Ring Forcing: Towards Precise Long-Term Memory for Autoregressive Video Diffusion

Bowen Xue[1], Brandon Y. Feng[2], Chenguo Lin[3], Yuchen Lin[3], Yujia Zeng[4], Lvmin Zhang[1], Maneesh Agrawala[1], Honglei Yan[5], and Panwang Pan[5⋆]

[1] Stanford University, Stanford, CA, USA
bowenxue2005@gmail.com, {lvmin,maneesh}@cs.stanford.edu
[2] Massachusetts Institute of Technology, Cambridge, MA, USA
brandon.fengys@gmail.com
[3] Peking University, Beijing, China
{chenguolin,linyuchen}@stu.pku.edu.cn
[4] University of California, Berkeley, Berkeley, CA, USA
yujiazng@gmail.com
[5] ByteDance, Beijing, China
yanhonglei@bytedance.com, paulpanwang@gmail.com

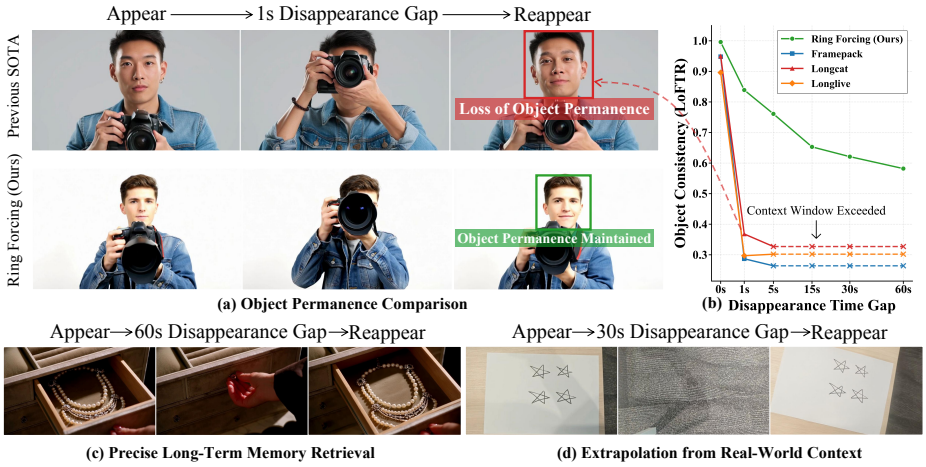

Fig. 1: Ring Forcing endows autoregressive video diffusion models with precise long-term memory and robust object permanence. It preserves exact object identity through prolonged occlusions and disappearances, significantly outperforming SOTA baselines **(a, b)**. Even across an extreme 60-second gap, our method faithfully recovers complex high-frequency details from distant history **(c)**. This effectively overcomes the limitations of short context windows and generalizes robustly to unconstrained real-world video histories **(d)**.

Abstract. Scaling video generation to long durations reveals a critical bottleneck: current models lack robust long-term memory. This deficiency can be studied along two critical aspects: *object permanence*, the

⋆ Project lead and corresponding author.

ability to precisely reproduce the appearance of objects upon re-entry; and *memory capacity*, the ability to process ultra-long context and use information from distant history. Robust long-term memory requires both: object permanence without sufficient context handling limits the temporal scope, while long context length without permanence fails to maintain identity. To address this, we present **Ring Forcing**, an autoregressive video diffusion framework designed to robustly construct and precisely utilize long-term memory. Our ring-structured training strategy enforces retrieval from distant history, effectively reconciling the trade-off between strict historical adherence and generative diversity. To expand memory capacity, we introduce a compression and timestep composition strategy. Under fixed sequence length constraints, this method extends the effective historical span to minutes-long durations and achieves a comprehensive receptive field over the entire history. Furthermore, we present a sparse RoPE mechanism to enable flexible, scalable memory adaptation while fully exploiting pre-trained priors. Extensive experiments demonstrate that Ring Forcing achieves superior minutes-long coherence and object permanence, significantly outperforming state-of-the-art methods.

Keywords: Long Video Generation · Autoregressive Video Diffusion · Long-term Memory

1 Introduction

Diffusion Transformers have recently advanced video generation, producing short clips with striking realism [11, 22, 33, 34, 44]. However, the research frontier is no longer confined to a few seconds: emerging applications increasingly demand minutes-long rollouts with sustained narrative coherence [19, 38, 41, 51]. At this horizon, the central obstacle is not local smoothness but *long-term memory*: the ability to preserve and reuse information that may disappear from view and return much later. Autoregressive video models [19, 27, 33, 50] still struggle with *object permanence*. A canonical failure occurs when an object exits the camera frustum and later re-enters: despite being observed earlier, the model frequently regenerates it with a different identity (Fig. 1a). Crucially, this behavior is not solely explained by limited context length; even when long histories are provided, models often treat distant observations as weak evidence. In effect, they learn to *continue* the immediate visual stream but fail to *retrieve* decisive information from the distant past.

We identify the root cause of this limitation to a ***misalignment between training objectives and inference requirements***. Standard video training data typically presents a "linear narrative" bias, where objects rarely exit and reappear within a short context window. Constrained by this bias, models tend to learn "myopic" transition probabilities (i.e., inferring x_{t+1} solely from x_t), with little incentive to learn how to retrieve information from distant history **(as visualized in Fig. 2a)**. In essence, the model learns to smoothly *continue* the video sequence but fails to learn how to *retrieve* answers from history, rendering long-term context effectively useless during inference even when provided.

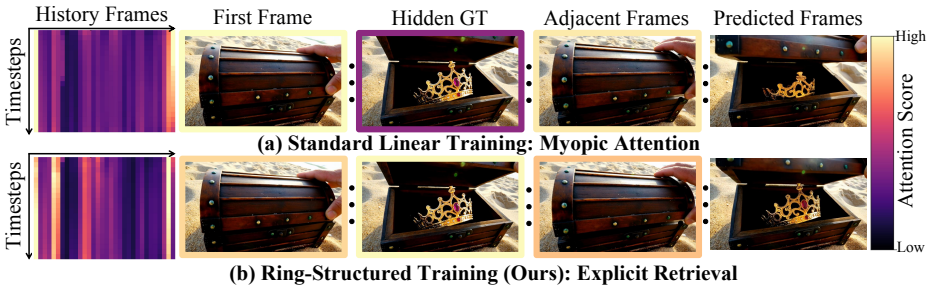

Fig. 2: Attention allocation of Predicted Frames to History Frames. (a) Standard Linear Training suffers from myopic attention bias, failing to retrieve effective information from distant history. **(b) Ring-Structured Training (Ours)** enables the model to extract more information from the past, achieving precise retrieval of historical frames.

Beyond the training paradigm, another fundamental bottleneck lies in the inability to precisely utilize distant historical information. Most existing long-video solutions operate at the frame or token level, creating an inherent dilemma: expanding the receptive field to cover long history requires increasing the sequence length, which incurs prohibitive computational costs; conversely, reducing the sequence length inevitably shrinks the receptive field. This trade-off fundamentally limits the construction of long-term memory required for consistent long-video generation.

To address both bottlenecks, we propose **Ring Forcing**, an autoregressive video diffusion framework designed for robust long-term memory construction and precise utilization. Our core insight is to create training instances where the supervision for the current target is anchored in the distant past. Specifically, we introduce a ring-structured data construction where "the future becomes the history," embedding the ground-truth target clip into historical frames. This turns long-range retrieval from an optional behavior into a necessity for minimizing training error, **yielding a highly precise retrieval focus on the distant target (Fig. 2b)**. To prevent degenerate shortcuts and to balance strict history adherence with open-ended continuation, we further employ a random head-cropping and context-drop strategy that controls whether the history contains direct supervisory evidence.

Ring Forcing also targets scalability. Leveraging the empirical prior that diffusion models emphasize global structure at high noise levels and fine details at low noise levels, we propose compression and timestep composition to expose complementary views of history across timesteps. This design extends the effective history length to cover minutes-long durations under a fixed sequence-length budget while maintaining full historical coverage. Finally, to maximize transfer from pretrained priors under variable compression and history length, we introduce sparse Rotary Positional Embedding (RoPE) that anchors compressed tokens to their physical spatiotemporal coordinates. Experiments show that Ring Forcing improves object permanence and enables faithful reuse of minutes-long

history (Fig. 1), establishing a strong baseline for long-term memory in video generation.

Our main contributions are summarized as follows:

- We introduce Ring Forcing, an autoregressive video diffusion framework that equips long-duration generation with precise long-term memory and improved object permanence.
- We devise ring-structured training that embeds targets into distant history and enforces explicit long-range retrieval, together with a cropping and context-drop mechanism that balances fidelity and diversity.
- We propose timestep-composed long-context conditioning integrated with sparse rotary positional encoding. This approach achieves a significant expansion in effective context under fixed sequence length constraints, enabling the faithful reproduction of minutes-long history and significantly outperforming SOTA methods.

2 Related Work

2.1 Autoregressive Video Diffusion

While bidirectional video diffusion models such as Wan and HunyuanVideo [20, 21,33,44] excel at short-clip generation, scaling them to long videos is constrained by the quadratic cost of bidirectional attention. Consequently, autoregressive video diffusion has become the prevailing framework for long video generation. Recent works, including SkyReels-V2, Magi-1, and CausVid [19, 38, 41, 51], have achieved impressive visual quality. However, autoregressive generation inherently suffers from error accumulation and challenges in long-sequence context modeling, hindering the generation of longer and higher-quality videos.

To tackle this issue, various strategies have been proposed [3, 12, 15, 23, 42]. One line of work focuses on training paradigms: FramePack [57] introduces a planned anti-drifting mechanism, while LongLive [48] employs attention sinks coupled with a "train long, test long" strategy. Other approaches enhance robustness by simulating inference errors during training. For instance, Rolling Forcing [27] proposes a joint denoising scheme; Self-Forcing++ [6] utilizes local teacher distillation on self-generated sequences; and SVI [24] alongside Resampling Forcing [13] exposes the model to synthesized or imperfect history to learn error correction capabilities. In parallel, some works focus on efficient video generation architectures. For instance, SANA-Video [4] incorporates linear attention into video generation, while others explore the use of more efficient VAEs [5,14] or optimize attention mechanisms for long sequences [8, 9, 25, 54–56]. Furthermore, training-free methods have also been explored [30,31], such as Deep Forcing [50], which utilizes a "Deep Sink" mechanism, and Infinity-RoPE [49], which adjusts Rotary Positional Embeddings to constrain the generation process closer to the pre-trained distribution.

2.2 Context Modeling for Long Video Generation

Global consistency in long video generation relies on efficient context management, generally categorized into retrieval-based and compression-based methods. **(1) Retrieval-based approaches** aim to select the most relevant historical information to extend the effective memory horizon. For instance, Context-as-Memory [52] and WorldMem [46] incorporate Field-of-View-based retrieval mechanisms within world models, while Pack-and-Force [45] employs a contextual semantic retriever. Memory Forcing [18] persists memory through 3D point cloud reconstruction, whereas Deep Forcing [50] utilizes attention mechanisms to retrieve relevant KV caches. RELIC [16] uses time reversal to construct revisit data for spatial memory in video world models. Recent advancements also focus on hybrid architectures and state-space mechanisms; for example, VideoSSM [53] utilizes a hybrid State-Space Memory, and long-context state-space video world models [36] have been developed to handle autoregressive long video generation efficiently. Furthermore, Mixture-of-Contexts [1] learns attention routing to identify critical historical segments across multi-clip generation tasks. **(2) Compression-based methods** aim to condense historical information into compact representations. FramePack [57] compresses prior frames into fixed-size latent "packs". WorldPack [35] improves spatial consistency in world models via history packing. TTTVideo [7] and LaCT [60] introduce learnable parameters as memory representations that are updated during inference (Test-Time Optimization). Similarly, TinyHistory [58] pre-trains a context compression model via the reconstruction loss. Nevertheless, while these approaches successfully extend the input context window, they leave out direct training objectives that explicitly enforce the utilization and reproduction of historical information.

3 Method

We propose Ring Forcing, an autoregressive video diffusion framework for long-term memory construction and precise utilization. It integrates three designs: ring-structured training for explicit long-range retrieval (Sec. 3.1), timestep-composed history compression under a fixed context budget (Sec. 3.2), and sparse RoPE for consistent spatiotemporal scaling under variable compression (Sec. 3.3). We train the model with a rectified-flow objective [10, 28] that unifies these components (Sec. 3.4).

3.1 Ring-Structured Training Strategy

As shown in Fig. 3, we introduce a training strategy based on *ring-structured* sequences. The key idea is to embed the ground-truth target clip into the distant history, turning long-term retrieval into a self-supervised learning signal.

Ring-Based Sequence Construction Formally, let $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ denote a raw video sequence of N frames. We construct a closed-loop reference video $\mathcal{V}_{ring}$ by concatenating the original sequence with its reversed counterpart $\mathcal{V}_{rev} = \{v_N, v_{N-1}, \ldots, v_1\}$:

$$\mathcal{V}_{ring} = [\mathcal{V}, \mathcal{V}_{rev}]. \tag{1}$$

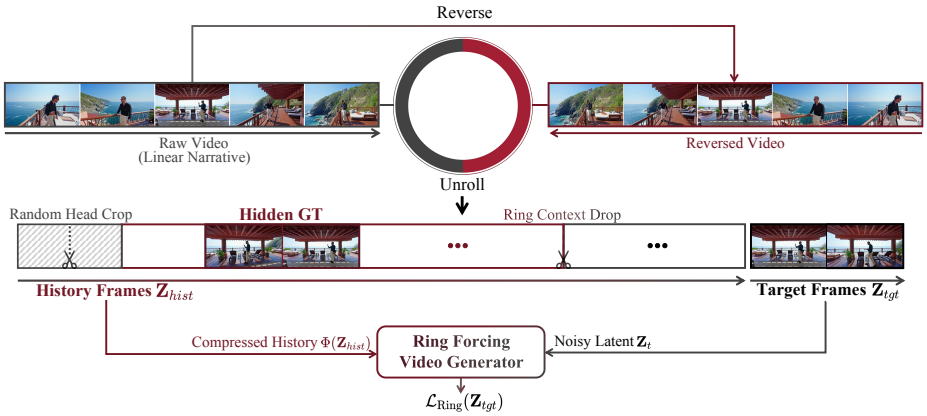

Fig. 3: Overview of the Ring-Structured Training Strategy. We construct a sequence ring by concatenating a linear raw video with its reversed counterpart. By unrolling the ring into a conditioning sequence, the target frames are naturally embedded into the distant history as the Hidden GT, which essentially forces the autoregressive generator to learn explicit long-range retrieval. To prevent trivial shortcuts, a Random Head Crop is applied to truncate boundary information leakage, while a Random Context Drop is used to balance history adherence and generation diversity.

History-Target Decomposition We strictly sample the target video clip $\mathbf{x}_{tgt}$ from the forward part $\mathcal{V}$ to preserve the arrow of time. Based on the target sequence of length L_{tgt}, denoted as $\mathbf{x}_{tgt} = \{v_t, \ldots, v_{t+L_{tgt}-1}\}$, we decompose $\mathcal{V}_{ring}$ into three consecutive segments: (1) Original history $\mathbf{h}_{org} = \{v_1, \ldots, v_{t-1}\}$, (2) Target clip $\mathbf{x}_{tgt} = \{v_t, \ldots, v_{t+L_{tgt}-1}\}$, and (3) Synthetic history $\mathbf{h}_{syn}$ is the remaining segment in the ring, which contains the reversed target clip as a distant subsequence.

To construct the conditioning context, we place the synthetic history *before* the original history:

$$\mathbf{C}_{full} = [\mathbf{h}_{syn}, \mathbf{h}_{org}]. \tag{2}$$

This ring-based construction ensures visual continuity throughout the reference sequence by leveraging the temporal symmetry of the reversed video.

Leakage Prevention via Random Head Crop In this constructed history $\mathbf{C}_{full}$, a trivial shortcut (information leakage) exists at the head of the sequence. The first frame of $\mathbf{h}_{syn}$ is $v_{t+L_{tgt}}$, which is temporally adjacent and visually similar to the last frame of the target clip ($v_{t+L_{tgt}-1}$). If left unaddressed, the model can infer the end of the target simply by looking at the beginning of the history, ignoring the context (the reversed target clip $\mathbf{x}^{rev}_{tgt}$) embedded later in $\mathbf{h}_{syn}$. To prevent the leakage shortcut while forcing the model to leverage the full context, we apply a Random Head Cropping strategy. Concretely, we randomly remove a prefix of the synthetic history while preserving the reversed target clip inside $\mathbf{h}_{syn}$, yielding the cropped context $\mathbf{C}_{crop}$, which removes the shortcut while keeping the distant answer retrievable.

Balancing Adherence and Diversity via Ring Context Drop To regulate the trade-off between strict history adherence (when the target is retrievable from the past) and open-ended generative diversity (when it is not), we introduce a **Ring Context Drop** mechanism. Rather than always enforcing the ring structure, we randomly drop the entire synthetic history during training. Specifically, we sample a boolean variable $b \sim \mathit{Bernoulli}(1 - p_{\text{drop}})$ and formulate the final conditioning context as:

$$\mathbf{C}_{final} = \begin{cases} \mathbf{C}_{crop}, & \text{if } b = 1 \quad \text{(Ring Mode)}, \\ \mathbf{h}_{org}, & \text{if } b = 0 \quad \text{(Standard Mode)}. \end{cases} \tag{3}$$

When $b = 0$, the ring-constructed context is completely dropped, thereby preventing the model from over-relying on the synthetic history and forcing it to learn unconstrained progression solely from $\mathbf{h}_{org}$. When $b = 1$, the model learns to explicitly leverage the embedded target for precise long-term consistency.

3.2 Compression and Timestep Composition

While the Ring-Structured strategy provides the essential supervision for long-range retrieval, directly attending to minute-long raw history is computationally prohibitive due to the quadratic complexity of attention. To resolve this, we propose **Timestep-Composed Compression**, which maps a long history $\mathbf{Z}_{hist}$ into a bounded conditioning sequence $\tilde{\mathbf{C}}$ under a fixed token budget $\mathcal{B}_{max}$. This design ensures efficient memory utilization while preserving both global structure and local details.

Token Budget and Compression Operator We define a spatiotemporal compression operator $\Psi(\mathbf{Z}; r_h, r_w, r_t)$ that downsamples history in space and time to fit $\mathcal{B}_{max}$.

Timestep-Dependent Composition Operator Uniform spatiotemporal downsampling inevitably sacrifices high-frequency details. However, we leverage the intrinsic generative nature of diffusion models: *high-noise intervals primarily determine semantic structure and global motion, whereas low-noise intervals focus on refining local textures and edges.* Motivated by this, we define a **Composite Condition Operator** $\Phi(\mathbf{Z}_{hist}, t, \delta)$ that dynamically transforms the history representation based on the diffusion timestep t:

$$\Phi(\mathbf{Z}_{hist}, t, \delta) = \begin{cases} \Psi(\mathbf{Z}_{hist}; r_h^{\text{g}}, r_w^{\text{g}}, r_t^{\text{g}}), & t > \tau \\ \mathcal{S}\left(\Psi(\mathbf{Z}_{hist}; r_h^{\text{d}}, r_w^{\text{d}}, r_t^{\text{d}}), \delta\right), & t \leq \tau \end{cases} \tag{4}$$

Here, the first branch provides a temporally dense context for recovering global structure, while the second branch preserves high-frequency spatial details with temporally sparse sampling. $\mathcal{S}(\cdot, \delta)$ is a cyclic shift over the temporal sampling grid, and varying δ ensures full coverage over long histories.

3.3 Sparse RoPE Strategy

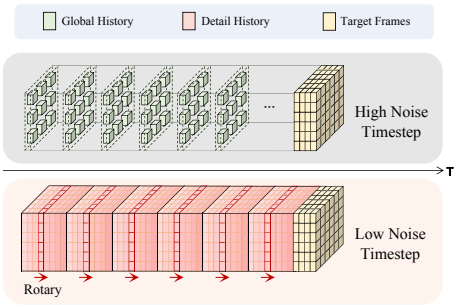

Fig. 4: Timestep-composed long-context conditioning.

To maximize the use of pretrained video generation priors (i.e., the model's perception of relative space-time structures), and to accommodate historical contexts of varying compression rates and lengths, we introduce a sparse RoPE strategy. As shown in Fig. 4, instead of encoding the compressed history features using their discrete indices in the latent space, we map them back to the original continuous spatiotemporal coordinate system used during pre-training.

Formally, we treat the target clip as the reference coordinate system with origin $(0, 0, 0)$. For a compressed history token at index (t, h, w) with spatial downsampling factor s_s and temporal downsampling factor s_t, we map it back to physical coordinates $(\tilde{t}, \tilde{h}, \tilde{w})$ in the original grid:

$$\tilde{h} = h \cdot s_s + 0.5 \cdot (s_s - 1),$$
$$\tilde{w} = w \cdot s_s + 0.5 \cdot (s_s - 1), \tag{5}$$
$$\tilde{t} = -\left((L_{hist} - 1 - t) \cdot s_t + 0.5 \cdot (s_t - 1) + 0.5\right), \tag{6}$$

where L_{hist} is the number of compressed history steps along time. The spatial terms center the token within its corresponding patch, and the temporal term places the history strictly in the negative relative-time domain. We then compute 3D RoPE using $(\tilde{t}, \tilde{h}, \tilde{w})$ to preserve consistent relative spatiotemporal scales under variable compression.

3.4 Training Objective

We adopt the rectified flow framework and train the generator G_θ to predict the velocity field derived from the clean target $\mathbf{Z}_{tgt}$. The training objective unifies the ring-structured strategy (controlled by b) and the compression strategy (encapsulated by Φ) into a holistic loss function. Let $\mathbf{Z}_{hist}^{(b)}$ denote the history latents constructed by the ring strategy, where $b \sim \text{Bernoulli}(1 - p_{\text{drop}})$ determines whether to use the ring context ($b = 1$) or standard context ($b = 0$). The final conditioning input is computed as $\tilde{\mathbf{C}} = \Phi(\mathbf{Z}_{hist}^{(b)}, t, \delta)$. The formal optimization objective is:

$$\mathcal{L}_{\text{Ring}} = \mathbb{E}_{t, \mathbf{Z}_{tgt}, \epsilon, b, \delta}\left[\left\|(\epsilon - \mathbf{Z}_{tgt}) - G_\theta(\mathbf{Z}_t, t, \tilde{\mathbf{C}})\right\|_2^2\right]. \tag{7}$$

In this formulation, $t \sim \mathcal{U}(0, 1)$ represents the sampling timestep, and $\mathbf{Z}_t = (1 - t)\mathbf{Z}_{tgt} + t\epsilon$ denotes the noisy latent mixed with Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

The joint expectations over b, t, and δ optimize the model to balance precise history retrieval with generative diversity, while efficiently capturing both global structures and local details across varying noise levels.

4 Experiments

In this section, we present a comprehensive evaluation of Ring Forcing. We begin by detailing the implementation and experimental setup. Next, we compare Ring Forcing against state-of-the-art long video generation models to demonstrate its superiority in maintaining object permanence and long-term consistency. Finally, we conduct in-depth ablation studies to validate the efficacy of our core contributions: the Timestep Composition strategy, the ring-structured training trade-offs, and the sparse RoPE mechanism.

4.1 Implementation Details

Dataset Construction We utilize the UltraVideo-Long dataset [47]. To isolate the challenge of temporal consistency within continuous shots, we filter the dataset to retain only single-shot videos. We curate a training set of 10,000 long-duration videos, each with a duration ranging from 10 to 240 seconds. All video data is standardized to a resolution of 480×832 and a frame rate of 15 FPS to match the training requirements of the base models.

Model Instantiation We instantiate Ring Forcing using two backbones: Wan2.1 T2V 1.3B and Wan2.2 T2V A14B [44]. All ablation studies are conducted on the Wan2.1 1.3B variant, while the final comparative evaluation employs the Wan2.2 A14B model. To ensure parameter efficiency, all models are fine-tuned using Low-Rank Adaptation (LoRA) [17] with a rank of $r = 128$. Training for the Wan2.2 A14B model was distributed across a cluster of 32 NVIDIA H800 GPUs for 7,000 steps, taking approximately 30 hours. For inference, we utilize a single NVIDIA H800 GPU. By aligning the effective sequence length with the standard generation length via compression, our method maintains computational costs virtually identical to the original model. Consequently, *Ring Forcing is deployable on any hardware supporting the base Wan model, incurring negligible overhead in VRAM or compute.*

4.2 Comparison with Baselines

Baselines We compare Ring Forcing against three leading long-video generation models: LongLive [48], LongCat [33], and FramePack [57].

Evaluation Protocol Object permanence is a core manifestation of a video generation model's long-term memory capacity. To rigorously quantify this capability, we specifically design a benchmark based on an "Appear-Disappear-Reappear" (A-D-R) logic. Additionally, to evaluate the model's long video generation capabilities, we conduct a targeted assessment of its general performance in generating 1-minute coherent long videos.

Table 1: Quantitative results on the A-D-R Benchmark. Metrics are categorized into Consistency and General Performance. **I.F.** denotes Instruction Following. For $\Delta t_{gap} > 5s$, baselines are excluded because the "appear" segment falls completely outside their maximum context window. Best results are highlighted in bold, and second-best are underlined.

Gap (Δt_{gap})	Model	Consistency			General Performance		
		Texture	Geometry	Semantic	Aesthetic	I.F.	Dynamics
0s	LongLive	0.578	0.896	0.886	5.032	<u>24.117</u>	<u>1.615</u>
	LongCat	<u>0.737</u>	<u>0.950</u>	**0.925**	5.214	24.090	1.493
	FramePack	0.605	0.949	0.913	**5.222**	23.162	**1.662**
	Ours	**0.742**	**0.997**	<u>0.918</u>	<u>5.216</u>	**24.642**	1.550
1s	LongLive	0.181	0.297	0.707	5.080	23.803	1.694
	LongCat	0.236	<u>0.368</u>	<u>0.712</u>	5.138	<u>23.894</u>	<u>1.990</u>
	FramePack	<u>0.252</u>	0.287	0.692	**5.220**	23.040	1.093
	Ours	**0.723**	**0.839**	**0.823**	<u>5.176</u>	**24.601**	**2.617**
5s	LongLive	<u>0.290</u>	0.303	<u>0.730</u>	5.065	<u>23.933</u>	1.555
	LongCat	0.203	<u>0.328</u>	0.722	5.148	23.859	<u>1.640</u>
	FramePack	0.250	0.264	0.701	**5.229**	22.795	0.996
	Ours	**0.515**	**0.761**	**0.831**	<u>5.163</u>	**23.940**	**2.371**
15s	**Ours**	0.509	0.653	0.802	5.104	23.803	2.689
30s	**Ours**	0.421	0.621	0.822	5.055	23.590	2.946
60s	**Ours**	0.396	0.582	0.793	5.014	22.872	1.809

A-D-R Benchmark To rigorously and fairly measure the model's memory retrieval capability across temporal dimensions, we meticulously construct an A-D-R prompt library comprising 64 test cases. Each test case consists of four core elements: an appear prompt, a disappear prompt, a reappear prompt, and a precise subject keyword. During video generation, the durations of the "appear" and "reappear" segments are fixed at 5 seconds each. Furthermore, to prevent the model from exploiting shortcuts by relying solely on the initial frame, the first frame is intentionally devoid of the target subject, and its appearance is randomly triggered within the first 5 seconds, thereby better approximating realistic dynamic generation scenarios. By dynamically adjusting the duration of the intermediate "disappear" segment, we systematically investigate the model's capability bounds in maintaining subject identity across varying temporal spans. Specifically, we set the disappearance durations to 0 (where the entire video is evaluated as a control group), 1, 5, 15, 30 and 60 seconds.

To maximally isolate the consistency evaluation from background interference, we employ SAM 3 [2] combined with subject keywords to perform precise zero-shot instance segmentation on the appear and reappear segments. We also perform subject detection on the disappear segment to exclude cases where

Fig. 5: Qualitative results on the Appear–Disappear–Reappear (A-D-R) benchmark. Our model demonstrates robust object permanence by retrieving identity information from distant history, ensuring consistent subject reconstruction even after long temporal gaps where baseline models typically fail.

the subject fails to vanish from consistency calculations, while our meticulous prompt design ensures that most cases successfully adhere to the A-D-R logic. For both segments, we extract the keyframe with the largest subject mask area, crop the subject, and normalize it against a pure white background. For the extracted subject images before and after disappearance, we comprehensively employ SIFT [29], LoFTR [43], and DINOv3 [40] to evaluate local texture consistency, geometric structure consistency, and high-level semantic consistency, respectively. This multi-dimensional feature matching strategy allows us to thoroughly quantify the preservation of object identity. Furthermore, general metrics are evaluated on the reappear segments: we utilize the Improved Aesthetic Predictor [39] to quantify the aesthetic quality of the videos, employ X-CLIP [32] to measure instruction following capabilities, and apply Optical Flow-based Motion Magnitude to quantify the amplitude of video dynamics.

General Video Generation Benchmark To comprehensively evaluate the model's generalized capability in regular scenarios, we construct a General Benchmark comprising 64 diverse prompts. These prompts are utilized to generate 60-second continuous videos where the explicit "disappear-reappear" logic is absent, and the primary subjects maintain a persistent presence. We adopt the same automated metrics used in the A-D-R benchmark to evaluate aesthetic quality, prompt adherence, and motion magnitude, and further include *Motion Smoothness* by calculating the normalized frame interpolation error via AMT [26] to quantify temporal fluidity. Furthermore, to assess complex perceptual qualities that automated metrics struggle to capture, we introduce the Large Multimodal Model Qwen3-VL [37] to conduct a comprehensive 5-point scale evaluation. This evaluation focuses on three critical dimensions: *physical logic rationality* to ensure realistic object interactions, *spatiotemporal consistency* to prevent semantic

Table 2: Quantitative comparison on the General long-video generation benchmark (1 minute).

Model	Objective Metrics					User Study	
	Aesthetic	Dynamics	Naturalness	Motion Smoothness	Instruction Following	Consistency	Overall Quality
LongLive	5.622	1.056	3.031	**0.991**	24.805	<u>4.12</u>	2.25
LongCat	<u>5.808</u>	<u>2.196</u>	<u>3.875</u>	**0.991**	<u>25.341</u>	3.59	<u>4.19</u>
FramePack	5.728	1.790	3.656	**0.991**	25.014	3.71	4.04
Ours	**5.822**	**2.261**	**3.922**	<u>0.986</u>	**25.471**	**4.35**	**4.22**
Base model†	5.825	1.974	4.016	0.984	25.136	*N/A*	

† *Wan2.2-T2V-A14B (5.4s).*

collapse over time, and *visual naturalness* to penalize unnatural flickering and morphological distortions.

Human Evaluation To further validate our results through human perception, we conducted a Human Evaluation involving 21 participants from diverse backgrounds. We randomly sampled 10 video pairs from the General Benchmark (pairing our generated videos with those from baseline models). Under a strict blind-test protocol, evaluators rated each video on a 0-5 scale with a primary focus on "Overall Video Quality."

Results Analysis Comprehensive evaluations demonstrate that Ring Forcing achieves state-of-the-art long-term memory and general video generation quality. First, on the A-D-R benchmark (Tab. 1, Fig. 5), Ring Forcing exhibits robust object permanence by precisely retrieving distant history. Conversely, baselines suffer catastrophic forgetting when targets temporarily disappear, restricted by short-sighted attention biases from linear training. Second, Ring Forcing fundamentally breaks the "consistency vs. dynamics" trade-off. Baselines like LongLive over-rely on the initial frame as an attention sink, degenerating into static or repetitive outputs (low Dynamics). In contrast, Ring Forcing yields rich, fluid motions while maintaining strict spatiotemporal consistency. Finally, General Benchmark results (Tab. 2, Fig. 6) show that Ring Forcing preserves the backbone's generative priors while extending generation to 60-second videos. Among long-video baselines, it achieves the best aesthetics, naturalness, and overall human preference, demonstrating that our long-term memory mechanism improves temporal coherence without substantially degrading visual quality.

4.3 Ablation Studies

Impact of Compression and Timestep Composition We first investigate the impact of spatiotemporal compression strategies on the model's ability to reconstruct history. We construct a test set of 128 unseen samples where the "answer" (ground truth target) is naturally embedded within the history using our Ring-Structured construction. We evaluate reconstruction fidelity using photometric and perceptual metrics [59].

Fig. 6: Qualitative results of 60-second long video generation. The visualization demonstrates the model's capability to maintain overall consistency and visual quality throughout a 1-minute duration in general scenarios.

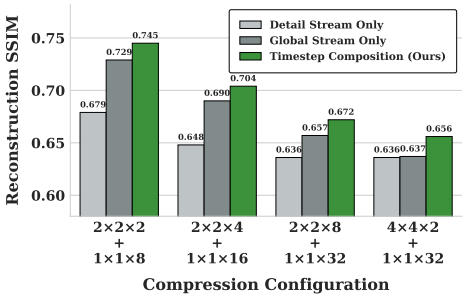

Fig. 7: Effectiveness of Timestep Composition Strategy.

Compression Trade-offs and Composition Effectiveness As illustrated in Fig. 7, we first analyze the single-stream baselines (gray bars) to understand the trade-offs between spatial and temporal information. We observe that the "Global Stream Only" baseline (high spatial compression, intact temporal density) generally yields superior reconstruction SSIM compared to the "Detail Stream Only" baseline (aggressive temporal compression). This suggests that for precisely leveraging long-term history, retaining temporal density is often more critical than spatial resolution. Building on this insight, we evaluate our proposed timestep composition strategy. As shown by the green bars in Fig. 7, our method dynamically synergizes the structural guidance from the global stream with the fine-grained refinement from the detail stream. Remarkably, without increasing the sequence length compared to single-stream baselines, the combined strategy consistently outperforms both standalone baselines across all metrics and compression settings, compellingly validating the effectiveness of our dual-stream design.

To balance reconstruction quality and long-video capacity, we select $2 \times 2 \times 8 + 1 \times 1 \times 32$ as our default. Under this configuration, the inference overhead for

Table 3: Ablation study on Ring Context Drop probability p_{drop}.

p_{drop}	Ring Mode (Reconstruction)			Standard Mode
	PSNR ↑	SSIM ↑	LPIPS ↓	Diversity Score ↑
0.0	**25.02**	**0.73**	**0.12**	3.71
0.2	<u>24.67</u>	<u>0.72</u>	**0.12**	3.93
0.4	24.22	0.71	<u>0.13</u>	4.11
0.6	24.02	0.71	<u>0.13</u>	4.13
0.8	23.02	0.68	0.16	<u>4.21</u>
1.0	21.86	0.65	0.21	**4.23**

Table 4: Comparison of reconstruction capabilities between standard index-based RoPE and our sparse RoPE.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Standard RoPE	19.05	0.58	0.27
Sparse RoPE (Ours)	**24.22**	**0.71**	**0.13**

140s of history is virtually identical to the original Wan model's 5.4s generation, enabling efficient long-video synthesis without increasing hardware requirements.

Trade-offs of Adherence and Diversity To determine the optimal balance between object permanence and generative diversity, we conduct an ablation on the Ring Context Drop probability p_{drop} (which controls the Bernoulli variable b). We employ a dual-evaluation protocol that measures reconstruction fidelity on 128 unseen samples with embedded answers (Ring Mode), and generates 15-second video continuations on 128 samples without embedded answers (Standard Mode). In this mode, we employ Qwen3-VL to assess the diversity of the generated content relative to the input history on a 5-point Likert scale, and we report the average score across all samples. As shown in Tab. 3, we identify $p_{drop} = 0.4$ as the "sweet spot", where the model maintains high reconstruction fidelity without collapsing into mode dropping.

Efficacy of Sparse RoPE To validate whether our Sparse RoPE strategy effectively leverages pre-trained priors, we compare the reconstruction capabilities of two models trained with identical settings: one using standard index-based RoPE and the other using our Sparse RoPE. As presented in Tab. 4, experimental results demonstrate that sparse RoPE better utilizes priors to achieve faster convergence and superior reconstruction quality. This confirms that mapping compressed tokens back to their physical coordinates significantly aids the model in understanding relative spatiotemporal relationships.

5 Conclusion

We address the challenge of establishing robust long-term memory in autoregressive video generation, specifically targeting the critical bottlenecks of object per-

manence and memory capacity. We proposed Ring Forcing, a unified framework that enables the robust construction and precise utilization of long-term memory. Through our ring-structured training strategy, we successfully extracted rich information from long-term history, effectively reconciling the trade-off between historical adherence and generative diversity. Furthermore, our compression and timestep composition mechanism expanded the effective historical span to support minutes-long generation under fixed sequence length constraints, achieving a comprehensive receptive field over the entire history. Together with the sparse RoPE mechanism, our approach enables the precise utilization and reproduction of information spanning minutes-long history. Ring Forcing establishes a foundational paradigm for future research into infinite-context generative modeling, paving the way for consistent, long-duration video synthesis.

References

1. Cai, S., Yang, C., Zhang, L., Guo, Y., Xiao, J., Yang, Z., Xu, Y., Yang, Z., Yuille, A.L., Guibas, L.J., Agrawala, M., Jiang, L., Wetzstein, G.: Mixture of contexts for long video generation. ArXiv preprint (2025)
2. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. ArXiv preprint (2025)
3. Chen, B., Monso, D.M., Du, Y., Simchowitz, M., Tedrake, R., Sitzmann, V.: Diffusion forcing: Next-token prediction meets full-sequence diffusion. In: Conference and Workshop on Neural Information Processing Systems (2024)
4. Chen, J., Zhao, Y., Yu, J., Chu, R., Chen, J., Yang, S., Wang, X., Pan, Y., Zhou, D., Ling, H., Liu, H., Yi, H., Zhang, H., Li, M., Chen, Y., Cai, H., Fidler, S., Luo, P., Han, S., Xie, E.: Sana-video: Efficient video generation with block linear diffusion transformer. ArXiv preprint (2025)
5. Chen, J., Cai, H., Chen, J., Xie, E., Yang, S., Tang, H., Li, M., Han, S.: Deep compression autoencoder for efficient high-resolution diffusion models. In: International Conference on Learning Representations (2025)
6. Cui, J., Wu, J., Li, M., Yang, T., Li, X., Wang, R., Bai, A., Ban, Y., Hsieh, C.: Self-forcing++: Towards minute-scale high-quality video generation. ArXiv preprint (2025)
7. Dalal, K., Koceja, D., Xu, J., Zhao, Y., Han, S., Cheung, K.C., Kautz, J., Choi, Y., Sun, Y., Wang, X.: One-minute video generation with test-time training. In: Conference on Computer Vision and Pattern Recognition (2025)
8. Dao, T.: FlashAttention-2: Faster attention with better parallelism and work partitioning. In: International Conference on Learning Representations (2024)
9. Dao, T., Fu, D.Y., Ermon, S., Rudra, A., Ré, C.: FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In: Conference and Workshop on Neural Information Processing Systems (2022)
10. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: International Conference on Machine Learning (2024)
11. Google DeepMind: Veo 3.1 (2025), `https://deepmind.google/models/veo/`, accessed: 2026-06-30

12. Gu, Y., Mao, W., Shou, M.Z.: Long-context autoregressive video modeling with next-frame prediction. ArXiv preprint (2025)
13. Guo, Y., Yang, C., He, H., Zhao, Y., Wei, M., Yang, Z., Huang, W., Lin, D.: End-to-end training for autoregressive video diffusion via self-resampling. ArXiv preprint (2025)
14. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., Panet, P., Weissbuch, S., Kulikov, V., Bitterman, Y., Melumian, Z., Bibi, O.: Ltx-video: Realtime video latent diffusion. ArXiv preprint (2024)
15. Henschel, R., Khachatryan, L., Poghosyan, H., Hayrapetyan, D., Tadevosyan, V., Wang, Z., Navasardyan, S., Shi, H.: Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In: Conference on Computer Vision and Pattern Recognition (2025)
16. Hong, Y., Mei, Y., Ge, C., Xu, Y., Zhou, Y., Bi, S., Hold-Geoffroy, Y., Roberts, M., Fisher, M., Shechtman, E., Sunkavalli, K., Liu, F., Li, Z., Tan, H.: RELIC: Interactive video world model with long-horizon memory. ArXiv preprint (2025)
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9
18. Huang, J., Hu, X., Han, B., Shi, S., Tian, Z., He, T., Jiang, L.: Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft. ArXiv preprint (2025)
19. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. In: Conference and Workshop on Neural Information Processing Systems (2025)
20. Hunyuan Foundation Model Team: Hunyuanvideo: A systematic framework for large video generative models. ArXiv preprint (2024)
21. Hunyuan Foundation Model Team: Hunyuanvideo 1.5 technical report. ArXiv preprint (2025)
22. Kling Team: Kling-omni technical report. ArXiv preprint (2025)
23. Kodaira, A., Hou, T., Hou, J., Tomizuka, M., Zhao, Y.: Streamdit: Real-time streaming text-to-video generation. ArXiv preprint (2025)
24. Li, W., Pan, W., Luan, P., Gao, Y., Alahi, A.: Stable video infinity: Infinite-length video generation with error recycling. ArXiv preprint (2025)
25. Li, X., Li, M., Cai, T., Xi, H., Yang, S., Lin, Y., Zhang, L., Yang, S., Hu, J., Peng, K., Agrawala, M., Stoica, I., Keutzer, K., Han, S.: Radial attention: O(n log n) sparse attention with energy decay for long video generation. In: Conference and Workshop on Neural Information Processing Systems (2025)
26. Li, Z., Zhu, Z., Han, L., Hou, Q., Guo, C., Cheng, M.: AMT: all-pairs multi-field transforms for efficient frame interpolation. In: Conference on Computer Vision and Pattern Recognition (2023)
27. Liu, K., Hu, W., Xu, J., Shan, Y., Lu, S.: Rolling forcing: Autoregressive long video diffusion in real time. ArXiv preprint (2025)
28. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: International Conference on Learning Representations (2023)
29. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision (2004)

30. Lu, Y., Liang, Y., Zhu, L., Yang, Y.: Freelong: Training-free long video generation with spectralblend temporal attention. In: Conference and Workshop on Neural Information Processing Systems (2024)
31. Lu, Y., Yang, Y.: Freelong++: Training-free long video generation via multi-band spectralfusion. ArXiv preprint (2025)
32. Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., Ji, R.: X-CLIP: End-to-end multi-grained contrastive learning for video-text retrieval. In: ACM International Conference on Multimedia (2022). https://doi.org/10.1145/3503161.3547910
33. Meituan LongCat Team: Longcat-video technical report. ArXiv preprint (2025)
34. OpenAI: Sora 2 is here (2025), https://openai.com/index/sora-2/, accessed: 2026-06-30
35. Oshima, Y., Iwasawa, Y., Suzuki, M., Matsuo, Y., Furuta, H.: WorldPack: Compressed memory improves spatial consistency in video world modeling. ArXiv preprint (2025)
36. Po, R., Nitzan, Y., Zhang, R., Chen, B., Dao, T., Shechtman, E., Wetzstein, G., Huang, X.: Long-context state-space video world models. In: International Conference on Computer Vision (2025)
37. Qwen Team: Qwen3-vl technical report. ArXiv preprint (2025), https://arxiv.org/abs/2511.21631
38. Sand AI: MAGI-1: autoregressive video generation at scale. ArXiv preprint (2025)
39. Schuhmann, C.: Improved aesthetic predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor (2022), accessed: 2026-06-30
40. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. ArXiv preprint (2025)
41. SkyReels Team: Skyreels-v2: Infinite-length film generative model. ArXiv preprint (2025)
42. Song, K., Chen, B., Simchowitz, M., Du, Y., Tedrake, R., Sitzmann, V.: History-guided video diffusion. In: International Conference on Machine Learning (2025)
43. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: Conference on Computer Vision and Pattern Recognition (2021)
44. Wan Team: Wan: Open and advanced large-scale video generative models. ArXiv preprint (2025)
45. Wu, X., Zhang, G., Xu, Z., Zhou, Y., Lu, Q., He, X.: Pack and force your memory: Long-form and consistent video generation. ArXiv preprint (2025)
46. Xiao, Z., Lan, Y., Zhou, Y., Ouyang, W., Yang, S., Zeng, Y., Pan, X.: WORLDMEM: long-term consistent world simulation with memory. In: Conference and Workshop on Neural Information Processing Systems (2025)
47. Xue, Z., Zhang, J., Hu, T., He, H., Chen, Y., Cai, Y., Wang, Y., Wang, C., Liu, Y., Li, X., et al.: Ultravideo: High-quality uhd video dataset with comprehensive captions. In: Conference and Workshop on Neural Information Processing Systems (2025)
48. Yang, S., Huang, W., Chu, R., Xiao, Y., Zhao, Y., Wang, X., Li, M., Xie, E., Chen, Y., Lu, Y., Han, S., Chen, Y.: Longlive: Real-time interactive long video generation. ArXiv preprint (2025)
49. Yesiltepe, H., Meral, T.H.S., Akan, A.K., Oktay, K., Yanardag, P.: Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. ArXiv preprint (2025)

50. Yi, J., Jang, W., Cho, P.H., Nam, J., Yoon, H., Kim, S.: Deep forcing: Training-free long video generation with deep sink and participative compression. ArXiv preprint (2025)
51. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast autoregressive video diffusion models. In: Conference on Computer Vision and Pattern Recognition (2025)
52. Yu, J., Bai, J., Qin, Y., Liu, Q., Wang, X., Wan, P., Zhang, D., Liu, X.: Context as memory: Scene-consistent interactive long video generation with memory retrieval. ArXiv preprint (2025)
53. Yu, Y., Wu, X., Hu, X., Hu, T., Sun, Y., Lyu, X., Wang, B., Ma, L., Ma, Y., Wang, Z., et al.: Videossm: Autoregressive long video generation with hybrid state-space memory. ArXiv preprint (2025)
54. Zhang, J., Huang, H., Zhang, P., Wei, J., Zhu, J., Chen, J.: Sageattention2: Efficient attention with thorough outlier smoothing and per-thread int4 quantization. In: International Conference on Machine Learning (2025)
55. Zhang, J., Wei, J., Zhang, P., Xu, X., Huang, H., Wang, H., Jiang, K., Zhu, J., Chen, J.: Sageattention3: Microscaling fp4 attention for inference and an exploration of 8-bit training. ArXiv preprint (2025)
56. Zhang, J., Wei, J., Zhang, P., Zhu, J., Chen, J.: Sageattention: Accurate 8-bit attention for plug-and-play inference acceleration. In: International Conference on Learning Representations (2025)
57. Zhang, L., Cai, S., Li, M., Wetzstein, G., Agrawala, M.: Frame context packing and drift prevention in next-frame-prediction video diffusion models. In: Conference and Workshop on Neural Information Processing Systems (2025)
58. Zhang, L., Cai, S., Li, M., Zeng, C., Lu, B., Rao, A., Han, S., Wetzstein, G., Agrawala, M.: Tinyhistory: Lightweight video history embeddings via two-stage context learning. In: European Conference on Computer Vision (2026)
59. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Conference on Computer Vision and Pattern Recognition (2018)
60. Zhang, T., Bi, S., Hong, Y., Zhang, K., Luan, F., Yang, S., Sunkavalli, K., Freeman, W.T., Tan, H.: Test-time training done right. ArXiv preprint (2025)