ARMEM: AUTOREGRESSIVE LATENT MEMORY FOR STREAMING VIDEO GENERATION

Anonymous authors
Paper under double-blind review

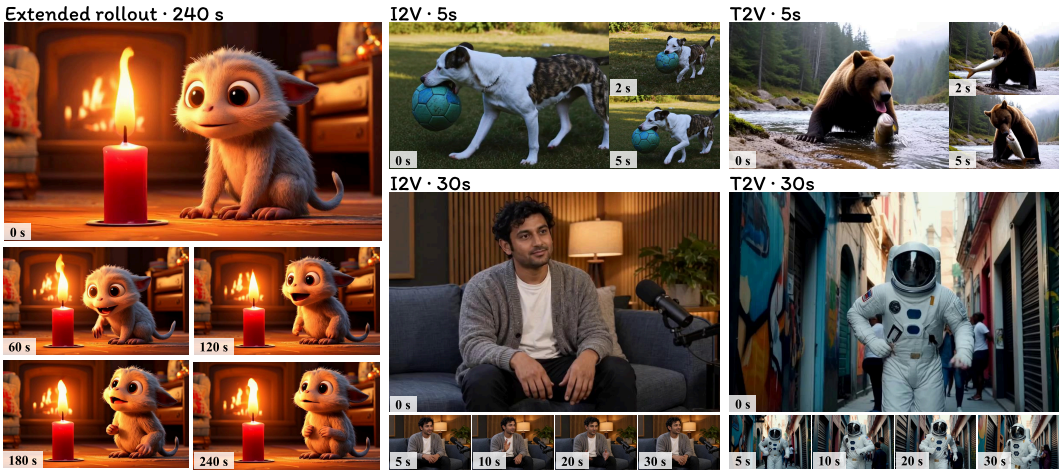


Figure 1: Video generation with ARMem. We showcase 5- and 30-second text-to-video (T2V) and image-to-video (I2V) generation, together with an extended 240-second autoregressive rollout.

ABSTRACT

Long-horizon autoregressive (AR) video generation requires compact memory that preserves persistent information while adapting to evolving content. Existing methods typically preserve history through post-hoc selection or compression of previously generated visual representations and reuse them as static context. While effective for preserving past information, these static memories cannot adapt to evolving video dynamics and may lead to quality degradation during long rollouts. In this work, we introduce **ARMem**, an autoregressive memory paradigm that extends AR modeling from video generation to memory evolution. Specifically, memory is represented as compact latent tokens that are jointly denoised with video tokens in a shared diffusion transformer. At each step, the previous memory state guides video synthesis, while newly generated content is incorporated into the next memory state to support subsequent generation. To mitigate optimization conflicts between video generation and memory evolution, we introduce Dual-LoRA Self-Attention, which decouples their parameter adaptations. We further introduce a Historical Readout Task that explicitly supervises the evolving memory to retain future-useful information, such as subject identity. Extensive experiments on LTX and Wan demonstrate improved short-video quality and long-horizon performance under bounded memory constraints.

1 INTRODUCTION

Bidirectional video diffusion models (Wan et al., 2025; HaCohen et al., 2026) excel at generating visually compelling short video clips. Autoregressive (AR) video diffusion models (Chen et al., 2025; Teng et al., 2025; Yin et al., 2025; Huang et al., 2025a) extend this progress toward long-horizon streaming generation, continually producing new content conditioned on the past. A central chal-

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.

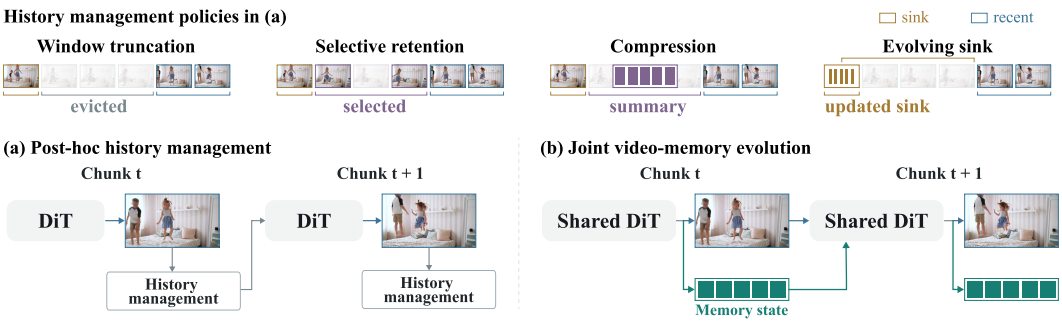


Figure 2: Comparison of memory mechanisms. Top: existing policies for managing history in (a), which select, compress, or aggregate previously generated visual representations for reuse. (a) Post-hoc historical context management, where generation and memory update are separate. (b) Our joint video-memory evolution, where video and memory tokens are jointly generated and carried forward.

lenge is to maintain long-range consistency while sustaining natural motion and scene progression. This requires memory that preserves persistent information and integrates new content.

Existing methods maintain memory through explicit reuse of previously generated visual representations, using selection (Ji et al., 2025; Yi et al., 2026; Zhao et al., 2026b), compression (Zhu et al., 2025; Bian et al., 2026; Jiang et al., 2026), or aggregation (Lu et al., 2026; Kim et al., 2026) to construct the context from historical frames. However, such reuse may encourage the model to reproduce earlier content, limiting scene progression and leading to quality degradation over long rollouts. As illustrated in Fig. 2 (a), these methods follow a post-hoc design: the generator produces video, while a separate, often heuristic history-management policy determines what information to carry forward. *We take a different view of memory and ask: can the model learn to evolve its memory autoregressively as it generates the video?*

Building on this insight, we introduce **ARMem**, which models video generation and memory evolution as a joint autoregressive process (Fig. 2 (b)). At its core, ARMem represents memory as a compact set of latent tokens jointly denoised with video tokens in a shared diffusion transformer (DiT). At each autoregressive step, video generation reads the previous memory, while the next memory state integrates the previous state with the newly generated content. The updated memory is then carried forward to support subsequent generation. Memory transitions are thus learned within the video generator itself, without a separate update network or post-hoc history retention.

Jointly optimizing video generation and memory evolution with fully shared parameters can introduce competing gradients between immediate generation quality and long-term memory preservation. We therefore introduce Dual-LoRA Self-Attention to provide task-specific parameter adaptations while preserving interaction within a shared backbone. Furthermore, the generation objective trains memory to support video synthesis, but does not explicitly specify what information should survive successive updates. We therefore propose a Historical Readout Task that requires the model to reconstruct historical information, such as subject identity, conditioned on the evolving memory, directly supervising what memory should preserve.

We implement ARMem on Wan (Wan et al., 2025) and LTX (HaCohen et al., 2026), with experiments covering text-to-video and image-to-video generation. As shown in Fig. 1, we support both short-video synthesis and extended autoregressive rollouts reaching 240 seconds. Extensive experiments across multiple tasks and horizons demonstrate that our method delivers stronger visual quality and temporal consistency under a bounded memory budget.

In summary, our contributions are as follows:

- We introduce ARMem, which unifies video synthesis and memory evolution in a joint autoregressive process, without a separate post-hoc memory-update stage.
- We design Dual-LoRA Self-Attention to preserve interactions between video generation and memory evolution while mitigating optimization conflicts, and a Historical Readout Task to explicitly supervise long-term identity retention.

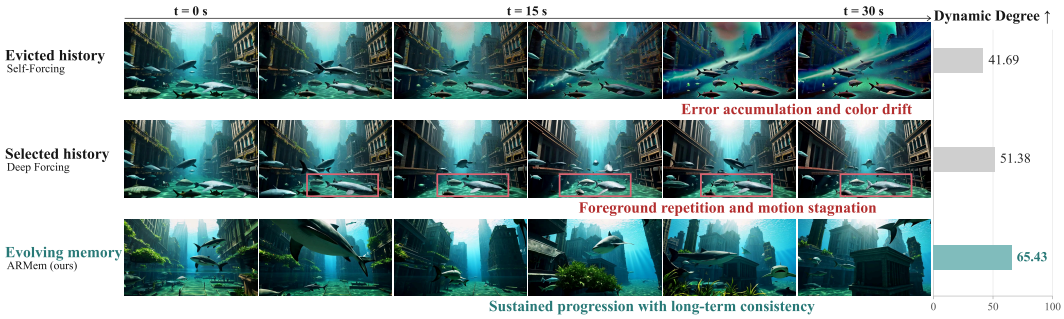


Figure 3: Evicting history risks visual drift, while reusing selected history can lead to repetition and motion stagnation. ARMem evolves memory alongside video generation, preserving long-term consistency while sustaining content progression.

- Extensive experiments on Wan and LTX demonstrate improved visual quality and long-horizon consistency across text-to-video and image-to-video generation.

2 Related Work

2.1 Autoregressive Video Diffusion

Bidirectional video diffusion models (HaCohen et al., 2024; Yang et al., 2025; Wan et al., 2025; HaCohen et al., 2026) achieve strong generation quality through joint spatio-temporal modeling, but require full-sequence access and substantial computation. Recent work converts these models into causal autoregressive generators with few-step distillation, enabling efficient chunk-wise synthesis (Yin et al., 2025; Huang et al., 2025a; Liu et al., 2026). A representative work, Self-Forcing (Huang et al., 2025a), trains on self-generated rollouts to reduce the mismatch between training and inference. Subsequent methods (Zhu et al., 2026; Cui et al., 2026; Zhao et al., 2026a; Zhuang et al., 2026) further refine this training paradigm and improve generation quality.

2.2 Memory for Long-Horizon Generation

Long-horizon autoregressive generation typically uses persistent initial sinks and a sliding window of recent frames (Yang et al., 2026a). To preserve important long-range history between them, selection-based methods retain a subset of historical frames based on hand-crafted rules (Ji et al., 2025; Chen et al., 2026a; Yi et al., 2026; Zhao et al., 2026b; Yang et al., 2026b), while compression-based methods summarize history into compact representations (Zhu et al., 2025; Jiang et al., 2026; Bian et al., 2026). However, reusing static historical representations may not adequately reflect changes in video content as generation progresses. Recent work shifts from memory retention to memory evolving through EMA-based updates (Lu et al., 2026; Kim et al., 2026), dynamic sink retrieval (Ye et al., 2026), and recurrent latent anchors (Dalva & Yanardag, 2026). Building on this direction, we introduce autoregressive memory that evolves jointly with video content, unifying memory evolution and video generation within a single diffusion transformer.

3 Method

3.1 Problem Formulation

Autoregressive video diffusion models generate videos chunk by chunk, where each future chunk is conditioned on previously generated content. A chunk may contain one or more frames and serves as the basic autoregressive unit. Let x_t denote the clean latent of the t-th video chunk and x_t^τ denote its noisy latent at diffusion timestep τ. Omitting text or image conditioning for simplicity, conventional autoregressive generation can be formulated as:

$$\hat{x}_t = G_\theta(x_t^\tau, \tau; C_{t-1}), \tag{1}$$

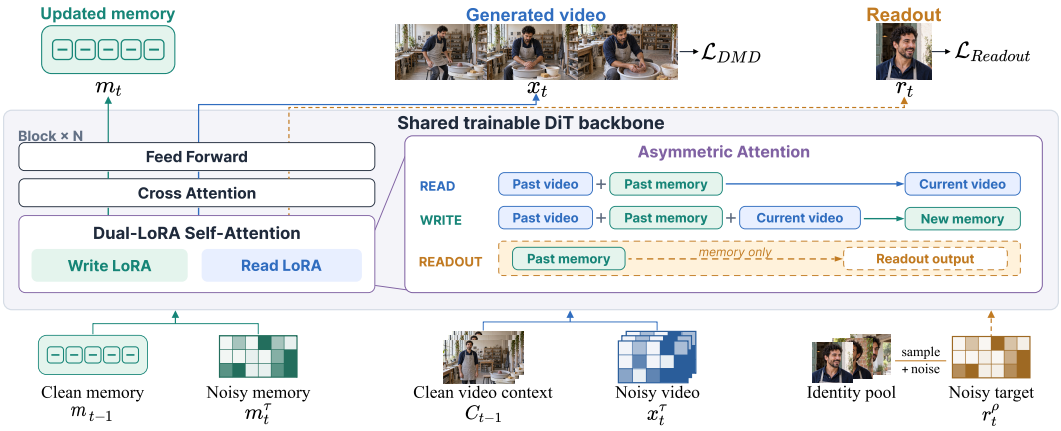


Figure 4: Overview of the proposed framework for joint autoregressive video generation and recurrent latent memory updating. We employ Dual-LoRA Self-Attention across token groups, enabling interactions between memory reading (video generation) and memory writing while mitigating gradient conflicts. The Historical Readout Task reconstructs identity targets from memory alone, providing explicit supervision for long-term identity retention.

where $C_{t-1} = \{\hat{x}_0, \ldots, \hat{x}_{t-1}\}$ denotes the generated video chunks.

As the generation horizon grows, conditioning on the full history becomes increasingly costly. Existing methods commonly structure the historical context as:

$$C_{t-1} = [\ \underbrace{X_{\text{sink}}}_{\text{persistent sink}}\ ;\ \underbrace{X_{\text{hist}}}_{\text{managed history}}\ ;\ \underbrace{X_{\text{recent}}}_{\text{recent window}}\ ]. \tag{2}$$

The initial-frame sink offers a stable visual anchor, and the recent window supports continuity across adjacent segments. Beyond these, long video generation needs to maintain long-term history within a limited memory. The management of X_{hist} is mostly heuristic, using hand-crafted rules to select, compress, or rearrange past visual representations post-hoc. As Fig. 3 shows, this can lead to repetitive content, restrict scene progression, and worsen error accumulation.

To address these limitations, we present *ARMem*, an extension of autoregressive modeling from video generation to memory evolution. Instead of post-hoc management of historical information from old frames, ARMem introduces a compact recurrent latent memory state m_t, which is recursively updated and carried forward across chunks. At each autoregressive step, a shared diffusion transformer jointly denoises the current noisy video chunk and the live memory latent:

$$(\hat{x}_t, \hat{m}_t) = G_\theta(x_t^\tau, m_t^\tau, \tau; C_{t-1}, m_{t-1}), \tag{3}$$

where m_{t-1} is the previous memory state, m_t^τ represents noisy live version at timestep τ, and $\hat{m}_t$ is the denoised output. The explicit context is reduced to $C_{t-1} = [X_{\text{sink}}; X_{\text{recent}}]$. After denoising, $\hat{x}_t$ is appended to the video stream and updates the recent context window, while $\hat{m}_t$ replaces m_{t-1} as the memory state for the next generation step. This integrates video generation and memory evolution into a single autoregressive process.

3.2 Joint Denoising with Asymmetric Attention

As illustrated in Fig. 4, ARMem realizes joint video generation and memory updating within a shared diffusion transformer (DiT). We describe its unified token interface and the asymmetric attention constraints that govern information exchange.

Unified token interface. We propose a unified token interface consisting of video, memory, and optional readout tokens. At each denoising step, these are represented as:

$$x_t^\tau \in \mathbb{R}^{N_v \times d}, \quad m_t^\tau \in \mathbb{R}^{N_m \times d}, \quad r_t^\rho \in \mathbb{R}^{N_r \times d},$$

respectively, where N_v, N_m, and N_r are the token counts and d is the shared embedding dimension. The video and memory tokens share timestep τ, while readout tokens use a separately sampled

timestep ρ and are included only during training. During training, the clean conditions and noisy inputs are packed into a single sequence:

$$Z_t^{\tau,\rho} = [m_{t-1}; m_t^\tau; C_{t-1}; x_t^\tau; r_t^\rho]. \tag{4}$$

New video and live memory latents are initialized from independent Gaussian noise at the start of each chunk. Readout tokens are constructed by adding noise to a ground-truth reconstruction target at timestep ρ, supervising what information should be retained in memory as detailed in Section 3.4.

Asymmetric attention. We employ an asymmetric attention mask to manage information flow between token groups. Current video queries attend to the retained visual context and the previous memory state, allowing generation to use both explicit visual references and accumulated historical information. Live memory queries also consider current noisy video tokens, allowing memory updates to include new content. Conversely, live video queries cannot access live memory tokens, preventing circular dependencies between video generation and memory writing. For readout queries, they attend only to the previous memory state, ensuring that historical reconstruction relies solely on the stored memory and avoiding information leakage from current video tokens.

3.3 Decoupled Memory Read–Write Learning

Video generation and memory updating impose different functional requirements on the shared DiT. Memory reading retrieves past information for current video chunks, while memory writing integrates new content and maintains long-term information. Optimizing both simultaneously with shared parameters may cause interference, hindering effective memory retrieval and state evolution.

Dual-LoRA self-attention. To address this issue, we introduce separate LoRA (Hu et al., 2022) adapters for memory reading and writing within the self-attention layers:

$$W_r = W + \Delta W_r, \quad W_w = W + \Delta W_w, \tag{5}$$

where ΔW_r adapts memory reading and ΔW_w adapts memory writing. This allows the two roles to specialize while retaining a shared DiT backbone.

For generation supervision, we optimize the autoregressive rollout with Distribution Matching Distillation (DMD) (Yin et al., 2024a). Given a self-generated video $\tilde{x} = G_\theta(\epsilon)$, the DMD gradient is estimated by the difference between the teacher and fake scores:

$$\nabla_\theta \mathcal{L}_{\text{DMD}} \approx -\mathbb{E}_{\tau,\epsilon}\left[\left(s_{\text{real}}(\tilde{x}_\tau, \tau) - s_{\text{fake}}(\tilde{x}_\tau, \tau)\right) \frac{\partial \tilde{x}}{\partial \theta}\right], \tag{6}$$

Since the updated memory state m_t is used by subsequent chunks, its usefulness for generation must be learned through future outputs. This necessitates propagating gradients from later video-generation losses back through earlier memory updates to train memory writing. We achieve this by adapting the two-pass training scheme of Self Gradient Forcing (Zhuang et al., 2026), which restores context-writing gradients blocked by detached historical KV caches in Self Forcing (Huang et al., 2025a). The first pass performs a no-gradient autoregressive rollout and records the generated context and noisy video and memory inputs at a sampled denoising exit step. To accommodate recurrent memory, we replace SGF's parallel second-pass reconstruction with sequential replay, passing each recomputed memory state to the next chunk without detaching it. Consequently, subsequent video-generation losses can propagate through earlier memory updates and train the write LoRA.

3.4 Historical Readout Supervision

Generation supervision encourages memory to support future video synthesis, but does not directly specify which information should remain recoverable after repeated updates. We therefore introduce historical readout supervision to encourage the retention of persistent information. Through the unified token interface, readout tokens reconstruct selected targets using the previous memory state as their only source of historical conditioning. The reconstruction targets thus specify what information the evolving memory is encouraged to preserve.

Inspired by Memento (Wei et al., 2026), we instantiate this supervision through identity reconstruction. Using ConsisID (Yuan et al., 2025) data, we construct identity pools containing multiple crops

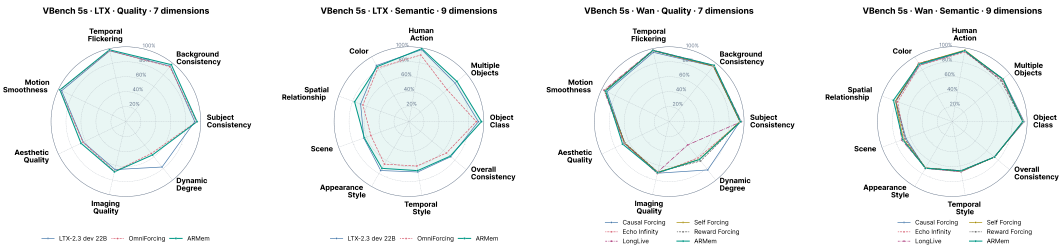


Figure 5: Per-dimension VBench results on 5-second text-to-video generation. From left to right, the radar plots show seven quality dimensions and nine semantic dimensions for the LTX-based and Wan-based baselines, respectively.

Table 1: VBench results for 5-second text-to-video generation. The comparison includes representative open-source models. $\times 2$ accounts for classifier-free guidance (CFG).

Method	Generation	NFE ↓	Total ↑	Quality ↑	Semantic ↑
Wan-based					Resolution: 832×480
Wan2.1	Bidirectional	50×2	84.36	85.17	81.12
Self-Forcing (NeurIPS'25)	Causal AR	4	83.84	84.53	81.06
LongLive (ICLR'26)	Causal AR	4	83.01	83.55	80.85
Causal Forcing (ICML'26)	Causal AR	4	84.28	85.21	80.57
Reward Forcing (CVPR'26)	Causal AR	4	83.85	84.79	80.08
Echo-Infinity (arXiv'26)	Causal AR	4	83.73	84.62	80.17
Ours	Causal AR	4	**84.57**	**85.35**	**81.47**
LTX-based					Resolution: 768×512
LTX-2.3	Bidirectional	30×2	82.96	84.07	78.51
OmniForcing (ECCV'26)	Causal AR	4	80.94	83.39	71.13
Ours	Causal AR	4	**83.87**	**84.97**	**79.49**

of the same subject across different poses and viewpoints. For each training sequence, we condition the model on a ground-truth frame and perform an autoregressive self-rollout. At subsequent chunks, we sample crops from the corresponding identity pool as reconstruction targets and add noise at timestep ρ to obtain readout tokens r_t^ρ. The shared DiT denoises these tokens conditioned on the carried memory state m_{t-1}. Formally,

$$\mathcal{L}_{\text{readout}} = \mathbb{E}_{r_t,\rho,\epsilon}\left[\|r_t - \hat{r}_t(m_{t-1}, r_t^\rho, \rho)\|_2^2\right], \tag{7}$$

where r_t denotes the clean reconstruction target and $\hat{r}_t$ the model prediction. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{DMD}} + \lambda \mathcal{L}_{\text{readout}}, \tag{8}$$

where λ controls its contribution. Reconstructing the same subject from successive memory states encourages identity information to remain recoverable as memory evolves.

4 Experiments

4.1 Experimental Settings

Implementation Details. We instantiate our framework on LTX-2.3-22B-Dev (HaCohen et al., 2026) and Wan2.1-T2V-1.3B (Wan et al., 2025), with LTX-2.3-22B-Dev and Wan2.1-T2V-14B as their respective teachers. Both models use chunk-wise autoregressive generation with 16 memory tokens, one sink latent frame, and six recent latent frames. LTX is jointly trained on T2V and I2V. We further apply 1,000 steps of Streaming Long Tuning following LongLive. Initialization procedures and detailed training hyperparameters are provided in Appendix B.

Baselines. We compare against representative autoregressive video generators covering diverse historical context and memory organization strategies. For long-horizon generation, these include recent-window context (Self Forcing (Huang et al., 2025a)), fixed sinks with recent context

Table 2: VBench I2V results for 5-second image-to-video generation.

Method	Generation	Resolution	NFE ↓	I2V Subject ↑	I2V Background ↑	Dynamic Degree ↑	Quality ↑	I2V ↑	Total ↑
LTX-2.3	Bidirectional	768×512	30×2	98.91	99.12	18.94	79.12	98.77	88.95
Ours	**Causal AR**	**768×512**	**4**	**98.96**	**99.20**	**29.84**	**80.18**	**98.85**	**89.51**

Table 3: VBench-Long results for 30- and 60-second video generation. The top three scores within each duration and backbone group are highlighted as **1st**, 2nd, and 3rd, with the best in bold.

Method	Context/Memory	Subject Cons. ↑	Background Cons. ↑	Motion Smooth. ↑	Dynamic Degree ↑	Aesthetic Qual. ↑	Imaging Qual. ↑	Temporal Flick. ↑	Overall Cons. ↑	Average ↑
30 seconds										
WAN-based										
Self-Forcing (NeurIPS'25)	Recent	96.45	96.12	98.09	41.69	59.07	67.70	97.24	24.12	72.56
LongLive (ICLR'26)	Fixed Sink + Recent	**97.57**	**96.60**	**98.80**	42.18	61.63	68.74	97.70	**24.84**	73.51
Reward Forcing (CVPR'26)	EMA Sink + Recent	96.43	95.85	98.33	63.45	60.86	68.82	96.71	24.83	75.66
Deep Forcing (ICML'26)	Deep Sink + Selected + Recent	96.42	95.82	98.11	51.38	60.30	68.10	96.90	24.62	73.96
MemRoPE (ECCV'26)	Fixed Sink + Dual-EMA Memory + Recent	96.77	95.80	97.72	61.77	58.73	66.95	**98.80**	24.47	75.13
Echo-Infinity (arXiv'26)	Fixed Sink + Evolving Memory Queries + Recent	96.85	95.89	98.26	60.75	59.81	**70.84**	98.77	24.79	75.75
Ours	Fixed Sink + Learnable Recurrent Memory + Recent	97.13	95.97	98.13	**65.43**	**61.80**	69.37	98.78	24.67	**76.41**
LTX-based										
OmniForcing (ECCV'26)	Fixed Sink + Recent	**97.61**	96.10	99.13	**56.43**	57.67	61.70	98.62	19.61	73.36
Ours	Fixed Sink + Learnable Recurrent Memory + Recent	97.37	**97.00**	**99.25**	47.13	**61.21**	**63.97**	**99.42**	**24.41**	**73.72**
60 seconds										
WAN-based										
Self-Forcing (NeurIPS'25)	Recent	96.38	96.29	97.84	35.51	56.07	65.74	97.25	21.84	70.86
LongLive (ICLR'26)	Fixed Sink + Recent	**97.35**	**96.43**	**98.76**	42.70	**61.25**	68.40	97.66	**25.04**	73.45
Reward Forcing (CVPR'26)	EMA Sink + Recent	96.33	95.73	98.30	66.15	60.10	68.37	98.42	24.66	76.01
Deep Forcing (ICML'26)	Deep Sink + Selected + Recent	96.80	96.00	98.28	46.01	60.07	67.70	97.26	24.72	73.36
MemRoPE (ECCV'26)	Fixed Sink + Dual-EMA Memory + Recent	95.40	95.18	97.38	65.66	57.41	66.17	98.66	23.87	74.97
Echo-Infinity (arXiv'26)	Fixed Sink + Evolving Memory Queries + Recent	96.75	95.68	98.20	62.56	59.16	**70.85**	98.61	24.84	75.83
Ours	Fixed Sink + Learnable Recurrent Memory + Recent	96.88	95.87	98.11	**68.34**	60.27	68.37	**98.74**	24.58	**76.39**
LTX-based										
OmniForcing (ECCV'26)	Fixed Sink + Recent	**97.96**	96.30	99.08	**54.20**	56.73	61.11	98.56	17.81	72.72
Ours	Fixed Sink + Learnable Recurrent Memory + Recent	97.81	**97.37**	**99.41**	49.81	**61.35**	**66.23**	**99.31**	**24.04**	**74.42**

(LongLive (Yang et al., 2026a) and OmniForcing (Su et al., 2026)), EMA-updated sinks (Reward Forcing (Lu et al., 2026)), selective history retention (Deep Forcing (Yi et al., 2026)), dual-EMA memory (MemRoPE (Kim et al., 2026)), and evolving memory queries (Echo-Infinity (Bian et al., 2026)). These comparisons evaluate our learned recurrent memory against alternative mechanisms for retaining and updating historical information. For short-horizon generation, we additionally include the original bidirectional backbones and Causal Forcing (Zhu et al., 2026) to assess generation quality. We organize all comparisons by backbone, separating Wan-based and LTX-based models.

Evaluation. We evaluate short- and long-horizon video generation. For 5-second T2V, we evaluate all 16 VBench (Huang et al., 2024) dimensions and report quality, semantic, and total scores using the official normalization and weighting scheme. For 5-second I2V, we report VBench-I2V (Huang et al., 2025b) subject and background consistency, dynamic degree, and aggregate quality, I2V, and total scores. For long-horizon video, we generate 30- and 60-second videos using the first 128 prompts from MovieGen (Polyak et al., 2024). Following prior works (Bian et al., 2026; Cui et al., 2026), these videos were evaluated using six VBench-Long (Huang et al., 2025b) metrics: aesthetic quality, background consistency, dynamic degree, imaging quality, motion smoothness, and subject consistency. We also report overall consistency for video-text alignment, as in Deep Forcing (Yi et al., 2026), and temporal flickering for stability, as in Self Gradient Forcing (Zhuang et al., 2026).

4.2 Main Results

Short Video Generation. We first evaluate ARMem on 5-second text-to-video and image-to-video generation. As shown in Tab. 1, ARMem achieves the highest total, quality, and semantic scores within both backbone groups using only four denoising steps, outperforming both the bidirectional backbones and existing autoregressive baselines. On Wan2.1, it achieves a total score of 84.57; on LTX-2.3, it surpasses OmniForcing by 2.93 points in total score and 8.36 points in semantic score. For image-to-video generation (Tab. 2), ARMem improves both quality and total scores over the bidirectional teacher. Notably, dynamic degree increases from 18.94 to 29.84, while subject and background consistency with the input image also improves slightly. These results demonstrate strong short-video quality and conditional alignment under efficient causal generation. Fig. 5

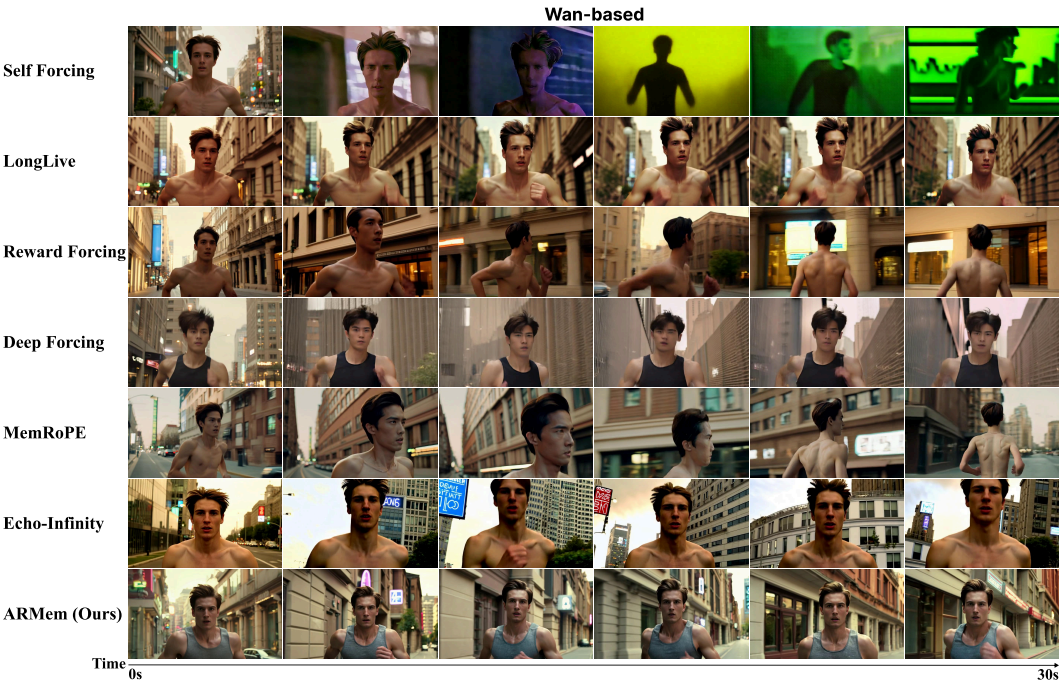


Figure 7: Qualitative comparison of ARMem with Wan-based baselines for long-horizon generation.

presents the per-dimension VBench results, complementing the aggregate scores, and additional qualitative comparisons are provided in the Appendix.

Long-Horizon Video Generation. Tab. 3 reports results for 30- and 60-second generation, where ARMem achieves the highest average score within both backbone groups at both horizons. On Wan2.1, its average score remains nearly unchanged as the duration doubles (76.41 vs. 76.39), with the highest dynamic degree on both horizons and competitive consistency and visual quality. On LTX-2.3, it outperforms OmniForcing in six of eight metrics on both horizons, with greater gains in imaging quality and overall consistency at 60 seconds. The qualitative results in Fig. 7 and Fig. 8 illustrate the ability of ARMem to preserve visual coherence over extended generation. The baseline outputs exhibit color drift, identity drift, and structural degradation as generation progresses, whereas ARMem maintains a more consistent subject appearance and scene structure. The image-to-video results also show sustained fidelity to the conditioning image.

User Study. We conduct a pairwise user study on 30-second videos generated from 20 prompts sampled from MovieGen (Polyak et al., 2024), comparing ARMem with three baselines initialized from the same Self Forcing checkpoint. Participants indicate their preference based on visual quality, temporal consistency, and text alignment, with a no-preference option. Fig. 6 shows a consistent preference for ARMem.

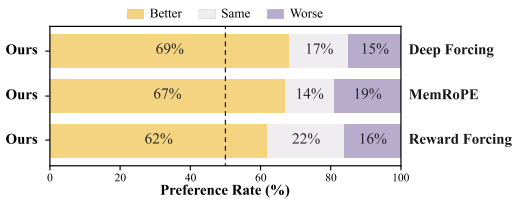


Figure 6: User preference for ARMem.

4.3 Ablation Studies

We ablate key ARMem designs on LTX, training all variants without Streaming Long Tuning. We evaluate generation at 5 and 30 seconds. For 30-second evaluation, Quality aggregates seven VBench-Long metrics, and overall consistency measures video-text alignment.

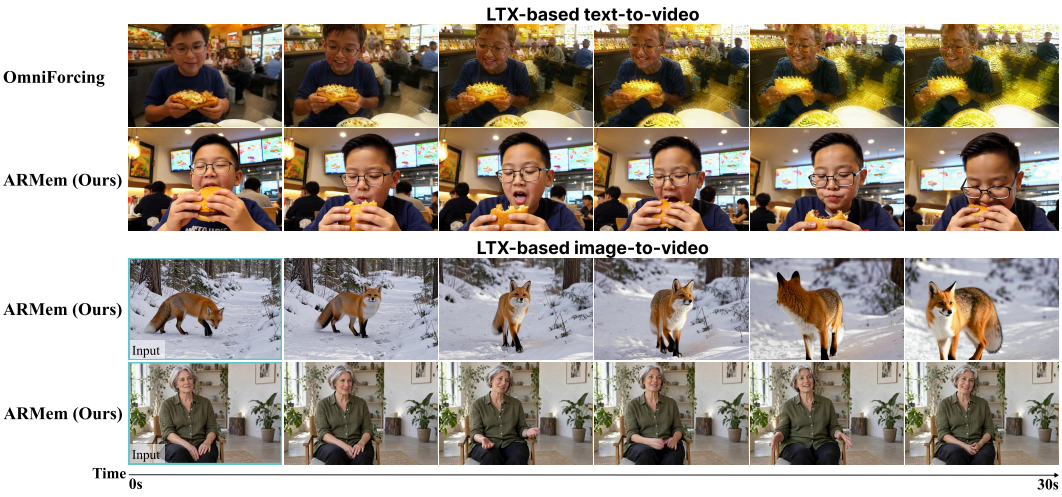


Figure 8: Qualitative comparison of ARMem with LTX-based baselines for long-horizon generation.

Table 4: Ablation study of the proposed components for 5-second and 30-second (6× the training horizon) video generation.

	5s			30s	
Variant	Total ↑	Quality ↑	Semantic ↑	Quality ↑	Overall ↑
w/o Memory	83.24	84.75	77.18	**82.24**	24.14
w/o W/R LoRA	83.09	84.22	78.58	81.94	24.49
w/o Readout Sup.	82.91	83.85	**79.15**	81.62	24.56
Full	**83.71**	**84.88**	79.03	82.14	**24.79**

Table 5: Ablation study on the number of memory tokens for 5-second and 30-second (6× the training horizon) video generation.

	5s			30s	
# Memory Tokens	Total ↑	Quality ↑	Semantic ↑	Quality ↑	Overall ↑
4	83.13	84.31	78.39	82.10	24.71
8	83.27	84.40	78.75	82.23	24.34
16	**83.71**	**84.88**	79.03	82.14	**24.79**
32	83.30	84.25	**79.52**	**82.34**	24.76

Effect of Individual Components. Tab. 4 compares the full model with three variants: without memory tokens, without read/write LoRA adapters, and without historical readout supervision. Removing memory tokens causes the largest declines in the 5-second semantic score (from 79.03 to 77.18) and 30-second overall consistency. This suggests that evolving memory provides critical long-range semantic preservation beyond the sink and recent context. Removing W/R LoRA degrades every reported metric, indicating that task-specific adaptation helps mitigate interference between memory evolution and video generation. Lastly, explicit historical readout supervision encourages memory to retain information relevant to long-term consistency, better supporting continued generation and improving quality at both horizons.

Effect of Memory Capacity. We further study the effect of memory size in Tab. 5. Increasing memory capacity improves semantic performance, with short-video quality also improving from 4 to 16 tokens. Further expanding to 32 tokens slightly reduces 5-second quality while improving 30-second quality. Meanwhile, the 16- and 32-token configurations achieve comparable 30-second overall consistency. We therefore adopt 16 tokens as a compact memory configuration that supports continued generation while maintaining competitive performance at both horizons.

5 Conclusion

We presented ARMem, which models long-term memory as an autoregressive latent state that evolves jointly with video generation. Our framework integrates video synthesis and memory updating within a shared diffusion transformer. Dual-LoRA Self-Attention separates parameter adaptations for the two tasks, while historical readout supervision encourages persistent information to remain recoverable as memory evolves. Experiments on LTX and Wan demonstrate improved temporal consistency and long-horizon generation quality under bounded memory constraints. These results demonstrate the effectiveness of evolving memory and offer a new perspective on long-horizon video generation: a shift from post-hoc history preservation to online memory evolution.

AI USE STATEMENT

In this work, we used generative AI tools for code development, literature review, and writing refinement. All AI-assisted content was reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

A comprehensive description of the implementation is provided in Section 4.1 and Appendix B. The source code and additional comparison videos are included in the supplementary material.

REFERENCES

Yuxuan Bian, Zeyue Xue, Songchun Zhang, Shiyi Zhang, Weiyang Jin, Yaowei Li, Junhao Zhuang, Haoran Li, Jie Huang, Haoyang Huang, et al. Echo-infinity: Learning evolving memory for real-time infinite video generation. *arXiv preprint arXiv:2606.04527*, 2026.

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.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context Forcing: Consistent autoregressive video generation with long context. *Proceedings of the 43rd International Conference on Machine Learning*, 2026a.

Siyi Chen, Shaowei Liu, Yixuan Jia, Zian Wang, Huan Ling, Qing Qu, and Jun Gao. Data-forcing distillation: Restoring diversity and fidelity in few-step video generation. *arXiv preprint arXiv:2606.18478*, 2026b.

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.

Yusuf Dalva and Pinar Yanardag. Adastate: Self-evolving anchors for streaming video generation. *arXiv preprint arXiv:2605.30349*, 2026.

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. In *European Conference on Computer Vision*, pp. 324–344. Springer, 2026.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2022.

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, 2025a.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b.

Sihui Ji, Xi Chen, Shuai Yang, Xin Tao, Pengfei Wan, and Hengshuang Zhao. Memflow: Flowing adaptive memory for consistent and efficient long video narratives. *arXiv preprint arXiv:2512.14699*, 2025.

Jiaxiu Jiang, Wenbo Li, Jingjing Ren, Yuping Qiu, Renjing Pei, Fenglong Song, Yong Guo, Xiaogang Xu, Han Wu, and Wangmeng Zuo. Lovic: Efficient long video generation with context compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 4022–4034, 2026.

Youngrae Kim, Qixin Hu, C-C Jay Kuo, and Peter Beerel. Memrope: Training-free infinite video generation via evolving memory tokens. In *European Conference on Computer Vision*, pp. 244–261. Springer, 2026.

Haobo Li, Yanhong Zeng, Yunhong Lu, Jiapeng Zhu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Yujun Shen, and Zhipeng Zhang. Aad-1: Asymmetric adversarial distillation for one-step autoregressive video generation. *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

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.

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

Xiaofeng Mao, Shaohao Rui, Kaining Ying, Bo Zheng, Chuanhao Li, Mingmin Chi, and Kaipeng Zhang. Packforcing: Short video training suffices for long video sampling and long context inference. In *European Conference on Computer Vision*, pp. 576–593. Springer, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

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. *arXiv preprint arXiv:2410.13720*, 2024.

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

Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 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.

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

Xuan Wei, Longbin Ji, Guan Wang, Xiangrui Liu, Zhenyu Zhang, Shuohuan Wang, Yu Sun, and Qingqi Hong. Memento: Reconstruct to remember for consistent long video generation. *arXiv preprint arXiv:2606.14667*, 2026.

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. *International Conference on Learning Representations*, 2026a.

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. *European Conference on Computer Vision*, 2026b.

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*, volume 2025, pp. 83048–83077, 2025.

Bo Ye, Xinyu Cui, Jian Zhao, Tong Wei, and Min-Ling Zhang. Dysink: Dynamic frame sinks for autoregressive long video generation. *arXiv preprint arXiv:2605.21028*, 2026.

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. *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

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

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

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

Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12978–12988. IEEE, 2025.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. *Advances in Neural Information Processing Systems*, 38:30546–30566, 2025.

Lvmin Zhang, Shengqu Cai, Muyang Li, Chong Zeng, Beijia Lu, Anyi Rao, Song Han, Gordon Wetzstein, and Maneesh Agrawala. Tinyhistory: Lightweight video history embeddings via two-stage context learning. *European Conference on Computer Vision*, 2026.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation. *arXiv preprint arXiv:2605.15141*, 2026a.

Zengqun Zhao, Yanzuo Lu, Ziquan Liu, Jifei Song, Jiankang Deng, and Ioannis Patras. Relax forcing: Relaxed kv-memory for consistent long video generation. *British Machine Vision Conference*, 2026b.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Tianrui Zhu, Shiyi Zhang, Zhirui Sun, Jingqi Tian, and Yansong Tang. Memorize-and-generate: Towards long-term consistency in real-time video generation. *arXiv preprint arXiv:2512.18741*, 2025.

Junhao Zhuang, Shiyi Zhang, Yuxuan Bian, Yaowei Li, Yawen Luo, Yijun Liu, Weiyang Jin, Songchun Zhang, Xianglong He, Xuying Zhang, et al. Self gradient forcing: Native long video extrapolation. *arXiv preprint arXiv:2607.20368*, 2026.

A Related Works

A.1 From Bidirectional to Autoregressive Video Diffusion

Bidirectional video diffusion models (Peebles & Xie, 2023; HaCohen et al., 2024; Yang et al., 2025; Wan et al., 2025; HaCohen et al., 2026) have demonstrated remarkable generation quality by letting all frames jointly attend to each other during denoising. While this design enables strong spatio-temporal modeling, it requires access to the entire video sequence and incurs substantial computational cost, making it unsuitable for real-time streaming and interactive generation. To address this limitation, recent studies (Yin et al., 2025; Huang et al., 2025a; Zhu et al., 2026; Liu et al., 2026) have converted bidirectional video diffusion models into causal autoregressive (AR) generators and accelerated them with DMD-based distillation (Yin et al., 2024a; Chen et al., 2026b; Yin et al., 2024b), generating future frames or chunks conditioned on previously synthesized content.

CausVid (Yin et al., 2025) first demonstrates that a powerful bidirectional video diffusion model can be distilled into a causal generator, enabling real-time chunk-wise video synthesis. Self-Forcing (Huang et al., 2025a) further identifies the train–test discrepancy in Diffusion Forcing (Chen et al., 2024), where training relies on ground-truth histories while inference depends on self-generated content, and introduces self-rollout training to better align training and inference distributions. Causal Forcing (Zhu et al., 2026) and Causal Forcing++ (Zhao et al., 2026a) further investigate causal distillation by improving causal teacher construction and initialization strategies. Beyond distribution alignment, Self Gradient Forcing (Zhuang et al., 2026) analyzes the optimization dynamics of autoregressive training and reveals the importance of properly supervising historical state updates during self-generated rollouts. Despite these advances, current autoregressive video models primarily establish the capability of efficient streaming generation. Extending them from seconds to minute- or hour-scale horizons raises a key challenge: how to maintain a compact memory that preserves persistent information while adapting to evolving content as generation unfolds.

A.2 Memory Management in Long-Horizon Autoregressive Generation

As autoregressive video generation scales to longer horizons, maintaining historical context under limited memory becomes increasingly challenging. LongLive (Yang et al., 2026a) introduces a sink-and-window strategy, where persistent early-frame sinks provide stable global references and a FIFO recent window maintains local temporal continuity. However, tokens evicted from the recent window usually contain valuable long-range information. Subsequent works address this limitation from two perspectives: selection and compression. Selection-based methods (Ji et al., 2025; Chen et al., 2026a; Yi et al., 2026; Zhao et al., 2026b; Yang et al., 2026b; Mao et al., 2026) aim to retain informative historical KVs beyond the recent window. For example, Deep Forcing (Yi et al., 2026) preserves historical tokens based on their contribution to attention computation, while Relaxed Forcing (Zhao et al., 2026b) selects complementary historical frames that align with sink tokens while reducing redundancy with recent context. Compression-based methods (Zhu et al., 2025; Jiang et al., 2026; Bian et al., 2026; Zhang et al., 2025; 2026) instead summarize evicted history into compact representations. For instance, Echo-Infinity (Bian et al., 2026) compresses discarded history through learnable query-based aggregation.

While these methods improve the utilization of historical context, they still rely on static historical representations, which may limit scene evolution and motion dynamics. Recent works therefore explore dynamic memory mechanisms that update memory over time. Reward Forcing (Lu et al., 2026) updates sink tokens by incorporating evicted history through exponential moving averages, whereas MemRoPE (Kim et al., 2026) retains fixed sink tokens and compresses evicted history into separate long- and short-term memory tokens via dual-rate EMA. DySink (Ye et al., 2026) further introduces dynamic sink retrieval to adapt memory anchors during generation. AdaState (Dalva & Yanardag, 2026) takes a further step by introducing a recurrent latent anchor that evolves during denoising, demonstrating the potential of learned latent memory states for long-horizon generation.

Instead of relying on external memory update rules or separate memory modules, our framework learns memory transition directly within the diffusion transformer, enabling autoregressive memory evolution alongside video generation.

B Implementation Details

Backbones and Initialization. We instantiate ARMem on LTX-2.3-22B-Dev (HaCohen et al., 2026) and Wan2.1-T2V-1.3B (Wan et al., 2025), using LTX-2.3-22B-Dev and Wan2.1-T2V-14B as their respective teachers. For Wan, we initialize the student from the ODE initialization checkpoint released by Self Forcing (Huang et al., 2025a). For LTX, we obtain a multi-step autoregressive initialization through 2,000 steps of teacher forcing followed by 2,000 steps of resampling forcing (Guo et al., 2026).

Chunking and Memory Configuration. Both backbones use chunk-wise autoregressive generation with 16 memory tokens, one sink latent frame, and six recent latent frames. Wan uses three latent frames per chunk, following the chunking convention of prior forcing-based methods (Huang et al., 2025a; Yang et al., 2026a). For LTX, we follow OmniForcing (Su et al., 2026) and use four latent frames in the first chunk and three in each subsequent chunk, yielding a 4–3–3–3–3 partition for a 5-second clip. The memory state is carried forward across chunks to provide historical information alongside the sink and recent context.

Training Hyperparameters. For LTX, the student and fake-score model both use a learning rate of 2×10^{-5}. For Wan, their learning rates are 2×10^{-6} and 4×10^{-7}, respectively. The read and write LoRA adapters use a learning rate of 10^{-4} for both backbones. We set the historical readout loss weight to 0.2. For LTX, we first perform 300 warm-up steps using only text-to-video generation, followed by joint text-to-video and image-to-video training with a task sampling ratio of 0.7:0.3. We use the filtered and LLM-extended VidProM (Wang & Yang, 2024) prompt set released by Self Forcing (Huang et al., 2025a) as training prompts, containing approximately 250k prompts.

C Evaluation Details

Baseline Configurations. We use the official implementations and released checkpoints of the baselines, and organize comparisons by backbone. Wan-based methods use Wan2.1-T2V-1.3B; our LTX model uses LTX-2.3. MemRoPE (Kim et al., 2026) uses its Self-Forcing-checkpoint configuration. OmniForcing (Su et al., 2026) uses the released five-second causal checkpoint based on LTX-2. For long-horizon evaluation, we extend its autoregressive rollout using a rolling KV cache with the same five-second context budget of 16 video latent frames and 126 audio latent tokens. Distilled autoregressive models use four denoising steps with their respective noise schedules. The original Wan backbone uses 50-step UniPC sampling with shift 8 and guidance scale 6; the original LTX-2.3 backbone uses 30 steps and guidance scale 3, with STG and guidance rescaling disabled.

Short-Video Evaluation. For five-second T2V generation, we use the extended VBench prompts released with prior work (Huang et al., 2025a), shared across methods, and generate five samples per unique prompt with replica seeds. The 946-entry list contains 944 unique prompts, yielding 4,720 videos per model. Wan-based models generate 81 frames at 832×480 and 16 fps; LTX-based models generate 121 frames at 768×512 and 24 fps. We evaluate all 16 VBench (Huang et al., 2024) dimensions using the official benchmark metadata, normalization, and weighting to compute quality, semantic, and total scores. The complete per-dimension results are provided in the following section. For I2V, we use 355 official VBench-I2V (Huang et al., 2025b) image–prompt pairs and generate five samples per pair. Images are cropped to 3:2 using the official cropping tool, with the same references used for scoring. We report image–video subject and background consistency, dynamic degree, quality, I2V, and total scores.

Long-Video Evaluation. We use the first 128 MovieGen (Polyak et al., 2024) prompts in their original order, adopting the extended prompts released with prior work (Huang et al., 2025a) for all methods. We generate five samples per prompt at both 30- and 60-second horizons, yielding 640 videos per model per horizon, with the same resolutions, frame rates, and replica seeds as above. At these nominal horizons, Wan baselines produce 477/957 frames, our Wan model produces 481/961 frames, and LTX-based models produce 721/1,441 frames. Following prior work (Bian et al., 2026; Cui et al., 2026), we report six VBench-Long (Huang et al., 2025b) metrics: aesthetic quality, background consistency, dynamic degree, imaging quality, motion smoothness, and subject

consistency. We additionally report overall consistency to assess video–text alignment, following Deep Forcing (Yi et al., 2026), and temporal flickering to assess temporal stability, following Self Gradient Forcing (Zhuang et al., 2026).

D Additional Results

Five-Second Text-to-Video Generation. Fig. 5 presents the per-dimension VBench results for both LTX-based and Wan-based comparisons, complementing the aggregate scores in the main paper. The seven quality dimensions and nine semantic dimensions provide a detailed breakdown of visual quality, temporal consistency, and adherence to the text prompt. Fig. 9 and Fig. 10 further provide qualitative comparisons with the corresponding baselines.



Figure 9: Qualitative comparison of ARMem with wan-based baselines for 5-second generation.



Figure 10: Qualitative comparison of ARMem with LTX-based baselines for 5-second generation.

Five-Second Image-to-Video Generation. Figure 11 compares ARMem with LTX-2.3 across six video-quality dimensions and two image–video consistency dimensions, following the evaluation setting of AAD-1 (Li et al., 2026). These results complement the aggregate scores by separately assessing the quality of the generated videos and their fidelity to the conditioning images. Figure 12 provides qualitative examples, with each conditioning image followed by frames from the generated video.

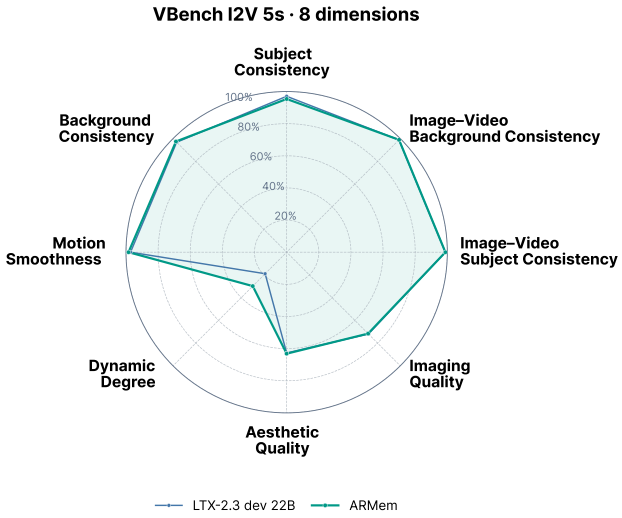


Figure 11: Per-dimension VBench-I2V comparison between ARMem and LTX-2.3 on 5-second image-to-video generation. We report the eight dimensions of the AAD-1 (Li et al., 2026) evaluation setting: six video-quality dimensions and two image–video consistency dimensions, excluding camera motion.

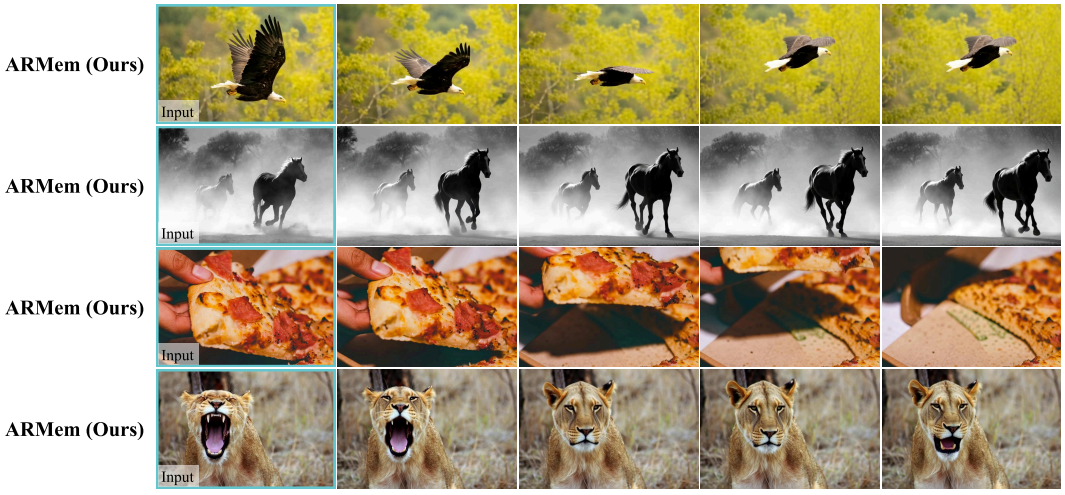


Figure 12: Image-to-video generation results of ARMem on Vbench-I2V. The first column is the image condition input.

Long-Horizon Generation. Fig. 13 presents additional qualitative comparisons with Wan-based baselines for 60-second text-to-video generation. The sampled frames illustrate how subject appearance, scene content, and visual quality evolve over extended generation, complementing the VBench-Long results reported in the main paper.

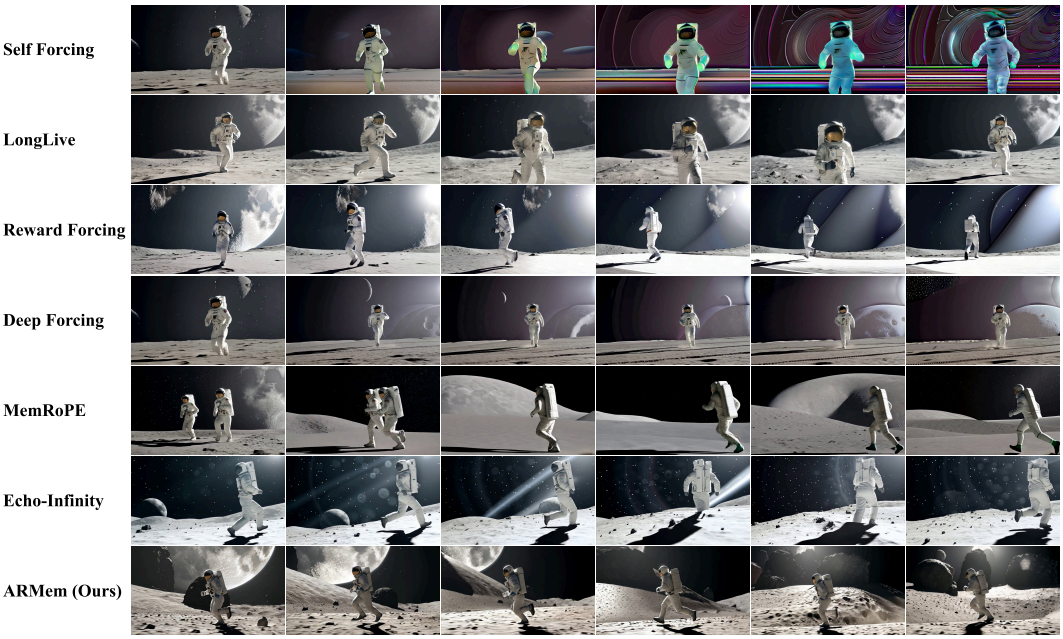


Figure 13: Qualitative comparison of ARMem with Wan-based baselines for 60-second text-to-video generation.

E User Study Details

We conduct a pairwise user study using 20 prompts sampled from MovieGen (Polyak et al., 2024), comparing ARMem with three baselines built on Self Forcing (Huang et al., 2025a). For each prompt, the video generated by ARMem is paired with the corresponding video from each baseline, yielding $3 \times 20 = 60$ video pairs. We recruit 15 participants, each of whom evaluates all 60 pairs, resulting in 900 pairwise judgments in total. Participants are asked to indicate their overall preference based on visual quality and overall consistency. Method identities are hidden, and both the presentation order of the pairs and the left–right placement of the videos are randomized independently for each participant. For each pair, they can prefer either video or indicate that the two are of similar quality.