaluation

We evaluate offline and streaming settings separately. The offline comparison includes InfiniteTalk Yang et al. (2025a) and DeForcing-Bi, while the streaming comparison includes LiveTalk Chern et al. (2025), LiveAvatar Huang et al. (2025), SoulX-FlashTalk Shen et al. (2025), DeForcing, and DeForcing-H3. All methods are evaluated at 480p using their official inference settings. We also evaluate the SekoBench-Short ground-truth videos as reference values; SekoBench-Long has no corresponding ground truth. We report five metrics for identity preservation (ID-Sim), audio–visual synchronization (Sync-C and Sync-D), perceptual quality (IQA), and aesthetics (ASE). Metric definitions and known biases are discussed in Appendix A.4.2. Results are reported in Tab. 1.

Main results. DeForcing provides the strongest overall streaming trade-off on both benchmarks: it leads LiveAvatar on all five Short metrics and four of five Long metrics, while remaining competitive with SoulX-FlashTalk on synchronization and identity. Relative to the bidirectional teacher DeForcing-Bi, it improves perceptual, aesthetic, and synchronization scores with comparable identity similarity.

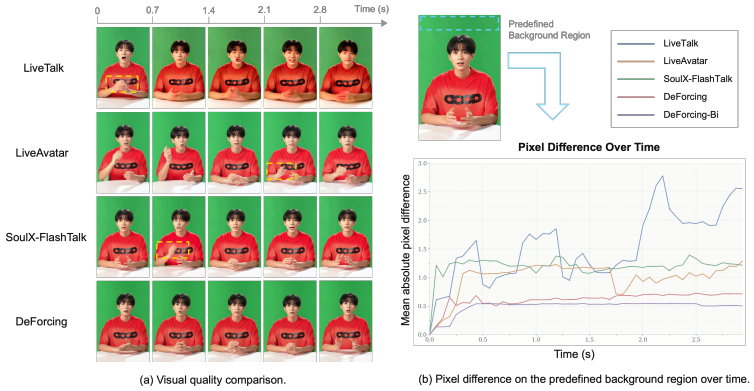
Cross-backbone transfer. Applying the same pipeline to the independent MiniMax H3 Ref2VA backbone, DeForcing-H3 achieves the best IQA and the second-best results on the other four Long

metrics. Compared with LiveAvatar, it improves ID-Sim from 0.651 to 0.790, Sync-C from 5.340 to 6.384, and Sync-D from 9.306 to 8.457. Relative to Wan-based DeForcing, H3 provides higher IQA, Sync-C, and ID-Sim, while DeForcing remains better in ASE and Sync-D. Metric limitations are discussed in Appendix A.4.2. We complement the quantitative results with the qualitative comparison below.

Table 1: **Quantitative comparison on SekoBench.** Short and Long denotes SekoBench-Short and SekoBench-Long. Bold values indicate the best results within each setting, while underlined values indicate the second-best results among streaming methods. GT is reported as a reference.

Benchmark	Model	Metrics				
		IQA ↑	ASE ↑	Sync-C ↑	Sync-D ↓	ID-Sim ↑
Short	GT	0.905	5.054	4.993	8.535	0.855
	Offline Generation					
	InfiniteTalk Yang et al. (2025a)	0.839	4.997	**5.276**	**8.463**	0.792
	DeForcing-Bi	**0.867**	**5.057**	4.386	9.035	**0.800**
	Streaming Generation					
	LiveAvatar Huang et al. (2025)	0.865	5.033	5.120	8.932	0.788
	LiveTalk Chern et al. (2025)	**0.883**	4.968	2.078	11.695	0.629
	SoulX-FlashTalk Shen et al. (2025)	0.842	4.956	**6.236**	**8.427**	**0.869**
	DeForcing-H3	0.859	5.093	5.347	8.687	0.831
	DeForcing	0.875	**5.096**	5.144	8.472	0.794
Long	*Streaming Generation*					
	LiveAvatar Huang et al. (2025)	0.947	5.343	5.340	9.306	0.651
	SoulX-FlashTalk Shen et al. (2025)	0.926	5.258	**6.404**	8.718	**0.792**
	DeForcing-H3	**0.948**	5.358	6.384	8.457	0.790
	DeForcing	0.941	**5.366**	6.107	**8.445**	0.769

4.3 Qualitative Comparisons



(a) Visual quality comparison. (b) Pixel difference on the predefined background region over time.

Figure 6: Qualitative comparison with baseline methods. Frames and background-difference curves are shown for 2.8 s clips; lower and smoother curves indicate stronger temporal stability.

Fig. 6 compares representative methods using visual samples and background-difference curves. Our teacher and distilled model show the lowest and smoothest curves in this comparison, indicating stable appearance across the clip.

4.4 Inference Efficiency Analysis

We focus on three deployment properties: the speed–quality schedule, partition-invariant execution, and long-horizon streaming. Full backend measurements and additional qualitative results are provided in Appendix A.4.4.

Inference Schedule and Real-Time Throughput. We compare the four-step model used for quality evaluation with a two-step variant under prefill, MTP, and AR execution. As shown in Tab. 2, larger chunks improve total throughput while smaller chunks expose the first result earlier. The two-step FP8-accelerated AR configuration generates a 5 s, 81-frame clip in 4.230 s, with a first-chunk latency of 0.305 s and 1.182× real-time DiT throughput.

Table 2: DiT rollout time, first-chunk latency, and real-time throughput at 480p for generating 21 temporal latent units (81 frames, 5 s) on a single H100 GPU. We compare four- and two-step models with prefill (21 units/pass), MTP (3 units/pass), and AR (1 unit/pass). Unless otherwise noted, measurements use BF16 with native PyTorch; the accelerated AR row uses FP8 linear layers and `torch.compile`. Results are medians over 20 runs after 5 warmup iterations. RTF is the generated video duration divided by total DiT rollout time; values above 1 indicate faster-than-real-time throughput. First-chunk latency measures the time to produce the first temporal chunk.

Inference Schedule	**Total Rollout Time (s)↓**		**First Chunk Latency (s)↓**		**RTF↑**	
	4 Steps	**2 Steps**	**4 Steps**	**2 Steps**	**4 Steps**	**2 Steps**
Prefill (21 units/pass)	11.944	6.006	–	–	0.419	0.833
MTP (3 units/pass)	15.948	7.713	2.625	1.312	0.314	0.648
AR rollout (1 unit/pass)	19.479	9.795	1.419	0.716	0.257	0.510
AR rollout (FP8 linear + `torch.compile`)	**8.315**	**4.230**	**0.606**	**0.305**	**0.601**	**1.182**

Partition-Invariant Execution. With identical noise and conditioning, prefill, AR, and MTP generate nearly identical 5 s clips; the small discrepancies arise from finite-precision accumulation and attention kernels. Representative samples are provided in Appendix A.4.4.

Long-Horizon Streaming. With hybrid attention and attention sinks, the rollout remains visually stable through frame 9080 (roughly 9.5 minutes at 16 fps). The long-horizon samples, prompt-travel results, and additional action sequences are provided in Appendix A.4.4.

4.5 Ablation Studies

Causal Generator Initialization. The three initialization variants yield closely matched distilled generators, with causal warm-up providing small improvements in identity or perceptual quality. The complete comparison, including the Self-Forcing ODE control, is reported in Appendix A.4.5.

Hybrid Local–Global Attention with Matched Attention Sinks. We compare matched Hybrid and all-$W21$ causal configurations. Hybrid uses the $[W2, W2, W2, W2, W21]$ attention schedule with corresponding $[S1, S1, S1, S1, S2]$ sinks, whereas the baseline uses all-$W21$ attention with sink size 2. Across ID-Sim and mean optical flow (Mean-OF), Hybrid improves identity similarity while reducing motion intensity. See detailed values in Appendix A.4.6.

In addition, the independent-patchification ablation is reported in Appendix A.4.6, while representative attention maps across heads and transformer depths are provided in Appendix B.

5 Conclusions

DeForcing shows that, under noisy-cache inference, noise-aligned history and a fixed block-causal mask make denoising invariant to temporal partitioning. Prefill, autoregressive, and arbitrary chunk-wise execution therefore share the same computation, enabling rollout-free distillation, parallel DMD back-simulation, and a single checkpoint that trades latency for throughput by changing only its inference partition. Applied to audio-driven avatars, DeForcing supports both low-latency streaming and high-throughput generation. Beyond avatars, the same principle may extend to causal diffusion world models, where noise-aligned historical context could enable rollout-free distillation and flexible temporal execution for long-horizon generation.

AI USE STATEMENT

Generative AI tools were used to assist with literature discovery and summarization; and polising writing, including its structure, title, references, and LaTeX formatting. Generative AI tools were not used to produce the reported numerical results. All AI-suggested text, claims, citations, equations, and revisions were reviewed and verified by the authors against primary sources, derivations, the implementation, and experimental outputs. The authors made all final scientific decisions and take responsibility for the complete content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

Audio-driven avatar generation can be misused for non-consensual image manipulation, impersonation, and misleading synthetic media. We encourage the use of this technology only with the consent of the depicted individuals and the relevant rights holders, and with clear disclosure that generated content is synthetic. Any release or deployment of the model should include appropriate provenance, access control, and misuse monitoring, particularly for identity-sensitive or public-facing applications.

REPRODUCIBILITY STATEMENT

The paper organizes the information needed to reproduce our method and experiments across the main text and Supplementary Material. Section 3 describes the noisy-cache formulation, the rolling-free training and distillation pipeline, the avatar teacher design, and the flexible inference schedules. Section 4 specifies the datasets, evaluation benchmarks, comparison protocols, efficiency measurements, and ablation studies.

Supplementary Appendix A.4 provides the detailed dataset composition, benchmark construction, training schedules, optimization settings, and implementation details. Appendix A.2 and Appendix A.3 provide the full flow-matching, DMD, and causal CFG-distillation formulations. Appendix A.6 describes the causal audio–visual alignment and audio encoder configuration. Additional metric analysis, inference-backend measurements, inference-schedule comparisons, and ablations are provided in Appendix A.4.2, Appendix A.4.4, and Appendix A.4.6.

We will release the inference configurations and pretrained model weights needed to reproduce the reported generation results.

REFERENCES

Hmrishav Bandyopadhyay, Xuanchi Ren, Zijian Huang, Jay Zhangjie Wu, Tianshi Cao, Ruilong Li, Bryan Chu, Sanja Fidler, Yi-Zhe Song, and Zian Wang. Context-matched distillation: Teacher causality for autoregressive video distillation, 2026. URL https://arxiv.org/abs/2608.13391.

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

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, October 2022. ISSN 1941-0484. doi: 10.1109/jstsp.2022.3188113. URL http://dx.doi.org/10.1109/JSTSP.2022.3188113.

Ethan Chern, Zhulin Hu, Bohao Tang, Jiadi Su, Steffi Chern, Zhijie Deng, and Pengfei Liu. Livetalk: Real-time multimodal interactive video diffusion via improved on-policy distillation. *arXiv preprint arXiv:2512.23576*, 2025.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Asian Conference on Computer Vision*, pp. 251–263. Springer, 2016.

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

Jiaxing 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. In *International Conference on Learning Representations*, volume 2026, pp. 85802–85822, 2026.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, et al. Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Xiangyu Fan, Jiaqi Li, Zhiqian Lin, Weiye Xiao, and Lei Yang. Unitalker: Scaling up audio-driven 3d facial animation through a unified model, 2024. URL https://arxiv.org/abs/2408.00762.

Xiangyu Fan, Zesong Qiu, Zhuguanyu Wu, Fanzhou Wang, Zhiqian Lin, Tianxiang Ren, Dahua Lin, Ruihao Gong, and Lei Yang. Phased dmd: Few-step distribution matching distillation via score matching within subintervals, 2026. URL https://arxiv.org/abs/2510.27684.

Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Roziere, David Lopez-Paz, and Gabriel Synnaeve. Better & faster large language models via multi-token prediction. *arXiv preprint arXiv:2404.19737*, 2024.

Jonatas Grosman. Fine-tuned XLSR-53 large model for speech recognition in English. https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english, 2021.

Driss Guessous, Reuben Stern, Markus Hoehnerbach, Fung Xie, Ted Zadouri, Jay Shah, and Tri Dao. FlexAttention + FlashAttention-4: Fast and flexible. https://pytorch.org/blog/flexattention-flashattention-4-fast-and-flexible/, March 2026. PyTorch Blog, accessed August 30, 2026.

Pengcheng Guo and Shixing Liu. Chinese speech pretrain, 2022. URL https://github.com/TencentGameMate/chinese_speech_pretrain.

Yuwei Guo, Ceyuan Yang, Hao He, Yang Zhao, Meng Wei, Zhenheng Yang, Weilin Huang, and Dahua Lin. End-to-end training for autoregressive video diffusion via self-resampling. *arXiv preprint arXiv:2512.15702*, 2025.

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

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Jianwen Jiang, Weihong Zeng, Zerong Zheng, Jiaqi Yang, Chao Liang, Wang Liao, Han Liang, Yuan Zhang, and Mingyuan Gao. Omnihuman-1.5: Instilling an active mind in avatars via cognitive simulation, 2025. URL https://arxiv.org/abs/2508.19209.

Zhe Kong, Feng Gao, Yong Zhang, Zhuoliang Kang, Xiaoming Wei, Xunliang Cai, Guanying Chen, and Wenhan Luo. Let them talk: Audio-driven multi-person conversational video generation. *Advances in Neural Information Processing Systems*, 38:70990–71013, 2026.

Haodong Li, Shaoteng Liu, Zhe Lin, and Manmohan Chandraker. Rolling sink: Bridging limited-horizon training and open-ended testing in autoregressive video diffusion. *arXiv preprint arXiv:2602.07775*, 2026.

Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, et al. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. *arXiv preprint arXiv:2412.00115*, 2024.

Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, and Chao Liang. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models, 2025. URL https://arxiv.org/abs/2502.01061.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations*, volume 2026, pp. 91177–91196, 2026.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

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

Yifeng Ma, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yingya Zhang, and Zhidong Deng. Dreamtalk: When emotional talking head generation meets diffusion probabilistic models, 2024. URL https://arxiv.org/abs/2312.09767.

MiniMaxAI. MiniMax H3. Hugging Face model card, 2026. URL https://huggingface.co/MiniMaxAI/MiniMax-H3. Accessed: 2026-09-24.

Q-Future. Q-ReAlign: Lightweight human-aligned multimodal judges built on modern vision-language models. https://github.com/Q-Future/Q-ReAlign, 2026. Accessed: 2026-08-05.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. *Advances in Neural Information Processing Systems*, 37:68658–68685, 2024.

Le Shen, Qian Qiao, Tan Yu, Ke Zhou, Tianhang Yu, Yu Zhan, Zhenjie Wang, 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*, 2025.

Mingzhen Sun, Weining Wang, Gen Li, Jiawei Liu, Jiahui Sun, Wanquan Feng, Shanshan Lao, SiYu Zhou, Qian He, and Jing Liu. Ar-diffusion: Asynchronous video generation with auto-regressive diffusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7364–7373. IEEE, 2025.

Kling Team, Jialu Chen, Yikang Ding, Zhixue Fang, Kun Gai, Yuan Gao, Kang He, Jingyun Hua, Boyuan Jiang, Mingming Lao, et al. Klingavatar 2.0 technical report. *arXiv preprint arXiv:2512.13313*, 2025.

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.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In *International Conference on Machine Learning (ICML)*, 2024.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13941–13958, 2023. doi: 10.1109/TPAMI.2023.3298645.

Shaoshu Yang, Zhe Kong, Feng Gao, Meng Cheng, Xiangyu Liu, Yong Zhang, Zhuoliang Kang, Wenhan Luo, Xunliang Cai, Ran He, and Xiaoming Wei. Infinitetalk: Audio-driven video generation for sparse-frame video dubbing, 2025a. URL https://arxiv.org/abs/2508.14033.

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.

Shurong Yang, Huadong Li, Juhao Wu, Minhao Jing, Linze Li, Renhe Ji, Jiajun Liang, and Haoqiang Fan. Megactor: Harness the power of raw video for vivid portrait animation. *arXiv preprint arXiv:2405.20851*, 2024.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. IEEE, 2025.

Tao Zewei and Huang Yunpeng. Magiattention: A distributed attention towards linear scalability for ultra-long context, heterogeneous mask training. https://github.com/SandAI-org/MagiAttention/, 2025.

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

A.1 Training and Distillation Design Space

Having defined the clean-cache and noisy-cache inference paradigms, we characterize the corresponding design space of causal initialization and distribution matching distillation. Fig. 7 organizes the alternatives by inference paradigm, student initialization, and teacher choice.

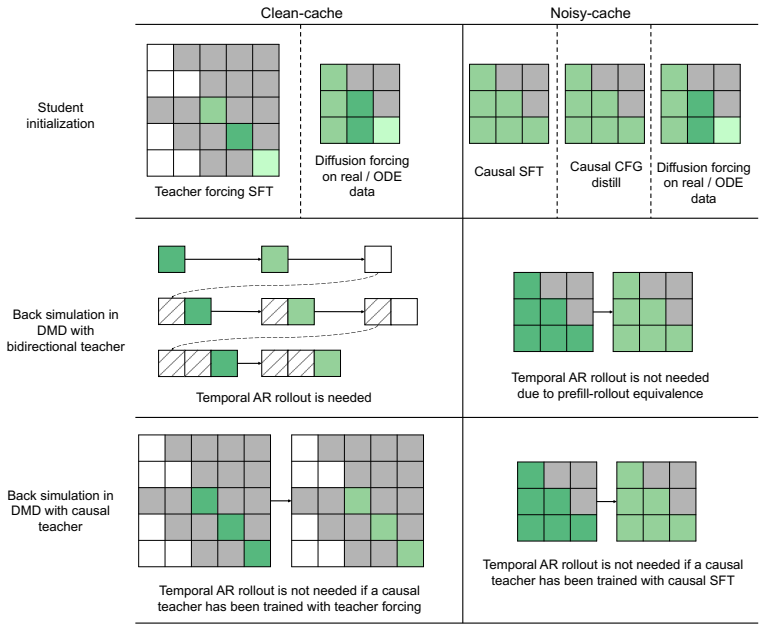


Figure 7: A holistic view of training and distillation schemes for clean-cache (left column) and noisy-cache (right column) AR inference. The top row shows student initialization, while the middle and bottom rows show DMD with bidirectional and causal teachers, respectively. **Top row: student initialization.** Diffusion Forcing (Chen et al., 2024), applied to either real video clips or ODE-rollout synthetic data, is the generic option for both paradigms but is sample-inefficient because its per-unit independent flow times rarely match the deployed cache state. The targeted alternative is *Teacher Forcing SFT* for clean-cache, whereas noisy-cache admits two alternative choices, *Causal SFT* and *Causal CFG distillation*; the latter uses the guided prediction of the bidirectional teacher as its target. **Middle row: Distribution matching distillation against a bidirectional teacher.** Back-simulation under clean-cache requires explicit temporal autoregressive rollout of the student, whereas under noisy-cache it reduces to a single parallel forward pass per flow step by virtue of the prefill–rollout equivalence. **Bottom row: Distribution matching distillation against a causal teacher.** A causal teacher can support teacher-forced score-model evaluation without temporal rollout. This variant is distinct from the noisy-cache equivalence used by our main training pipeline.

Causal Model Initialization. A generic initialization strategy applicable to both inference paradigms is Diffusion Forcing (Chen et al., 2024), which assigns independent flow times to different temporal latent units and trains them in parallel with a causal mask. Although flexible, this strategy spends training capacity on context and flow-time configurations that are not encountered by the deployed model. For clean-cache inference, Teacher Forcing SFT instead conditions each

unit on preceding clean latents and converts the bidirectional DiT into a causal model. For noisy-cache inference, we consider Causal SFT and Causal CFG distillation while preserving the latent and conditioning inputs, the flow-time assignment, and the regression-loss form of bidirectional DiT training. Each historical context latent is assigned the same flow time as the current target unit. Causal SFT retains the standard flow-matching target, whereas Causal CFG distillation replaces it with the classifier-free-guided prediction of the bidirectional teacher. Relative to the bidirectional DiT, the only architectural change in both cases is the use of causal temporal attention in the model being trained.

Distillation with a Bidirectional Teacher. With a bidirectional teacher, multi-step DMD back-simulates video latents along the student's inference-time trajectory before computing the score-based DMD update in Eq. (9). Under clean-cache inference, reproducing the inference-time trajectory requires explicit temporal autoregressive rollout. A parallel teacher-forcing pass instead conditions each temporal latent unit on ground-truth history and therefore does not reproduce the joint distribution induced by inference-time rollout. Under noisy-cache inference, the prefill–rollout equivalence established in Sec. 3.2.2 allows each denoising step to be back-simulated with a single parallel forward pass over the temporal dimension. In both cases, the back-simulated latents are re-noised and evaluated by the bidirectional teacher and fake score estimators to compute the score discrepancy used by DMD.

Distillation with a Causal Teacher. When a causal model has been obtained through Teacher Forcing SFT or Causal SFT, a frozen copy of its weights can serve as the teacher in the subsequent DMD stage. For noisy-cache inference, replacing the bidirectional teacher with a causal teacher leaves the back-simulation procedure unchanged, while the teacher and fake score estimator use block-wise causal attention.

For clean-cache inference, a causal teacher enables teacher-forced DMD without autoregressive rollout along the video temporal dimension. Here, teacher forcing refers only to the construction of temporal context: backward simulation along the diffusion trajectory remains necessary, while all temporal latent units can follow their denoising trajectories in parallel under ground-truth clean history. For each temporal latent unit, the generator back-simulates the current unit conditioned on its preceding ground-truth clean latents. The back-simulated clean latent is then re-noised and evaluated by the causal teacher and causal fake score estimator, both conditioned on the preceding ground-truth clean latents. Because the historical context does not depend on preceding student outputs, all temporal latent units can be processed in parallel with block-wise causal attention. This formulation follows the autoregressive factorization $p(\mathbf{x}_0^{1:n}) = \prod_{i=1}^{n} p(\mathbf{x}_0^i \mid \mathbf{x}_0^{1:(i-1)})$ by matching the student and teacher conditional distributions for each temporal latent unit under ground-truth clean history. It therefore does not require a temporally autoregressive student rollout during distillation. Teacher-forced causal DMD has also been validated in our prior work (Diao et al., 2026). Compared with this alternative, Causal-rCM (Zheng et al., 2026) still uses self-forcing DMD with temporally autoregressive student rollouts, while Context-Matched Distillation (Bandyopadhyay et al., 2026) evaluates causal scores on student-generated prefixes obtained through on-policy rollout.

Our Design Choice. We adopt noisy-cache inference and distillation with a bidirectional teacher. Clean-cache inference recursively conditions each temporal latent unit on previously generated clean outputs, making local errors more likely to accumulate during long-horizon continuation. Noisy-cache inference instead supports long-horizon continuation with an attention sink and permits arbitrary inference chunk sizes through the prefill–rollout equivalence. Within this noisy-cache formulation, both causal and bidirectional teachers are compatible with rolling-free DMD, while our experiments find that distillation with a bidirectional teacher yields better generation quality. Our final pipeline therefore applies either Causal SFT or Causal CFG distillation for initialization and then performs rolling-free DMD with a frozen bidirectional teacher and a bidirectional fake score estimator.

A.2 Flow Matching and DMD Details

This section collects the standard flow-matching and distribution-matching derivations omitted from the main text.

Flow matching. Let $\mathbf{x}_0 \sim p_{\text{data}}$ be a clean latent and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ be Gaussian noise. For $t \in [0, 1]$, we use the linear interpolation

$$\mathbf{x}_t = (1 - t)\mathbf{x}_0 + t\boldsymbol{\epsilon}, \tag{5}$$

whose conditional velocity is

$$\frac{d\mathbf{x}_t}{dt} = \boldsymbol{\epsilon} - \mathbf{x}_0. \tag{6}$$

Sampling t from a predefined distribution, such as a logit-normal distribution (Esser et al., 2024), gives the standard flow-matching objective

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}, t}\left[\left\|\boldsymbol{\psi}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) - (\boldsymbol{\epsilon} - \mathbf{x}_0)\right\|_2^2\right]. \tag{7}$$

At inference time, the learned ordinary differential equation is integrated backward from Gaussian noise at $t = 1$ to the data endpoint at $t = 0$.

Distribution matching distillation. DMD matches the generated and real distributions at intermediate flow states. Let $p_{\text{real},t}$ and $p_{\text{fake},t}$ denote these distributions. The reverse-KL objective is

$$\mathcal{L}_{\text{DMD}} = \mathbb{E}_t\left[D_{\text{KL}}\left(p_{\text{fake},t} \,\|p_{\text{real},t}\right)\right]. \tag{8}$$

For a generated sample $\mathbf{x}_0 = \boldsymbol{G}_{\boldsymbol{\phi}}(z)$, we construct $\mathbf{x}_t$ using Eq. (5). The frozen teacher and fake score estimators approximate the scores of $p_{\text{real},t}$ and $p_{\text{fake},t}$, respectively. Differentiating the reverse-KL objective with respect to the generator gives the score-difference estimator

$$\nabla_{\boldsymbol{\phi}}\mathcal{L}_{\text{DMD}} \approx -\mathbb{E}_{z,t,\boldsymbol{\epsilon}}\left[w_t\left(\boldsymbol{T}_{\hat{\boldsymbol{\theta}}}(\mathbf{x}_t, t) - \boldsymbol{F}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\right)^{\top}\frac{\partial \boldsymbol{G}_{\boldsymbol{\phi}}(z)}{\partial \boldsymbol{\phi}}\right], \tag{9}$$

where w_t is a timestep-dependent weighting factor. The fake score estimator is updated to track the generator distribution, while the generator is updated using the teacher–fake score discrepancy.

A.3 Causal CFG Distillation Details

For Causal CFG distillation, let $P_\emptyset$ denote the negative-prompt embedding used by the unconditional branch and let s_{cfg} be the classifier-free guidance scale. The frozen bidirectional teacher provides the guided target

$$\boldsymbol{T}^{\text{cfg}}_{\hat{\boldsymbol{\theta}}} = \boldsymbol{T}_{\hat{\boldsymbol{\theta}}}\left(\mathbf{x}_t, t, I, P_\emptyset\right) + s_{\text{cfg}}\left[\boldsymbol{T}_{\hat{\boldsymbol{\theta}}}\left(\mathbf{x}_t, t, I, P\right) - \boldsymbol{T}_{\hat{\boldsymbol{\theta}}}\left(\mathbf{x}_t, t, I, P_\emptyset\right)\right]. \tag{10}$$

The causal student uses the same noise-aligned history and causal mask as in Causal SFT, but regresses to this guided teacher prediction:

$$\mathcal{L}_{\text{CausalCFGDistill}} = \mathbb{E}_{\mathbf{x}_0, t, \boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\psi}_{\boldsymbol{\phi}}\left(\mathbf{x}_t, t, I, P; M\right) - \boldsymbol{T}^{\text{cfg}}_{\hat{\boldsymbol{\theta}}}\right\|_2^2\right]. \tag{11}$$

A.4 Experimental Details and Additional Analyses

A.4.1 Datasets and Implementation Details

The training set comprises 1.1K hours of human-centric videos collected from public datasets, including Hallo3 Cui et al. (2024) and OpenHumanVid Li et al. (2024), as well as internally collected data. Each video is paired with audio and a prompt. The training videos are sampled at 16 fps and include both 480p and 720p resolutions.

For evaluation, we construct SekoBench with two subsets. None of the benchmark videos or subjects appears in the training data, and each test case contains a reference image, a driving audio clip, and its corresponding prompt. SekoBench-Short contains 100 cases at 480p with audio clips of up to 20 seconds, whereas SekoBench-Long contains 25 cases at 480p with four-minute driving audio clips. The benchmark covers close-up, half-body, and full-body references and multiple languages.

We first adapt the pretrained Wan2.1-I2V-720p model into a bidirectional R2V backbone for 6,000 iterations on 64 NVIDIA H100 GPUs, training at both 480p and 720p. Although the bidirectional backbone supports both resolutions, we retain only 480p generation during causal adaptation. With

the R2V backbone frozen, we train the teacher audio adapter for 13,000 iterations on 32 H100 GPUs. We separately train the CFG distillation and Causal SFT causal-backbone variants for 8,000 iterations each on 32 H100 GPUs, then fine-tune the student audio adapter for 10,000 iterations on 32 H100 GPUs. Finally, we perform 600 generator updates during asymmetric DMD on 16 H100 GPUs with a batch size of 16. We also apply the pipeline to the MiniMax H3 Ref2VA base model MiniMaxAI (2026), yielding DeForcing-H3.

A.4.2 Metric Definitions and Interpretation

We use ArcFace Deng et al. (2019) to extract identity embeddings from the input reference image and generated frames, and compute identity similarity (ID-Sim) as their mean cosine similarity. For audio–visual synchronization, we use SyncNet Chung & Zisserman (2016): Sync-C measures alignment confidence and Sync-D measures the temporal offset between speech and lip movements. For perceptual and aesthetic quality, we use Q-ReAlign Q-Future (2026), which follows the visual scoring formulation of Q-Align Wu et al. (2024), to obtain IQA and ASE scores.

No single metric fully captures generation quality, so the scores should be interpreted jointly. ID-Sim is sensitive to motion magnitude: large pose or expression changes can reduce similarity to the reference even when identity remains perceptually consistent. Consistent with this limitation, the ground truth videos achieve an ID-Sim of only 0.855 rather than 1.0 on SekoBench-Short. Conversely, a nearly static face can obtain a high ID-Sim without exhibiting useful motion.

Sync-C can be increased by raising the classifier-free guidance (CFG) scale of the audio condition, which amplifies mouth movements. This may improve the score while producing exaggerated motion that is less natural. On SekoBench-Short, SoulX-FlashTalk produces pronounced mouth articulation, which may partly explain why its Sync-C score of 6.236 exceeds the ground-truth score of 4.993. We therefore omit audio-condition dropout when training our audio adapter and do not apply audio CFG during distillation or inference of DeForcing. DeForcing consequently obtains Sync-C and Sync-D scores of 5.144 and 8.472, close to the ground-truth values of 4.993 and 8.535.

IQA is a frame-level metric and does not directly capture temporal color consistency. Although LiveTalk achieves the highest IQA on SekoBench-Short, its outputs exhibit noticeable color drift, suggesting that IQA may favor saturated or high-contrast frames. These examples show why quantitative metrics provide only a partial characterization of generated video quality.

A.4.3 Cross-Backbone Comparison

Across both SekoBench subsets, DeForcing and DeForcing-H3 achieve broadly comparable scores in identity preservation, audio–visual synchronization, and frame-level quality, with each variant leading on different metrics in Tab. 1. In our qualitative inspection, DeForcing-H3 follows prompts more faithfully and generates more natural human motion than Wan-based DeForcing. We fine-tune only the H3-Base Ref2VA checkpoint MiniMaxAI (2026), which may limit the visual-quality scores of DeForcing-H3. Applying H3-Regenerate-2K to its outputs could further improve these scores.

A.4.4 Additional Inference Efficiency Results

Unless otherwise noted, efficiency measurements are performed at 480p on a single H100 GPU and report the median over 20 runs after five warmup iterations. The rollout benchmark generates 21 temporal latent units, corresponding to 81 frames or 5 s of video, with the hybrid schedule $[W2, W2, W2, W2, W21]$. The schedule reduces allocated KV-cache memory by approximately 69.1% relative to all-$W21$ attention. Reported latency covers DiT rollout only and excludes audio encoding, VAE decoding, data transfer, and other application overhead. First-chunk latency measures the time to produce the first temporal chunk.

Attention Backend Benchmark. The attention-backend comparison is an implementation benchmark rather than a model-design ablation. We benchmark two H100 execution paths for the $W21$ component of the hybrid schedule. MagiAttention Zewei & Yunpeng (2025) uses `flex_flash_attn_func`, while PyTorch 2.13 uses its built-in `flex_attention` implementation Guessous et al. (2026). Both execute the same block-wise causal mask, while dense FlashAttention3 Shah et al. (2024) serves as an unmasked reference.

Table 3: Kernel-level efficiency comparison of two implementations of the block-wise causal $W21$ pattern on a single NVIDIA H100 GPU. MagiAttention (Zewei & Yunpeng, 2025) uses `flex_flash_attn_func`, while PyTorch 2.13 uses its built-in `flex_attention` implementation (Guessous et al., 2026). Dense bidirectional FlashAttention3 (Shah et al., 2024) serves as a compute-throughput reference. We report forward and backward latency and compute throughput using a batch size of 1, 40 attention heads, a head dimension of 128, BF16 precision, one reference block, and 21 target blocks. Each measurement is averaged over 20 runs after four warmup iterations.

Resolution	Backend	Time (ms) ↓			Throughput (TFLOPs/s) ↑		
		Fwd	Bwd	Total	Fwd	Bwd	Total
480p, 22f (S=34,320)	FlashAttention3 (dense)	31.30	95.18	126.49	770.6	506.9	572.1
	MagiAttention (causal-W21)	20.23	68.88	89.12	623.2	366.1	424.5
	PyTorch 2.13 (causal-W21)	18.91	55.70	74.62	666.7	452.7	507.0
720p, 22f (S=79,200)	FlashAttention3 (dense)	179.56	495.77	675.33	715.4	518.2	570.7
	MagiAttention (causal-W21)	107.63	319.20	426.83	623.9	420.8	472.0
	PyTorch 2.13 (causal-W21)	106.50	292.83	399.33	630.5	458.6	504.5

Both causal backends execute the structured $W21$ pattern efficiently at 480p and 720p. The detailed forward/backward latency and throughput comparison is reported in Tab. 3.

Attention-sink cache update. Following StreamLLM Xiao et al. (2024), we maintain separate persistent cache entries for the $W21$ and $W2$ layers. $W21$ layers retain the reference latent and the first generated temporal unit as attention sinks, whereas $W2$ layers retain only the reference latent. The remaining temporal units form a FIFO rolling tail. We cache unrotated keys with their values and assign temporal RoPE indices according to the current cache slots, while keeping sink indices fixed. This layer-specific update preserves the training-time visibility pattern as the cache advances.

Visual Comparison of Inference Schedules. The prefill, AR, and MTP schedules produce visually consistent samples under identical initial noise, reference image, text prompt, and audio conditioning.

Prompt Travel and Additional Action Gallery. The bounded-cache mechanism supports prompt travel during continuous streaming: users can change the text prompt without restarting the rollout, while the existing KV cache preserves preceding temporal context. The generated sequence follows successive motion prompts while maintaining visual continuity across prompt transitions, as shown in Fig. 10.

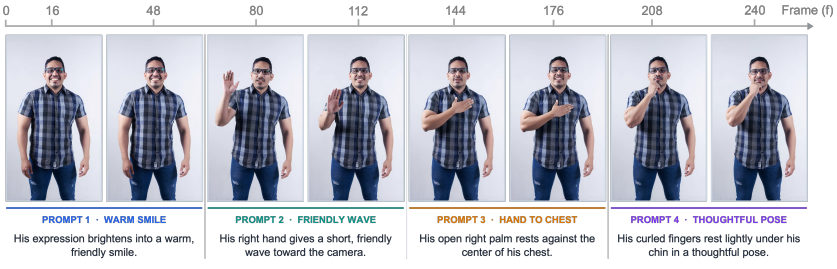


Figure 10: Qualitative demonstration of prompt travel during continuous streaming. The rollout transitions among four user-specified text prompts, producing a warm smile, a friendly wave, a hand-to-chest gesture, and a thoughtful pose while preserving the subject identity and scene appearance across prompt transitions.

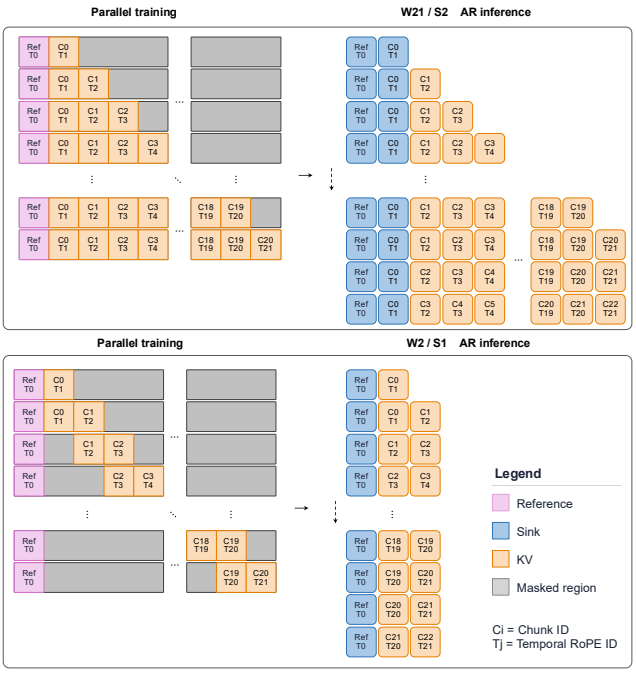


Figure 8: Schematic illustration of the training-time temporal attention patterns (left) and rolling AR inference procedures (right) for the $W21$ and $W2$ layers. During AR inference, both the KV-cache selection and temporal RoPE assignment follow the corresponding block-wise causal patterns used during training. The $W21$ layers retain the key-value states corresponding to the reference latent and the first generated temporal latent unit, whereas the $W2$ layers retain only the key-value states corresponding to the reference latent. Active cached keys are assigned window-relative RoPE indices according to their current cache slots, preserving the training-time positional layout as the cache rolls.

Long-horizon qualitative result. The long-horizon rollout remains visually stable through frame 9080, or roughly 568 s at 16 fps, using the layer-specific attention-sink configuration.



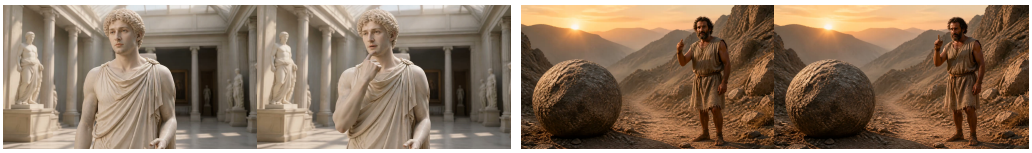
Figure 9: Qualitative comparison of three equivalent inference schedules on a representative talking-head example. The first, second, and third rows show frames generated with prefill, AR rollout, and MTP, respectively, sampled at 0.50, 1.50, 2.50, 3.50, and 4.50 s. All frames use identical initial noise, reference image, text prompt, and audio conditioning.



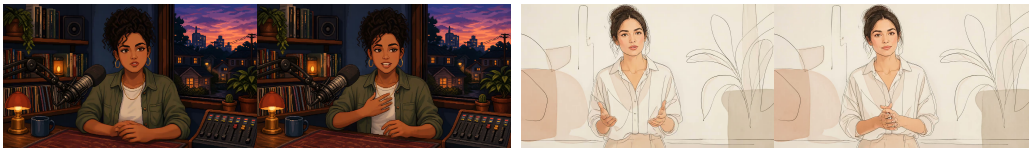
(a) Finger heart (b) Touch chin (c) Thumbs up (d) Clap



(e) Touch forehead (f) Touch forehead (g) Wave (h) Wave



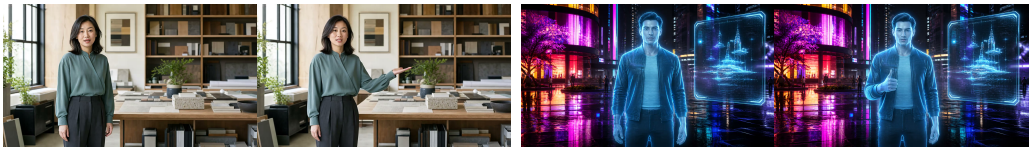
(i) Touch chin (j) Point upward



(k) Hand to chest (l) Clap



(m) Wave (n) Touch forehead



(o) Open palm forward (p) Thumbs up

Figure 11: Qualitative gallery of motion responses. Each subfigure shows the first frame on the left and the corresponding generated frame on the right. Portrait examples occupy the top two rows, followed by landscape examples in the bottom four rows, each subcaption identifies the action.

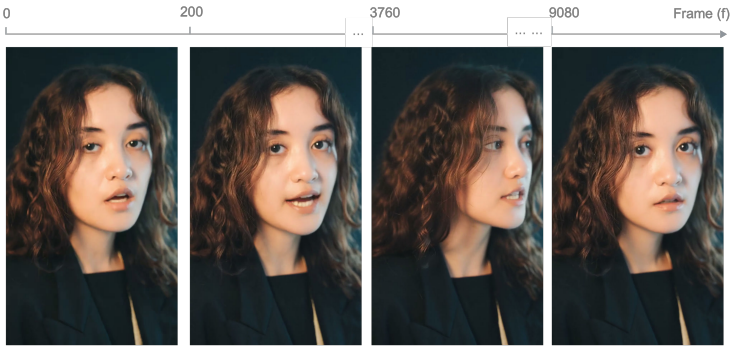


Figure 12: Qualitative results from a long-horizon streaming rollout using layer-specific attention sinks and a rolling KV cache. From left to right, we show generated frames at indices 0, 200, 3760, and 9080, corresponding to approximately 0, 12.5, 235, and 568 s at 16 fps.

A.4.5 Causal Initialization and Distillation Stability

We compare three initialization strategies for the causal adaptation described in Sec. 3.4 under the R2V setting. All variants use the same bidirectional R2V checkpoint, causal architecture, and downstream DMD configuration. The direct variant applies DMD without causal warm-up, whereas CFG distillation and Causal SFT regress to the guided teacher prediction and the data-derived flow target, respectively. The three variants yield closely matched distilled generators; CFG distillation obtains the highest ID-Sim, while Causal SFT obtains the highest IQA and ASE.

Table 4: Ablation of causal R2V generator initialization before asymmetric DMD. All variants are trained and evaluated without audio conditioning. "Direct" applies DMD to the bidirectional R2V checkpoint without causal warm-up, while CFG distillation and Causal SFT use alternative warm-up objectives.

Initialization	**ID-Sim↑**	**IQA↑**	**ASE↑**
Direct (DMD only)	0.798	0.892	5.129
CFG distillation warm-up	**0.809**	0.890	5.110
Causal SFT warm-up	0.808	**0.898**	**5.138**

We additionally compare Self-Forcing Huang et al. (2026) with and without ODE initialization in the official Wan T2V implementation. Removing ODE initialization produces broken frames in this control, whereas our DMD pipeline does not require ODE trajectory pairs. Direct distillation under noisy-cache inference remains stable, while the corresponding clean-cache setup collapses; causal initialization nevertheless provides a better starting point and is retained in our final pipeline.

A.4.6 Additional Ablations

Independent Patchification. We compare independent patchification modules for different token types with a single shared module. Both variants use Wan2.1-1.3B and are trained under the same R2V configuration described in Sec. 3.3.1 for 54,000 iterations on 16 H100 GPUs, without audio conditioning or causal adaptation. Independent patchification improves all three metrics, with the largest gain in ID-Sim from 0.738 to 0.771, suggesting that type-specific projections help the model distinguish reference, motion, and target tokens.

Table 5: Ablation of independent patchification in the bidirectional model. Both variants use Wan2.1-1.3B and are trained for R2V generation without audio conditioning or causal adaptation. Results are evaluated on SekoBench-Short.

Patchification	**ID-Sim↑**	**IQA↑**	**ASE↑**
Shared	0.738	0.817	4.916
Independent	**0.771**	**0.827**	**4.937**

Hybrid Local–Global Attention with Matched Attention Sinks. We compare the Hybrid and all-$W21$ configurations described in Sec. 4.5. Both use audio conditioning on SekoBench-Short; Hybrid uses the $[W2, W2, W2, W2, W21]$ schedule with $[S1, S1, S1, S1, S2]$ sinks, while the baseline uses all-$W21$ attention with sink size 2. Mean-OF is computed with UniMatch Xu et al. (2023); lower values indicate less frame-to-frame motion.

Table 6: Ablation of hybrid temporal attention and its corresponding attention-sink configuration for audio-conditioned causal generation. The Hybrid variant uses the $[W2, W2, W2, W2, W21]$ schedule with $[S1, S1, S1, S1, S2]$ sinks, whereas the baseline uses all-$W21$ causal attention with sink size 2.

Attention design	ID-Sim↑	Mean-OF↓
Hybrid + audio	**0.823**	0.437
All-$W21$ + audio	0.811	**0.528**

A.5 Bidirectional Avatar Teacher Implementation

This section provides the reference-conditioning and segment-continuation details of the teacher in Sec. 3.3. The visual backbone uses bidirectional attention within each segment; the causal audio-conditioning implementation is described separately in Appendix A.6.

A.5.1 Reference Injection and Sampling

We use two complementary reference-injection pathways. For low-level conditioning, following a design similar to OmniHuman-1.0 (Lin et al., 2025), we encode the reference image into VAE latents and concatenate its patched tokens with the noisy video tokens along the sequence dimension. Bidirectional attention can thus directly access facial structure, clothing, and background information. For high-level conditioning, CLIP image features (Radford et al., 2021) are injected through the image cross-attention pathway inherited from the Wan2.1 I2V model (Wan et al., 2025), retaining the pretrained pathway when adapting from first-frame conditioning to reference-image conditioning.

The reference-sampling location affects long-video behavior (Yang et al., 2025a; Jiang et al., 2025). Sampling within the target segment can encourage repeated copying of the reference into later segments; sampling outside it reduces repetition but can weaken identity or background consistency. During R2V fine-tuning, we randomly draw the reference from either the target segment or a nearby external segment. A learnable binary reference-state embedding is added to the reference tokens to indicate the source. At inference, the state can be set explicitly and alternated between the two modes to balance consistency and motion diversity.

A.5.2 Segment Continuation and Latent Processing

Motion context and encode-once processing. For continuation segments, motion frames from the preceding segment's tail provide temporal context, whereas the initial segment has no motion context. Motion tokens are concatenated with the current denoising-target tokens along the sequence dimension. During training, we set the number of motion frames to zero with probability 10%, allowing the same model to handle initial and continuation segments.

To avoid boundary artifacts in motion-frame latents, we construct a contiguous clip containing preceding context, motion, and target frames from the same source video and encode it in one 3D-VAE pass (Fig. 3(b)). We then discard the preceding-context latents and retain only the motion and target latents. Providing the motion frames with preceding temporal context during encoding reduces the mismatch with tail latents reused from a preceding generated segment. At inference, these denoised tail latents are directly reused as the next segment's motion context, avoiding the extra computation and reconstruction errors of an RGB decode–re-encode cycle.

Temporal positioning and independent patchification. For the temporal component of 3D RoPE, token positions follow the order motion, reference, and current denoising target. We use four independent patchification modules for reference latents, first-segment target latents, continuation target latents, and continuation motion latents. Their type-specific projections and biases provide implicit token-type embeddings, helping the transformer distinguish appearance conditioning, historical motion, and denoising targets. The independent-patchification ablation is reported in Appendix A.4.6.

A.6 AUDIO-VISUAL SYNCHRONIZATION DETAILS

Adapter architecture and timestep modulation. With the R2V visual backbone frozen, we train a lightweight audio adapter that projects frame-level speech features into compact tokens and injects them at selected transformer layers through residual cross-attention. The cross-attention module uses adaptive layer normalization: the diffusion-timestep embedding predicts the shift and scale of normalized hidden states and a gate for the cross-attention residual. This modulation allows the adapter to vary its conditioning across denoising steps, accommodating coarse motion guidance and fine lip synchronization.

Causal waveform windows. A bidirectional audio adapter would be compatible with the bidirectional visual backbone, but our implementation uses causal audio windows from the teacher stage onward to keep the audio-conditioning pathway compatible with streaming. Causal adaptation changes the adapter parameters, not this window design. For temporal latent $\mathbf{x}_t^i$, we extract a fixed-length waveform window $\mathbf{w}^i$ centered at the midpoint of its temporal support. Accessible audio is restricted to the latent's right boundary plus an optional look-ahead, which is set to zero in our implementation. Samples beyond this available prefix or outside the waveform boundaries are zero-padded before encoding. HuBERT (Guo & Liu, 2022) encodes each window, and its features are projected into aligned audio tokens A^i. Fig. 13 illustrates the window construction.

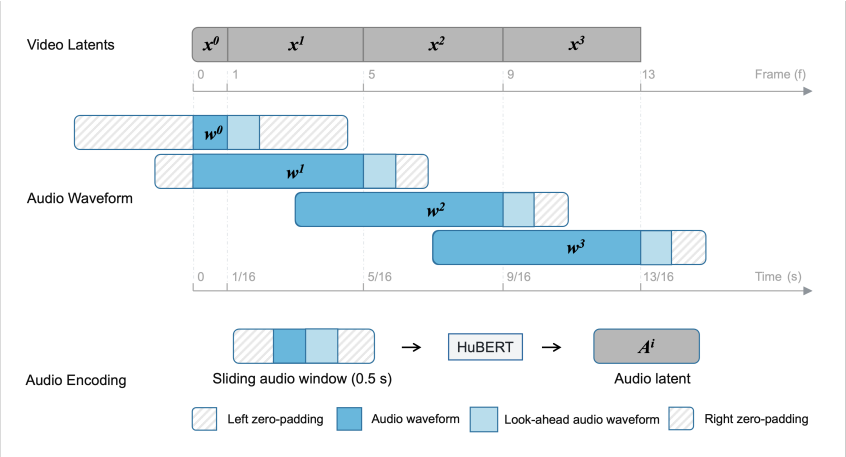


Figure 13: Causal sliding-window audio conditioning. The top row illustrates the $1 + 4n$ temporal compression pattern of the video VAE at 16 fps, where each video latent $\mathbf{x}_t^i$ is associated with a fixed-length waveform window $\mathbf{w}^i$ centered at the midpoint of its temporal support. Within this fixed-length window, waveform samples are accessible only up to the latent right boundary plus an optional look-ahead. The dark-blue region denotes currently available audio, the light-blue region denotes optional look-ahead audio, and the hatched regions denote zero-padding. With zero look-ahead, the light-blue region is absent, and samples beyond the latent right boundary or outside the waveform boundaries are zero-padded to preserve the fixed window length. Each waveform window is encoded by HuBERT (Guo & Liu, 2022) and then projected into the aligned audio tokens A^i. The 0.5-second window is used only for this schematic illustration, while a proportional visualization of the final 2-second configuration is provided in Fig. 14.

Temporal alignment and tensor configuration. Fig. 14 summarizes the causal audio–visual alignment used by the audio adapter. The four rows show the video-latent timeline, a fixed two-second audio window with zero look-ahead, the centered HuBERT feature span selected for each latent, and the resulting audio tokens injected into the matching video-latent token group. At 16 fps, each latent uses only audio available up to its right boundary, with unavailable samples zero-padded, so the conditioning remains causal. Audio features are projected independently for each latent and are applied through residual cross-attention only to the corresponding video tokens, excluding reference-image tokens.

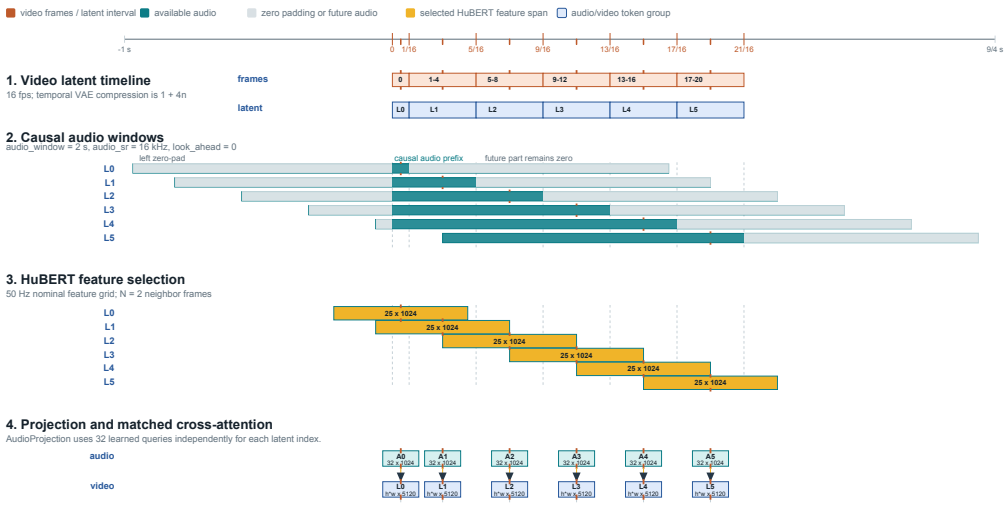


Figure 14: Temporal indexing and tensor shapes of the causal audio adapter. The video latent intervals follow the $1+4n$ temporal compression pattern at 16 fps. Each latent uses a 2-second audio window centered at its latent timestamp. With zero look-ahead, only the audio prefix available up to the latent right boundary is copied into the window, while the remaining portion is zero-padded. The selected HuBERT span contains 25 features of dimension 1024 for the configuration $N = 2$. Audio projection produces 32 tokens of dimension 1024 for each latent, which are matched to the corresponding video latent token group through residual cross-attention.

A.7 Audio Encoder Selection

The audio encoder serves as a frozen feature extractor that converts the input waveform into frame-level speech features, which the audio adapter then injects into the diffusion backbone. We compare four candidate encoders—Chinese HuBERT-large (Guo & Liu, 2022), WavLM-base-plus (Chen et al., 2022), wav2vec2-large-xlsr-53 (Grosman, 2021), and Unitalker-base (Fan et al., 2024)—under the same audio-adapter training setting. Chinese HuBERT-large achieves the best Sync-C score on both Chinese and English test sets. Qualitatively, it produces more accurate fine-grained mouth shapes and pronunciation details, particularly for Chinese speech. Based on this comparison, we adopt Chinese HuBERT-large in the final system and train only the audio projection and adapter layers.

B Attention Pattern Analysis

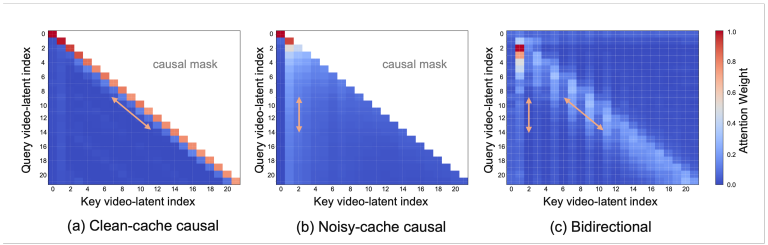


Figure 15: Temporal attention maps of clean-cache causal, noisy-cache causal, and bidirectional models, visualized at Block 29, Head 10. Similar tendencies are observed across the majority of examined attention heads.

We compare the temporal attention patterns of three model variants: clean-cache causal, noisy-cache causal, and bidirectional attention. All visualizations are extracted from Block 29, Head 10. The spatial attention weights are aggregated at the video-latent level, with the horizontal and vertical axes in Fig. 15 corresponding to key and query latent indices, respectively.

As shown in Fig. 15(a), the clean-cache causal model exhibits a pronounced diagonal-dominant pattern, indicating strong temporal locality between clean latents. Each query mainly attends to temporally corresponding or nearby cached latents, with substantially weaker responses to distant history. Such strong local dependency facilitates smooth temporal transitions between adjacent frames, but may also increase the risk of propagating local prediction errors over time. In our empirical evaluation, this tendency is associated with more noticeable error accumulation during long-horizon causal generation.

For the noisy-cache causal model in Fig. 15(b), the visualized latents are taken at the same diffusion timestep. Compared with clean-cache causal attention, the diagonal structure becomes substantially weaker, while multiple queries consistently attend to a subset of historical latents, resulting in a more vertically structured pattern. It suggests a reduced preference for strict temporal alignment and a stronger tendency to retrieve information from selected historical states.

The bidirectional model in Fig. 15(c) exhibits both local diagonal dependencies and broader nonlocal attention across the sequence. Notably, its attention pattern incorporates characteristics observed in both causal variants: the diagonal locality emphasized by clean-cache causal attention and the broader historical aggregation observed in noisy-cache causal attention.

These results suggest that the two causal formulations introduce different directional biases in temporal information aggregation. Clean-cache causal attention tends to overemphasize temporally adjacent states, whereas noisy-cache causal attention weakens temporal alignment and places relatively greater emphasis on selected historical states. In contrast, bidirectional attention provides a more balanced combination of these behaviors. Relative to the bidirectional model, either causal formulation therefore exhibits a preference, or degradation, along a particular temporal dependency pattern due to its restricted information flow. We note that Fig. 15 shows only a single attention head and should not be interpreted as the behavior of all heads. Nevertheless, similar tendencies are consistently observed in the majority of examined heads, and the selected visualization is representative of the dominant patterns of the corresponding models.

B.1 Additional Attention-Map Visualizations

To complement the representative single-head comparison in Fig. 15, we visualize additional attention heads at multiple transformer depths for the clean-cache causal, noisy-cache causal, and bidirectional models. We use the same samples and aggregation protocol as in the representative analysis above, so the patterns are directly comparable with the representative visualization.

B.2 Extended Related Work

B.2.1 Audio-Driven Talking Avatar Generation

Diffusion-based talking-avatar generation has substantially improved visual fidelity and motion diversity by formulating video synthesis as conditional denoising over latent representations Ma et al. (2024); Yang et al. (2024). Recent systems scale this paradigm with stronger video diffusion backbones and larger, more diverse training data, improving identity preservation, audio-motion alignment, and generalization across identities and motions Yang et al. (2024); Kong et al. (2026); Team et al. (2025); Yang et al. (2025a). Many of these models rely on bidirectional temporal modeling over predefined video clips, which supports globally coherent generation but is not naturally suited to open-ended interaction. Our work builds on this lineage by deriving a causal streaming avatar from a pretrained bidirectional model, rather than training causally from scratch.

B.2.2 Autoregressive and Causal Video Diffusion

Autoregressive diffusion models have attracted growing interest for video generation. Diffusion Forcing Chen et al. (2024) introduced a causal training paradigm in which different frames are assigned different noise levels, allowing the model to denoise partially observed trajectories while

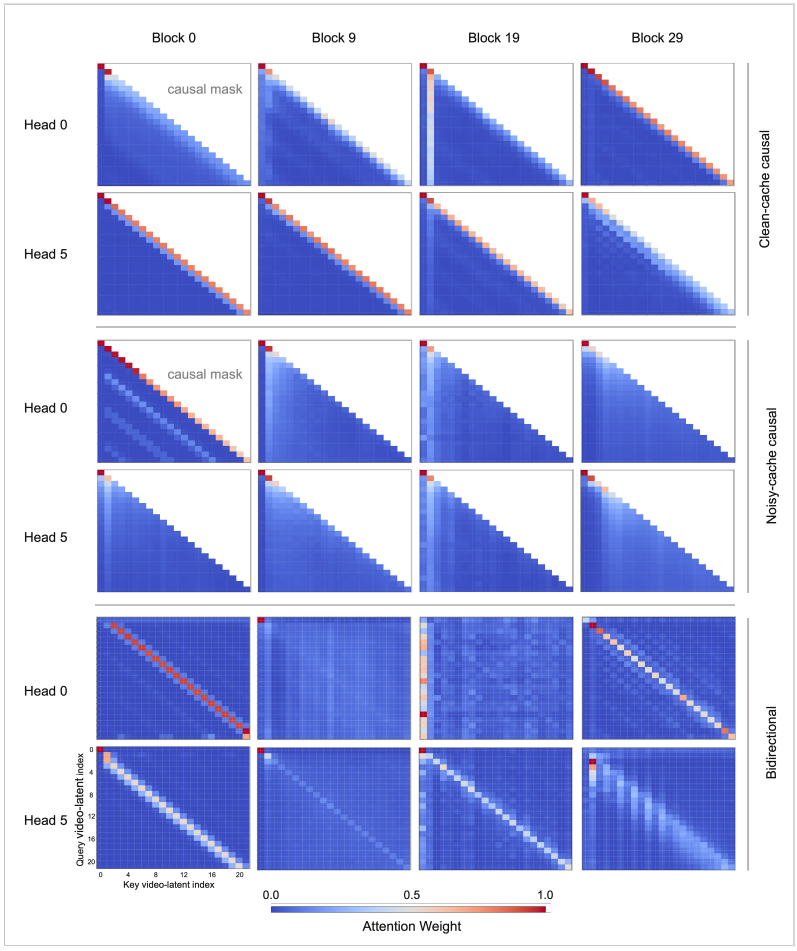


Figure 16: **Additional temporal attention maps across transformer blocks and attention heads.** The top, middle, and bottom groups correspond to the clean-cache causal, noisy-cache causal, and bidirectional models, respectively. Within each group, rows show Heads 0 and 5, and columns show Blocks 0, 9, 19, and 29. Spatial attention weights are aggregated at the video-latent level; the horizontal and vertical axes denote key and query video-latent indices, respectively. The blank upper-triangular regions in the two causal models reflect the causal mask. These results complement the representative Block-29, Head-10 visualization in Fig. 15 and show that its main qualitative tendencies recur across additional heads and network depths.

preserving temporal causality. AR-Diffusion Sun et al. (2025) further imposes a non-decreasing temporal noise schedule, so that earlier frames remain clearer than later ones and the space of timestep compositions is reduced. However, these generic timestep-sampling schemes are not fully aligned with the restricted configurations encountered in streaming inference. In clean-cache inference, cached past frames remain at $t{=}0$ while only the current frame is noisy; in noisy-cache inference, all frames in the active window share the same timestep t_k. Thus, deployment queries only a small subset of the joint timestep space $(t^1, t^2, \ldots, t^N)$. Independent per-frame timestep sampling consequently spends training capacity on configurations that are absent at deployment and exposes the model to cache/noise states that differ from its inference-time inputs. In contrast, target-aligned SFT schemes match the training configuration to the deployed cache and noise structure (see Sec. A.1). The corresponding schemes are Teacher Forcing SFT for clean-cache inference and Causal SFT for noisy-cache inference, thereby allocating the training budget more effectively.

CausVid Yin et al. (2025) is an early representative work that converts a pretrained bidirectional video diffusion transformer into a fast autoregressive generator through causal attention, ODE-trajectory-based student initialization, and asymmetric teacher-student distillation. However, its

DMD objective is optimized on states constructed by Diffusion Forcing rather than by the student's inference-time autoregressive rollout, creating a mismatch between the training and deployment distributions. Self-Forcing Huang et al. (2026) addresses this gap by performing autoregressive temporal rollout with KV caching during distillation, conditioning each frame on the model's previously generated frames rather than ground-truth context. Self-Forcing++ Cui et al. (2026) extends this paradigm beyond the teacher's short training horizon by applying distribution matching distillation to sampled windows from long self-generated rollouts, enabling generation up to 100 seconds. With increased training computation, it can generate videos longer than four minutes. Rolling Forcing Liu et al. (2026) extends autoregressive denoising from a single frame to a rolling multi-frame window with progressively increasing noise levels, enabling bidirectional refinement within the window to mitigate error accumulation. However, its window size is tied to the number of denoising steps, coupling the temporal receptive field with the sampling schedule and limiting design flexibility. Its distillation stage also relies on temporal autoregressive rollout during training.

In practice, self-forcing remains vulnerable to optimization instability and error accumulation over long horizons. LiveTalk Chern et al. (2025) reports that self-forcing produces flickering and black-frame artifacts under multimodal conditioning. Self-Resampling Guo et al. (2025) identifies progressive quality degradation and video collapse as characteristic failure modes of self-forcing on longer sequences, and proposes an end-to-end resampling strategy as an alternative.

Recent concurrent work further examines how teacher supervision should match the causal context available to an autoregressive student. Causal-rCM Zheng et al. (2026) combines teacher-forcing consistency-model initialization with self-forcing DMD refinement, retaining temporal autoregressive rollout in its final distillation stage. Context-Matched Distillation (CMD) Bandyopadhyay et al. (2026) instead uses a causal teacher and evaluates each target under the cached student-generated prefix that produced it; prefix corruption further stabilizes early training. These methods improve the alignment between supervision and the student's causal information boundary, but constructing the realized student prefix still requires on-policy temporal rollout.

The most closely related work to ours is LiveAvatar Huang et al. (2025), which also adopts a noisy-cache inference design for streaming avatar generation. Specifically, its History Corrupt strategy keeps generated-history KV states noisy, and timestep forcing aligns their noise level with that of the current block. LiveAvatar uses this cache formulation to improve long-horizon stability and to support timestep-forcing pipeline parallelism. Our contribution is therefore not noisy historical caching itself, but the execution property that follows when the same-noise context is combined with a fixed block-wise causal mask. LiveAvatar differs from our approach in two further aspects. First, its initial training stage follows diffusion forcing, where timesteps are sampled independently across temporal blocks. As discussed above, this sampling strategy is inefficient because it explores many timestep combinations that are not encountered during inference. Second, LiveAvatar still denoises one block at a time during self-forcing distillation and does not exploit equivalence across temporal execution partitions. Our method uses this equivalence to replace temporal rollout in DMD back-simulation with parallel prefill, avoiding the sequential training bottleneck while allowing the inference chunk size to be selected without retraining.

TalkingMachines Low & Wang (2025) also uses timestep-aligned noisy historical context for causal video generation. Our method differs in using the resulting partition invariance to support rollout-free distillation and flexible AR, MTP, and prefill inference.

B.2.3 Streaming Inference and Attention Sinks

StreamLLM Xiao et al. (2024) studies the deployment of autoregressive language models on infinite input streams with a bounded memory budget. It shows that naive window attention, which retains only the most recent KV states, can degrade once the initial tokens are evicted, because many attention heads consistently assign high attention weights to these early tokens regardless of their semantic content. The key idea is to retain a small set of initial attention-sink tokens together with a sliding window of recent tokens, enabling stable streaming without additional fine-tuning or unbounded KV-cache growth. This mechanism has since been adapted to streaming diffusion and multimodal generation in several forms. Rolling Forcing Liu et al. (2026) preserves the KV states of initial frames as global context anchors and adjusts their RoPE positions to avoid excessive positional drift. Rolling Sink Li et al. (2026) studies sink-based cache maintenance for open-ended autoregressive video diffusion and shows that it can extend generation beyond the training horizon.

LongLive Yang et al. (2025b) combines short-window attention with frame-level sinks and introduces KV re-caching for interactive prompt changes. LiveAvatar Huang et al. (2025) uses a rolling sink frame to preserve appearance cues from the initial reference frame during long-horizon generation. TalkingMachines Low & Wang (2025) adopts a sparse causal pattern that retains the current chunk, the preceding chunk, and the initial chunk as cached context. We apply the same attention-sink principle to causal diffusion avatars, where sink frames preserve identity and appearance cues while a sliding window tracks recent motion and audio context.