RESHAPING ROPE PHASE OF SINK FRAMES: TOWARDS LONG-HORIZON CONSISTENCY IN AUTOREGRESSIVE VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive video generation (AVG) generates video content chunk by chunk, enabling long-horizon generation. In AVG, long-horizon consistency is still hard to maintain without full context. One typical solution is to keep a KV cache of beginning frames (named sink frames) and force the following generation to attend to the sink frames to maintain consistency. These methods usually suffer from the sink collapse issue, *i.e.,* in certain steps, the model tends to unexpectedly replay the content of sink frames, rather than generating new continuous content. Typical previous works observe the correlation between RoPE encoding and the sink collapse, and propose to randomly disturb the original base of RoPE encoding. However, we observe that these methods cannot fully eliminate sink collapse. Moreover, the consequences are not stable due to the randomness and they may introduce artifacts. In this paper, we rethink the effect of RoPE on sink collapse and uncover two interesting key factors. Firstly, we carefully check the components of RoPE and find that only partial frequency pairs of RoPE dominate sink collapse phenomenon. Especially, the first pair (corresponding to the highest frequency component) is important but ignored by previous solutions. Secondly, we mathematically and empirically find that in nature, the query-sink correlation varies periodically resembling a cosine function. For different heads, the phases are largely aligned. Thus, at certain steps, all the heads may simultaneously pay high attention to the sink frames, causing sink collapse issue. Based on our findings, we propose to reshape the RoPE phase of sink frames to break the phase alignment of query-sink correlation across heads to mitigate sink collapse. Our operation is simple: we construct head-wise offsets and add them to the temporal RoPE index of sink frames when computing the query-sink attention. Consequently, the phases for different heads become different. Our method is training-free, deterministic and substantially reduces artifacts. Extensive experiments on long-horizon video generation show that our method substantially suppresses sink collapse, outperforming previous state-of-the-arts. Besides, our method largely maintains generation quality close to vanilla models.

1 INTRODUCTION

Recently, autoregressive video generation has become a popular paradigm which generates video content chunk by chunk and enables long-horizon generation. However, minute-long or even longer generation is not simply a matter of sampling more frames: each newly generated frame becomes part of the context for subsequent generation, allowing inconsistencies to accumulate over time. To maintain long-horizon consistency, models such as LongLive (Yang et al., 2025a) and Rolling Forcing (Liu et al., 2026) retain a small set of initial frames as persistent sink frames that provide a long-term anchor. However, models using this design exhibit a distinct failure mode: at certain moments during long-horizon generation, the content can abruptly revert to the appearance of the sink frames, severely disrupting temporal consistency. This unexpected phenomenon is known as *sink collapse*.

Previous typical work, such as LoL (Cui et al., 2026b), attributes sink collapse to collective synchronization of attention across heads, where multiple heads simultaneously assign strong attention to

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

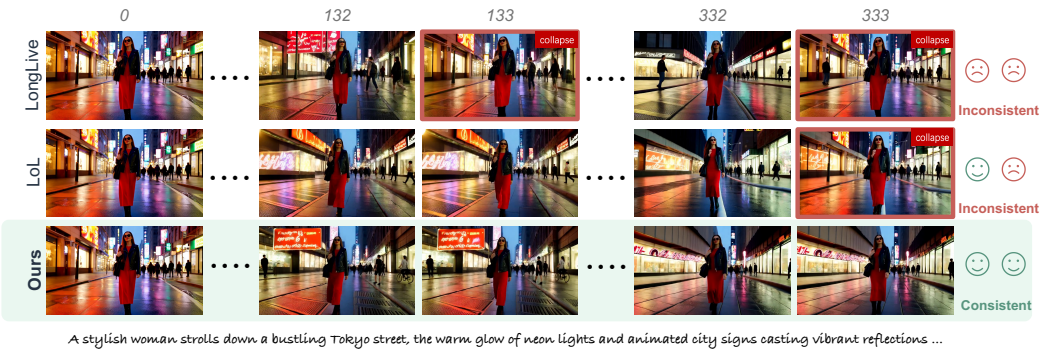


Figure 1: **RePhase RoPE mitigates sink collapse for temporally consistent video generation.** Autoregressive video models with sink frames can exhibit sink collapse at certain times over long horizons. Previous methods do not mitigate collapse across all these times. In contrast, RePhase RoPE avoids these collapses, preserving long-horizon temporal consistency. Numbers above the columns denote latent frame indices.

persistent sink frames. To disrupt this synchronization, LoL introduces multi-head RoPE-base jitter, assigning each attention head a RoPE base different from that used during pretraining and thereby inducing distinct temporal frequencies across heads.

However, several practical limitations remain. Although LoL suppresses many prominent failure cases, it does not fully eliminate sink collapse: collapse still occurs in multiple cases and may re-emerge at longer generation horizons. Its stochastic RoPE-base perturbation also introduces additional method-level randomness, leading to substantial variation in collapse suppression across different seeds. Moreover, the modified frequencies are applied not only to persistent sink keys but also to sliding-window keys throughout inference. In practice, we observe obvious degradation in generation quality compared to the base model, including flickering and local artifacts.

To better understand the mechanisms underlying sink collapse, we conduct a series of experiments and uncover two key findings. First, temporal RoPE pairs contribute unevenly to sink collapse. Targeted phase interventions identify the five highest-frequency pairs (pairs 0–4) as dominant: perturbing these pairs substantially alters the occurrence and timing of collapse, whereas interventions on the remaining pairs have little effect. Pair-wise query-sink response decomposition further shows that pair 0 exhibits the largest measured response among all temporal pairs. This exposes an important limitation of RoPE-base jitter: disturbing the base of RoPE will not affect the query-sink correlation strength in terms of pair 0.

Second, we mathematically and empirically find that in nature, the query-sink correlation varies periodically resembling a cosine function and for different heads, the correlation phase are largely aligned. Formally, the correlation phase consists of two terms. The first term is the difference of RoPE temporal phase between query and sink key, and the second term is the angle between query and sink key. The first term is the same across heads at certain steps, while for the second term, we empirically find it varies little across different heads and different generation steps. These two properties lead to strong phase alignment across heads. As the relative distance between sink frames and query frames increases, the first term varies, making the correlation strength vary periodically. At certain steps, all the heads may simultaneously pay high attention to the sink frames, which makes the query embedding largely determined by the sink frames and thus causes sink collapse issue.

Based on the two findings, we propose a new method, named RePhase RoPE, which aims to break the phase alignment of query-sink correlation through reshaping the RoPE phase of sink frames. Specifically, we construct head-wise offsets and add them to the temporal RoPE index of sink frames when computing query-sink correlation. We mathematically derive the optimal offsets to minimize the effect of phase alignment across heads, *i.e.,* evenly distribute the offsets within a certain interval. In practice, we set the interval to the largest cycle among our discovered dominant pairs of RoPE, *i.e.,* the cycle of the fifth pair of RoPE.

In a nutshell, our contributions can be summarzed as follows

- We uncover two key factors in the occurrence of sink collapse beyond previous work. Firstly, the sink collapse is highly correlated with a small set of RoPE frequency pairs, with the first pair (the highest frequency pair) ignored by previous solution. Secondly, in nature, the query-sink correlation varies periodically resembling a cosine function. For different heads, the phases are largely aligned. At certain steps, the heads may simultaneously pay high attention to the sink frames, causing sink collapse issue.

- We propose a new method named RePhase RoPE, which reshapes the RoPE phase of sink frames to break the phase alignment of query-sink correlation across heads to mitigate sink collapse. Specifically, we construct head-wise offsets and add them to the temporal RoPE index of sink frames when computing query-sink correlation. As a result, the phase for different heads becomes different. Our operation is training-free, deterministic and substantially reduces artifacts.

- Extensive experiments show that our method substantially suppresses sink collapse, outperforming previous state-of-arts. Considering the quality of generated content, our method can largely preserve generation quality compared to vanilla models. Ablation studies verify the effectiveness of each design in our method.

2 Related work

Video Diffusion Generation Early video diffusion models with temporal UNets were limited to short clips with simple motion (Ronneberger et al., 2015; Ho et al., 2022; Singer et al., 2022; Blattmann et al., 2023), while scaling DiT-based models with large datasets enabled high-fidelity and dynamic generation (Peebles & Xie, 2023; Lu et al., 2024; Ma et al., 2024; Gupta et al., 2024; Brooks et al., 2024; Seedance et al., 2026; Wan et al., 2025; Kong et al., 2024; Yang et al., 2025b; Team et al., 2025). However, their bidirectional attention over frame tokens scales quadratically with sequence length, making long-horizon generation prohibitively expensive.

Autoregressive Video Generation AR generation makes long videos tractable via KV caching, with per-step cost linear in cached context or constant with a sliding window, enabling indefinite generation. Early models predicted pixels or discrete tokens (Weissenborn et al., 2019; Yan et al., 2021; Ge et al., 2022), later continuous latents (Deng et al., 2025). Diffusion-based approaches achieve causality either during training (Chen et al., 2024; Ruhe et al., 2024) or by distilling teachers into causal students with DMD and related methods (Yin et al., 2024b;a; 2025; Huang et al., 2026; Cui et al., 2026a; Zhu et al., 2026; Feng et al., 2026). Because generated frames condition future predictions, errors accumulate over long rollouts (Huang et al., 2026; Yang et al., 2025a; Liu et al., 2026). Inspired by attention sinks (Xiao et al., 2024), video generators retain initial frames as persistent anchors (Yang et al., 2025a; Liu et al., 2026; Shin et al., 2026), which later work refines (Kim et al., 2026; Wu et al., 2026; Ye et al., 2026; Mao et al., 2026; Zhao et al., 2026; Tian et al., 2026; Yang et al., 2026; Lu et al., 2026; Yi et al., 2025).

Long-term Consistency Maintenance Video positional encoding inherits LLM designs: absolute encodings extrapolate poorly (Vaswani et al., 2017; Devlin et al., 2019; Radford et al., 2018; Dosovitskiy et al., 2020), whereas relative designs(Shaw et al., 2018; Su et al., 2024) are extended by frequency rescaling (Chen et al., 2023; bloc97, 2023; Peng et al., 2024; Ding et al., 2024). Video transformers use factorized 3D RoPE with a temporal axis, whose periodicity causes extrapolation failure: RIFLEx lowers intrinsic frequency in bidirectional DiTs to mitigate motion repetition (Zhao et al., 2025) and other work shows positional encoding is also problematic in AR generation (Li et al., 2026; Yesiltepe et al., 2026). Persistent sink frames exacerbate a failure mode, sink collapse, in which attention heads over-synchronize on the sink frames. Existing remedies specifically target this issue by acting on positional encoding: Deep Forcing and Helios make retained frames and the current window temporally adjacent (Yi et al., 2025; Yuan et al., 2026), a head-shared shift leaving cross-head relative phase unchanged; LoL attributes the failure to heads synchronizing on the sink frames and stochastically jitters per-head RoPE bases (Cui et al., 2026b), unable to reach the pair 0 and also perturbing local-context keys. We instead apply a fixed, head-wise phase offset only to persistent sink keys, spreading their responses across heads without perturbing pretrained frequencies or local-context keys.

3 METHOD

3.1 PRELIMINARIES: TEMPORAL ROPE AND SINK COLLAPSE

Temporal RoPE in video generation. Modern video generation models usually employ spatiotemporal rotary position embeddings (RoPE) (Su et al., 2024) to inject positional information into self-attention. We focus on the temporal component: all spatial tokens in the same latent frame share a temporal index. Temporal RoPE rotates each two-dimensional pair by an angle proportional to the frame index. If d dimensions are assigned to the temporal component, they form $P = d/2$ pairs. For pair $i \in \{0, \dots, P-1\}$ and RoPE base θ, the angular frequency is

$$\omega_i = \theta^{-i/P}. \tag{1}$$

A specific pair of query/key embeddings at temporal index t is rotated by $R(t\omega_i)$, *i.e.,*

$$\hat{q}_i = R(t\omega_i)q_i, \quad \hat{k}_i = R(t\omega_i)k_i \tag{2}$$

where

$$R(t\omega_i) = \begin{pmatrix} \cos t\omega_i & -\sin t\omega_i \\ \sin t\omega_i & \cos t\omega_i \end{pmatrix}. \tag{3}$$

Considering query pair $q_{i,t}$ at latent-frame index t and key pair $k_{i,s}$ at index s, the correlation value contributed by pair i can be represented as

$$\langle R(t\omega_i)q_{i,t}, R(s\omega_i)k_{i,s}\rangle = \big\langle q_{i,t}, R\big((s-t)\omega_i\big)k_{i,s}\big\rangle = \|q_{i,t}\|_2\|k_{i,s}\|_2 \cos\big(D\omega_i + \phi_{i,t,s}\big), \tag{4}$$

where $\phi_{i,t,s}$ is the angle between $q_{i,t}$ and ki, t. The $D = t - s$ denotes the distance between query and sink frames For a persistent sink key, s remains fixed as generation advances, so the query-sink distance D grows with t.

We denote the correlation phase in Eq. (4) by $\Phi_{i,t,s} = D\omega_i + \phi_{i,t,s}$. As discussed in Section 3.3, these phases are largely aligned across heads.

Sink frames and sink collapse. To provide context for current chunk generation, autoregressive video generation models often maintain KV cache of recent frames in a sliding window and those of several initial frames. The initial frames are known as sink frames, to give later chunks access to early visual context and help maintain long-term consistency. However, generation can unexpectedly replay the content of sink frame during long-horizon generation, a phenomenon known as sink collapse. LoL (Cui et al., 2026b) observes that sink collapse events coincide with multiple attention heads assigning high attention to sink frames. To mitigate these events, it proposes multi-head RoPE jitter to disperse the heads' responses to sink frames. Specifically, it samples a separate RoPE base for each head:

$$\theta_h = \theta(1 + \sigma_\theta \epsilon_h), \qquad \epsilon_h \sim \mathcal{U}[-1, 1], \qquad \omega_{h,i} = \theta_h^{-i/P}, \tag{5}$$

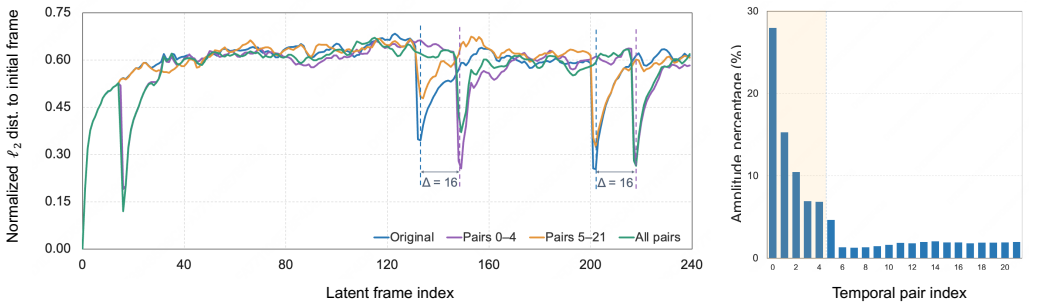
where θ is the original base and σ_θ controls the jitter range. The resulting head-specific frequencies not only modify attention to persistent sink frames but also unexpectedly change how queries attend to recent frames in the sliding window.

3.2 TEMPORAL ROPE PAIRS CONTRIBUTE UNEQUALLY TO SINK COLLAPSE.

We first ask whether different temporal RoPE pairs contribute equally to sink collapse. We conduct the pair-level and cross-head analysis on LongLive (Yang et al., 2025a), with three initial latent frames retained as persistent sink frames. To examine their contributions, we apply a same temporal frame offset to pairs 0–4, pairs 5–21 and all temporal pairs in the persistent sink keys respectively and compare the resulting generations with the vanilla model. Figure 2(a) shows the normalized ℓ_2 distance between generated frames and the initial frame, with troughs associated with the observed sink collapse. Applying the offset to pairs 0–4 delays the observed collapse by an amount equal to the applied offset, whereas applying it to pairs 5–21 produces no observable change. Extending the intervention to all temporal pairs yields a similar delay to that observed for pairs 0–4 alone. Together, these results identify pairs 0–4 as the dominant temporal RoPE pairs governing the observed shifts in sink collapse timing.

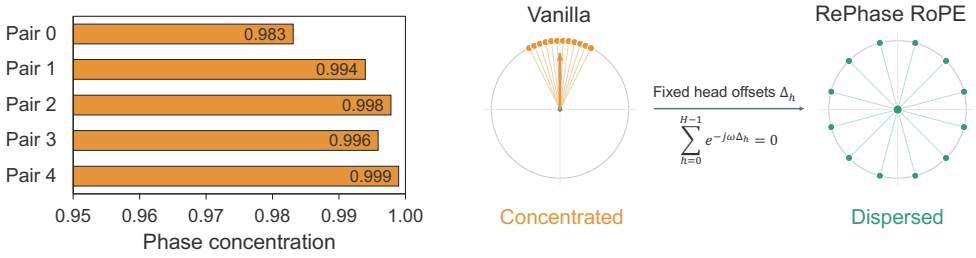
To further examine the contributions of these dominant pairs, we quantify the phase-independent response amplitude of each pair using query and persistent-sink key embeddings collected during generation:

$$A_i = \mathbb{E}\big[\|q_i\|_2\|k_i^S\|_2\big], \tag{6}$$



(a) Effect of temporal-offset interventions on sink collapse. (b) Per-pair amplitude percentage.

Figure 2: **Temporal-pair contributions to sink collapse.** (a) We apply a $+16$ latent frame offset to pairs 0–4, pairs 5–21 and all temporal RoPE pairs in persistent sink keys. Lower normalized ℓ_2 distance to the initial frame suggests more severe collapse. (b) Per-pair share of the query–sink response amplitude A. Despite comprising only $5/22$ of the temporal pairs, pairs 0–4 account for 67.51% of the total amplitude.



(a) Cross-head phase concentration (b) Dispersing phases across heads

Figure 3: **Cross-head phase alignment and dispersion.** (a) Phase concentration exceeds 0.98 for pairs 0–4 throughout generation. (b) Fixed head-wise offsets disperse the aligned phases.

where the expectation averages over query–sink token pairs within configurations sampled across prompts, layers, attention heads, denoising calls, and generation steps. Because RoPE preserves query and key norms, A_i measures the phase-independent amplitude scale associated with pair i.

Figure 2(b) reveals a strongly concentrated pair-wise response profile. Although pairs 0–4 constitute only five of the 22 temporal pairs, they account for a substantial majority of the overall response under A_i. Among these dominant pairs, pair 0 exhibits the largest measured response, making its behavior under RoPE-base jitter particularly important. Since $\omega_i = \theta^{-i/P}$, the frequency of pair 0 yields $\omega_0 = \theta_h^0 = 1$, which is independent of the jitter, regardless of whether it is applied to the RoPE base or the query–sink distance across all heads. Combined with its largest measured response among the temporal pairs, this exact invariance exposes a structural limitation of RoPE-base jitter: pair 0 cannot be jittered through the RoPE base.

3.3 Cross-Head Phase Alignment

We next examine why the dominant pairs produce large responses at particular query-sink distances and how their response phases align across attention heads.

Phase Alignment across Heads. The positional phase of each pair-wise query–sink response is determined by the query–sink distance D through $D\omega_i$. For a given pair, ω_i is shared across attention heads. For each head and temporal RoPE pair, we record the query–sink response over generation frames. At each frame, we normalize the query–sink response by the product of the query and sink-key norms. We average each head's query–sink response contributions over layers and denoising calls, then extract phases from Fourier coefficients at each pair's native RoPE frequency using sliding Hann windows. Figure 3(a) shows that the dominant pairs response phases are highly concentrated across heads. The result indicate that cross-head phase coordination persists during generation rather than arising only at collapse. As D increases, the shared frequency adds the same positional phase

increment to every head. Together with the observed phase concentration and stable relative offsets, this helps explain why responses within the same pair tend to rise and fall together across heads.

From pair-wise contributions to sink-attention peaks. As the query–sink distance increases, the dominant pairs evolve at different RoPE frequencies and can produce large positive contributions simultaneously at certain distances. Cross-head phase alignment allows these contributions to occur across multiple heads at similar distances, jointly increasing their query–sink responses. Consistent with this account, changes in sink attention mass exhibit strong positive cross-head alignment, indicating coordinated increases and decreases in attention to sink frames (Figure 4(a)). These observations motivate spreading the phases of persistent sink keys across heads to reduce simultaneous positive query–sink response peaks.

3.4 REPHASE ROPE

Motivated by the analysis above, we introduce RePhase RoPE, a training-free method that reshapes the temporal RoPE phases of persistent sink keys across heads by assigning a fixed temporal offset to each head. Different offsets produce different phase shifts across heads while preserving the original RoPE frequencies. Note that we only apply head-wise offset to the RoPE of sink frames, leaving current queries and keys in sliding window unchanged.

Persistent-sink phase shift. Figure 3(b) illustrates the overall motivation. We assign each attention head h a fixed temporal offset Δ_h, *i.e.,*

$$k^S_{h,i} \leftarrow R(\omega_i \Delta_h) k^S_{h,i}, \qquad i = 0, \ldots, P-1, \tag{7}$$

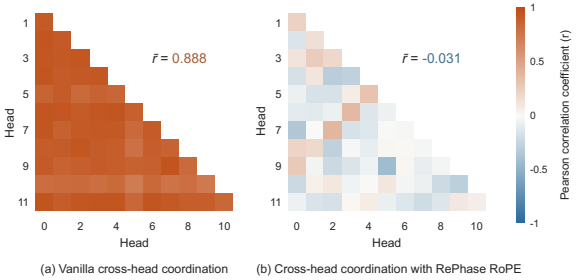
where ω_i is the original RoPE frequency of pair i. For a sink key at position s, this additional rotation is equivalent to computing its temporal RoPE using $s + \Delta_h$ instead of s. Its query–sink distance therefore changes from D to $D - \Delta_h$. Within each head, the same Δ_h is applied to all persistent sink frames and all temporal RoPE pairs. This preserves the temporal separations between sink frames, while leaving their spatial RoPE components unchanged. Algorithm 1 summarizes our method.

Reducing simultaneous response peaks. The dominant pairs exhibit aligned response phases across heads, as shown in Sec 3.3. This indicates that each dominant pair i shares a common content phase ϕ_i across heads. We keep ϕ_i fixed and vary the query–sink distance D, so that the phase change in each head comes from the term $\omega_i D$. To examine how the reshaped positional phases behave as generation continues, consider the cross-head alignment:

$$C_i(D) = \sum_{h=0}^{H-1} e^{j(\omega_i D + \phi_i)}. \tag{8}$$

Applying different offsets Δ_h in head h introduces different phase shifts $-\omega_i \Delta_h$, yielding

$$C_i^{\boldsymbol{\Delta}}(D) = \sum_{h=0}^{H-1} e^{j(\omega_i D + \phi_i - \omega_i \Delta_h)} = e^{j(\omega_i D + \phi_i)} \underbrace{\sum_{h=0}^{H-1} e^{-j\omega_i \Delta_h}}_{S_i(\boldsymbol{\Delta})}. \tag{9}$$



(a) Vanilla cross-head coordination (b) Cross-head coordination with RePhase RoPE

Figure 4: **Effect of RePhase RoPE on sink-attention changes.**

Algorithm 1 RePhase RoPE

Require: Frequencies $\{\omega_i\}_{i=0}^{P-1}$; H heads; sink-key cache k^S

Once per unmodified cache, after all sink frames leave the sliding window.

$T_4 \leftarrow 2\pi/\omega_4$
for $h = 0, \ldots, H-1$ **do**
  $\Delta_h \leftarrow \left(h - \frac{H-1}{2}\right) \frac{T_4}{H}$
  for $i = 0, \ldots, P-1$ **do**
    $k^S_{h,i} \leftarrow R(\omega_i \Delta_h) k^S_{h,i}$
  end for
end for
return k^S

where H is the number of attention heads. Writing the phase sum in polar form as $S_i(\mathbf{\Delta}) = |S_i(\mathbf{\Delta})|e^{j\alpha_i}$, we obtain

$$\operatorname{Re} C_i^{\Delta}(D) = |S_i(\mathbf{\Delta})| \cos(\omega_i D + \phi_i + \alpha_i). \tag{10}$$

As D varies continuously over one period $T_i = 2\pi/\omega_i$, the real part of the cross-head aggregate attains a maximum equal to its cosine amplitude $|S_i(\mathbf{\Delta})|$. Thus, minimizing the peak aggregate response is equivalent to minimizing the magnitude of the phase sum:

$$\min_{\mathbf{\Delta}} \max_{D\in[0,T_i]} \operatorname{Re} C_i^{\mathbf{\Delta}}(D) = \min_{\mathbf{\Delta}} \left| \sum_{h=0}^{H-1} e^{-j\omega_i \Delta_h} \right|. \tag{11}$$

Without intervention, the phase vectors add coherently, giving a maximum response of H. Any offsets satisfying $S_i(\mathbf{\Delta}) = 0$ reduce this maximum to its minimum possible value of zero. Moreover, Eq. 9 shows that this cancellation holds for every distance D, rather than only at a particular timestep.

Constructing head-wise offsets. To spread the positional phases across heads, we construct a fixed set of uniformly spaced temporal offsets, which satisfies $S_i(\mathbf{\Delta}) = 0$. We determine their scale from the dominant pairs identified in Sec. 3.2. Let $T_i = 2\pi/\omega_i$ denote the temporal period of pair i. Since pair 4 has the longest period among pairs 0–4, we select offsets over $[-T_4/2, T_4/2]$, an interval that spans at least one full period of every dominant pair. For the number of heads H, we place H offsets at regular intervals of T_4/H within this range:

$$\Delta_h = \left(h - \frac{H-1}{2}\right)\frac{T_4}{H}, \qquad h = 0, \ldots, H-1. \tag{12}$$

The offsets are assigned deterministically to the persistent sink keys in each attention head, using the same head-wise offsets across layers and prompts. For pair 4, these offsets place the positional phases of the H heads evenly around the circle, yielding

$$S_4(\mathbf{\Delta}) = e^{j\pi(H-1)/H} \sum_{h=0}^{H-1} e^{-j2\pi h/H} = 0. \tag{13}$$

Thus, the ideal offset grid gives $S_i(\mathbf{\Delta}) = 0$ for every query–sink distance D. As generation continues, these positional phases remain evenly spread across heads instead of reaching the positive peak of their cosine terms together. For the other dominant pairs, the same offsets produce phase shifts $-\omega_i\Delta_h$ according to their original frequencies, although their unit vectors need not sum exactly to zero.

4 Experiment

4.1 Experimental Setup

Backbone and evaluation set. We evaluate LongLive (Yang et al., 2025a) and Causal Forcing (Zhu et al., 2026) on the same 128 MovieGenBench (Polyak et al., 2024) prompts, generating 75- and 100-second videos with the first three latent frames retained as persistent sinks. All experiments use NVIDIA H100 80GB GPUs.

Compared methods. We compare existing methods including position interpolation (PI) (Chen et al., 2023), NTK-aware RoPE scaling(bloc97, 2023), YaRN (Peng et al., 2024), and RIFLEx (Zhao et al., 2025). The RoPE modifications and implementation details of the compared methods are provided in Appendix A. We implement LoL (Cui et al., 2026b) following its paper, using $\sigma = 0.8$ and 5 RoPE-base jitter seeds. Our method uses fixed offsets without additional method randomness.

Metrics. Following LoL (Cui et al., 2026b), we measure sink collapse using the largest consecutive drop in normalized distance to the first frame. We sample decoded frames at 4 Hz and define

$$d_n = \frac{\|I_n - I_0\|_2}{\|I_0\|_2}, \qquad \mathrm{SC}(V) = 100 \max_{0\le n<N-1}(d_n - d_{n+1}). \tag{14}$$

Larger scores indicate more abrupt returns toward the initial frame. We report the mean and maximum per-video scores across prompts as SC Avg and SC Max. To assess generation quality, we use



Figure 5: Qualitative comparison with baseline methods.

Table 1: Comparison with sink-collapse mitigation methods in terms of sink collapse mitigation and generation quality. All results are reproduced under a unified benchmark. SC denotes Sink Collapse score.

Method	Sink Collapse			Generation Quality							
	SC Avg ↓	SC Max ↓	Dynamic Degree↑	Temporal Quality ↑	Text Alignment ↑	CLIP Score ↑	Subject Consistency↑	Background Consistency↑	Motion Smoothness ↑	Imaging Quality ↑	Overall Consistency ↑
Results on LongLive											
75 seconds											
Baseline	27.13	112.09	40.16	89.58	29.28	33.18	97.84	96.66	98.94	69.64	25.38
PI	1.78	17.59	0.76	84.70	26.35	30.76	98.74	97.74	99.59	60.01	21.94
NTK	15.73	52.52	33.45	88.75	29.26	33.15	98.04	96.82	99.03	69.59	25.37
YaRN	3.20	18.89	1.90	84.89	29.39	33.49	99.04	97.63	99.50	70.21	25.30
RIFLEx	27.50	69.69	40.05	89.56	29.32	33.20	97.82	96.68	98.94	69.59	25.44
LoL	5.15 ± 1.03	23.05 ± 14.02	38.88 ± 2.32	89.36±0.28	29.14±0.06	33.09±0.05	97.77±0.09	96.60±0.06	98.95±0.04	69.29±0.27	25.19±0.08
LoL (best of 5)	4.30	11.16	37.63	89.22	29.08	33.04	97.84	96.63	98.98	69.14	25.12
RePhase RoPE (ours)	3.26	7.87	39.44	89.43	29.17	33.12	97.74	96.59	98.96	69.38	25.22
100 seconds											
Baseline	30.40	148.15	39.58	89.49	29.27	33.18	97.83	96.66	98.95	69.59	25.36
PI	2.05	18.46	0.69	84.69	25.99	30.39	98.75	97.76	99.55	58.41	21.60
NTK	15.76	52.52	32.95	88.67	29.22	33.10	98.03	96.83	99.02	69.56	25.33
YaRN	4.37	23.04	1.80	84.86	29.33	33.40	99.02	97.61	99.49	70.10	25.25
RIFLEx	30.00	124.88	39.05	89.42	29.32	33.21	97.82	96.68	98.95	69.59	25.44
LoL	6.31 ± 2.15	29.83 ± 11.51	37.87 ± 1.86	89.21±0.22	29.09±0.08	33.04±0.07	97.77±0.09	96.59±0.05	98.95±0.04	69.21±0.28	25.14±0.10
LoL (best of 5)	4.67	17.65	36.61	89.07	29.02	32.98	97.83	96.62	98.99	69.01	25.05
RePhase RoPE (ours)	3.41	7.87	38.98	89.35	29.09	33.02	97.73	96.56	98.96	69.23	25.15
Results on Causal Forcing											
75 seconds											
Baseline	35.38	96.23	72.45	92.59	26.81	30.89	95.27	95.35	97.21	68.04	22.74
PI	4.79	28.31	15.77	84.77	17.86	24.19	95.60	96.59	96.63	47.25	11.54
NTK	22.70	60.17	69.70	92.00	26.24	30.41	94.96	95.29	96.91	66.09	22.07
YaRN	5.87	35.50	26.52	86.17	21.43	26.85	95.17	96.15	97.01	55.43	16.02
RIFLEx	36.12	104.24	71.88	92.49	26.79	30.92	95.24	95.35	97.19	67.83	22.66
LoL	10.51±0.45	86.13±27.13	66.40±2.26	91.44±0.30	24.71±0.29	29.16±0.22	94.73±0.08	95.29±0.04	96.83±0.06	64.88±0.37	20.25±0.36
LoL (best of 5)	10.18	117.86	66.83	91.56	24.35	28.89	94.84	95.32	96.90	64.25	19.82
RePhase RoPE (ours)	8.67	27.97	67.46	91.55	24.42	28.94	94.69	95.28	96.73	65.05	19.89
100 seconds											
Baseline	38.14	96.23	71.20	92.39	26.61	30.72	95.26	95.36	97.14	67.62	22.49
PI	5.38	28.31	17.33	84.82	16.50	23.16	95.35	96.53	96.33	44.39	9.84
NTK	23.06	60.17	66.67	91.43	25.20	29.57	94.82	95.31	96.55	63.75	20.84
YaRN	9.11	54.58	28.09	86.23	19.84	25.63	94.84	96.09	96.84	50.91	14.05
RIFLEx	38.43	104.24	70.28	92.24	26.57	30.75	95.23	95.37	97.10	67.54	22.40
LoL	11.76±0.56	96.13±12.65	62.91±3.07	90.92±0.42	23.96±0.40	28.56±0.29	94.69±0.05	95.34±0.05	96.72±0.04	63.54±0.58	19.36±0.50
LoL (best of 5)	10.97	117.86	63.33	91.01	23.46	28.19	94.75	95.36	96.78	62.62	18.73
RePhase RoPE (ours)	10.18	68.52	63.63	90.94	23.50	28.18	94.60	95.32	96.55	63.56	18.82

Red and Green colors indicate the presence of severe repetition or substantially reduced motion.

VBench-Long (Huang et al., 2024; 2025) to evaluate subject consistency, background consistency, motion smoothness, dynamic degree, aesthetic quality, and imaging quality. Following VBench's weighting scheme, Temporal Quality aggregates subject consistency, background consistency, motion smoothness, and dynamic degree with weights $1 : 1 : 1 : 0.5$, while Framewise Quality averages aesthetic and imaging quality. We additionally evaluate text–video alignment using CLIP Score (Hessel et al., 2021) and Overall Consistency, reporting their average as Text Alignment.

4.2 Main Results

Sink collapse mitigation. Table 1 compares sink collapse at 75 and 100 seconds. RePhase RoPE achieves lower SC Avg and SC Max than all five tested LoL seeds at both durations. Compared with the selected best LoL seed, it reduces the 100-second SC Avg from 4.67 to 3.41 and SC Max from 17.65 to 7.87. Although PI achieves a lower SC Avg, both PI and YaRN substantially reduce Dynamic Degree, indicating reduced motion. RIFLEx retain large collapse scores, while NTK-aware scaling provides only a partial reduction. Figure 5 provides qualitative comparisons.

Further comparison with other models. We further compare the generation quality of our method with other autoregressive video generation models. For Self Forcing (Huang et al., 2026) and Causal

Table 2: Generation-quality comparison with other autoregressive video generation models.

Method	Results on 75 seconds ↑					Results on 100 seconds ↑				
	Subject Consistency	Background Consistency	Text Alignment	CLIP Score	Framewise Quality	Subject Consistency	Background Consistency	Text Alignment	CLIP Score	Framewise Quality
Self-Forcing	96.38	95.75	28.42	32.37	62.42	96.42	95.78	28.28	32.25	62.33
Causal Forcing	95.19	95.40	26.72	30.92	61.19	95.19	95.40	26.49	30.73	60.76
Rolling Forcing	97.00	96.05	27.49	31.56	62.61	96.97	96.04	27.83	31.85	62.45
LongLive	97.84	96.66	29.28	33.18	64.84	97.83	96.66	29.27	33.18	64.77
MemRoPE	97.82	96.69	28.76	32.63	64.80	97.82	96.67	28.66	32.51	64.64
Reward Forcing	96.86	96.07	28.91	32.98	63.98	96.84	96.05	28.89	32.95	63.89
RePhase RoPE	97.74	96.59	29.17	33.12	64.48	97.73	96.56	29.09	33.02	64.28

Table 3: Generation quality of LongLive with our method versus an inference-time relative-RoPE substitution.

Method	75 seconds				
	Subject Consistency↑	Background Consistency↑	Text Alignment↑	CLIP Score↑	Framewise Quality↑
Moving Relative RoPE	97.55	96.49	28.42	32.40	63.96
Relative RoPE	97.61	96.55	28.39	32.38	64.08
RePhase RoPE (ours)	**97.74**	**96.59**	**29.17**	**33.12**	**64.48**
Method	100 seconds				
	Subject Consistency↑	Background Consistency↑	Text Alignment↑	CLIP Score↑	Framewise Quality↑
Moving Relative RoPE	97.51	96.46	28.23	32.23	63.78
Relative RoPE	97.58	96.53	28.26	32.25	63.89
RePhase RoPE (ours)	**97.73**	**96.56**	**29.09**	**33.02**	**64.28**

Table 4: Effect of different offsets. For, random non-uniform offsets, we report mean and standard deviation over three seeds.

Method	75 seconds		100 seconds	
	SC Avg ↓	SC Max ↓	SC Avg ↓	SC Max ↓
RePhase RoPE (0.5× offset magnitude)	3.68	13.33	3.98	13.33
RePhase RoPE (2× offset magnitude)	3.95	11.46	4.19	12.69
Random non-uniform offsets	4.12±0.28	17.90±6.58	4.36±0.20	17.90±6.58
RePhase RoPE (default)	**3.26**	**7.87**	**3.41**	**7.87**

Forcing (Zhu et al., 2026), we enable the sink frame with the default sink size. For Rolling Forcing (Liu et al., 2026) and Reward Forcing (Lu et al., 2026) we enable their officially supported attention-sink option with three sink frames. We also evaluate MemRoPE Kim et al. (2026) applied to LongLive. These results complement Table 1, showing that sink collapse is mitigated without affecting overall generation quality.

4.3 Ablation studies

Effect on sink attention. RePhase RoPE reduces cross-head correlation of sink-mass changes from 0.888 to -0.031 in the illustrated example (Fig. 4(b)). The mean peak increase in head-averaged sink mass drops by 87.5%, while the mean sink mass drops by only 2.5%, suggesting reduced synchrony while sink attention is largely preserved.

Comparision with relative RoPE Table 3 compares RePhase RoPE with Relative RoPE (Yuan et al., 2026) and Moving Relative RoPE (Yi et al., 2025), both applied to LongLive. These two variants also reduce sink collapse, but show some degradation in overall generation quality compared with RePhase RoPE. Implementation details and qualitative results are provided in Appendix B.

Effect of different offsets. We further examine the two key design choices of the phase offsets: its overall phase span and its deterministic placement across attention heads. We scale the default offsets by 0.5 and 2 or replace them with independently sampled offsets within the same span. As shown in Table 4, the default configuration achieves the best performance for mitigating sink collapse.

5 Conclusion

In this work, we investigated the effect of RoPE on sink collapse in autoregressive video generation. We identified two key factors: a small set of high-frequency temporal RoPE pairs dominate the observed collapse behavior, and query–sink correlations exhibit cosine-like variation with phases largely aligned across attention heads. In particular, pair 0 exhibits the largest measured response amplitude, yet its frequency remains unchanged under RoPE-base jitter. Based on these findings, we proposed RePhase RoPE, a training-free, deterministic method that reshapes the RoPE phase of sink frames to break the phase alignment of query–sink correlations across heads. By adding fixed head-wise offsets to the temporal RoPE indices of persistent sink keys, our method preserves the original RoPE frequencies and leaves current queries and sliding-window keys unchanged. Experiments on LongLive and Causal Forcing demonstrate substantial suppression of sink collapse, while results on LongLive show that generation quality remains close to the vanilla model. These findings provide insight into the mechanisms underlying sink collapse and a simple, effective approach to improving long-horizon consistency in autoregressive video generation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation and language polishing. We have not used generative AI tools for formulate mathematical claims and provide critical ingredients for proving mathematical claims. Additionally, we used generative AI tools to summarize or analyse existing literature, source/search for information, and identify relevant literature. We have reviewed all AI-assisted work. All AI-assisted text was reviewed and edited by the authors before inclusion. Generating synthetic data sets is not applicable to this work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have taken several steps to ensure the reproducibility of our work. The complete theoretical derivations and the specification of our method are provided in Sec. 3.1 3.2 3.3 3.4, with pseudocode for the proposed algorithm given in Sec. 3.4. The experimental setup, including the compared algorithms and evaluation metrics, is described in Sec. 4.1. Detailed implementation specifics, such as hyperparameter settings, are provided in Appendix A and Appendix B. We believe these materials provide sufficient information to reproduce our results.

REFERENCES

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

bloc97. NTK-Aware Scaled RoPE allows LLaMA models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation., 2023. URL https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have/.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

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.

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.

Jiaxing Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. In *International Conference on Learning Representations*, volume 2026, pp. 85802–85822, 2026a.

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*, 2026b.

Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. In *International Conference on Learning Representations*, volume 2025, pp. 44730–44745, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jiaqi Feng, Justin Cui, Yuanhao Ban, and Cho-Jui Hsieh. One-forcing: Towards stable one-step autoregressive video generation. *arXiv preprint arXiv:2605.23458*, 2026.

Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European conference on computer vision*, pp. 102–118. Springer, 2022.

Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. In *European Conference on Computer Vision*, pp. 393–411. Springer, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

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

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. doi: 10.1109/TPAMI.2025.3633890.

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

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.

Jia Li, Xiaomeng Fu, Xurui Peng, Weifeng Chen, Youwei Zheng, Tianyu Zhao, Jiexi Wang, Fangmin Chen, Xing Wang, and Hayden Kwok-Hay So. Train short, inference long: Training-free horizon extension for autoregressive video generation. *arXiv preprint arXiv:2602.14027*, 2026.

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

Haoyu Lu, Guoxing Yang, Nanyi Fei, Yuqi Huo, Zhiwu Lu, Ping Luo, and Mingyu Ding. Vdt: General-purpose video diffusion transformers via mask modeling. In *International Conference on Learning Representations*, volume 2024, pp. 19259–19286, 2024.

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

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

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

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

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 31932–31951, 2024.

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.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling diffusion models. *arXiv preprint arXiv:2402.09470*, 2024.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 464–468, 2018.

Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. Motionstream: Real-time video generation with interactive motion controls. In *International Conference on Learning Representations*, volume 2026, pp. 7381–7406, 2026.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

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

Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, et al. Longcat-video technical report, 2025. *URL https://arxiv. org/abs/2510.22200*, 2025.

Jiahao Tian, Yiwei Wang, Gang Yu, and Chi Zhang. Head forcing: Long autoregressive video generation via head heterogeneity. *arXiv preprint arXiv:2605.14487*, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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.

Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. *arXiv preprint arXiv:1906.02634*, 2019.

Mingqiang Wu, Weilun Feng, Zhefeng Zhang, Haotong Qin, Yuqi Li, Guoxin Fan, Xiaokun Liu, Zhulin An, Libo Huang, Yongjun Xu, et al. Echo-forcing: A scene memory framework for interactive long video generation. *arXiv preprint arXiv:2605.16003*, 2026.

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

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

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025a.

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

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025b.

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

Hidir Yesiltepe, Tuna Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40256–40265, 2026.

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

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

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

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

Shenghai Yuan, Yuanyang Yin, Zongjian Li, Xinwei Huang, Xiao Yang, and Li Yuan. Helios: Real real-time long video generation model. *arXiv preprint arXiv:2603.04379*, 2026.

Min Zhao, Guande He, Yixiao Chen, Hongzhou Zhu, Chongxuan Li, and Jun Zhu. Riflex: A free lunch for length extrapolation in video diffusion transformers. *arXiv preprint arXiv:2502.15894*, 2025.

Zengqun Zhao, Yanzuo Lu, Ziquan Liu, Jifei Song, Jiankang Deng, and Ioannis Patras. Relax forcing: Relaxed kv-memory for consistent long video generation. *arXiv preprint arXiv:2603.21366*, 2026.

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.

A Details of Compared Methods

Common settings. We implement all compared methods as inference-time modifications of temporal RoPE in LongLive (Yang et al., 2025a) and Causal Forcing (Zhu et al., 2026), without additional training. Each attention head has $d_t = 44$ temporal dimensions, forming $P = 22$ two-dimensional pairs. The original temporal frequencies are

$$\omega_i = \theta^{-i/P}, \qquad \theta = 10000, \qquad i = 0, \ldots, P-1. \tag{15}$$

All methods retain the original absolute temporal positions and spatial RoPE. The temporal frequencies specified below are used when rotating both queries and keys, including keys subsequently stored in the KV cache. For PI (Chen et al., 2023), NTK-aware scaling (bloc97, 2023), YaRN (Peng et al., 2024), and RIFLEx (Zhao et al., 2025), the frequencies are shared across attention heads and layers and remain fixed throughout generation. LoL (Cui et al., 2026b) instead uses a fixed set of head-specific frequencies.

Reference and target lengths. Following LoL, we use a reference length of $L = 132$ latent frames based on the first sink-collapse location. For length-dependent modifications, we set the target length to $L' = 402$ latent frames and use the scaling factor

$$s = \frac{L'}{L} = \frac{402}{132}. \tag{16}$$

The generated sequence decodes to 1605 frames, from which we retain the first 1600 frames for 100-second evaluation at 16 FPS. The 75-second evaluation uses the first 1200 frames of the same generated video. Thus, both durations use the same frequency configuration with $L' = 402$.

Position interpolation (PI). PI uniformly reduces all temporal frequencies by the length scaling factor:

$$\omega_i^{\mathrm{PI}} = \frac{\omega_i}{s}. \tag{17}$$

This is equivalent to using the temporal position t/s in the original RoPE rotation. The scaling is applied only to temporal RoPE.

NTK-aware RoPE scaling. We use the original power-law scaling of the temporal RoPE base:

$$\theta_{\mathrm{NTK}} = \theta s^{d_t/(d_t-2)} = 10000\, s^{44/42}, \qquad \omega_i^{\mathrm{NTK}} = \theta_{\mathrm{NTK}}^{-i/P}. \tag{18}$$

The exponent uses the temporal dimension $d_t = 44$. The resulting base is fixed for the full generation.

YaRN. We use cycle-based interpolation between the original and uniformly scaled temporal frequencies. For each pair, let

$$r_i = \frac{L\omega_i}{2\pi} \tag{19}$$

denote the number of rotations over the reference length. With thresholds $\beta_{\mathrm{slow}} = 1$ and $\beta_{\mathrm{fast}} = 32$, the interpolation weight is

$$\gamma_i = \begin{cases} 0, & r_i \le 1, \\ (r_i - 1)/31, & 1 < r_i < 32, \\ 1, & r_i \ge 32. \end{cases} \tag{20}$$

The modified frequency is

$$\omega_i^{\mathrm{YaRN}} = \gamma_i \omega_i + (1 - \gamma_i)\frac{\omega_i}{s}. \tag{21}$$

Pairs with fewer rotations receive more uniform scaling, while pairs with more rotations retain a larger contribution from their original frequency. Our implementation uses this linear interpolation in cycle count and does not apply additional attention-temperature scaling or rescale Q/K magnitudes.

RIFLEx. We fix the one-based pair index to $k = 8$, corresponding to pair $i = 7$ under the zero-based indexing used in this paper. Its original period is approximately 118 latent frames, the closest among the temporal pairs to the reference length $L = 132$. We replace only this pair's frequency:

$$\omega_i^{\mathrm{RIFLEx}} = \begin{cases} 2\pi/L', & i = 7, \\ \omega_i, & i \ne 7. \end{cases} \tag{22}$$

This sets its period to the target length $L' = 402$. The frequency multiplier is 1.0, and the selected pair is fixed across prompts.

LoL. We implement head-wise RoPE-base jitter following the description in LoL. For each attention head h, we sample

$$\epsilon_h \overset{\text{i.i.d.}}{\sim} \mathcal{U}(-1, 1), \qquad \theta_h = \theta(1 + \sigma\epsilon_h), \qquad \sigma = 0.8. \tag{23}$$

The corresponding temporal frequencies are

$$\omega_{h,i}^{\text{LoL}} = \theta_h^{-i/P}. \tag{24}$$

Jitter is applied to all $H = 12$ attention heads. The bases are sampled once at model initialization and shared across layers, denoising steps, and prompts within each run. Spatial RoPE remains unchanged. Since $\omega_{h,0}^{\text{LoL}} = 1$ for every head, RoPE-base jitter leaves pair 0 unchanged.

We use a dedicated CPU random-number generator for jitter sampling, separate from the generation random stream. The five jitter seeds are 8042, 9322, 10971, 23973 and 76995. Seed 76995 attains the lowest 100-second Sink-Collapse Avg on the evaluated subset and is used for the best-seed results at both durations and for the corresponding quality comparisons. We also conduct a per-prompt comparision across LoL seeds. For each seed, we compare the two methods on the same prompts and generation seed, using the first 100 seconds of each video. A strict win means that RePhase RoPE achieves a lower Collapse score than LoL for the corresponding prompt. As shown

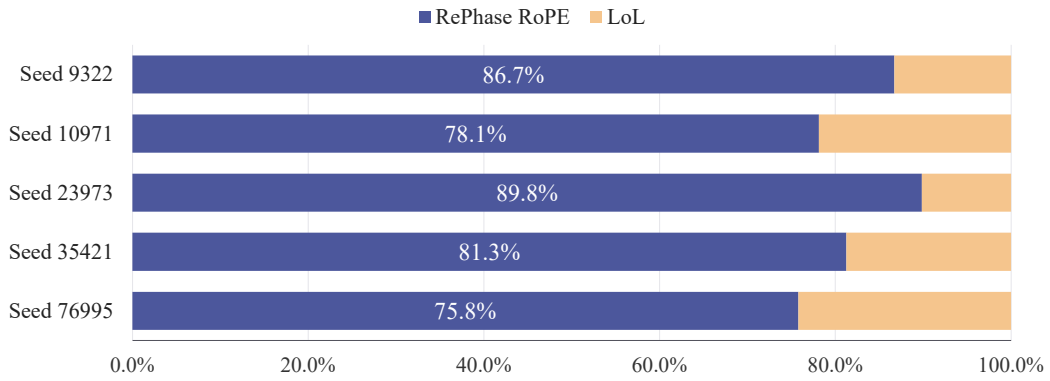


Figure 6: Per-prompt Collapse comparison between RePhase RoPE and LoL on 128 prompts. Each row corresponds to a LoL seed. Blue and orange segments show the fractions of prompts on which RePhase RoPE and LoL achieve a strictly lower Collapse score, respectively.

in Figure 6, RePhase RoPE wins on 75.8%–89.8% of prompts against each LoL seed, with no ties. These results show that its lower Collapse scores extend to a majority of prompts for every evaluated seed.

B Relative-RoPE Comparison

A natural alternative for reshaping persistent-sink RoPE phases is to prevent the query–sink positional distance from growing altogether. We test this by replacing LongLive's native temporal RoPE with relative-RoPE variants at inference time while leaving the pretrained model unchanged. Since this changes the positional geometry seen by the pretrained model, we focus on whether generation quality is preserved rather than treating it as another collapse-mitigation method. Table 3 compares our method (RePhase RoPE, which utilizes standard RoPE during inference) with the relative-RoPE configurations at both 75 and 100 seconds, using the same evaluation protocol as in Sec. 4.1.

The difference between Relative RoPE and Moving Relative RoPE is as follows: Relative RoPE uses a fixed, cache-relative temporal coordinate that does not change during generation, whereas Moving Relative RoPE, by contrast, uses a moving temporal coordinate that changes as generation proceeds.

Relative RoPE. We enable LongLive's `use_infinite_attention` option to evaluate its block-relative RoPE implementation. The model weights, original RoPE frequencies, and spatial RoPE remain unchanged. The attention cache contains $W = 12$ latent frames, including $S = 3$ persistent sink frames, and generation proceeds in blocks of $B = 3$ latent frames.

Unlike the native implementation, which caches keys after applying RoPE at their absolute temporal positions, the Relative RoPE variant caches keys before applying RoPE. At each attention call, cached keys receive temporal indices according to their current cache positions. The sink frames occupy indices $0, 1, 2$, followed by the retained recent frames. Queries use the indices assigned to the current block in the same coordinate system. In this sense, the temporal coordinate is relative: it always starts from 0, does not advance with the global generation step, and keeps the sink frames at fixed indices, while the current window slides forward and is appended after the sink frames.

Before the cache is full, indices increase sequentially with the number of cached frames. Once the cache is full, current queries use indices 9, 10, and 11, while persistent sink keys retain indices 0, 1, and 2. As generation continues, older non-sink frames are evicted and the remaining keys are assigned their updated cache-relative indices when RoPE is applied. The same indexing rule is used across attention heads. This preserves relative temporal distances within the contiguous recent context while changing the distances between that context and the persistent sink frames relative to the native absolute-RoPE implementation. No additional rotation or scaling is applied beyond this temporal re-indexing.

In short, Relative RoPE keeps the temporal coordinate fixed and cache-relative: the assigned indices do not advance with the global generation step.

Moving Relative RoPE. For Moving Relative RoPE (Yi et al., 2025), the temporal coordinate is moving rather than block-relative. At each denoising step for the current frame at temporal position t, the query is anchored at t, and the visible keys follow the moving window $[t-18, t+2]$: the query attends to 18 preceding frames, the current frame, and 2 following frames. The persistent sink frames remain in the attention cache, but their temporal RoPE indices are re-anchored to the neighborhood of the current generation block of $B = 3$ latent frames, even though they originate from much earlier positions in the sequence. In other words, the sink frames' temporal indices are shifted by an offset so that they lie near the current query block, rather than retaining their original early indices. Their cache positions and spatial RoPE indices remain unchanged; only the temporal component is shifted. This keeps the query–sink temporal distance small without modifying the cache contents. No additional rotation or scaling is applied beyond this temporal re-indexing. Before the cache is full, the window is truncated to the available frames.

In short, Moving Relative RoPE keeps the temporal coordinate moving with generation: the sink frames' effective temporal positions are placed near the current block.

Comparison protocol. We use the same prompts, model weights, generation seed, and generation random stream as the native LongLive and RePhase RoPE runs. All variants retain three persistent sink frames and use the same attention budget. We evaluate sink collapse and generation quality using the protocol described in the main text.

C Pair 0 intervention under LoL

Additionally, we examine how pair 0 contributes to the continued occurrence of collapse under LoL. As shown in Fig. 7, parts of the initial-frame content reappear in later frames under LoL. For these events, we rotate only the pair 0 components of the persistent sink keys around these events, leaving all other pairs and local-context keys unchanged. This intervention reduces the partial reappearance of initial-frame content, indicating that pair 0, which is left unchanged by LoL's RoPE-base jitter, contributes to these collapse events.

D Results on Ultra-Long-Horizon Generation

We further evaluate RePhase RoPE on 64 prompts over a 255-second generation horizon. Table 5 reports the average and maximum sink collapse (SC) scores, together with the per-prompt win rate of our method against the comparison methods. Our method achieves an average SC score of 5.20,



Figure 7: Qualitative effect of explicitly intervening pair 0. Compared with the original LoL generation, intervening on pair 0 suppresses the reappearance of initial-frame content around the collapse event.

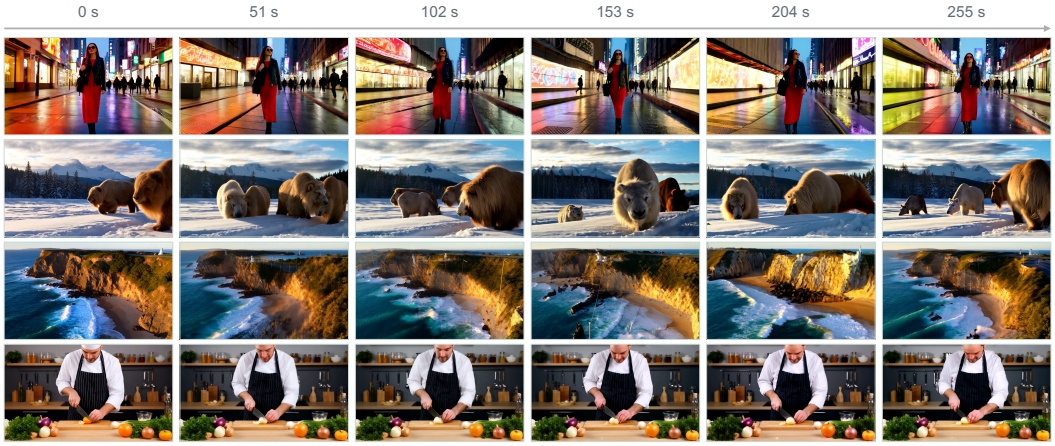


Figure 8: Ultra-long-horizon qualitative examples of RePhase RoPE. Each row corresponds to one prompt, with six frames sampled across a 255-second generation horizon.

Table 5: Quantitative comparison on sink collapse and per-prompt win rate over 255- second generation horizon on 64 prompts.

Method	SC Avg ↓	SC Max ↓	Ours Win Rate ↑
LongLive	30.48	67.89	98.44% (63/64)
LoL (3 seeds)	9.36 ± 1.65	55.88 ± 6.48	90.10% (173/192)
RePhase RoPE	**5.20**	**25.17**	–

compared with 30.48 for the vanilla model and 9.36 for LoL, while also obtaining the lowest reported maximum SC score. It achieves a lower SC score than the vanilla model on 63 of 64 prompts and outperforms LoL in 173 of 192 prompt–seed comparisons across three LoL seeds.

Figure 8 complements this evaluation with four qualitative examples. From top to bottom, the prompts describe a stylish woman walking along a neon-lit Tokyo street, wooly mammoths moving through a snowy meadow, a drone view of waves crashing against coastal cliffs in Big Sur, and a chef chopping onions in a kitchen.