# MemRoPE: Training-Free Infinite Video Generation via Evolving Memory Tokens

Youngrae Kim*, Qixin Hu*, C.-C. Jay Kuo, and Peter Beerel

University of Southern California, USA

Project Page: `https://memrope.github.io`

**Abstract.** Autoregressive diffusion enables real-time frame streaming, yet existing sliding-window caches discard past context, causing fidelity degradation, identity drift, and motion stagnation over long horizons. Current approaches preserve a fixed set of early tokens as attention sinks, but this static anchor cannot reflect the evolving content of a growing video. We introduce MemRoPE, a training-free framework with two co-designed components. Memory Tokens continuously compress all past keys into dual long-term and short-term streams via exponential moving averages, maintaining both global identity and recent dynamics within a fixed-size cache. Online RoPE Indexing caches unrotated keys and applies positional embeddings dynamically at attention time, ensuring the aggregation is free of conflicting positional phases. These two mechanisms are mutually enabling: positional decoupling makes temporal aggregation well-defined, while aggregation makes fixed-size caching viable for unbounded generation. Extensive experiments validate that MemRoPE outperforms existing methods in temporal coherence, visual fidelity, and subject consistency across minute- to hour-scale generation.

**Keywords:** Long Video Generation · Memory Aware · Training Free

## 1 Introduction

Recent video diffusion models [26,31,40,50] excel at producing cinematic-quality clips, but are inherently limited to fixed lengths in a single forward pass. Beyond mere short-clip synthesis, the broader objective is to simulate evolving visual worlds that maintain persistent identity and temporal coherence over minutes to hours. This long-form generation is essential for powering advanced applications such as continuous world simulation [2,13,19,38,49], cinematic long takes [12,35], and synthetic data generation [20, 36]. Extending them to achieve arbitrarily long, coherent videos requires a fundamentally different generation paradigm.

Autoregressive video diffusion generates frames sequentially from a pretrained model, naturally enabling variable-length generation. CausVid [54] distills a bidirectional DiT into a causal generator via DMD [53] for real-time synthesis, and

* Equal contribution.


**Fig. 1: Training-free long-horizon video generation.** MemRoPE requires ***no additional training*** and supports arbitrary-duration generation with a fixed-size KV cache. We demonstrate a continuous one-hour generation that maintains subject identity and visual fidelity substantially better than existing training-free baselines.

Self Forcing [21] closes the train-inference gap by conditioning on self-generated frames. LongLive [48] introduces streaming long tuning that enables long video training to align training and inference. While these models produce high-quality frames in real-time, they are limited to a finite temporal horizon. $\infty$-RoPE [37,51] resolves the resulting positional extrapolation by block-relative RoPE, enabling generation beyond the 1024-frame limit. However, FIFO eviction (Fig. 2(a)) in the sliding-window KV cache still discards past context as generation proceeds, leading to progressive error accumulation [21, 48, 51, 52].

To retain past context during extended generation, several approaches adapt sliding windows by leveraging the attention sink phenomenon [44], preserving initial frames as fixed anchors [30, 48, 51, 52]. Deep Forcing [52] goes further with Participative Compression, selecting cached tokens by cumulative attention score (Fig. 2(b)). However, as shown in Fig. 3, the selected set rapidly converges to long-persisted tokens, and newly admitted tokens carry disproportionately high scores, causing abrupt visual shifts at each cache update.

Current approaches either evict past context entirely or converge to a stagnant token set that shifts abruptly whenever the cache updates. *None maintains a smoothly evolving representation that adapts as the video unfolds.*

We introduce **MemRoPE**, a training-free infinite long video generation framework with two co-designed components. **Memory Tokens** continuously compress all past keys into dual long-term and short-term streams via exponential moving averages (EMA), maintaining both persistent identity and recent dynamics within a fixed-size cache. This aggregation requires keys free of positional encoding, since merging keys from different timesteps would otherwise mix incompatible rotary phases. **Online RoPE Indexing** enables this by storing keys without positional embedding (Fig. 2(c)) and applying block-relative indices at attention time, which also resolves positional extrapolation. MemRoPE is entirely *training-free* and maintains a constant-size KV cache regardless of the generated duration. As demonstrated by the continuous one-hour example in Fig. 1, this design substantially mitigates long-horizon degradation and preserves temporal consistency over the evaluated duration. Our contributions are as follows:

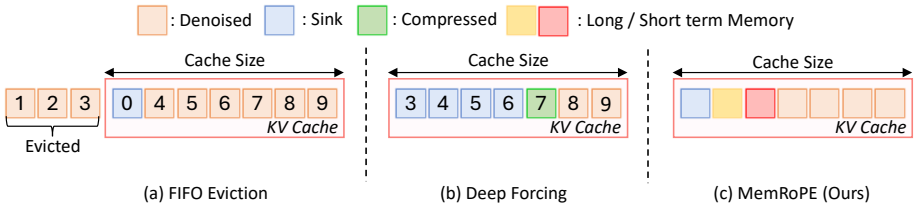



**Fig. 2: KV cache structures for long video generation. (a)** FIFO eviction [21,48,51,54] maintains an initial sink frame while discarding the oldest remaining frames when the cache is full, losing distant context. **(b)** Deep Forcing [52] dedicates over half the cache to static sink tokens and selects a small number of compressed tokens via attention-based importance scoring, which often causes temporal instability in the generated sequence (see Fig. 3). **(c)** MemRoPE preserves a sink frame and manages distinct long- and short-term **memories** (Sec. 4.1). By storing all keys **without RoPE**, it prevents positional interference from corrupting the stored features, thereby enabling stable memory aggregation (Sec. 4.2).

- We propose **MemRoPE**, a training-free framework for long-horizon video generation that jointly addresses context retention and positional extrapolation, enabling hour-scale generation with a duration-independent KV-cache footprint.
- We introduce **Memory Tokens**, an EMA-based dual memory mechanism that continuously compresses generation history into the long-term and short-term representations within a fixed-size cache.
- We present **Online RoPE Indexing**, which decouples content from position in the KV cache, simultaneously enabling clean temporal aggregation and resolving positional extrapolation without post-hoc corrections.
- We demonstrate that MemRoPE outperforms state-of-the-art training-free methods in visual fidelity and consistency across video generation tasks ranging from minute-scale to hour-long sequences.

## 2 Related Work

### 2.1 Autoregressive Video Generation

Autoregressive video generation spans several paradigms: token-based methods [11, 25, 46, 56, 57] quantize video for next-token prediction; chunk-level diffusion [5,23,39] denoises multi-frame chunks; and FramePack [59] and StreamDiT [24] reduce memory via compressed contexts and window attention.

Most relevant to our work is frame-level autoregressive diffusion [8, 21, 28, 30, 48, 54, 64]. CausVid [54] distills a bidirectional DiT into a causal generator via DMD, and Self Forcing [21] closes the train-test gap by conditioning on self-generated frames. Self-Forcing++ [8] extends this to four minutes with teacher-guided error correction, Rolling Forcing [30] introduces a rolling denoising window, and Causal Forcing [64] improves distillation via ODE initialization.

LongLive [48] aligns training with inference-time rollout for 240-second generation. FAR [16] compresses distant frames via aggressive patchification and proposes FlexRoPE for 16× temporal extrapolation; while conceptually related to our dual-rate memory, FAR requires training from scratch, whereas MemRoPE is training-free.

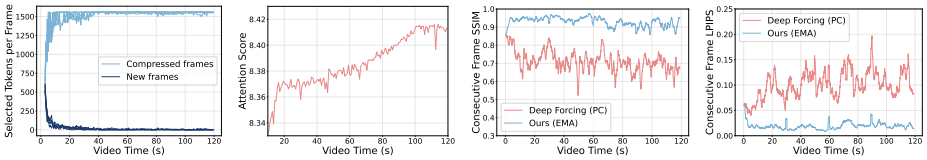
### 2.2 Long-Horizon Context in Long Video Generation

Extending autoregressive generation beyond the training horizon raises two challenges: retaining useful past context and maintaining valid positional encoding.

**Context retention and memory mechanism.** KV cache compression has been widely studied for LLMs through attention-based token selection [29, 62], token merging [42, 61], dynamic budget allocation [41], and correlation-aware eviction [14], but these techniques operate within a fixed-length cache and cannot recover information once evicted [59]. EMA-based temporal aggregation has also been explored as an architectural primitive [32,33], and test-time training layers expand context with expressive hidden states [10], but these require architectural modifications and additional training. Our memory tokens instead operate on the KV cache of an unmodified pretrained model.

To retain context beyond the window, StreamingLLM [44] identified the *attention sink* phenomenon, and video generation methods adopted a similar strategy of preserving initial frames as static anchors [30, 48]. Deep Forcing [52] goes beyond static anchors with Participative Compression, but its selected tokens quickly collapse to a fixed set during generation (Fig. 3). Other approaches introduce long-context streaming tuning [7, 17], memory modules in diffusion-forcing pipelines [4, 18, 60], compressed video histories [3, 43, 58, 59], or geometric priors from 3D representations [27, 45, 55]. These methods highlight the value of long-horizon context, but often rely on additional training, learned modules, growing histories, or domain-specific priors. In contrast, MemRoPE keeps the generator frozen and maintains a fixed-size, continuously evolving memory inside the causal autoregressive KV cache, adapting to new content without increasing memory cost regardless of the total generated video length.

**Positional extrapolation.** Extending RoPE [37] beyond its trained range has been studied in LLMs through position interpolation [6], frequency scaling [1], and their combinations [34], but LoL [9] shows these induce a quality-dynamics tradeoff in video diffusion. In the video domain, $\infty$-RoPE [51] confines indices to the trained range via a block-relative coordinate system, and RIFLEx [63] targets bidirectional settings. However, these methods store keys with RoPE already applied and re-rotate them upon eviction; since RoPE does not distribute over addition (Eq. (5)), averaging such keys from different timesteps produces invalid representations, preventing temporal aggregation (Sec. 4.2). Our Online RoPE Indexing instead stores keys without positional encoding and applies block-relative RoPE for the first time at each attention step, which makes EMA-based aggregation well-defined while resolving positional extrapolation.




**(a)** Token selection **(b)** Attention score **(c)** Consecutive SSIM **(d)** Consecutive LPIPS

**Fig. 3: Failure mode of Participative Compression. (a)** Participative Compression (PC), proposed in Deep Forcing [52], rapidly converges to retaining the same long-persisted tokens in the compressed frames, discarding most newly arriving tokens. **(b)** The few newly admitted tokens carry high attention scores, so each rare cache update exerts a disproportionately strong influence on generation. **(c, d)** This causes frame-to-frame instability: consecutive SSIM drops and LPIPS spikes indicate abrupt visual shifts whenever the cache content changes. Our EMA memory evolves continuously, maintaining smooth transitions.

## 3 Background

**Autoregressive inference.** Autoregressive video diffusion generates video one frame chunk at a time [8,9,21,30,48,52,54]. Given the KV cache of all previously generated frames $\{x_1, \ldots, x_{t-1}\}$, the transformer $\mathcal{G}_\theta$ denoises a noisy input $x_t^{(\sigma)}$ through a small number of diffusion steps:

$$\hat{x}_t = \mathcal{G}_\theta\big(x_t^{(\sigma)},\ \sigma,\ \mathbf{K}_{<t},\ \mathbf{V}_{<t}\big). \tag{1}$$

The resulting key-value pairs $(W_K \hat{x}_t,\ W_V \hat{x}_t)$ are then appended to the KV cache as conditioning for the next step. Because the model uses causal attention, keys and values of past frames can be cached and reused in subsequent generation steps, avoiding redundant computation. Since storing all past KV states is infeasible for long video generations, a sliding window retains only the most recent frames, evicting older ones via FIFO (Fig. 2(a)) [8, 9, 21, 30, 48, 52, 54].

**Rotary Position Embedding.** Positional ordering is often encoded via 3D-RoPE in video generation [37, 40], which incorporates the temporal and spatial coordinates of query and key tokens using the following relation:

$$\tilde{q}_i = R_i\, q_i, \qquad \tilde{k}_j = R_j\, k_j, \tag{2}$$

where $R_i \in \mathbb{R}^{d\times d}$ is a rotation matrix encoding the spatiotemporal coordinates of token $i$. The key property is that the dot product $\tilde{q}_i^\top \tilde{k}_j = q_i^\top R_{i-j}\, k_j$ depends only on the *relative* distance $i-j$, enabling the model to learn position-invariant attention patterns. In standard practice, RoPE is applied at cache time: each key is stored as $\tilde{k}_j = R_j\, k_j$, permanently binding it to its exact spatiotemporal position $j$. Value states are not rotated and are cached as-is.

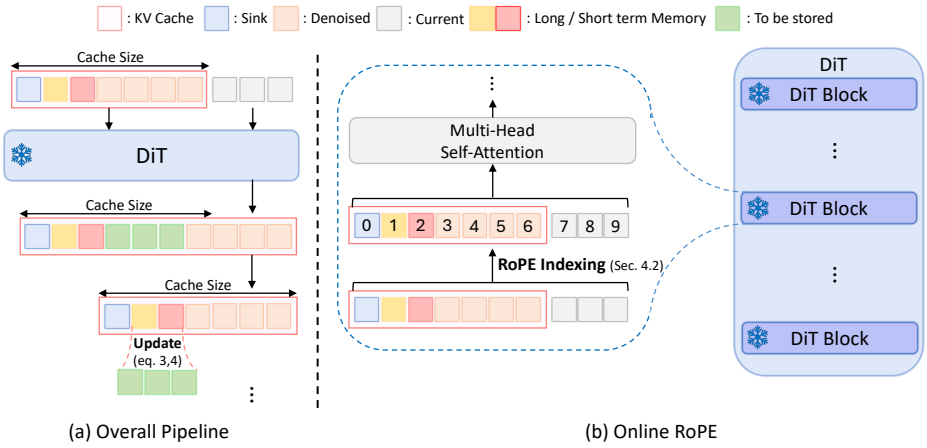



**Fig. 4: Overview of MemRoPE. (a)** At each autoregressive step, the three-tier KV cache (sink, memory, and recent tokens) is concatenated with the current noisy chunk and fed into the DiT. When the local window is full, the oldest frames are absorbed into the long- and short-term memory tokens via dual EMA updates. **(b)** All cached keys are stored without RoPE, enabling temporal aggregation for memory tokens. At attention time, contiguous block-relative indices are assigned to the full sequence, and RoPE is applied on the fly, ensuring that indices never exceed the training range.

## 4 Method

We propose **MemRoPE**, a training-free framework that is implemented through two co-designed mechanisms: **Memory Tokens** (Sec. 4.1) and **Online RoPE Indexing** (Sec. 4.2). Fig. 4 provides a comprehensive visual overview.

### 4.1 Memory Tokens

**Why existing cache management fails.** FIFO eviction irrecoverably discards past frames, causing identity drift, scene inconsistency, and motion stagnation over long horizons [21,48,51,52]. To mitigate this, several methods [30,48,52] retain the initial frames as *attention sinks* alongside the sliding window, inspired by the attention sink phenomenon in LLMs [44]. This is well motivated, as early frames lie within the training horizon and tend to be the highest quality in DMD-distilled models; however, these sinks are frozen at generation start and cannot reflect later changes in scene content or subject identity.

Deep Forcing [52] extends this with *Participative Compression* (PC), which dynamically selects cached tokens by cumulative attention score. However, the selection mechanism suffers from a self-reinforcing bias: tokens that have persisted in the cache accumulate higher attention scores, making them increasingly likely to be retained regardless of their current relevance. As shown in Fig. 3(a), the selected set rapidly converges to a stagnant, fixed group of long-persisted to-

kens. Moreover, when a new token is finally admitted, it carries a disproportionately high attention score relative to the entrenched set (Fig. 3(b)), producing abrupt, highly noticeable visual shifts at each cache update (Fig. 3(c,d)).

Across all existing methods, preserved context is either a static snapshot of early frames or subject to discrete token selection that converges to a stale set. We instead propose a *continuously evolving* representation.

**Dual EMA memory.** We propose Memory Tokens: a fixed-size buffer evolving throughout generation via exponential moving averages (EMA). As chunks exit the window, we apply spatial average pooling over the $H \times W$ spatial token grid of each frame to produce a single key or value vector. The pooled key $\bar{k}_{\text{new}}$ then updates two parallel streams:

$$\mu_L^{(t)} = (1 - \alpha_L)\, \mu_L^{(t-1)} + \alpha_L\, \bar{k}_{\text{new}}, \tag{3}$$

$$\mu_S^{(t)} = (1 - \alpha_S)\, \mu_S^{(t-1)} + \alpha_S\, \bar{k}_{\text{new}}, \tag{4}$$

where $\alpha_L \ll \alpha_S$: the long-term stream accumulates the full generation history into a stable representation, while the short-term stream tracks recent dynamics. The same update applies to value states. Memory size is constant regardless of video length, allowing the generation duration to extend without increasing the KV-cache budget.

### 4.2 Online RoPE Indexing

The memory token updates in Eqs. (3) and (4) average keys from different timesteps. However, conventional KV caches store keys with RoPE already applied: $\tilde{k}_j = R_j k_j$. Averaging such keys mixes incompatible rotary phases:

$$\alpha\, R_j k_j + (1{-}\alpha)\, R_{j'} k_{j'} \;\neq\; R_\phi\big(\alpha\, k_j + (1{-}\alpha)\, k_{j'}\big) \tag{5}$$

for any rotation $R_\phi$, since RoPE does not distribute over addition. As a consequence, the averaged key no longer corresponds to any valid temporal position, breaking the relative-position structure that attention relies on.

We resolve this aggregation barrier together with the positional extrapolation problem through a single design change: *store all keys without RoPE, and dynamically assign block-relative temporal indices at each attention step.*

**Position-free caching.** Every individual key is computed in its raw, unrotated form without positional embeddings before entering the cache:

$$k_j^{\text{cache}} = k_j \qquad \text{(no RoPE)}. \tag{6}$$

EMA updates (Eqs. (3) and (4)) aggregate these position-free keys in the cache, and the resulting memory tokens operate as valid key vectors that can dynamically receive any positional index at attention time.

**Block-relative index assignment.** At each generation step, we treat the cache as a single block and assign a contiguous index map $\phi$ starting from zero:

$$\underbrace{[0, \ldots, S{-}1]}_{\text{sink}} \ \underbrace{[S, \ldots, S{+}2M{-}1]}_{\text{memory}} \ \underbrace{[S{+}2M, \ldots, S{+}2M{+}L{-}1]}_{\text{local}}, \tag{7}$$

where $S$, $M$, $L$ are the sink, memory, and local window sizes. The current query receives indices starting from $S{+}2M{+}L$. RoPE is then applied on the fly: $\tilde{k}_j = R_{\phi(j)}\, k_j^{\text{cache}}$ for every cached key, and $\tilde{q}_i = R_{\phi(i)}\, q_i$ for every current query. Value states are not rotated. Since all indices are recomputed from zero at every generation step, no position ever exceeds the total cache size $C = S{+}2M{+}L$, which remains within the range seen during training.

**Relationship to block-relative RoPE.** $\infty$-RoPE [51] re-anchors cached keys backward at each step to keep the newest block at the maximum trained index, requiring per-step phase rotations on keys already stored with RoPE. Our Online RoPE Indexing instead stores keys without RoPE and applies positional encoding once at attention time, eliminating per-step rotations and enabling temporal aggregation via EMA, which is ill-defined over keys stored with conflicting rotary phases (Eq. (5)). We map the cache compactly to $[0, C{-}1]$ rather than $[f_{\text{limit}} - C, f_{\text{limit}}]$, where $f_{\text{limit}}$ is the maximum position index seen during pre-training, so that the sink tokens always receive the lowest indices and memory tokens occupy consistent positions in the sequence.

**Relationship to Dynamic RoPE.** Rolling Forcing [30] stores raw keys for its sink tokens and reapplies RoPE dynamically at attention time, an approach conceptually close to our Online RoPE Indexing. However, this is limited to the static sink frames; keys in the rolling window are still stored with monotonically increasing RoPE indices, which both precludes temporal aggregation and reintroduces positional extrapolation as the generated sequence grows. We generalize position-free storage to the entire cache, including the sink, memory, and local window alike. This makes EMA aggregation well-defined across all cache slots while keeping all indices within the training range.

### 4.3 Three-Tier Cache

The complete MemRoPE cache at step $t$ is:

$$\mathbf{K}^{(t)} = \big[\, \underbrace{\mathbf{K}_{\text{sink}}}_{S} \,\|\, \underbrace{\mathbf{K}_{\text{mem}}}_{2M} \,\|\, \underbrace{\mathbf{K}_{\text{local}}}_{L} \,\big], \tag{8}$$

with the value cache $\mathbf{V}^{(t)}$ structured identically. This ordering mirrors the temporal structure of the video: $\mathbf{K}_{\text{sink}}$ anchors the earliest high-quality frames, $\mathbf{K}_{\text{mem}} = [\mu_L \| \mu_S]$ summarizes the evolving history via dual EMA, and $\mathbf{K}_{\text{local}}$ captures the recent sliding window. All keys are stored position-free (Fig. 2); block-relative RoPE indices (Eq. (7)) are applied at every attention call.

The full inference procedure is given in Algorithm 1.

**Algorithm 1** MemRoPE Inference

**Require:** Diffusion Transformer $\mathcal{G}_\theta$, denoising steps $N$, EMA rates $\alpha_L, \alpha_S$
**Notation:** $x_t^{(s)}$: latent of chunk $t$ after $s$ denoising steps; $\hat{x}_t = x_t^{(N)}$: final denoised chunk

$\hat{x}_0 \leftarrow \mathcal{G}_\theta(x_0^{(0)}, N)$; $\mathbf{K}_{\text{sink}}, \mathbf{V}_{\text{sink}} \leftarrow W_K\hat{x}_0,\ W_V\hat{x}_0$ ▷ no RoPE
$\mu_L^k, \mu_L^v, \mu_S^k, \mu_S^v \leftarrow \mathbf{0}$; $\mathbf{K}_{\text{local}}, \mathbf{V}_{\text{local}} \leftarrow \emptyset$
**for** $t = 1, 2, \ldots$ **do** ▷ chunk loop
  **for** $s = 0, \ldots, N-1$ **do** ▷ denoising steps
    $q_s, k_s, v_s \leftarrow W_Q x_t^{(s)},\ W_K x_t^{(s)},\ W_V x_t^{(s)}$
    $\mathbf{K} \leftarrow [\mathbf{K}_{\text{sink}} \| \mu_L^k \| \mu_S^k \| \mathbf{K}_{\text{local}} \| k_s]$ ▷ all without RoPE
    $\mathbf{V} \leftarrow [\mathbf{V}_{\text{sink}} \| \mu_L^v \| \mu_S^v \| \mathbf{V}_{\text{local}} \| v_s]$
    $\phi_K \leftarrow [0, \ldots, |\mathbf{K}|-1]$; $\phi_Q \leftarrow [|\mathbf{K}|-|q_s|, \ldots, |\mathbf{K}|-1]$
    $x_t^{(s+1)} \leftarrow \text{DiT-Block}(R_{\phi_Q} q_s,\ R_{\phi_K}\mathbf{K},\ \mathbf{V})$ ▷ Online RoPE
  **end for**
  $\hat{x}_t \leftarrow x_t^{(N)}$ ▷ denoised chunk
  Append $W_K\hat{x}_t$, $W_V\hat{x}_t$ to $\mathbf{K}_{\text{local}}$, $\mathbf{V}_{\text{local}}$ ▷ no RoPE
  **if** $|\mathbf{K}_{\text{local}}| > L$ **then** ▷ evict & update memory
    $\bar{k}, \bar{v} \leftarrow \text{SpatialPool}(\mathbf{K}_{\text{local}}[0]),\ \text{SpatialPool}(\mathbf{V}_{\text{local}}[0])$ ▷ average over $H \times W$
    $\mu_L^k \leftarrow (1-\alpha_L)\,\mu_L^k + \alpha_L\,\bar{k}$; $\mu_L^v \leftarrow (1-\alpha_L)\,\mu_L^v + \alpha_L\,\bar{v}$
    $\mu_S^k \leftarrow (1-\alpha_S)\,\mu_S^k + \alpha_S\,\bar{k}$; $\mu_S^v \leftarrow (1-\alpha_S)\,\mu_S^v + \alpha_S\,\bar{v}$
    Evict oldest chunk from $\mathbf{K}_{\text{local}}$, $\mathbf{V}_{\text{local}}$
  **end if**
**end for**

## 5 Experiments

### 5.1 Setup

**Implementation details.** We build on the Wan2.1-T2V-1.3B architecture [40] and evaluate MemRoPE on two base models: Self-Forcing [21] and LongLive [48]. Since MemRoPE is training-free, it can be plugged into any autoregressive video diffusion model without modification. Video is generated autoregressively in chunks of 3 latent frames with a 4-step denoising schedule at timesteps $\{1000, 750, 500, 250\}$. The three-tier cache consists of $S = 3$ sink tokens, $M = 1$ memory token per stream (long-term and short-term), and a local window of $L = 4$ frames, with EMA decay rates $\alpha_L = 0.01$ and $\alpha_S = 0.1$. The KV cache is updated only after the final denoising step of each chunk, ensuring that only fully denoised keys and values enter the cache. Deep Forcing [52] and $\infty$-RoPE [51] are also training-free methods; we apply them on both base models under the same protocol for fair comparison.

**Evaluation protocol.** We generate videos from text prompts sampled from MovieGenBench [35] at $480 \times 832$ resolution and 16 fps. For 120 and 240 seconds, we use 128 prompts; for 480 seconds, 20 randomly sampled prompts; for 1 hour, 10 randomly sampled prompts; and for ablation studies, 20 randomly sampled prompts at 60 seconds. Following Deep Forcing [52] and Self-Forcing [21], all

**Table 1: Quantitative comparison on long video generation.** We report VBench-Long [22] metrics. Within each base model group, best results are bold with darker background and second best are with lighter background. $\Delta$ denotes the average improvement over the respective base model.

| Method | Aesthetic Quality ↑ | Background Consistency ↑ | Imaging Quality ↑ | Motion Smoothness ↑ | Subject Consistency ↑ | Temporal Flickering ↑ | Average ↑ | $\Delta$ |
|---|---|---|---|---|---|---|---|---|
| *120 seconds* | | | | | | | | |
| *Results on Self-Forcing* [21] | | | | | | | | |
| Self-Forcing [21] | 55.11 | 95.41 | 65.95 | 97.17 | 95.35 | 96.53 | 84.25 | – |
| Deep Forcing [52] | **56.86** | 94.58 | 65.02 | 97.07 | 94.17 | 95.33 | 83.84 | −0.41 |
| ∞-RoPE [51] | 52.82 | **95.66** | 61.07 | **98.48** | 96.30 | **97.50** | 83.64 | −0.61 |
| **MemRoPE (Ours)** | 56.77 | 95.54 | **68.44** | 97.90 | **96.37** | 96.33 | **85.23** | **+0.98** |
| *Results on LongLive* [48] | | | | | | | | |
| LongLive [48] | 57.21 | 95.96 | 67.13 | 98.56 | 97.06 | 97.12 | 85.51 | – |
| Deep Forcing [52] | 59.16 | 96.13 | 68.07 | 98.51 | 96.79 | 97.38 | 86.01 | +0.50 |
| ∞-RoPE [51] | 57.87 | **96.46** | 64.84 | **99.00** | 97.29 | **98.00** | 85.58 | +0.07 |
| **MemRoPE (Ours)** | **59.25** | 96.29 | **69.43** | 98.73 | **97.54** | 97.47 | **86.45** | **+0.94** |
| *240 seconds* | | | | | | | | |
| *Results on Self-Forcing* [21] | | | | | | | | |
| Self-Forcing [21] | 51.00 | 95.01 | 61.52 | 98.18 | 95.83 | 96.72 | 83.04 | – |
| Deep Forcing [52] | 52.20 | 93.91 | 59.50 | 96.72 | 92.28 | 95.38 | 81.66 | −1.38 |
| ∞-RoPE [51] | 50.51 | **95.52** | 58.81 | **98.47** | 96.24 | **97.48** | 82.84 | −0.20 |
| **MemRoPE (Ours)** | **55.54** | 95.45 | **67.77** | 97.93 | **96.30** | 96.39 | **84.89** | **+1.85** |
| *Results on LongLive* [48] | | | | | | | | |
| LongLive [48] | 56.70 | 95.80 | 66.90 | 98.52 | 97.02 | 97.04 | 85.33 | – |
| Deep Forcing [52] | 57.75 | 96.03 | 66.48 | 98.50 | 96.65 | 97.47 | 85.48 | +0.15 |
| ∞-RoPE [51] | 57.47 | 96.34 | 63.38 | **98.99** | 97.11 | **97.94** | 85.21 | −0.12 |
| **MemRoPE (Ours)** | **58.90** | **96.39** | **68.93** | 98.59 | **97.37** | 97.13 | **86.22** | **+0.89** |

prompts are refined using Qwen2.5-7B-Instruct [47] before generation. We report metrics from VBench-Long [22], and further validate with a user study following the 2AFC protocol [52] and VLM-based evaluation using Gemini 3.1-Pro [15].

### 5.2 Results

**Quantitative results.** Tab. 1 summarizes the quantitative comparison at 120 and 240 seconds on two base models (*i.e.*, Self-Forcing, LongLive). MemRoPE achieves the highest average score across all settings, consistently improving over both base models. In contrast, Deep Forcing and ∞-RoPE degrade the Self-Forcing base at both durations, with Deep Forcing falling 1.38 points below at 240 seconds. On LongLive, Deep Forcing provides moderate gains at 120 seconds but nearly vanishes at 240 seconds, while ∞-RoPE remains near or slightly below the baseline. MemRoPE maintains stable improvement at both durations. Per-metric analysis reveals that MemRoPE leads in Aesthetic Quality, Imaging Quality, and Subject Consistency—the dimensions most sensitive to long-range context retention—while ∞-RoPE tends to score higher on Motion Smoothness and Temporal Flickering, which primarily measure local frame-to-frame stability. This pattern suggests that ∞-RoPE preserves short-range smoothness but lacks the evolving memory needed to maintain visual fidelity and identity over longer horizons.


**Fig. 5: Qualitative comparison on 2-minute video generation**. MemRoPE maintains subject identity and background consistency throughout, whereas baselines exhibit progressive degradation, including structural collapse and color corruption.

**Qualitative results.** Fig. 5 presents frame-by-frame comparisons over 2-minute generations on both base models. On Self-Forcing, the base model degrades to the point where subjects become unrecognizable by the end of generation. Deep Forcing retains rough shapes but loses all fine detail, while ∞-RoPE suffers from severe color instability with no meaningful preservation of the original appearance. On LongLive, both the base model and Deep Forcing fail to maintain subject consistency—the number of puppies fluctuates from three to four or five across frames, and Deep Forcing additionally shifts the overall color tone. ∞-RoPE exhibits the most aggressive error accumulation, with the background collapsing into unrecognizable forms. MemRoPE preserves subject count, identity, and background consistency throughout on both base models, producing the most temporally stable results across both prompts.

**Table 2: Ultra-long video generation.** We report VBench-Long [22] metrics. Within each base model group, best results are bold with darker background.

| Method | Aesthetic Quality ↑ | Background Consistency ↑ | Imaging Quality ↑ | Motion Smoothness ↑ | Subject Consistency ↑ | Temporal Flickering ↑ | Average ↑ |
|---|---|---|---|---|---|---|---|
| *480 seconds* | | | | | | | |
| *Results on Self-Forcing* [21] | | | | | | | |
| ∞-RoPE [51] | 48.35 | 95.45 | 56.00 | **98.59** | 96.46 | **97.71** | 82.09 |
| **MemRoPE (Ours)** | **56.12** | **95.65** | **68.23** | 97.96 | **96.87** | 96.42 | **85.21** |
| *Results on LongLive* [48] | | | | | | | |
| ∞-RoPE [51] | 56.77 | **96.25** | 61.32 | **99.05** | 97.16 | **97.95** | 84.75 |
| **MemRoPE (Ours)** | **57.96** | 96.12 | **67.52** | 98.66 | **97.37** | 97.23 | **85.81** |
| *1 hour* | | | | | | | |
| *Results on LongLive* [48] | | | | | | | |
| ∞-RoPE [51] | 60.87 | 96.42 | 70.42 | 99.01 | 97.71 | 97.62 | 87.01 |
| **MemRoPE (Ours)** | **63.05** | **96.77** | **72.71** | **99.07** | **98.18** | **98.14** | **87.99** |

**Ultra-long generation.** Tab. 2 extends the comparison to 480 seconds and 1 hour. Besides context retention, ultra-long video generation is also constrained by the length of RoPE. The maximum generation length of both Self-Forcing and LongLive is 4 minutes and 15 seconds, limited by the 1024-frame latent sequence. Among our baselines, only ∞-RoPE resolves this through a block-relative coordinate system that confines indices to the trained range. Online RoPE Indexing similarly confines indices to the trained range, enabling both MemRoPE and ∞-RoPE to generate beyond this limit.

At 480 seconds, MemRoPE outperforms ∞-RoPE by over 3 points on Self-Forcing and 1 point on LongLive in average score. Consistent with shorter durations, the gap is driven by Aesthetic Quality and Imaging Quality, while ∞-RoPE scores higher on Motion Smoothness and Temporal Flickering on both base models. At 1 hour on LongLive, MemRoPE leads across all six metrics with an average of 87.99 versus 87.01, demonstrating that the dual EMA memory scales gracefully to hour-length generation.

**Trends across durations.** We compare VBench-Long averages of MemRoPE and ∞-RoPE [51] on Self-Forcing [21] at 30, 60, 120, 240, and 480 seconds, using the 20-prompt subset from the 480-second evaluation throughout. As shown in Fig. 6, both degrade with duration, but MemRoPE's slope is significantly flatter and the gap widens monotonically, confirming that the dual-stream EMA preserves long-range context that ∞-RoPE discards.

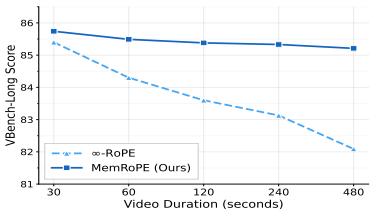



**Fig. 6: Score vs. duration on Self-Forcing [21].** The gap widens as videos grow longer.

**Table 3: (a)** User study results following the 2AFC protocol [52]. Values represent percentage of votes favoring MemRoPE over each baseline. **(b)** Visual stability scored by Gemini 3.1-Pro [15] on 120-second videos.

**(a)** User study (% favoring Ours).

| Baseline | Color Cons. | Subject Cons. | Bg. Cons. | Text Align. | Motion Smo. | Overall Pref. |
|---|---|---|---|---|---|---|
| Self-Forcing [21] | 93.0 | 93.0 | 97.4 | 91.6 | 95.7 | 98.3 |
| Rolling Forcing [30] | 90.4 | 77.4 | 79.1 | 80.7 | 88.7 | 82.6 |
| *Results on LongLive* [48] | | | | | | |
| LongLive [48] | 81.6 | 70.2 | 83.3 | 66.1 | 71.9 | 71.1 |
| Deep Forcing [52] | 79.0 | 75.2 | 81.9 | 70.5 | 72.4 | 72.6 |
| ∞-RoPE [51] | 81.3 | 76.4 | 77.6 | 74.5 | 73.8 | 70.1 |

**(b)** Visual stability (VLM).

| Method | Visual Stability |
|---|---|
| Self-Forcing [21] | 1.55 |
| Rolling Forcing [30] | 3.40 |
| *Results on LongLive* [48] | |
| LongLive [48] | 4.10 |
| Deep Forcing [52] | 3.90 |
| ∞-RoPE [51] | 4.05 |
| **MemRoPE (Ours)** | **4.15** |

**Table 4: Ablation studies.** Combining both components, and both memory streams, yields the highest overall average. Both are from independent runs.

**(a)** Component isolation.

| Memory Tokens | Online RoPE Indexing | Aesthetic Quality ↑ | Imaging Quality ↑ | Subject Consistency ↑ | Avg ↑ |
|---|---|---|---|---|---|
| | | 57.93 | 67.41 | 96.10 | 85.71 |
| | ✓ | 58.09 | 65.29 | **97.40** | 85.48 |
| ✓ | | 58.50 | 68.45 | 97.35 | 86.08 |
| ✓ | ✓ | **58.76** | **69.01** | 97.34 | **86.42** |

**(b)** Memory-stream ablation.

| Long-term Memory ($\mu_L$) | Short-term Memory ($\mu_S$) | Aesthetic Quality ↑ | Imaging Quality ↑ | Subject Consistency ↑ | Avg ↑ |
|---|---|---|---|---|---|
| | | 58.09 | 65.29 | 97.40 | 85.48 |
| ✓ | | 58.69 | 66.31 | **97.62** | 85.84 |
| | ✓ | 58.73 | 66.63 | 97.61 | 85.95 |
| ✓ | ✓ | **58.76** | **67.09** | 97.61 | **85.97** |

**User study.** To complement the automated metrics, we conducted a user study with 30 participants following the 2AFC protocol [52]; all data was collected anonymously and handled in accordance with ethical guidelines. Each participant compared 20 pairs of 120-second videos (MemRoPE vs. a baseline) across six perceptual dimensions listed in Tab. 3(a). Participants consistently preferred MemRoPE across all aspects, corroborating the quantitative results.

**VLM evaluation.** We further assess long-horizon visual stability using the advanced multimodal vision-language model Gemini 3.1-Pro [15]. Following the protocol of Self-Forcing++ [8] and Deep Forcing [52], we prompt the VLM to score each 120-second generated video in terms of exposure stability and degradation. The prompt and corresponding examples can be found in the supplementary material. As shown in Tab. 3(b), MemRoPE achieves the highest stability scores, consistent with both the automated metrics and the user study.

**Ablation studies.** Tab. 4(a) isolates the two components. Online RoPE Indexing alone improves subject consistency, while Memory Tokens alone provide stronger gains in aesthetic and imaging quality. Combining the two yields the best average score, indicating that position-free indexing and evolving memory are most effective when used together. Full results are reported in the supplementary material.

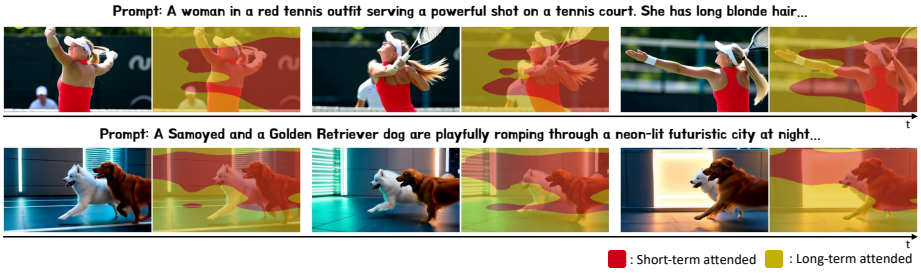



**Fig. 7: Spatial attention distribution between the dual memory streams.** Yellow indicates higher attention to long-term memory. Red indicates higher attention to short-term memory. Temporally consistent regions attend more to long-term memory, while rapidly changing regions attend more to short-term memory.

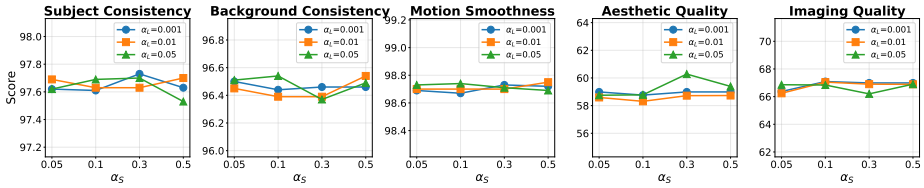



**Fig. 8:** Sensitivity to EMA decay rates. Each subplot shows one VBench metric as a function of short-term decay $\alpha_S$ for three distinct long-term decay values $\alpha_L$. All metrics remain stable across the entire evaluated parameter range, with the average score varying by less than 0.7 across all tested configurations.

Tab. 4(b) further isolates the long- and short-term memory streams. Any memory configuration improves over the no-memory baseline, with the largest gain observed in Imaging Quality (+1.8 points). Combining both streams yields the best Imaging Quality, although the margin over short-term memory alone is small. This suggests that the two streams capture complementary information: $\mu_L$ summarizes the stable aspects of the scene while $\mu_S$ reflects recent changes. Fig. 7 supports this: temporally consistent regions attend more to long-term memory, while rapidly changing regions rely more on short-term memory.

**EMA decay rate sensitivity.** MemRoPE introduces two hyperparameters: the long-term decay $\alpha_L$ and the short-term decay $\alpha_S$. Fig. 8 plots five VBench metrics across all 12 combinations of $\alpha_L \in \{0.001, 0.01, 0.05\}$ together with $\alpha_S \in \{0.05, 0.1, 0.3, 0.5\}$. The average score varies by less than 0.7 across the entire grid, indicating that MemRoPE is robust to the choice of EMA hyperparameters. We use $\alpha_L$=0.01 and $\alpha_S$=0.1 for all other experiments.

**EMA vs. attention-based cache management.** Deep Forcing's participative compression [52] is the only other training-free method that goes beyond

**Table 5: Inference overhead on an NVIDIA A6000 GPU.** Latency is measured per denoising step, and memory denotes peak GPU memory usage.

| Method | Latency (ms/step) | Peak Memory (GB) |
|---|---|---|
| LongLive [48] | 669.60 | 20.97 |
| + MemRoPE | +3.83 | +0.07 |

static sinks for cache management. As shown in Fig. 3(c,d), its discrete token selection produces a cache that rarely updates yet causes abrupt visual shifts when it does, whereas our continuous EMA aggregation absorbs every evicted frame smoothly, maintaining stable transitions throughout generation.

**Efficiency.** Since MemRoPE operates entirely on the existing KV cache without modifying the base model, its computational footprint remains minimal. The additional cost consists of elementwise RoPE rotations on cached keys and dual EMA updates during cache eviction. On a single NVIDIA A6000 GPU, this amounts to only 3.83 ms per denoising step and 0.07 GB of peak GPU memory over LongLive, corresponding to approximately 0.57% latency and 0.33% memory overhead (Tab. 5).

## 6 Conclusion

We present MemRoPE, a training-free framework for long video generation from autoregressive diffusion models. Memory Tokens smoothly preserve evolving context into dual EMA streams, while Online RoPE Indexing makes this well-defined by applying relative positional indices at attention time, enabling unbounded generation within a fixed-size cache. Comprehensive experiments consistently validate the effectiveness of MemRoPE across all tested durations and evaluation protocols. Our results suggest that the key to long-horizon coherence lies not in retaining more frames, but in remembering them better.

**Limitations.** As a training-free method built on a frozen checkpoint, MemRoPE's per-frame quality is bounded by the base model. The EMA aggregation is also lossy by design, which may limit precise recall of distant content. Incorporating learned memory compression could address this in future work.

## References


1. bloc97: NTK-aware scaled RoPE allows LLaMA models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation. Reddit post (2023), https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/
2. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al.: Video generation models as world simulators. OpenAI Blog **1**(8), 1 (2024)
3. Cai, S., Yang, C., Zhang, L., Guo, Y., Xiao, J., Yang, Z., Xu, Y., Yang, Z., Yuille, A., Guibas, L., Agrawala, M., Jiang, L., Wetzstein, G.: Mixture of contexts for long video generation. In: ICLR (2026)
4. Chen, B., Martí Monsó, D., Du, Y., Simchowitz, M., Tedrake, R., Sitzmann, V.: Diffusion forcing: Next-token prediction meets full-sequence diffusion. NeurIPS (2024)
5. Chen, G., Lin, D., Yang, J., Lin, C., Zhu, J., Fan, M., Zhang, H., Chen, S., Chen, Z., Ma, C., et al.: Skyreels-v2: Infinite-length film generative model. arXiv preprint arXiv:2504.13074 (2025)
6. Chen, S., Wong, S., Chen, L., Tian, Y.: Extending context window of large language models via positional interpolation. arXiv preprint arXiv:2306.15595 (2023)
7. Chen, S., Wei, C., Sun, S., Nie, P., Zhou, K., Zhang, G., Yang, M.H., Chen, W.: Context forcing: Consistent autoregressive video generation with long context. arXiv preprint arXiv:2602.06028 (2026)
8. Cui, J., Wu, J., Li, M., Yang, T., Li, X., Wang, R., Bai, A., Ban, Y., Hsieh, C.J.: Self-forcing++: Towards minute-scale high-quality video generation. arXiv preprint arXiv:2510.02283 (2025)
9. Cui, J., Wu, J., Li, M., Yang, T., Li, X., Wang, R., Bai, A., Ban, Y., Hsieh, C.J.: Lol: Longer than longer, scaling video generation to hour. arXiv preprint arXiv:2601.16914 (2026)
10. Dalal, K., Koceja, D., Xu, J., Zhao, Y., Han, S., Cheung, K.C., Kautz, J., Choi, Y., Sun, Y., Wang, X.: One-minute video generation with test-time training. In: CVPR (2025)
11. Deng, H., Pan, T., Diao, H., Luo, Z., Cui, Y., Lu, H., Shan, S., Qi, Y., Wang, X.: Autoregressive video generation without vector quantization. In: ICLR (2025)
12. Elmoghany, M., Rossi, R., Yoon, S., Mukherjee, S., Bakr, E.M., Mathur, P., Wu, G., Lai, V.D., Lipka, N., Zhang, R., et al.: A survey on long-video storytelling generation: architectures, consistency, and cinematic quality. In: ICCV (2025)
13. Feng, R., Zhang, H., Yang, Z., Xiao, J., Shu, Z., Liu, Z., Zheng, A., Huang, Y., Liu, Y., Zhang, H.: The matrix: Infinite-horizon world generation with real-time moving control. arXiv preprint arXiv:2412.03568 (2024)
14. Ghadia, R., Kumar, A., Jain, G., Nair, P., Das, P.: Dialogue without limits: Constant-sized kv caches for extended responses in llms. arXiv preprint arXiv:2503.00979 (2025)
15. Google DeepMind: Gemini 3.1 pro model card (2026), https://deepmind.google/models/model-cards/gemini-3-1-pro/
16. Gu, Y., Mao, W., Shou, M.Z.: Long-context autoregressive video modeling with next-frame prediction. arXiv preprint arXiv:2503.19325 (2025)
17. Guo, Y., Yang, C., Yang, Z., Ma, Z., Lin, Z., Yang, Z., Lin, D., Jiang, L.: Long context tuning for video generation. In: ICCV (2025)
18. Henschel, R., Khachatryan, L., Poghosyan, H., Hayrapetyan, D., Tadevosyan, V., Wang, Z., Navasardyan, S., Shi, H.: Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In: CVPR (2025)

19. Hong, Y., Mei, Y., Ge, C., Xu, Y., Zhou, Y., Bi, S., Hold-Geoffroy, Y., Roberts, M., Fisher, M., Shechtman, E., et al.: Relic: Interactive video world model with long-horizon memory. arXiv preprint arXiv:2512.04040 (2025)
20. Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., Corrado, G.: Gaia-1: A generative world model for autonomous driving. arXiv preprint arXiv:2309.17080 (2023)
21. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. In: NeurIPS (2025)
22. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: CVPR (2024)
23. Jin, Y., Sun, Z., Li, N., Xu, K., Xu, K., Jiang, H., Zhuang, N., Huang, Q., Song, Y., MU, Y., Lin, Z.: Pyramidal flow matching for efficient video generative modeling. In: ICLR (2025)
24. Kodaira, A., Hou, T., Hou, J., Georgopoulos, M., Juefei-Xu, F., Tomizuka, M., Zhao, Y.: Streamdit: Real-time streaming text-to-video generation. arXiv preprint arXiv:2507.03745 (2025)
25. Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Schindler, G., Hornung, R., Birodkar, V., Yan, J., Chiu, M.C., et al.: Videopoet: A large language model for zero-shot video generation. In: ICML (2024)
26. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
27. Li, R., Torr, P., Vedaldi, A., Jakab, T.: Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In: ICCV (2025)
28. Li, W., Pan, W., Luan, P.C., Gao, Y., Alahi, A.: Stable video infinity: Infinite-length video generation with error recycling. In: ICLR (2026)
29. Li, Y., Huang, Y., Yang, B., Venkitesh, B., Locatelli, A., Ye, H., Cai, T., Lewis, P., Chen, D.: Snapkv: Llm knows what you are looking for before generation. NeurIPS (2024)
30. Liu, K., Hu, W., Xu, J., Shan, Y., Lu, S.: Rolling forcing: Autoregressive long video diffusion in real time. In: ICLR (2026)
31. Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., Yuan, Z., Huang, Y., Sun, H., Gao, J., et al.: Sora: A review on background, technology, limitations, and opportunities of large vision models. arXiv preprint arXiv:2402.17177 (2024)
32. Ma, X., Yang, X., Xiong, W., Chen, B., Yu, L., Zhang, H., May, J., Zettlemoyer, L., Levy, O., Zhou, C.: Megalodon: Efficient llm pretraining and inference with unlimited context length. NeurIPS (2024)
33. Ma, X., Zhou, C., Kong, X., He, J., Gui, L., Neubig, G., May, J., Zettlemoyer, L.: Mega: Moving average equipped gated attention. In: ICLR (2023)
34. Peng, B., Quesnelle, J., Fan, H., Shippole, E.: Yarn: Efficient context window extension of large language models. In: ICLR (2024)
35. Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720 (2024)
36. Ren, X., Lu, Y., Cao, T., Gao, R., Huang, S., Sabour, A., Shen, T., Pfaff, T., Wu, J.Z., Chen, R., et al.: Cosmos-drive-dreams: Scalable synthetic driving data generation with world foundation models. arXiv preprint arXiv:2506.09042 (2025)
37. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)

38. Sun, W., Zhang, H., Wang, H., Wu, J., Wang, Z., Wang, Z., Wang, Y., Zhang, J., Wang, T., Guo, C.: Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. arXiv preprint arXiv:2512.14614 (2025)
39. Teng, H., Jia, H., Sun, L., Li, L., Li, M., Tang, M., Han, S., Zhang, T., Zhang, W., Luo, W., et al.: Magi-1: Autoregressive video generation at scale. arXiv preprint arXiv:2505.13211 (2025)
40. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
41. Wan, Z., Wu, X., Zhang, Y., Xin, Y., Tao, C., Zhu, Z., Wang, X., Luo, S., Xiong, J., Zhang, M.: D2o: Dynamic discriminative operations for efficient generative inference of large language models. arXiv preprint arXiv:2406.13035 (2024)
42. Wang, Z., Jin, B., Yu, Z., Zhang, M.: Model tells you where to merge: Adaptive kv cache merging for llms on long-context tasks. arXiv preprint arXiv:2407.08454 (2024)
43. Wu, R., He, X., Cheng, M., Yang, T., Zhang, Y., Kang, Z., Cai, X., Wei, X., Guo, C., Li, C., et al.: Infinite-world: Scaling interactive world models to 1000-frame horizons via pose-free hierarchical memory. arXiv preprint arXiv:2602.02393 (2026)
44. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. In: ICLR (2024)
45. Xiao, Z., Lan, Y., Zhou, Y., Ouyang, W., Yang, S., Zeng, Y., Pan, X.: Worldmem: Long-term consistent world simulation with memory. arXiv preprint arXiv:2504.12369 (2025)
46. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157 (2021)
47. Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al.: Qwen2.5 technical report. arXiv preprint arXiv:2412.15115 (2025)
48. Yang, S., Huang, W., Chu, R., Xiao, Y., Zhao, Y., Wang, X., Li, M., Xie, E., Chen, Y.C., Lu, Y., Han, S., Chen, Y.: Longlive: Real-time interactive long video generation. In: ICLR (2026)
49. Yang, Y., Lv, Z., Pan, T., Wang, H., Yang, B., Yin, H., Li, C., Liu, Z., Si, C.: Stableworld: Towards stable and consistent long interactive video generation. arXiv preprint arXiv:2601.15281 (2026)
50. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: ICLR (2025)
51. Yesiltepe, H., Meral, T., Akan, A.K., Oktay, K., Yanardag, P.: Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. In: CVPR (2026)
52. Yi, J., Jang, W., Cho, P.H., Nam, J., Yoon, H., Kim, S.: Deep forcing: Training-free long video generation with deep sink and participative compression. arXiv preprint arXiv:2512.05081 (2025)
53. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: CVPR (2024)
54. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast autoregressive video diffusion models. In: CVPR (2025)
55. Yu, J., Bai, J., Qin, Y., Liu, Q., Wang, X., Wan, P., Zhang, D., Liu, X.: Context as memory: Scene-consistent interactive long video generation with memory retrieval. In: SIGGRAPH Asia (2025)

56. Yu, L., Cheng, Y., Sohn, K., Lezama, J., Zhang, H., Chang, H., Hauptmann, A.G., Yang, M.H., Hao, Y., Essa, I., et al.: Magvit: Masked generative video transformer. In: CVPR (2023)
57. Yu, L., Lezama, J., Gundavarapu, N.B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Gupta, A., Gu, X., Hauptmann, A.G., Gong, B., Yang, M.H., Essa, I., Ross, D.A., Jiang, L.: Language model beats diffusion - tokenizer is key to visual generation. In: ICLR (2024)
58. Yu, Y., Wu, X., Hu, X., Hu, T., Sun, Y., Lyu, X., Wang, B., Ma, L., Ma, Y., Wang, Z., et al.: Videossm: Autoregressive long video generation with hybrid state-space memory. arXiv preprint arXiv:2512.04519 (2025)
59. Zhang, L., Agrawala, M.: Packing input frame context in next-frame prediction models for video generation. arXiv preprint arXiv:2504.07458 (2025)
60. Zhang, L., Cai, S., Li, M., Zeng, C., Lu, B., Rao, A., Han, S., Wetzstein, G., Agrawala, M.: Pretraining frame preservation in autoregressive video memory compression. arXiv preprint arXiv:2512.23851 (2025)
61. Zhang, Y., Du, Y., Luo, G., Zhong, Y., Zhang, Z., Liu, S., Ji, R.: Cam: Cache merging for memory-efficient llms inference. In: ICML (2024)
62. Zhang, Z., Sheng, Y., Zhou, T., Chen, T., Zheng, L., Cai, R., Song, Z., Tian, Y., Ré, C., Barrett, C., et al.: H2o: Heavy-hitter oracle for efficient generative inference of large language models. NeurIPS (2023)
63. Zhao, M., He, G., Chen, Y., Zhu, H., Li, C., Zhu, J.: RIFLEx: A free lunch for length extrapolation in video diffusion transformers. In: ICML (2025)
64. Zhu, H., Zhao, M., He, G., Su, H., Li, C., Zhu, J.: Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. arXiv preprint arXiv:2602.02214 (2026)