MEMORY FORCING: ATTENDABLE MID-HORIZON HISTORY FOR FEW-STEP STREAMING VIDEO

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive video diffusion realizes causal streaming generation without a bidirectional pass over the full clip. Those histories are stored as keys and values in a KV cache, which later denoising steps attend without re-encoding the past. To keep attention cost from growing with clip length, most current systems keep only a recent FIFO of fixed length and drop the oldest keys when the window is full, and we call this *mid-horizon forgetting*, where lost events sit between the start of the clip and the recent representations that the FIFO still holds. We present *Memory Forcing*, a few-step streaming method that focuses on those lost events rather than only the recent window. To this end, we introduce *Archive & Working Banks* so those events remain attendable, local motion stays in the recent window, and later queries can still retrieve a subject or scene after it leaves and returns. Further, to give those banks distinct places on a time axis whose indices otherwise grow outside the range seen in training, we introduce *Bank-aware RoPE* so later queries read the start, the recent representations, and mid-horizon events in the right place. Extensive experiments show that Memory Forcing outperforms existing methods on long-horizon clips and stays more stable as generation extends from 5s to 60s. To move past the public 1.3B ceiling on this few-step forcing line, we also introduce, to our knowledge, the first public 5B streaming model.

1 INTRODUCTION

Modern video diffusion models generate short clips with rich detail and coherent motion by denoising the full sequence with bidirectional attention. These offline generators set a high quality bar but cannot emit a frame until the clip is finished. Streaming applications instead need a causal generator with low latency and temporal consistency over long horizons. CausVid (Yin et al., 2025) and Self-Forcing (Huang et al., 2025) distill a bidirectional teacher into a few-step causal student and can produce frames on the fly. Their strictly causal prediction lets each frame inherit errors from its predecessors, so defects compound into drift. Rolling Forcing (Liu et al., 2025b) mitigates that in-window drift with rolling-window joint denoising, so frames that share a window stay aligned.

Streaming video generation conditions each new chunk on frames already produced. Those frames are stored as keys and values in a KV cache, which later denoising steps attend without re-encoding the past. This cache keeps attention cost from growing with clip length. Most current systems manage this cache as a recent FIFO of fixed length. New chunks are appended, and the oldest keys are dropped when the window is full. Local motion is preserved, but events that have left the window are gone. A subject that leaves and later returns, or a scene that cuts away and comes back, is no longer addressable. We call this *mid-horizon forgetting*, where lost events sit between the opening of the clip and the recent FIFO window, so later queries have no slot to retrieve them.

To keep those mid-horizon events attendable, we present *Memory Forcing*, a few-step streaming method that keeps rolling and DMD. The opening of a clip, the recent motion, and the events in between are different kinds of history. A single FIFO keeps only the recent keys, so intermediate events have nowhere to go. We introduce **Archive & Working Banks** and partition the KV cache into *sink*, *archive*, and *working* banks. New chunks enter working. When working is full, Similarity writes a chunk into archive, and a full archive drops the oldest chunk by FIFO. Mid-horizon events remain attendable, local motion stays in the recent window, and the cache does not grow.

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



Figure 1: Qualitative samples from Memory Forcing. Each row is one generated clip, frames in time order. Subjects, motion, and scene layout stay consistent across the clip.

Those banks must be placed on a time axis so later queries can tell the opening, the recent window, and mid-horizon events apart. Previous methods write keys with *Absolute* time τ^{abs}, the frame index at write time. On a long video those indices grow outside the trained range, so later queries read stored events in the wrong place. The same rule also fails a three-bank cache. Archive is older than the recent window, and the three roles need different places on the axis, which one growing index cannot give them. We introduce **Bank-aware RoPE**, store raw keys for the whole window, and re-assign a temporal index τ before every attention. The assigned indices stay in the trained range, the three banks remain distinguishable, and later queries read stored events in the right place.

We evaluate the same method at Wan2.1-T2V-1.3B and Wan2.2 5B (Wang et al., 2025a). 1.3B matches prior few-step models and covers the main table and ablations that turn off the partition or τ. To our knowledge, the 5B checkpoint is the first public 5B on this forcing line and holds up better on crowded multi-subject scenes. Experiments show Memory Forcing tops longer clips, drops least from 5s to 60s among methods reporting all four lengths, and still retrieves leave-and-return.

The contributions of this work are three. First, we introduce Archive & Working Banks, a sink / archive / working cache that keeps mid-horizon events attendable, with Similarity as the default write into archive. Second, we introduce Bank-aware RoPE, which stores raw keys for the whole window and re-assigns a temporal index before every attention. The three banks can then be told apart, and later queries read stored events in the right place. Third, experiments show Memory Forcing tops longer clips, drops least from 5s to 60s among methods reporting all four lengths, and still retrieves leave-and-return. To our knowledge, we also introduce the first public 5B.

2 Related Work

Bidirectional video diffusion. Video diffusion models denoise the full clip with bidirectional attention. Latent U-Nets such as SVD (Blattmann et al., 2023a) and VideoLDM (Blattmann et al., 2023b) set that recipe. DiT-scale generators set the quality bar: Sora (Brooks et al., 2024), MovieGen (Polyak et al., 2024), CogVideoX (Yang et al., 2025b), HunyuanVideo (Kong et al., 2024), and Wan (Wang et al., 2025a), but cannot emit a causal stream. Autoregressive models such as NOVA (Deng et al., 2025), SANA-Video (Chen et al., 2025), and InfinityStar (Liu et al., 2025a) emit frames in time order and can grow a clip without a bidirectional pass over the full sequence.

Forcing, from causal distillation to a two-bank cache. CausVid (Yin et al., 2025) distills a bidirectional teacher with DMD (Yin et al., 2024). Self-Forcing (Huang et al., 2025) trains on self-generated history. CausalForcing (Zhu et al., 2026) uses a causal teacher ODE. SkyReels-V2 (SkyReels Team, 2025) follows Diffusion Forcing (Chen et al., 2024). Rolling Forcing (Liu et al., 2025b) adds rolling-window repair, an attention sink, and a recent FIFO L_{tem}. LongLive (Yang

et al., 2025a) is scored in the same table. Rolling Forcing holds in-window drift down, and the KV cache is still a sink plus L_{tem}, so mid-horizon events fall out of the FIFO.

Long-range memory and position. RewardForcing (Lu et al., 2025), DeepForcing (Yi et al., 2025), and RollingSink (Li et al., 2026) strengthen only the sink, so intermediate history still leaves the window. Infinity-RoPE (Yesiltepe et al., 2025) and MemRoPE (Kim et al., 2026) rewrite time or fold the past into fixed tokens, with no archive. LongLive (Yang et al., 2025a) and AnchorForcing (Yang et al., 2026) recache on prompt changes, not leave-and-return. DummyForcing (Guo et al., 2026) and Forcing-KV (Ji et al., 2026) compress the window, not mid-horizon slots. These still leave mid-horizon events without a slot, so we use Archive & Working Banks to keep them attendable and Bank-aware RoPE so later queries read them in the right place.

3 METHOD

Memory Forcing keeps the Rolling Forcing (Liu et al., 2025b) training recipe, its rolling schedule, and its DMD objective. What changes is the KV cache later steps attend and the temporal index used once that cache is split by role. Sections 3.1–3.2 state the inherited generator, the DMD gradient, and rolling-window joint denoising. Sections 3.3–3.6 specify the sink, archive, and working banks, Bank-aware RoPE, Aligned cache insertion, and the same recipe on Wan2.2 5B.

3.1 PRELIMINARIES

Autoregressive video diffusion. Autoregressive video diffusion combines temporal causality with iterative denoising, so each new frame may look only at history that has already been produced. For a clip $x_{1:N} = (x_1, \ldots, x_N)$ the joint therefore factorizes as

$$p(x_{1:N}) = \prod_{i=1}^{N} p(x_i \mid x_{<i}), \tag{1}$$

and each factor is realized by denoising Gaussian noise conditioned on previously generated frames. Streaming systems generate these frames in small chunks, reuse a temporal key–value cache of recent latents and a small sink of the opening frames, and condition on the current text embedding (Yin et al., 2025; Yang et al., 2026; Huang et al., 2025; Lu et al., 2025). One step may emit a chunk rather than a singleton, still called a frame unless size matters.

Those conditionals depend on how history is constructed. Two choices are Teacher Forcing (TF) and Diffusion Forcing (DF) (Chen et al., 2024). Under TF the conditional for the ith frame at noise level t_j is $p(x_i^{t_j} \mid x_{<i}^{0})$, so every history frame is clean. Under DF it is $p(x_i^{t_j} \mid x_{<i}^{t\geq 0})$, so the histories are ground-truth frames corrupted with independent noise. Training sees ground-truth prefixes while inference feeds the model's own predictions, known as exposure bias (Schmidt, 2019). Self-Forcing (Huang et al., 2025) lets G_θ denoise each noisy frame given rolled-out $\hat{x}_{<i}^{0}$,

$$p(\hat{x}_i^{t_j} \mid \hat{x}_{<i}^{0}), \qquad \hat{x}_i^{t_{j-1}} = \Psi\big(G_\theta(\hat{x}_i^{t_j}, t_j, \hat{x}_{<i}^{0}), t_{j-1}\big), \tag{2}$$

with $\hat{x}_i^{t_T} \sim \mathcal{N}(0, I)$. Here Ψ injects Gaussian noise at the next lower noise level. Training and inference then see the same exposure, so the student denoises on top of its own errors rather than a ground-truth prefix. Relative to TF and DF, the train-time conditionals stay closer to a real rollout.

Distribution matching distillation. A self-rollout has no paired ground-truth frame, so a pixel denoising loss cannot be formed. CausVid (Yin et al., 2025) distills a bidirectional multi-step teacher into a few-step causal student with Distribution Matching Distillation (DMD) (Yin et al., 2024). DMD minimizes a reverse KL between the student's output distribution $p_{\text{gen},t}$ and a smoothed data distribution $p_{\text{data},t}$ at sampled times t, as the difference of two scores.

$$\begin{aligned} \nabla_\theta \mathcal{L}_{\text{DMD}} &= \mathbb{E}_t\big[\nabla_\theta \text{KL}(p_{\text{gen},t} \,\|\, p_{\text{data},t})\big] \\ &\approx -\mathbb{E}_t \int \big(s_{\text{data}}(\Psi(G_\theta(\epsilon), t), t) - s_{\text{gen}}(\Psi(G_\theta(\epsilon), t), t)\big) \frac{\mathrm{d}G_\theta(\epsilon)}{\mathrm{d}\theta}\, \mathrm{d}\epsilon, \end{aligned} \tag{3}$$

where Ψ is the forward process, ϵ is Gaussian noise, G_θ is the generator, and s_{data} and s_{gen} score the data and the generator's outputs. The data score comes from a frozen teacher, the student score from a separate network. Then equation 2 trains in few steps with this DMD and flow critic.

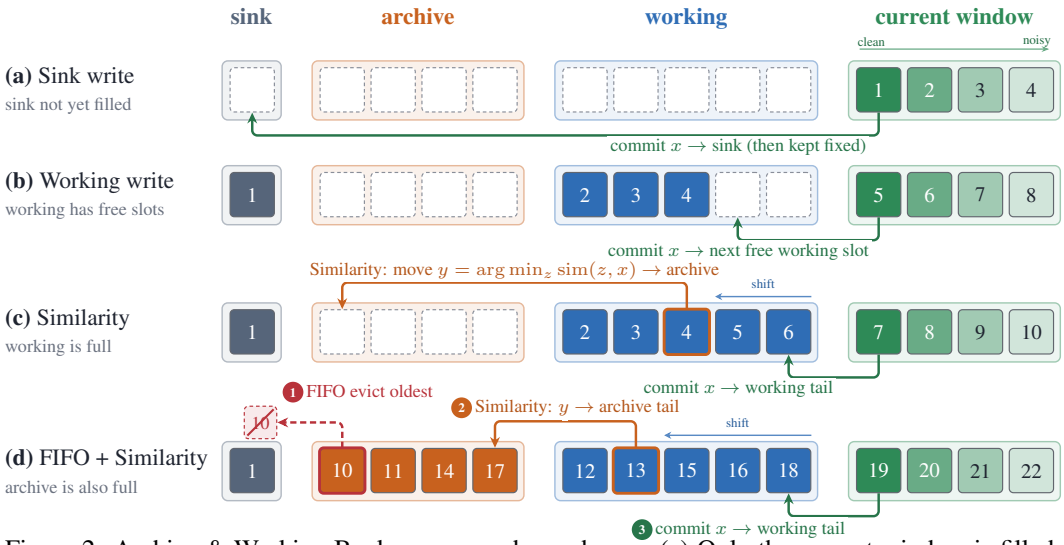

Figure 2: Archive & Working Banks, same cache each row. (a) Only the current window is filled. (b) The first chunk goes into sink and stays. (c) Later chunks enter working while slots remain. (d) When working is full, Similarity writes into archive. (e) FIFO then drops the oldest archive chunk.

3.2 Rolling-window joint denoising

Self-Forcing with DMD aligns train and inference history through equation 2. Each step still denoises one frame against a clean prefix, and that frame never attends back into its history. A local error is written into the next prefix and cannot be revised, so later frames inherit and compound it.

Rolling Forcing (Liu et al., 2025b) replaces that singleton with a window of T frames that share one forward pass and rise in noise from old to new. Neighbors revise a local error before commit, while each step emits one clean chunk. Starting at frame i, one roll updates that window by

$$p_\theta(\hat{x}^{t_{0:T-1}}_{i:i+T-1} \mid \hat{x}^{t_{1:T}}_{i:i+T-1}, \hat{x}^0_{<i}) = \Psi\big(G_\theta(\hat{x}^{t_{1:T}}_{i:i+T-1}, t_{1:T}, \hat{x}^0_{<i}), t_{0:T-1}\big). \tag{4}$$

The front is almost clean after several steps, and the back is still nearly Gaussian as it has just entered. The window then rolls forward and a new Gaussian chunk is appended. Causality is relaxed inside the window, so frames stay in order. Errors can be revised while noisy, so less enters it.

3.3 Archive & Working Banks

Rolling Forcing keeps a global sink and a recent FIFO. The sink anchors palette and exposure, while the FIFO carries local motion for recent chunks. After a chunk leaves the FIFO it is in neither, so a person who leaves and later returns, or a scene that cuts away and comes back, is no longer addressable. The rolling window can still repair neighbors in one pass, but it does not store what happened in between, the mid-horizon gap we change. We close that gap inside the KV cache the rolling model already writes. A few-step generator already commits clean keys as it rolls, so the question is not whether to cache but how those keys are kept after they have left the recent window. We do not add a side store, and we do not cut the first frames into a longer prefix that pretends mid-horizon content is still nearby.

Algorithm 1 Archive and Working Step

Require: $C = [\text{sink} \mid \text{archive} \mid \text{working}]$
Require: clean chunk x and query window Q
if sink is not full **then**
  write $\mathrm{KV}(x)$ to sink
else
  if working is full **then**
    $y \leftarrow \arg\min_{z \in \text{working}} \mathrm{sim}(z, x)$ ▷ equation 5
    move y from working to archive
    if archive is full **then**
      FIFO-evict the oldest archive chunk
    end if
  end if
  append $\mathrm{KV}(x)$ to the working tail
end if
$K_{\text{ctx}} \leftarrow [\text{sink}; \text{archive}; \text{working}; \text{current}]$
apply RoPE ▷ §3.4
return $\mathrm{Attend}(Q, K_{\text{ctx}})$

As shown in Figure 2, the cache is kept as $C = [\text{sink} \mid \text{archive} \mid \text{working}]$ and split into three banks by role, not by a time cut of the first frames. Sink still holds the opening chunk and stays fixed, so the origin does not drift. Working holds recent history so the tail stays continuous. Archive holds intermediate events that working later had to give up, so mid-horizon stays diverse.

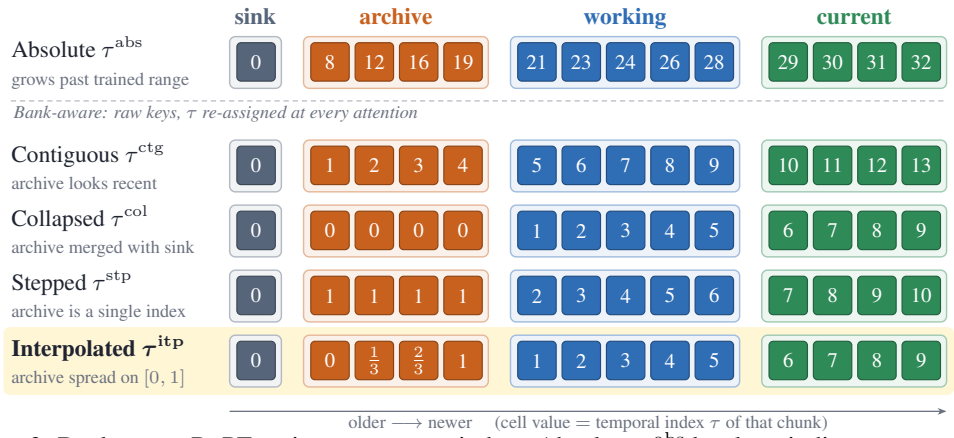

Figure 3: Bank-aware RoPE assigns τ on one window. Absolute τ^{abs} lets later indices grow past the trained range. Contiguous τ^{ctg} numbers from 0. Collapsed τ^{col} puts sink and archive at 0. Stepped τ^{stp} puts sink at 0, archive at 1, working from 2. Interpolated τ^{itp} places archive in $(0, 1)$.

A new clean chunk is written into working, except while the sink is still being filled, and a full sink does not move. The working tail keeps the newest keys next to the next frames, which local motion needs. If working is full, dropping an arbitrary recent chunk would break that continuity, and archive should not collect near-duplicates of x. We therefore move

$$y = \arg \min_{z \in \text{working}} \text{sim}(z, x) \tag{5}$$

from working into archive, where sim compares working keys to x. That promotion makes archive a diverse mid-horizon store. If archive is full, FIFO drops the oldest chunk to bound the cache, and the new chunk sits at the working tail. Algorithm 1 does this write and returns $\text{Attend}(Q, K_{\text{ctx}})$.

3.4 Bank-aware RoPE

Video transformers encode order with three-dimensional rotary embeddings (Su et al., 2024). We retarget the time axis. A query at index τ and a key at index τ' are rotated by $\tilde{q} = R(\tau)\, q$, $\tilde{k} = R(\tau')\, k$, where R is the rotary block, and $\tilde{q}^\top \tilde{k} = q^\top R(\tau - \tau')\, k$ depends only on the relative offset, so attention can learn shift-invariant patterns. In the usual cache rotation is applied at write time, so the stored key is already $\tilde{k} = R(\tau)\, k$. Later steps reuse that key and do not apply a new τ. Once written, the phase stays fixed until eviction and the key stays bound to the write-time index. Values stay unrotated, so moving that key later means storing it raw and rotating it at attention time.

Rolling Forcing makes one exception for the sink: those keys are stored raw and rotated only when replayed just before L_{tem}, and the rest still follows write-time rotation. Every other key is written with **Absolute** time τ^{abs}. Past the training horizon these τ^{abs} indices grow without bound and leave the trained range, so later keys attend under untrained phases and cannot be moved, and only the sink can be placed again. That binding is worse after three banks: archive keys are older than L_{tem}, so their τ^{abs} indices drift farther, and re-encoding only the sink leaves archive on stale indices. They also need a mid-horizon band that τ^{abs} cannot provide, so that RoPE cannot be inherited as is.

Therefore, we store raw keys for the whole window and re-assign τ at every attention. Relative time encodings are prior tools (Yesiltepe et al., 2025), and attaching them to the three-bank cache is the contribution. We still have to decide how sink, archive, working, and current sit on the axis. As shown in Figure 3, we design the four assignments (N_a and N_w count archive and working frames).

- **Contiguous** τ^{ctg}. The aim is to keep every index inside the trained range as one relative clip. We concatenate sink, archive, working, and current in time order, number them from the left, and fix the sink at 0 with no gap, as if the three banks were one FIFO.

$$\tau^{\text{ctg}} = (0, 1, \ldots, N{-}1). \tag{6}$$

  The indices never leave the trained range and the rule is cheap to apply. The three banks then share one scale, so archive sits next to the sink and looks like a moment ago.

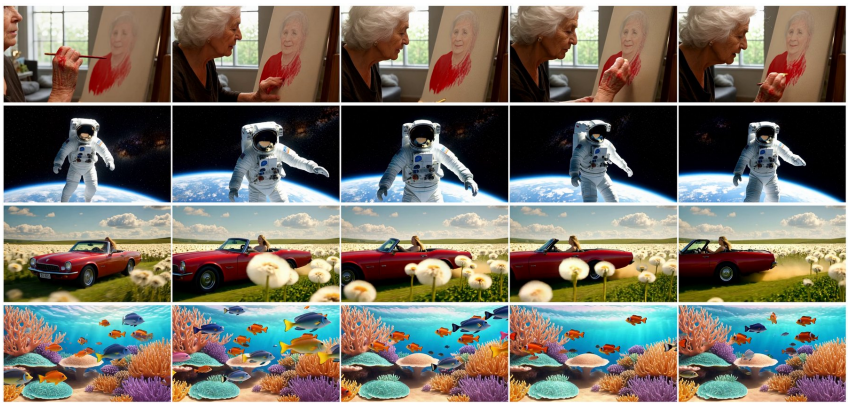

Figure 4: Further qualitative samples from Memory Forcing. Each row is one generated clip, frames in time order. Fine action and wide scenes keep identity and layout across the clip.

- **Collapsed** τ^{col}. The aim is to pin the opening shot and park mid-horizon keys on that origin. Every sink and archive token is given index 0, working starts at 1 and increases by one frame, and current continues after working. Archive has no internal order on the axis.

$$\begin{cases} \tau^{\mathrm{col}}(\mathrm{Sink}) = 0, \\ \tau^{\mathrm{col}}(\mathrm{Archive}_a) = 0 \quad (a = 1, \ldots, N_a), \\ \tau^{\mathrm{col}}(\mathrm{Working}_w) = w \quad (w = 1, \ldots, N_w). \end{cases} \tag{7}$$

  The origin stays fixed, working still has a local order, and all indices remain small. A mid-horizon event cannot be told from the opening shot because both sit at 0.

- **Stepped** τ^{stp}. The aim is to mark three timescales without growing indices. The sink sits at 0, archive at 1, and working at $2, 3, \ldots$, with current after working. Neighboring banks differ by one and do not share a slot, but archive is still a single index.

$$\begin{cases} \tau^{\mathrm{stp}}(\mathrm{Sink}) = 0, \\ \tau^{\mathrm{stp}}(\mathrm{Archive}_a) = 1 \quad (a = 1, \ldots, N_a), \\ \tau^{\mathrm{stp}}(\mathrm{Working}_w) = w{+}1 \quad (w = 1, \ldots, N_w). \end{cases} \tag{8}$$

  The banks no longer collide and the indices stay small. Archive is still a single point, so frames inside archive cannot be ordered on the axis.

- **Interpolated** τ^{itp}. The aim is to put the sink at 0, working near the present, and archive in between with its frames distinguishable. The N_a archive frames are spread evenly on $[0, 1]$, so neighboring archive keys differ by $1/(N_a{-}1)$ and collapse to 0 if $N_a{=}1$. Working increases from 1 as a near-present chain, and current follows working.

$$\begin{cases} \tau^{\mathrm{itp}}(\mathrm{Sink}) = 0, \\ \tau^{\mathrm{itp}}(\mathrm{Archive}_a) = \dfrac{a{-}1}{N_a{-}1} \quad (a = 1, \ldots, N_a,\ 0 \text{ if } N_a{=}1), \\ \tau^{\mathrm{itp}}(\mathrm{Working}_w) = w \quad (w = 1, \ldots, N_w). \end{cases} \tag{9}$$

  The sink stays at the origin, working stays near the present, and archive occupies the mid-horizon band with its frames distinguishable.

3.5 Training

We keep Rolling Forcing's distillation loop, with DMD and a flow critic on non-overlapping windows (§3.1, §3.2). A rolling-only rollout assembles the clean video from frames that left different noise levels, so the DMD clip is uneven in clarity. Training mixes rolling and self-forcing windows with equal probability. Self-forcing windows regularize camera motion, and inference uses the rolling window alone. Algorithm 1 now selects RoPE and appends a key, so each step practices the three-bank write and Bank-aware RoPE, and a later clean chunk uses that same write.

If that insertion attends only the tokens already in the cache, the stored keys are formed under a different context than denoising. We therefore use **Aligned cache insertion**. The insertion forward looks at the sink, the extracted archive and working banks, and the current window, the same set that a denoising step attends. Keys that enter the cache then match the attention used for the loss.

3.6 Scaling to 5B

Public few-step autoregressive long-video models stop at Wan2.1-T2V-1.3B (Wang et al., 2025a), so we attach the same recipe to Wan2.2 5B (Wang et al., 2025a) to test whether the three-bank cache and Bank-aware RoPE scale on a stronger backbone. The banks, Interpolated τ^{itp}, mixed training, Aligned cache insertion, and DMD stay as in §3.3–3.5. The 5B cut, cache, and attention budget use that backbone's tokens per frame, and both scales are primary results in §4.

Wan2.2 5B begins as a bidirectional backbone. A forcing student is causal, so we first convert that backbone with the first two stages of CausalForcing (Zhu et al., 2026), **autoregressive diffusion training** and **causal ODE initialization**. The first stage trains a multi-step causal teacher under teacher forcing, so frame i at noise t sees a clean prefix $x^0_{<i}$ as in §3.1. The resulting PF-ODE is injective at the frame level. The second stage samples trajectories from that teacher and trains the few-step student to regress the clean frame from the noisy state on the same prefix,

$$\theta^* = \arg\min_\theta \mathbb{E}\left\|G_\theta(x_i^t, x^0_{<i}, t) - x_i^0\right\|_2^2, \tag{10}$$

where (x_i^t, x_i^0) lie on the teacher's PF-ODE. After these two stages the 5B student is causal. We then train it with the Memory Forcing recipe of §3.3–3.5.

Both stages supervise a next-frame map given clean history, so clip motion sets the target size. Low-motion clips make that map nearly the identity, so AR conversion is easy but the student copies the last frame. High-motion clips enlarge the residual and ODE-target variance, so the same regression is hard to fit. We therefore filter Koala-36M (Wang et al., 2025b) on quality and aesthetics, and mix low, mid, and high motion at 4:3:3, yielding $280{,}000$ clips for AR diffusion training and $36{,}000$ for causal ODE initialization, and the 5B result in §4 supports this mix and that the recipe scales.

4 Experiments

We report Memory Forcing at Wan2.1-T2V-1.3B and at Wan2.2 5B as two primary results. Ablations that isolate the partition, τ, or eviction stay on 1.3B.

4.1 Implementation Details

Model. We implement Memory Forcing with Wan2.1-T2V-1.3B (Wang et al., 2025a) as the default 5s, 16 FPS, 832×480 base, and with Wan2.2-TI2V-5B (Wang et al., 2025a) as a second primary result. We adopt Causal Forcing (Zhu et al., 2026) weights from two stages, autoregressive diffusion training and causal ODE initialization. Text prompts are drawn from the filtered and LLM-augmented VidProM (Wang & Yang, 2024) dataset. We set $T{=}4$, 3 latent frames per chunk, and denoising steps $[1000, 750, 500, 250]$. Training uses $1{,}000$ steps, batch size 32 on 32 A800 GPUs, and a window of 33 latent frames. The KV cache is partitioned into sink, archive, and working banks of 3, 6, and 12 frames. We use Bank-aware RoPE with default Interpolated τ^{itp}, and AdamW for both the generator G_θ (learning rate 1.5×10^{-6}) and the fake score s_{gen} (learning rate 4.0×10^{-7}). The generator is updated every 5 steps of fake score updates.

Evaluation. We evaluate on four prompt suites: official VBench (Huang et al., 2024) and MovieGen (Polyak et al., 2024), plus two harder sets, 20 VBench prompts as VBench-Complex and 30 LLM prompts as Gen-Complex. These fifty prompts emphasize multi-subject scenes, person-to-person interaction, and other mid-horizon difficulties. We generate each prompt at four durations, 5s (21 latent frames), 15s (63), 30s (120), and 60s (240), all at 16 FPS and 832×480. We score videos with the VBench toolkit, reporting Total, Quality, and Semantic on official VBench and Quality only on MovieGen, VBench-Complex, and Gen-Complex.

Table 1: Comparison on official VBench Total and Quality on VBench-Complex, Gen-Complex, and MovieGen at 5s, 15s, 30s, and 60s. $CF\ ODE$ and $SF\ ODE$ are the CausalForcing and Self-Forcing pretraining. Red is best and blue is second-best in each column.

Method	VBench (Total)					VBench-Complex (Quality)					Gen-Complex (Quality)					MovieGen (Quality)				
	5s	15s	30s	60s	$\Delta_{\downarrow}^{5s\to60s}$	5s	15s	30s	60s	$\Delta_{\downarrow}^{5s\to60s}$	5s	15s	30s	60s	$\Delta_{\downarrow}^{5s\to60s}$	5s	15s	30s	60s	$\Delta_{\downarrow}^{5s\to60s}$
CausVid	82.99	—	—	—	—	80.83	—	—	—	—	80.35	—	—	—	—	79.62	—	—	—	—
NOVA	78.54	74.78	—	—	—	73.32	69.18	—	—	—	74.86	69.41	—	—	—	74.08	68.86	—	—	—
SkyReels-V2	82.65	80.11	—	—	—	81.08	75.07	—	—	—	78.83	71.93	—	—	—	79.88	75.35	—	—	—
Self-Forcing	84.09	—	—	—	—	81.37	—	—	—	—	79.99	—	—	—	—	81.09	—	—	—	—
LongLive	83.23	82.95	82.10	82.21	-1.02	79.73	80.01	79.07	78.81	-0.92	81.51	80.59	79.98	80.36	-1.15	80.15	79.79	79.20	79.13	-1.02
LongLive $_{inf}$	82.97	82.84	82.43	81.77	-1.20	79.45	80.70	80.45	80.37	0.92	81.57	80.81	80.52	79.61	-1.96	80.14	79.93	79.52	78.82	-1.32
SANA-Video	83.18	83.17	82.80	82.22	-0.96	80.18	78.63	78.44	77.29	-2.89	83.27	80.37	80.46	78.59	-4.68	80.91	80.31	79.22	78.00	-2.91
RollingForcing $_{SF\ ODE}$	83.63	82.82	82.04	81.64	-1.99	80.23	80.83	79.00	78.59	-1.64	81.48	78.83	77.77	78.36	-3.12	80.21	79.07	78.61	77.85	-2.36
RollingForcing $_{CF\ ODE}$	83.43	80.33	78.61	77.24	-6.19	82.38	81.16	80.51	79.12	-3.26	80.18	77.95	77.88	76.75	-3.43	82.87	80.91	79.64	78.68	-4.19
InfinityStar	84.03	75.94	66.76	61.47	-22.56	80.34	71.45	62.31	58.86	-21.48	79.71	72.59	63.36	59.10	-20.61	80.03	70.75	62.19	60.19	-19.84
Infinity-RoPE $_{SF\ ODE}$	83.78	82.78	81.58	79.99	-3.79	81.04	80.59	79.74	79.18	-1.86	81.15	79.44	78.55	77.67	-3.48	81.09	79.70	78.65	77.07	-4.02
Infinity-RoPE $_{CF\ ODE}$	84.08	82.29	80.71	79.51	-4.57	81.65	79.29	77.63	77.40	-4.25	80.13	77.95	77.22	76.88	-3.25	82.42	80.30	79.21	78.11	-4.31
RewardForcing	84.21	83.63	83.03	82.06	-2.15	82.06	81.55	81.20	80.73	-1.33	81.26	79.77	79.47	79.36	-1.90	81.56	80.96	80.63	80.19	-1.37
DeepForcing $_{SF\ ODE}$	84.08	83.18	82.37	82.37	-1.71	81.05	80.54	80.29	79.89	-1.16	80.00	78.81	78.50	77.88	-2.12	81.05	79.93	78.57	78.17	-2.88
DeepForcing $_{CF\ ODE}$	84.05	82.42	80.55	80.24	-3.81	81.75	79.90	77.28	77.22	-4.53	80.39	78.02	75.70	75.44	-4.95	82.36	80.05	77.87	76.71	-5.65
CausalForcing	83.90	—	—	—	—	81.84	—	—	—	—	80.56	—	—	—	—	82.36	—	—	—	—
AnchorForcing	84.16	83.26	82.76	81.73	-2.43	80.92	81.19	80.74	80.41	-0.51	80.39	78.53	78.45	78.15	-2.24	81.30	80.97	80.47	79.81	-1.49
Memory Forcing	83.88	83.51	83.58	83.20	-0.68	82.41	82.13	81.89	81.99	-0.42	81.07	81.54	80.48	79.92	-1.15	82.28	81.55	80.98	80.66	-1.62
Memory Forcing (5B)	84.24	83.83	83.11	81.93	-2.31	81.34	80.69	80.43	79.77	-1.57	80.16	79.87	80.15	79.19	-0.97	81.34	80.63	79.91	79.16	-2.18

Table 2: Ablations on Wan2.1-T2V-1.3B (Wang et al., 2025a) scored on official VBench (Huang et al., 2024) at 30s and 60s. Every panel reports Total/Quality/Semantic at 16 FPS and 832×480.

(a) Memory banks.

Memory	Total	Quality	Semantic
Official	82.04	82.64	79.61
Causal-Forcing checkpoint	78.61	81.30	67.85
Working	82.41	83.46	78.22
Archive + working	83.58	84.57	79.63

(b) Bank-aware RoPE.

RoPE	Total	Quality	Semantic
Contiguous τ^{ctg}	81.31	82.37	77.10
Stepped τ^{stp}	82.56	83.37	79.29
Collapsed τ^{col}	81.50	82.41	77.87
Interpolated τ^{itp}	83.58	84.57	79.63

(c) Similarity eviction.

Write	Total	Quality	Semantic
FIFO (30s)	83.39	84.59	78.59
Similarity (30s)	83.58	84.57	79.63
FIFO (60s)	82.96	83.94	79.05
Similarity (60s)	83.20	84.11	79.57

4.2 Comparisons

Table 1 compares Memory Forcing with recent open streaming and long-video models on the four prompt suites and four durations in §4.1. The grid is meant to be read across settings, not as a single 5s contest. Where a method is publicly released on more than one initialization, we report both the Self-Forcing and CausalForcing weights so that a gap is not an artifact of the starting checkpoint. Although Memory Forcing is not the strongest row at 5s, it stays at the top of the table on the longer clips, and the drop from 5s to 60s is the smallest among methods that report all four lengths. That pattern is meant to show that mid-horizon history remains usable after the short window.

4.3 Ablations

Memory banks. Table 2a changes only the cache: official Rolling Forcing, a Causal-Forcing checkpoint, a working bank alone, and archive plus working. Archive plus working is the strongest row. Working alone is lower on Total, Quality, and Semantic, so the extra bank is a gain. Official Rolling Forcing sits below it, and the Causal-Forcing checkpoint is lower still, especially on Semantic. A dedicated mid-horizon bank therefore keeps events that a flat window drops.

Bank-aware RoPE. Table 2b freezes the three-bank cache and swaps only the index map: Contiguous τ^{ctg}, Stepped τ^{stp}, Collapsed τ^{col}, and default Interpolated τ^{itp}. Interpolated τ^{itp} is the strongest row. Stepped is second. Contiguous and Collapsed fall further, especially on Semantic. Archive history has to sit on $(0, 1)$ with frames that remain distinguishable; packing it with the present, collapsing it onto the sink, or stepping the whole bank onto one integer all weaken retrieval.

Similarity eviction. Table 2c freezes the three-bank cache and τ^{itp} and changes only which working chunk is written into archive. Similarity is the training default. FIFO is used only at inference, where it moves the oldest working chunk into archive. Similarity is the stronger of the two rows on Total, Quality, and Semantic. FIFO turns archive into a second recency queue, while Similarity moves the working chunk least like the incoming event so mid-horizon slots stay diverse.

4.4 Qualitative results

We use qualitative frames to show Memory Forcing on its own samples, then in a 1.3B comparison, and then at Wan2.2 5B on harder scenes. Figure 4 shows Memory Forcing clips on diverse prompts, with identity and layout held across each clip. Figure 5 compares Memory Forcing at 1.3B with four recent open-source models. Frames are aligned across time on the same prompt. The baselines drift

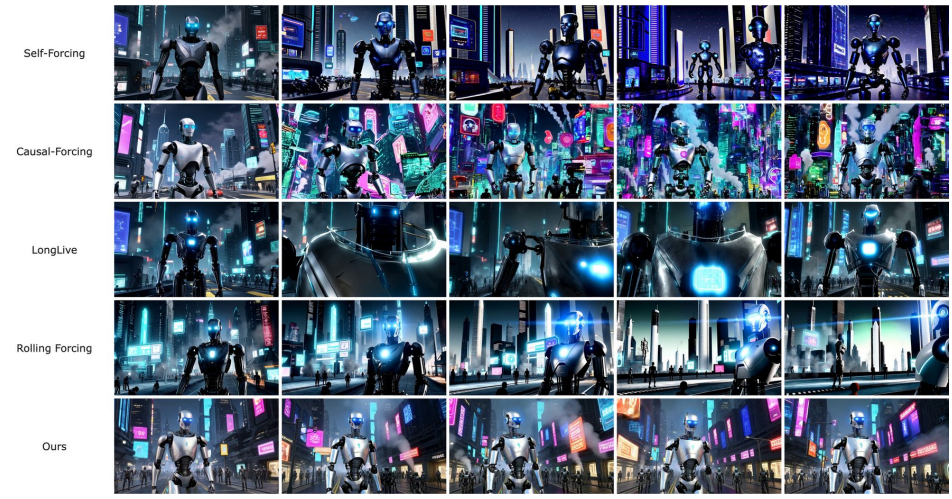

Figure 5: Qualitative comparison at 1.3B against four recent open-source models. Memory Forcing keeps the subject's identity and the scene layout, while the baselines drift or collapse.

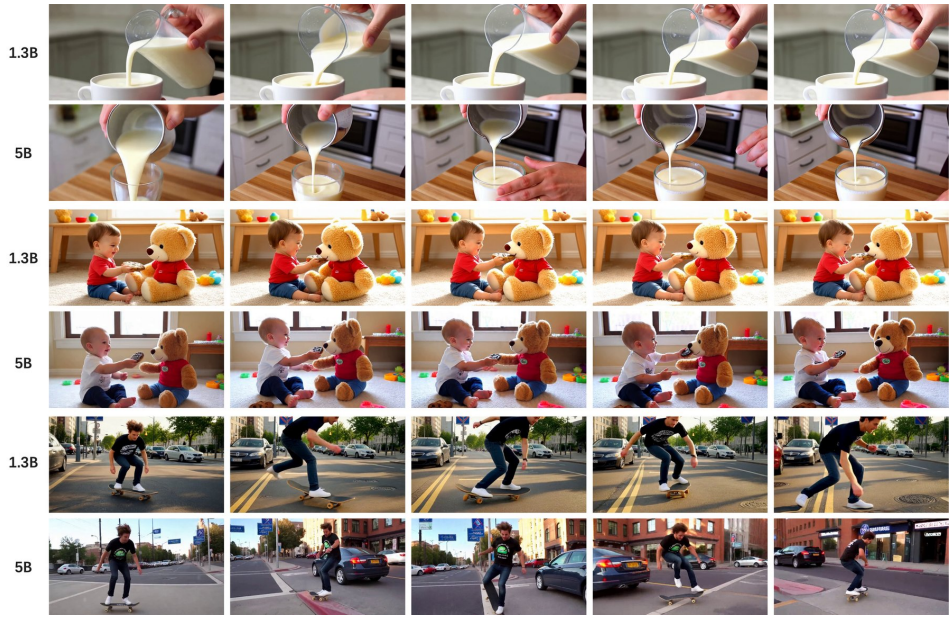

Figure 6: Memory Forcing at Wan2.1-T2V-1.3B and Wan2.2 5B on the same prompts. The 5B run is more physically plausible, more realistic, and more dynamic.

or collapse the subject's identity and the scene layout. Memory Forcing keeps both, which is the point of keeping an archive. Figure 6 then shows the same method at Wan2.2 5B. On crowded and multi-subject scenes the 5B run keeps objects and interactions that the 1.3B rollout starts to lose.

5 Conclusion

We present Memory Forcing, a few-step streaming method that addresses mid-horizon forgetting so later queries can still retrieve events that have left the window. To give intermediate history a place a single FIFO cannot hold, we introduce Archive & Working Banks, which keep mid-horizon events attendable without growing the cache. To give those banks distinct places on a time axis, we introduce Bank-aware RoPE, which lets later queries read stored events in the right place. The same method is a primary result at Wan2.1-T2V-1.3B and Wan2.2 5B, and to our knowledge the first public 5B on this forcing line. Extensive experiments show that Memory Forcing stays at the top of the table on the longer clips, that performance remains the most stable as inference length grows, and that qualitative frames still retrieve a subject or scene after it leaves and returns.

AI USE STATEMENT

In this work, we used generative AI tools to generate a synthetic evaluation prompt set and to implement parts of the method. We have not used generative AI tools to develop theoretical models, formulate mathematical claims, propose or refine hypotheses, design experiments, assist with translation, clean or reformat datasets, support qualitative analysis, or interpret results, and the writing of proofs is not applicable to this work. Additionally, we used generative AI tools to edit the manuscript for readability and to edit training and inference code. We have reviewed all AI-assisted work. The thirty Gen-Complex prompts were checked by the authors before scoring, and LLM-edited code was run and checked against the intended cache, write, and RoPE behavior. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies few-step streaming video generation. We do not recruit human subjects or collect a new personal-data corpus. Training text comes from public VidProM, and evaluation uses public VBench and MovieGen prompts together with a small author-checked set. Like other text-to-video models, Memory Forcing can be used to fabricate or extend footage of people and events, and we treat that misuse risk as the main ethical concern.

REPRODUCIBILITY STATEMENT

We have already described the training and inference details in the main text. Code and model weights will be released after the paper is accepted.

REFERENCES

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *CoRR*, abs/2311.15127, 2023a.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023b.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. OpenAI technical report, 2024.

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

Junsong Chen, Yuyang Zhao, Jincheng Yu, Ruihang Chu, Junyu Chen, Shuai Yang, Xianbang Wang, Yicheng Pan, Daquan Zhou, Huan Ling, Haozhe Liu, Hongwei Yi, Hao Zhang, Muyang Li, Yukang Chen, Han Cai, Sanja Fidler, Ping Luo, Song Han, and Enze Xie. SANA-Video: Efficient video generation with block linear diffusion transformer. *CoRR*, abs/2509.24695, 2025.

Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. In *International Conference on Learning Representations*, 2025.

Hang Guo, Zhaoyang Jia, Jiahao Li, Bin Li, Yuanhao Cai, Jiangshan Wang, Yawei Li, and Yan Lu. Efficient autoregressive video diffusion with dummy head. *CoRR*, abs/2601.20499, 2026.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Yicheng Ji, Zhizhou Zhong, Jun Zhang, Qin Yang, XiTai Jin, Ying Qin, Wenhan Luo, Shuiyang Mao, Wei Liu, and Huan Li. Forcing-KV: Hybrid KV cache compression for efficient autoregressive video diffusion models. *CoRR*, abs/2605.09681, 2026.

Youngrae Kim, Qixin Hu, C.-C. Jay Kuo, and Peter A. Beerel. MemRoPE: Training-free infinite video generation via evolving memory tokens. *CoRR*, abs/2603.12513, 2026.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. HunyuanVideo: A systematic framework for large video generative models. *CoRR*, abs/2412.03603, 2024.

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

Jinlai Liu, Jian Han, Bin Yan, Hui Wu, Fengda Zhu, Xing Wang, Yi Jiang, Bingyue Peng, and Zehuan Yuan. InfinityStar: Unified spacetime autoregressive modeling for visual generation. In *Advances in Neural Information Processing Systems*, 2025a.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *CoRR*, abs/2509.25161, 2025b.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, Yujun Shen, and Min Zhang. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. *CoRR*, abs/2512.04678, 2025.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *CoRR*, abs/2410.13720, 2024.

Florian Schmidt. Generalization in generation: A closer look at exposure bias. *CoRR*, abs/1910.00292, 2019.

SkyReels Team. SkyReels-V2: Infinite-length film generative model. *CoRR*, abs/2504.13074, 2025.

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

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. *CoRR*, abs/2503.20314, 2025a.

Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, Fei Yang, Pengfei Wan, and Di Zhang. Koala-36M: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Wenhao Wang and Yi Yang. VidProM: A million-scale real prompt-gallery dataset for text-to-video diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 65618–65642, 2024.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. LongLive: Real-time interactive long video generation. *CoRR*, abs/2509.22622, 2025a.

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

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025b.

Hidir Yesiltepe, Tuna Han Salih Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-RoPE: Action-controllable infinite video generation emerges from autoregressive self-rollout. *CoRR*, abs/2511.20649, 2025.

Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep forcing: Training-free long video generation with deep sink and participative compression. *CoRR*, abs/2512.05081, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems*, volume 37, pp. 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *CoRR*, abs/2602.02214, 2026.