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

Youngrae Kim* Qixin Hu* C.-C. Jay Kuo Peter A. Beerel
University of Southern California

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

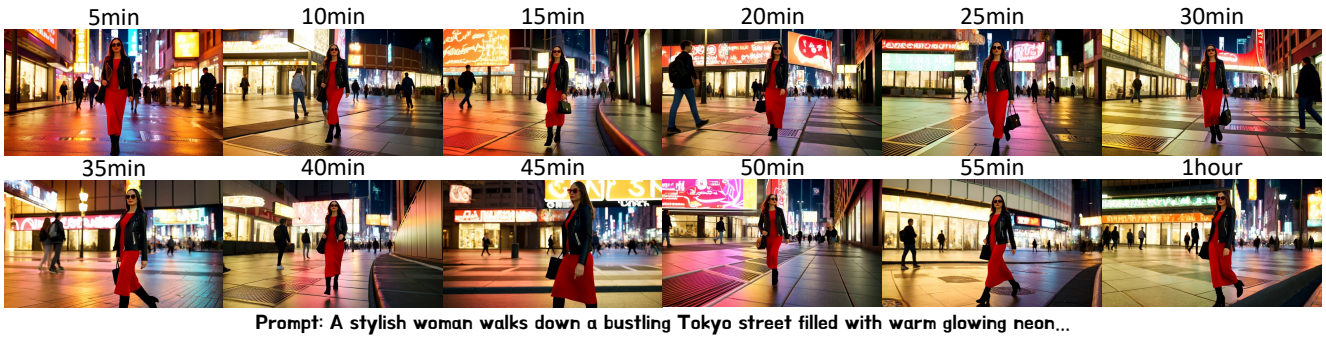


Figure 1. **Training-free unbounded video generation.** Our MemRoPE requires ***no additional training*** and enables ***unlimited generation*** with a fixed-size KV cache. We demonstrate a continuous one-hour generation process that perfectly preserves both subject identity and visual fidelity throughout.

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.

1. Introduction

Recent video diffusion models [24, 28, 37, 46] 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, 12, 17, 35, 45], cinematic long takes [11, 32], and synthetic data generation [18, 33]. 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 [50] distills a bidirectional DiT into a causal generator via DMD [49] for real-time synthesis, and Self Forcing [19] closes the train-inference gap by conditioning on self-generated frames. LongLive [44] 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

*Equal contribution.

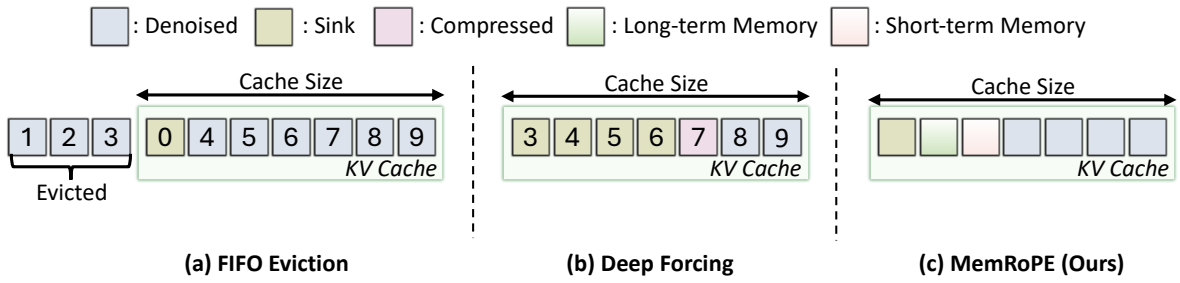


Figure 2. **KV cache structures for long video generation.** **(a)** FIFO eviction [19, 44, 47, 50] maintains an initial sink frame while discarding the oldest remaining frames when the cache is full, losing distant context. **(b)** Deep Forcing [48] dedicates over half the cache to static sink tokens 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).

a finite temporal horizon. ∞-RoPE [34, 47] 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 [19, 44, 47, 48].

To retain past context during extended generation, several approaches adapt sliding windows by leveraging the attention sink phenomenon [41], preserving initial frames as fixed anchors [27, 44, 47, 48]. Deep Forcing [48] 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 supports *unbounded generation* with constant memory; Fig. 1 demonstrates a continuous one-hour example that preserves temporal consistency throughout. Our contributions are as follows:

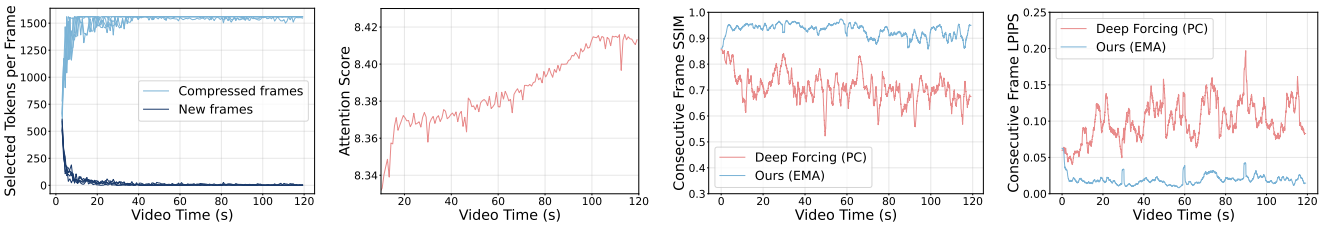
- We propose **MemRoPE**, a training-free framework for infinite-length video generation that jointly addresses context retention and positional extrapolation, enabling the generation of hours-long videos.
- We introduce **Memory Tokens**, an EMA-based dual memory mechanism that continuously compresses 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 [10, 23, 43, 52, 53] quantize video for next-token prediction; chunk-level diffusion [4, 21, 36] denoises multi-frame chunks; and FramePack [55] and StreamDiT [22] reduce memory via compressed contexts and window attention.

Most relevant to our work is frame-level autoregressive diffusion. CausVid [50] distills a bidirectional DiT into a causal generator via DMD, and Self Forcing [19] 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 [27] introduces a rolling denoising window, and Causal Forcing [60] improves distillation via ODE initialization. LongLive [44] aligns training with inference-time rollout for 240-second generation. FAR [15] compresses



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

Figure 3. **Failure mode of Participative Compression. (a)** Participative Compression (PC), proposed in Deep Forcing [48], 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.

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 [26, 58], token merging [39, 57], dynamic budget allocation [38], and correlation-aware eviction [13], but these techniques operate within a fixed-length cache and cannot recover information once evicted [55]. EMA-based temporal aggregation has also been explored as an architectural primitive [29, 30], but requires modifications to the model architecture and training from scratch. Our memory tokens instead operate on the KV cache of an unmodified pretrained model.

To retain context beyond the window, StreamingLLM [41] identified the *attention sink* phenomenon, and video generation methods adopted a similar strategy of preserving initial frames as static anchors [27, 44]. Deep Forcing [48] goes beyond this with Participative Compression, but the selected tokens quickly collapse to a fixed set during the generation process (Fig. 3). Other approaches introduce long-context streaming tuning [6], memory modules within the diffusion-forcing paradigm [3, 16], compressed video histories [40, 54–56], or geometric priors from 3D representations [25, 42, 51]. However, these methods either incur growing memory that becomes impractical at hour-length scales or rely on rigid priors that transfer poorly to open-ended synthesis.

In contrast, MemRoPE maintains a fixed-size, continuously evolving memory within a causal autoregressive KV cache, adapting to new content without increasing memory cost regardless of the total generated video length.

Positional extrapolation. Extending RoPE [34] beyond its trained range has been studied in LLMs through position interpolation [5], frequency scaling [1], and their combinations [31], but LoL [9] shows these induce a quality-dynamics tradeoff in video diffusion. In the video domain, ∞-RoPE [47] confines indices to the trained range via a block-relative coordinate system, and RIFLEx [59] 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.

3. Background

Autoregressive inference. Autoregressive video diffusion generates video one frame chunk at a time [8, 9, 19, 27, 44, 48, 50]. 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, 19, 27, 44, 48, 50].

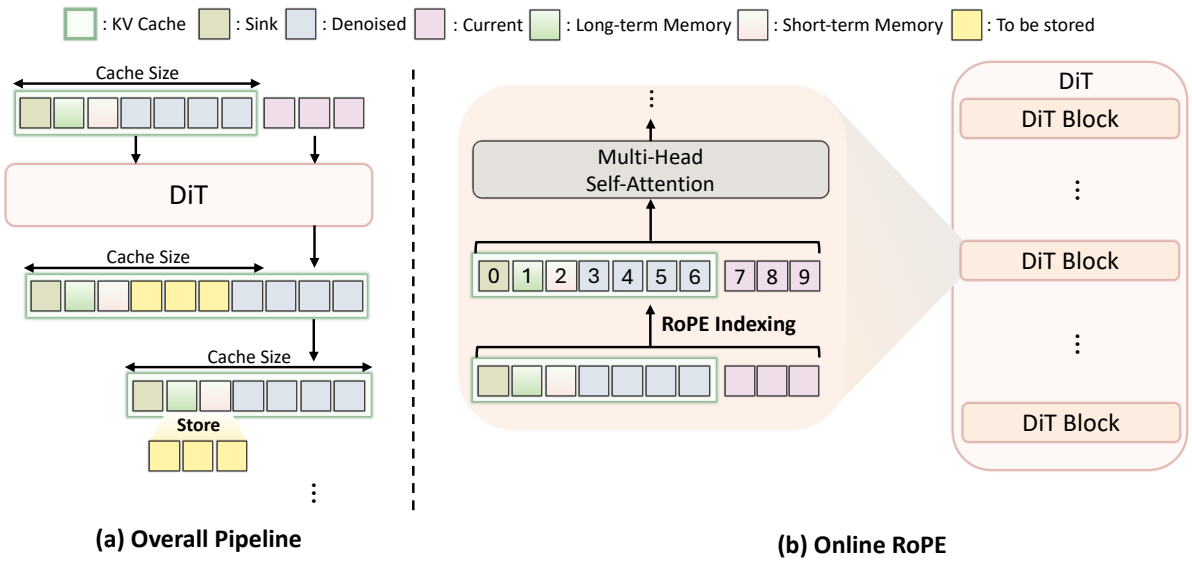


Figure 4. **Overview of MemRoPE. (a)** At each autoregressive step, the three-tier KV cache (sink, memory, and recent tokens) is concatenated with the current noisy chunk and fed into the DiT. When the local window is full, the oldest frames are absorbed into the long- and short-term memory tokens via 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.

Rotary Position Embedding. Positional ordering is often encoded via 3D-RoPE in video generation [34, 37], 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.

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 [19, 44, 47, 48]. To mitigate this, several methods [27, 44, 48] retain the initial frames as *attention sinks* alongside the sliding window, inspired by the attention sink phenomenon in LLMs [41]. 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 [48] 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 tokens. 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, their spatially pooled key $\bar{k}_{\text{new}}$ 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, enabling unbounded generation within a fixed 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_ϕ, 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 ϕ 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.

Algorithm 1 MemRoPE Inference

Require: Diffusion Transformer $\mathcal{G}_\theta$, denoising steps N, EMA rates α_L, α_S
Notation: $x_t^{(s)}$: noisy latent of chunk t at denoising step s; $\hat{x}_t = x_t^{(N)}$: final denoised chunk

$\hat{x}_0 \leftarrow \mathcal{G}_\theta(x_0^{(0)}, N)$
$\mathbf{K}_{\text{sink}}, \mathbf{V}_{\text{sink}} \leftarrow W_K\hat{x}_0, W_V\hat{x}_0$ ▷ w/o 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 $s = 1, \ldots, N$ **do** ▷ denoising steps
    $q_s, k_s, v_s \leftarrow W_Q x_t^{(s)}, W_K x_t^{(s)}, W_V x_t^{(s)}$
    $\mathbf{K} \leftarrow [\mathbf{K}_{\text{sink}} \| \mu_L^k \| \mu_S^k \| \mathbf{K}_{\text{local}} \| k_s]$ ▷ w/o RoPE
    $\mathbf{V} \leftarrow [\mathbf{V}_{\text{sink}} \| \mu_L^v \| \mu_S^v \| \mathbf{V}_{\text{local}} \| v_s]$
    $\phi \leftarrow [0, \ldots, |\mathbf{K}|{-}1]$ ▷ Online RoPE
    Apply R_ϕ to $q_s, \mathbf{K}$
    $x_t^{(s+1)} \leftarrow \text{DiT-Block}(R_\phi q_s, R_\phi \mathbf{K}, \mathbf{V})$
  end for
  $\hat{x}_t \leftarrow x_t^{(N)}$ ▷ denoised chunk
  Append $W_K\hat{x}_t, W_V\hat{x}_t$ to $\mathbf{K}_{\text{local}}, \mathbf{V}_{\text{local}}$ ▷ w/o RoPE
  if $|\mathbf{K}_{\text{local}}| > L$ **then** ▷ evict & update memory
    $\bar{k} \leftarrow \text{SpatialPool}(\mathbf{K}_{\text{local}}[0])$
    $\bar{v} \leftarrow \text{SpatialPool}(\mathbf{V}_{\text{local}}[0])$
    $\mu_L^k \leftarrow (1{-}\alpha_L)\mu_L^k + \alpha_L\bar{k}$
    $\mu_L^v \leftarrow (1{-}\alpha_L)\mu_L^v + \alpha_L\bar{v}$
    $\mu_S^k \leftarrow (1{-}\alpha_S)\mu_S^k + \alpha_S\bar{k}$
    $\mu_S^v \leftarrow (1{-}\alpha_S)\mu_S^v + \alpha_S\bar{v}$
    Evict oldest chunk from $\mathbf{K}_{\text{local}}, \mathbf{V}_{\text{local}}$
  end if
end for

Relationship to block-relative RoPE. ∞-RoPE [47] 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_{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 [27] 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

Table 1. **Quantitative comparison on long video generation.** We report VBench-Long [20] metrics. Within each base model group, best results are bold with darker background and second best are with lighter background. Δ denotes the average improvement over the respective base model.

Method	Aesthetic Quality ↑	Background Consistency ↑	Imaging Quality ↑	Motion Smoothness ↑	Subject Consistency ↑	Temporal Flickering ↑	Average ↑	Δ
120 seconds								
Results on Self-Forcing [19]								
Self-Forcing [19]	55.11	95.41	65.95	97.17	95.35	96.53	84.25	–
Deep Forcing [48]	**56.86**	94.58	65.02	97.07	94.17	95.33	83.84	−0.41
∞-RoPE [47]	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 [44]								
LongLive [44]	57.21	95.96	67.13	98.56	97.06	97.12	85.51	–
Deep Forcing [48]	59.16	96.13	68.07	98.51	96.79	97.38	86.01	+0.50
∞-RoPE [47]	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 [19]								
Self-Forcing [19]	51.00	95.01	61.52	98.18	95.83	96.72	83.04	–
Deep Forcing [48]	52.20	93.91	59.50	96.72	92.28	95.38	81.66	−1.38
∞-RoPE [47]	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 [44]								
LongLive [44]	56.70	95.80	66.90	98.52	97.02	97.04	85.33	–
Deep Forcing [48]	57.75	96.03	66.48	98.50	96.65	97.47	85.48	+0.15
∞-RoPE [47]	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**

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.

5. Experiments

5.1. Setup

Implementation details. We build on the Wan2.1-T2V-1.3B architecture [37] and evaluate MemRoPE on two base models: Self-Forcing [19] and LongLive [44]. 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 [48] and ∞-RoPE [47] 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 [32] at 480×832 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 [48] and Self-Forcing [19], all prompts are refined using Qwen2.5-7B-Instruct. We report metrics from VBench-Long [20], and further validate with a user study following the 2AFC protocol [48] and VLM-based evaluation using Gemini 3.1-Pro [7].

Figure 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.

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

Table 2. **Ultra-long video generation.** We report VBench-Long [20] 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 [19]							
∞-RoPE [47]	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 [44]							
∞-RoPE [47]	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 [44]							
∞-RoPE [47]	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**

fidelity and identity over longer horizons.

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.

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.

Table 3. User study (% favoring Ours).

Baseline	Color Cons.	Subject Cons.	Bg. Cons.	Text Align.	Motion Smo.	Overall Pref.
Self-Forcing [19]	93.0	93.0	97.4	91.6	95.7	98.3
Rolling Forcing [27]	90.4	77.4	79.1	80.7	88.7	82.6
Results on LongLive [44]						
LongLive [44]	81.6	70.2	83.3	66.1	71.9	71.1
Deep Forcing [48]	79.0	75.2	81.9	70.5	72.4	72.6
∞-RoPE [47]	81.3	76.4	77.6	74.5	73.8	70.1

User study. To complement the automated metrics, we conducted a user study with 30 participants following the 2AFC protocol [48]; 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. Participants consistently preferred MemRoPE across all aspects, corroborating the quantitative results.

Table 4. Visual stability (VLM).

Method	Stability
Self-Forcing [19]	1.55
Rolling Forcing [27]	3.40
Results on LongLive [44]	
LongLive [44]	4.10
Deep Forcing [48]	3.90
∞-RoPE [47]	4.05
MemRoPE (Ours)	**4.15**

VLM evaluation. We further assess long-horizon visual stability using the advanced multimodal vision-language model Gemini 3.1-Pro [14]. Following the protocol of Self-Forcing++ [8] and Deep Forcing [48], 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

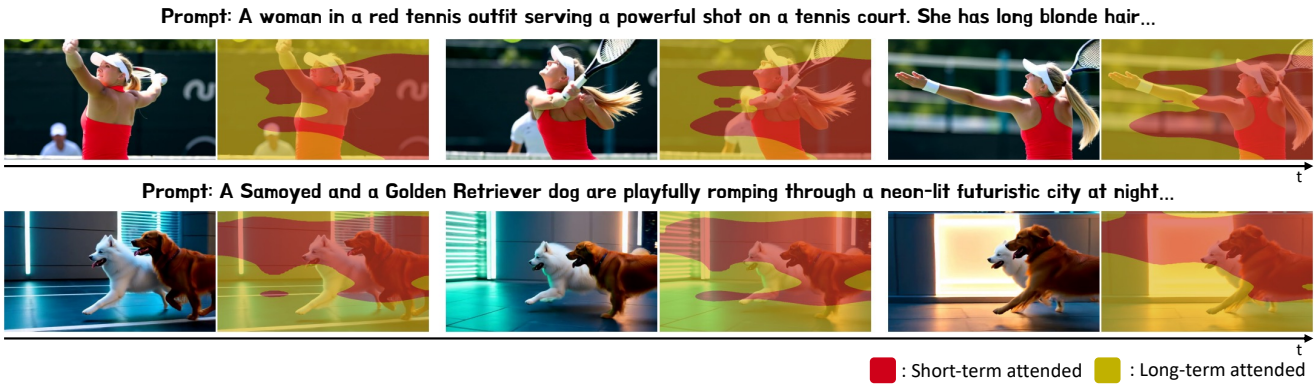


Figure 6. Spatial attention distribution between the dual memory streams. Yellow indicates higher attention to long-term memory. Red indicates higher attention to short-term memory. Temporally consistent regions attend more to long-term memory, while rapidly changing regions attend more to short-term memory.

Table 5. **Ablation on memory components.** While using only long-term memory (μ_L) achieves the highest Subject Consistency, combining both memories yields the highest scores in overall average.

Long-term Memory (μ_L)	Short-term Memory (μ_S)	Aesthetic Quality ↑	Imaging Quality ↑	Subject Consistency ↑	Avg ↑
Results on LongLive [44]					
		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**

Tab. 4, MemRoPE achieves the highest stability scores, consistent with both the automated metrics and the user study.

Memory component ablation. Tab. 5 isolates the contribution of each memory stream. 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 highest average and the best Imaging Quality among all configurations, suggesting that the two streams capture complementary information: μ_L summarizes the stable aspects of the scene while μ_S reflects recent changes. Fig. 6 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 α_L and the short-term decay α_S. Fig. 7 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 α_L=0.01 and α_S=0.1 for all other experiments.

EMA vs. attention-based cache management. Deep Forcing's participative compression [48] is the only other training-free method that goes beyond 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.

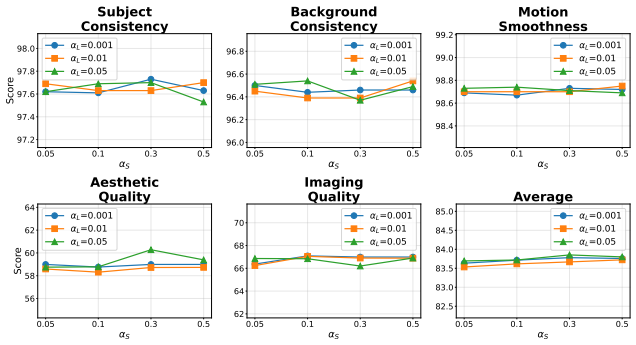


Figure 7. **Sensitivity to EMA decay rates.** All metrics remain stable across 12 combinations of α_L and α_S, with the average score varying by less than 0.7.

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. 3

[2] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024. 1

[3] 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. 3, 13

[4] 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. 2, 13, 14

[5] Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023. 3

[6] 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. *arXiv preprint arXiv:2602.06028*, 2026. 3

[7] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 6

[8] Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. *arXiv preprint arXiv:2510.02283*, 2025. 2, 3, 8, 17

[9] Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Lol: Longer than longer, scaling video generation to hour. *arXiv preprint arXiv:2601.16914*, 2026. 3

[10] Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 13, 14

[11] Mohamed Elmoghany, Ryan Rossi, Seunghyun Yoon, Subhojyoti Mukherjee, Eslam Mohamed Bakr, Puneet Mathur, Gang Wu, Viet Dac Lai, Nedim Lipka, Ruiyi Zhang, et al. A survey on long-video storytelling generation: architectures, consistency, and cinematic quality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7023–7035, 2025. 1

[12] Ruili Feng, Han Zhang, Zhantao Yang, Jie Xiao, Zhilei Shu, Zhiheng Liu, Andy Zheng, Yukun Huang, Yu Liu, and Hongyang Zhang. The matrix: Infinite-horizon world generation with real-time moving control. *arXiv preprint arXiv:2412.03568*, 2024. 1

[13] Ravi Ghadia, Avinash Kumar, Gaurav Jain, Prashant Nair, and Poulami Das. Dialogue without limits: Constant-sized kv caches for extended responses in llms. *arXiv preprint arXiv:2503.00979*, 2025. 3

[14] Google DeepMind. Gemini 3.1 pro model card, 2026. 8, 17, 18

[15] Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. *arXiv preprint arXiv:2503.19325*, 2025. 2

[16] Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2568–2577, 2025. 3

[17] Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, et al. Relic: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040*, 2025. 1

[18] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 1

[19] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1, 2, 3, 4, 6, 8, 13, 14, 15, 17

[20] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 6, 8, 13, 14

[21] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong MU, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[22] Akio Kodaira, Tingbo Hou, Ji Hou, Markos Georgopoulos, Felix Juefei-Xu, Masayoshi Tomizuka, and Yue Zhao. Streamdit: Real-time streaming text-to-video generation. *arXiv preprint arXiv:2507.03745*, 2025. 2

[23] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023. 2

[24] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 1

[25] Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 25690–25699, 2025. 3

[26] Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. *Advances in Neural Information Processing Systems*, 37:22947–22970, 2024. 3

[27] Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025. 2, 3, 4, 5, 8, 13, 14

[28] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024. 1

[29] Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Luke Zettlemoyer. Mega: Moving average equipped gated attention. *arXiv preprint arXiv:2209.10655*, 2022. 3

[30] Xuezhe Ma, Xiaomeng Yang, Wenhan Xiong, Beidi Chen, Lili Yu, Hao Zhang, Jonathan May, Luke Zettlemoyer, Omer Levy, and Chunting Zhou. Megalodon: Efficient llm pretraining and inference with unlimited context length. *Advances in Neural Information Processing Systems*, 37: 71831–71854, 2024. 3

[31] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023. 3

[32] 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. 1, 6

[33] Xuanchi Ren, Yifan Lu, Tianshi Cao, Ruiyuan Gao, Shengyu Huang, Amirmojtaba Sabour, Tianchang Shen, Tobias Pfaff, Jay Zhangjie Wu, Runjian Chen, et al. Cosmos-drive-dreams: Scalable synthetic driving data generation with world foundation models. *arXiv preprint arXiv:2506.09042*, 2025. 1

[34] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 2, 3, 4

[35] Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025. 1

[36] 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. 2, 13, 14

[37] 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. 1, 4, 6

[38] Zhongwei Wan, Xinjian Wu, Yu Zhang, Yi Xin, Chaofan Tao, Zhihong Zhu, Xin Wang, Siqi Luo, Jing Xiong, and Mi Zhang. D2o: Dynamic discriminative operations for efficient generative inference of large language models. *arXiv preprint arXiv:2406.13035*, 2, 2024. 3

[39] Zheng Wang, Boxiao Jin, Zhongzhi Yu, and Minjia Zhang. Model tells you where to merge: Adaptive kv cache merging for llms on long-context tasks. *arXiv preprint arXiv:2407.08454*, 2024. 3

[40] Ruiqi Wu, Xuanhua He, Meng Cheng, Tianyu Yang, Yong Zhang, Zhuoliang Kang, Xunliang Cai, Xiaoming Wei, Chunle Guo, Chongyi Li, et al. Infinite-world: Scaling interactive world models to 1000-frame horizons via pose-free hierarchical memory. *arXiv preprint arXiv:2602.02393*, 2026. 3

[41] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3, 4

[42] Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*, 2025. 3

[43] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 2

[44] Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Ying-Cong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. In *The Fourteenth International Conference on Learning Representations*, 2026. 1, 2, 3, 4, 6, 8, 9, 13, 14, 15, 16, 17

[45] Ying Yang, Zhengyao Lv, Tianlin Pan, Haofan Wang, Binxin Yang, Hubery Yin, Chen Li, Ziwei Liu, and Chenyang Si. Stableworld: Towards stable and consistent long interactive video generation. *arXiv preprint arXiv:2601.15281*, 2026. 1

[46] 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. *arXiv preprint arXiv:2408.06072*, 2024. 1

[47] Hidir Yesiltepe, Tuna Han Salih Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. *arXiv preprint arXiv:2511.20649*, 2025. 2, 3, 4, 5, 6, 8, 13, 14, 15, 16

[48] 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. *arXiv preprint arXiv:2512.05081*, 2025. 2, 3, 4, 6, 8, 9, 13, 14

[49] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6613–6623, 2024. 1

[50] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22963–22974, 2025. 1, 2, 3, 13, 14

[51] Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–11, 2025. 3

[52] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10459–10469, 2023. 2

[53] Lijun Yu, Jose Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A Ross, and Lu Jiang. Language model beats diffusion - tokenizer is key to visual generation. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[54] Yifei Yu, Xiaoshan Wu, Xinting Hu, Tao Hu, Yangtian Sun, Xiaoyang Lyu, Bo Wang, Lin Ma, Yuewen Ma, Zhongrui Wang, et al. Videossm: Autoregressive long video generation with hybrid state-space memory. *arXiv preprint arXiv:2512.04519*, 2025. 3

[55] Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. *arXiv e-prints*, pages arXiv–2504, 2025. 2, 3

[56] Lvmin Zhang, Shengqu Cai, Muyang Li, Chong Zeng, Beijia Lu, Anyi Rao, Song Han, Gordon Wetzstein, and Maneesh Agrawala. Pretraining frame preservation in autoregressive video memory compression. *arXiv preprint arXiv:2512.23851*, 2025. 3

[57] Yuxin Zhang, Yuxuan Du, Gen Luo, Yunshan Zhong, Zhenyu Zhang, Shiwei Liu, and Rongrong Ji. Cam: Cache merging for memory-efficient llms inference. In *Forty-first international conference on machine learning*, 2024. 3

[58] Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36: 34661–34710, 2023. 3

[59] Min Zhao, Guande He, Yixiao Chen, Hongzhou Zhu, Chongxuan Li, and Jun Zhu. RIFLEx: A free lunch for length extrapolation in video diffusion transformers. In *Forty-second International Conference on Machine Learning*, 2025. 3

[60] Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026. 2

Appendix

This appendix is organized as follows:

- Quantitative results on shorter videos (Appendix A)
- Ablation on position-free caching (Appendix B)
- VBench-Long score trends across durations (Appendix C)
- Additional qualitative comparisons (Appendix D)
- Inference latency measurements (Appendix E)
- VLM evaluation and user study details (Appendix F)

A. Quantitative Results on Shorter Videos

The main paper reports VBench-Long [20] results at 120, 240, 480 seconds and 1 hour, where cumulative degradation makes the differences between methods most pronounced. Here we provide complete results at 30 and 60 seconds to demonstrate that MemRoPE is effective even at shorter durations where degradation is minimal.

In addition to the Self-Forcing [19] and LongLive [44] base comparisons presented in the main paper, we include additional baselines for reference. CausVid [50] shares the same base model and 5-second training horizon as Self-Forcing, but suffers from a train-inference distribution mismatch—it optimizes against Diffusion Forcing [3] outputs rather than the actual inference-time distribution—leading to over-exposure artifacts that worsen at longer durations. We adopt Self-Forcing as a base model because it resolves this mismatch by simulating the actual inference process during training, providing a stronger foundation for long video generation. Rolling Forcing [27] performs additional distillation training and adopts a fundamentally different inference paradigm based on rolling-window joint denoising, making it incompatible as a base model for our plug-in method. NOVA [10], MAGI-1 [36], and SkyReels-V2 [4] use entirely different architectures. These were omitted from the main tables to keep the comparison focused on training-free methods under controlled base-model settings, but are included here for completeness.

MemRoPE achieves the highest average on both Self-Forcing and LongLive bases at both durations. Notably, even at 30 seconds where degradation has not yet accumulated significantly, MemRoPE already outperforms all competing methods within each base group, indicating that our memory mechanism improves generation quality from the outset rather than only compensating for long-horizon drift.

B. Ablation on Position-Free Caching

As discussed in Sec. 4.2 of the main paper, RoPE does not distribute over addition (Eq. (5)): averaging keys that carry different rotary phases produces representations that no longer correspond to any valid temporal position. A natural question is whether this theoretical concern has a measurable impact in practice. To answer this, we compare MemRoPE (position-free EMA) against an EMA variant that retains RoPE embeddings during aggregation (Aggregation w/ RoPE).

As shown in Tab. 7, both methods improve over their respective base models, confirming that temporal aggregation is broadly beneficial. However, MemRoPE consistently achieves a higher average improvement (Δ) on both Self-Forcing ($+0.84$ vs. $+0.71$) and LongLive ($+0.53$ vs. $+0.37$). While Aggregation w/ RoPE scores higher on certain individual metrics, its overall coherence is lower due to the conflicting rotary phases discussed in Eq. (5). This validates our design choice: decoupling positional information from cached keys before temporal aggregation is essential for well-defined memory compression.

C. VBench-Long Score Across Durations

We collect the VBench-Long averages of MemRoPE and ∞-RoPE [47] on Self-Forcing [19] across five durations: 30, 60, 120, 240, and 480 seconds. To ensure a consistent comparison, we report results using the 20-prompt subset used for the 480-second evaluation across all durations. As shown in Fig. 8, both methods degrade as duration increases, but MemRoPE exhibits a significantly flatter slope. The performance gap widens monotonically as duration increases, demonstrating that the benefits of our evolving memory tokens compound over time. This scaling behavior confirms that MemRoPE's dual-stream EMA effectively preserves long-range context that ∞-RoPE fully discards.

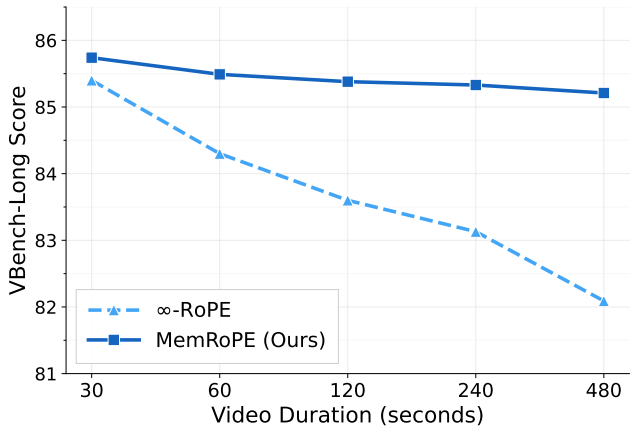


Figure 8. **VBench-Long average score vs. video duration on Self-Forcing [19].** MemRoPE maintains higher scores across all durations, with the gap widening as videos grow longer.

D. Qualitative Comparisons

2-Minute Generation. Fig. 9 compares MemRoPE against Self-Forcing [19], Deep Forcing [48], and ∞-RoPE [47]

Table 6. **Quantitative comparison on short video generation (30 and 60 seconds).** We report VBench-Long [20] metrics. Within each base model group, best results are bold with darker background and second best are with lighter background. Δ denotes the average improvement over the respective base model. *Results from [47].

Method	Aesthetic Quality ↑	Background Consistency ↑	Imaging Quality ↑	Motion Smoothness ↑	Subject Consistency ↑	Temporal Flickering ↑	Average ↑	Δ
30 seconds								
CausVid [50]	57.25	96.77	65.16	98.08	97.87	96.86	85.33	–
Rolling Forcing [27]	58.45	95.77	69.35	98.23	96.78	97.13	85.95	–
Results on Self-Forcing [19]								
Self-Forcing [19]	58.28	95.14	66.92	98.02	95.50	96.87	85.12	–
Deep Forcing [48]	57.83	94.86	66.29	97.53	94.98	95.71	84.53	−0.59
∞-RoPE [47]	57.13	**95.81**	65.10	**98.39**	**96.37**	**97.18**	85.00	−0.12
MemRoPE (Ours)	**58.48**	95.59	**68.73**	98.01	96.31	96.41	**85.59**	**+0.47**
Results on LongLive [44]								
LongLive [44]	58.57	96.22	67.85	98.65	97.21	97.32	85.97	–
Deep Forcing [48]	**59.87**	96.22	**69.41**	98.59	97.05	97.33	86.41	+0.44
∞-RoPE [47]	59.32	**96.40**	66.68	**98.94**	97.32	**97.93**	86.10	+0.13
MemRoPE (Ours)	59.31	96.34	69.27	98.68	**97.48**	97.40	**86.42**	**+0.45**
60 seconds								
NOVA* [10]	47.53	88.06	44.97	98.94	77.50	98.27	75.88	–
MAGI-1* [36]	52.10	87.76	54.54	99.26	79.46	98.48	78.60	–
SkyReels-V2* [4]	57.64	89.95	66.67	98.67	84.99	97.60	82.59	–
CausVid [50]	57.06	96.94	64.84	98.16	98.03	97.05	85.35	–
Rolling Forcing [27]	58.34	96.21	69.41	98.52	97.26	97.47	86.20	–
Results on Self-Forcing [19]								
Self-Forcing [19]	56.96	95.08	66.36	97.35	95.15	96.56	84.57	–
Deep Forcing [48]	57.30	94.59	65.67	97.22	94.36	95.38	84.09	−0.48
∞-RoPE [47]	54.97	**95.69**	62.81	**98.39**	96.26	**97.37**	84.25	−0.32
MemRoPE (Ours)	**57.77**	95.58	**68.54**	97.93	**96.29**	96.35	**85.41**	**+0.84**
Results on LongLive [44]								
LongLive [44]	57.93	96.10	67.41	98.57	97.10	97.15	85.71	–
Deep Forcing [48]	**59.63**	96.16	68.73	98.52	96.88	97.33	86.21	+0.50
∞-RoPE [47]	58.52	**96.33**	65.63	**98.95**	97.27	**97.95**	85.78	+0.07
MemRoPE (Ours)	58.76	96.27	**69.01**	98.67	**97.41**	97.34	**86.24**	**+0.53**

Table 7. **Ablation on position-free caching.** Comparison between MemRoPE (position-free EMA) and EMA with RoPE-rotated keys at 60 seconds. Δ denotes the average improvement over the respective base model.

Method	Aesthetic Quality ↑	Subject Consistency ↑	Imaging Quality ↑	Motion Smoothness ↑	Background Consistency ↑	Temporal Flickering ↑	Average ↑	Δ
Results on Self-Forcing [19]								
Self-Forcing [19]	56.96	95.08	66.36	97.35	95.15	**96.66**	84.57	–
Aggregation w/ RoPE	56.55	**96.44**	68.25	**98.19**	95.60	**96.66**	85.28	+0.71
MemRoPE (Ours)	**57.77**	95.58	**68.54**	97.93	**96.29**	96.35	**85.41**	**+0.84**
Results on LongLive [44]								
LongLive [44]	57.93	96.10	67.41	98.57	97.10	97.15	85.71	–
Aggregation w/ RoPE	58.50	**97.35**	68.45	**98.67**	96.18	97.33	86.08	+0.37
MemRoPE (Ours)	**58.76**	96.27	**69.01**	**98.67**	**97.41**	**97.34**	**86.24**	**+0.53**

on two prompts at 2-minute duration. On Self-Forcing, the base model degrades severely by the end of generation. Deep Forcing preserves the general structure but the main subject undergoes noticeable shape changes over time. ∞-RoPE loses subject consistency entirely, with colors collapsing across frames. On LongLive, the base model exhibits gradual subject appearance drift, Deep Forcing shifts the overall color tone, and ∞-RoPE suffers from both ap-

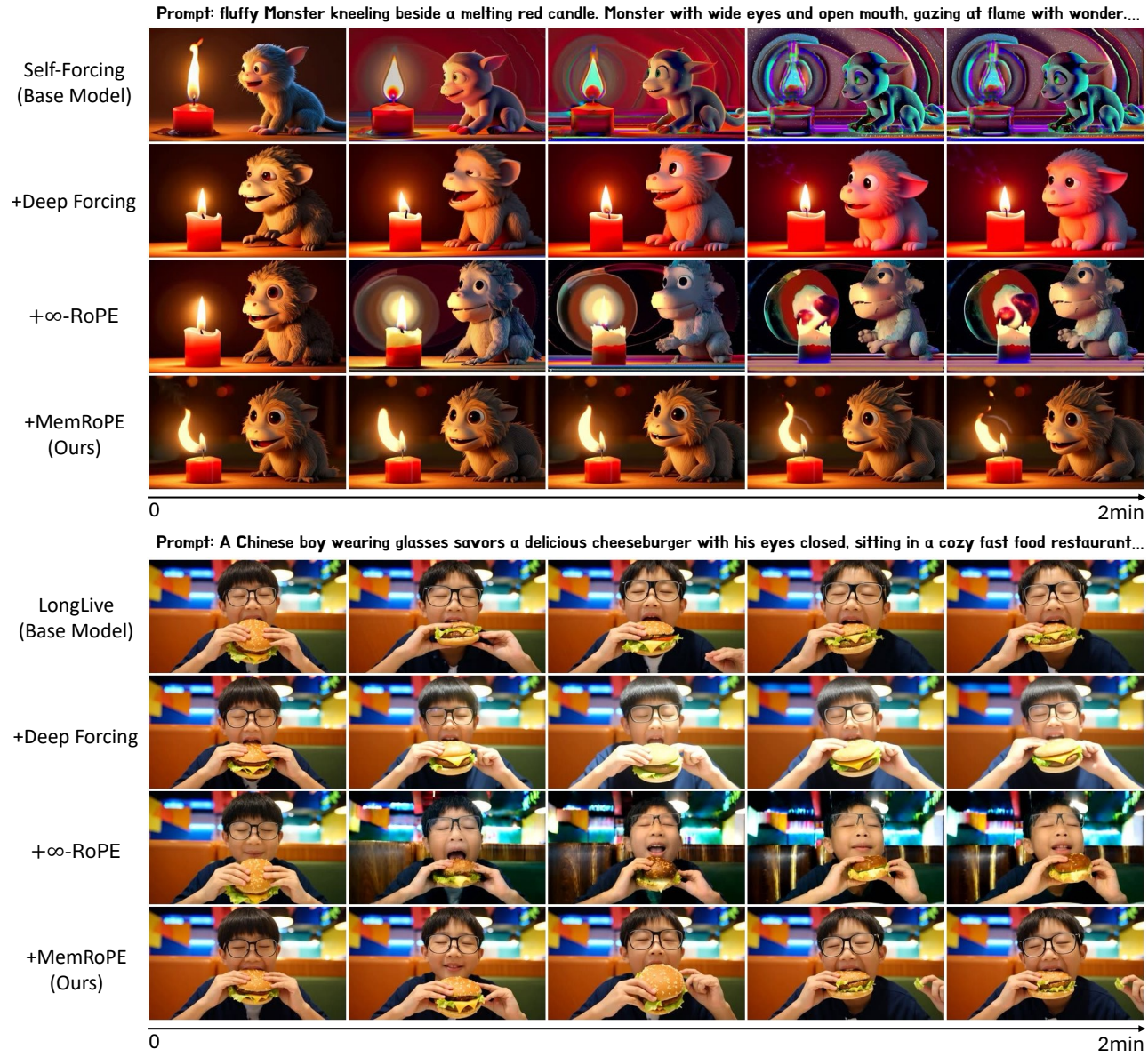


Figure 9. **Qualitative comparison at 2-minute generation on Self-Forcing [19] and LongLive [44].** MemRoPE preserves subject identity and visual fidelity more consistently than existing methods across both prompts.

pearance changes and color corruption. MemRoPE maintains consistent subject identity and visual fidelity throughout on both base models.

1-Hour Generation. Fig. 10 compares frames sampled at 12-minute intervals from hour-long videos generated by ∞-RoPE [47] and MemRoPE on LongLive [44]. Both methods maintain the general scene structure, but ∞-RoPE shows gradual inconsistency in the subject's facial features and hairstyle, along with increasingly saturated background colors in later segments. MemRoPE preserves more stable subject appearance and background color tone throughout the full hour.

Memory Component Ablation. Fig. 11 visualizes the effect of each memory component on LongLive [44] at 60 seconds. Without any memory, with long-term memory only, and with short-term memory only, background structures such as buildings gradually disappear as generation progresses. Only when both streams are combined does the full method preserve the background architecture and subject appearance throughout the sequence, suggesting that

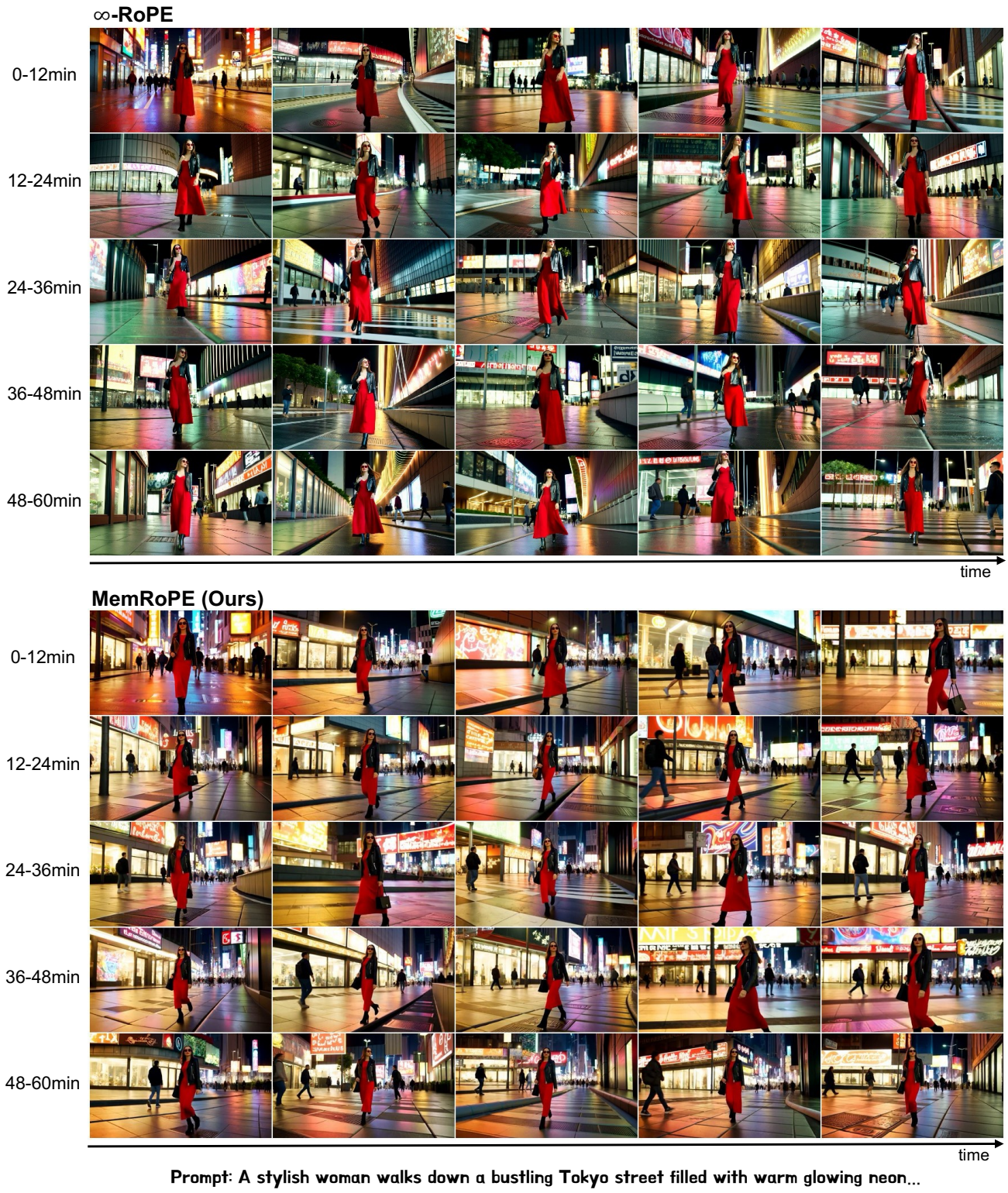


Figure 10. **1-hour generation comparison on LongLive [44].** Top: ∞-RoPE [47]. Bottom: MemRoPE (Ours). Frames sampled at 12-minute intervals. MemRoPE maintains more consistent subject appearance and background color tone over the full hour.

Figure 11. **Qualitative memory component ablation on LongLive [44].** From top to bottom: no memory, long-term memory only, short-term memory only, and both (ours). Only the full dual-stream configuration preserves background structures and subject appearance throughout.

both memory scales are jointly necessary for maintaining scene integrity.

E. Inference Latency

We measure inference speed on a single NVIDIA A6000 GPU after generating one initial chunk to warm up the model and GPU state. As shown in Tab. 8, MemRoPE adds negligible overhead at the same cache size (LongLive, both $C = 12$). On Self-Forcing, MemRoPE is *faster* because its compact three-tier cache ($C = 12$) replaces the original sliding window ($C = 21$), reducing the number of tokens attended to during each denoising step.

Table 8. **Inference speed.** Video FPS measured on a single A6000 GPU.

Method	Video FPS
Self-Forcing [19] ($C = 21$)	3.631
+ MemRoPE ($C = 12$)	5.115
LongLive [44] ($C = 12$)	4.409
+ MemRoPE ($C = 12$)	4.376

F. VLM Evaluation and User Study Details

VLM-Based Evaluation. In addition to the user study, we employ Gemini 3.1 Pro [14] as a VLM judge to evaluate generated videos on a 5-point exposure stability scale, following Self-Forcing++ [8]. The evaluation prompt is as follows:

VLM Evaluation Prompt

You are tasked with rating the exposure stability of a video. Assign a score according to the following scale:

0: Catastrophic Exposure. Nearly the entire frame is either blown out (pure white) or crushed (pure black), rendering the scene unreadable.
1: Severe Exposure Failure. Large portions are dominated by over- or under-exposure, substantially impairing visibility.
2: Noticeable Exposure Problems. Persistent clipping in highlights or shadows; significant areas lose detail.
3: Moderate Exposure Issues. Over-exposed highlights or under-exposed shadows occur but are limited in extent or duration.
4: Minor Exposure Flaws. Small regions are occasionally too bright or too dark, but do not meaningfully disrupt visibility.
5: Well-Exposed. Balanced lighting across the frame with no distracting over-exposure or darkening.

Fig. 12 shows example evaluations.

User Study Interface. Fig. 13 shows our user study interface. Participants evaluate 20 pairs of side-by-side videos generated from the same prompt by two anonymized meth-

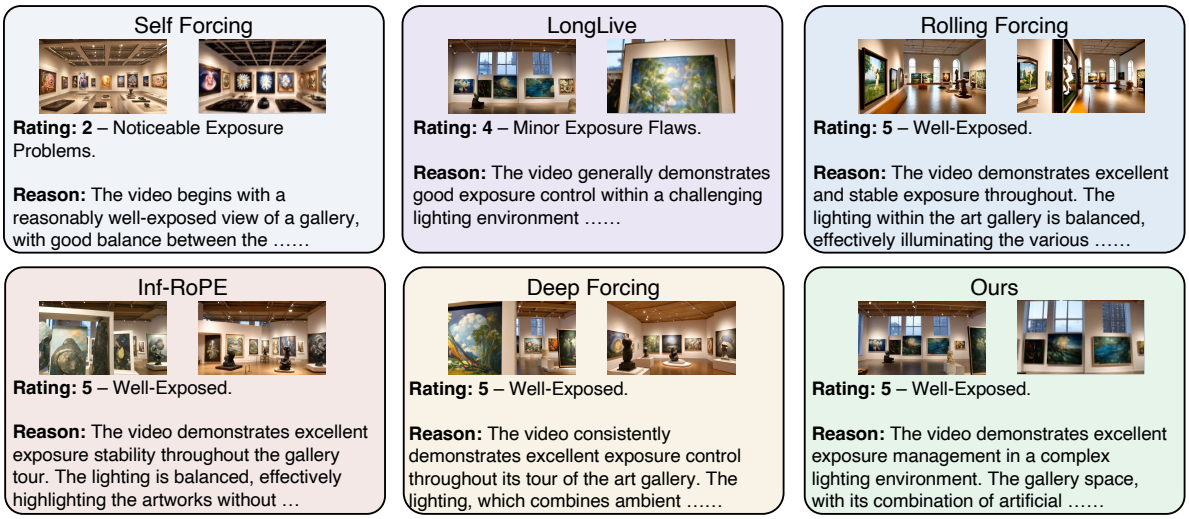
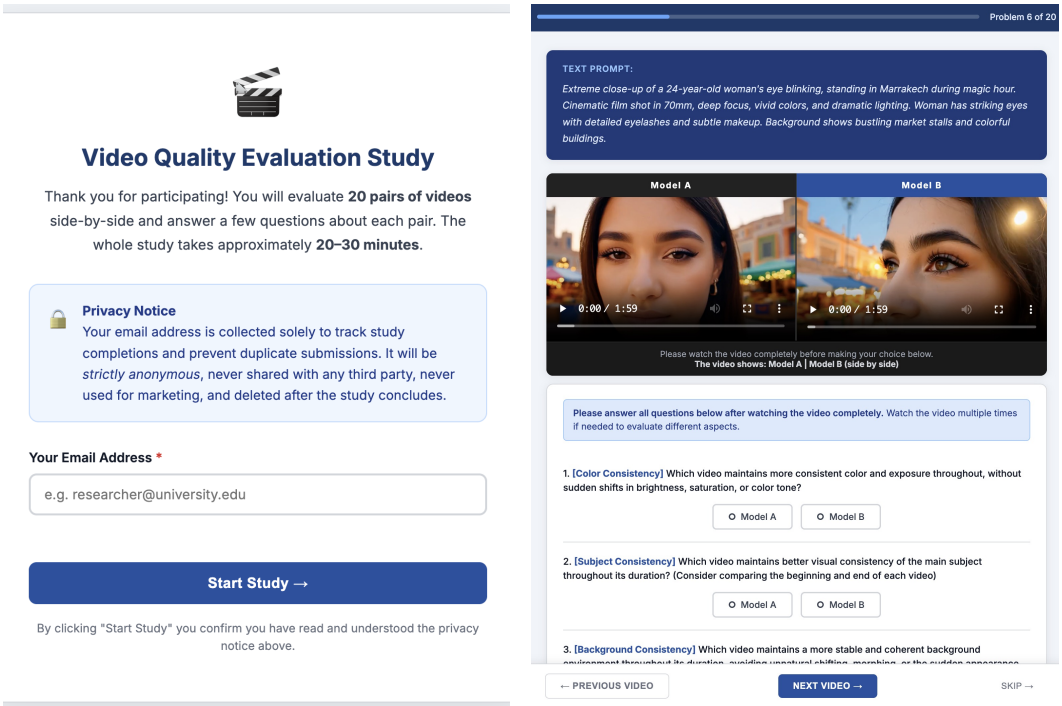


Figure 12. **Example VLM evaluation.** Gemini 3.1 Pro [14] rates each method on a 5-point scale with detailed reasoning.



(a) (b)

Figure 13. **User study interface.** (a) Welcome page with instructions. (b) Evaluation page showing side-by-side video comparison with per-dimension preference questions.

ods. For each pair, they answer six questions covering Color Consistency, Subject Consistency, Background Consistency, Text Alignment, Motion Smoothness, and Overall Preference.