RECENCY FORCING: SCAFFOLDING TEMPORAL DEPENDENCY FOR LONG AR VIDEO GENERATION

Anonymous authors
Paper under double-blind review



Figure 1: **Recency Forcing** improves long-horizon autoregressive video generation up to 60 seconds, far beyond the 5-second training horizon. Guided by the measured positional response of distilled AR video diffusion models, our method can be applied in a *training-free* setting on Self-Forcing or a *training-based* setting on Causal Forcing. Across both settings, Recency Forcing improves long-horizon video quality while preserving the efficiency of the underlying model.

ABSTRACT

Autoregressive (AR) video generation is causal by construction, yet how a generated frame depends on individual past context frames remains poorly understood, especially across the denoising steps of few-step distilled models. We introduce the **positional response** $R(\Delta, t_{\text{denoise}})$, a perturbation-based sensitivity measure quantifying how strongly predictions depend on a context frame at temporal distance Δ and denoising step t_{denoise}. Measured across distilled AR video diffusion models, R reveals a consistent two-axis structure: influence decays rapidly with distance, while later denoising steps draw on a broader temporal window than earlier ones. Motivated by this, we propose **Recency Forcing**, applying a timestep-dependent **Temporal Response Bias (TRB)** to attention logits that scaffolds attention to match the measured response. This also mitigates *KV eviction mismatch*, where distant frames must be evicted from the KV cache during long-horizon inference despite being available in training: since TRB already drives their influence toward zero, eviction has little impact, letting short-clip-trained models generate substantially longer videos. We further introduce **Biased Attention Reparameterization (BAR)**, an exact reformulation compatible with FlashAttention at negligible overhead. Recency Forcing supports both *training-free* and *training-based* settings, and experiments on VBench and VBench-Long show consistent long-horizon quality improvements with negligible inference overhead.

1 INTRODUCTION

Autoregressive (AR) video generation predicts each frame from past context, enabling streaming and interactive applications such as world modeling Mao et al. (2025); Sun et al. (2025); Hong et al. (2025), game simulation Valevski et al. (2024); Alonso et al. (2024); Zhang et al. (2025a); Yu et al. (2025a); Tang et al. (2025), and robotics Li et al. (2025); Yang et al. (2023). Unlike bidirectional video diffusion, which denoises all frames jointly, AR models condition each frame only on past

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.

context; training them with teacher-forced context Chen et al. (2025); Lei et al. (2023); Yin et al. (2025) leads to exposure bias, since the model never learns to handle its own imperfect rollouts at inference. Recent distillation-based methods, including Self-Forcing Huang et al. (2025a), Causal Forcing Zhu et al. (2026), and its extension Causal Forcing++ Zhao et al. (2026a), address this via self-rollout distribution matching distillation (DMD) Huang et al. (2025a), an on-policy procedure that trains on self-generated context with a gradient-based distribution-matching reward, substantially reducing multi-step drift and enabling fast, few-step causal generation.

Yet even as self-rollout DMD Huang et al. (2025a) closes the train-test *distribution* gap, a question remains open: how does a generated frame actually depend on individual context frames? In bidirectional diffusion, this is straightforward, since all frames attend to one another jointly; in few-step AR diffusion, generation is causal and denoising is compressed into a handful of steps, each of which may rely on past frames differently, and this dependency structure has not been directly measured. We address this question by introducing the **positional response** $R(\Delta, t_{\textbf{denoise}})$, a perturbation-based sensitivity measure that quantifies how strongly predictions depend on a context frame at distance Δ and denoising step t_{denoise}, by directly measuring the change in output when that frame is replaced, rather than reading attention weights. Measured on Causal Forcing Zhu et al. (2026), a Wan2.1-based distilled AR model, R reveals a consistent two-axis structure: influence decays rapidly with distance, while later denoising steps draw on a broader temporal window than earlier ones. This causal, perturbation-based characterization of temporal dependency is the main contribution of this work.

This raises a natural question: can this measured dependency be used to directly improve generation, rather than only serve as a diagnostic? We answer affirmatively with **Recency Forcing**, whose core component, **Temporal Response Bias (TRB)**, is a non-positive, timestep-dependent bias on attention logits that scaffolds attention toward the measured positional response. Because R concentrates in a narrow, step-dependent range of past frames, TRB provides an explicit temporal prior that directs attention toward this informative range, improving generation rather than merely reproducing the model's existing behavior.

This scaffolding has a further effect: it changes how these models tolerate long-horizon inference. Distilled AR models are trained on short clips (e.g., 5 seconds, 21 latent frames in Wan2.1), and extending inference beyond this horizon requires evicting distant frames from the KV cache to bound memory, a train–inference discrepancy we term **KV eviction mismatch**. Since TRB drives distant-frame influence toward zero well before eviction, their removal has little effect on predictions, allowing short-clip-trained models to generate substantially longer videos without added context length or retraining. To implement TRB efficiently, we introduce **Biased Attention Reparameterization (BAR)**, an exact reformulation under our query-grouping setting, compatible with FlashAttention Dao (2023) at negligible overhead.

We evaluate Recency Forcing on VBench Huang et al. (2023) and VBench-Long Huang et al. (2025b), generating videos up to 60 seconds, or $12\times$ beyond the training horizon. A *training-free* application to Self-Forcing Huang et al. (2025a) achieves a VBench-Long quality score of 82.63, and a *training-based* application to Causal Forcing Zhu et al. (2026) improves this to 84.02 after $\sim$12 hours of DMD retraining, in both cases improving long-horizon quality while preserving the base model's efficiency. Illustrative results are shown in Fig. 1.

Our contributions are as follows:

- **Analysis.** We introduce the positional response $R(\Delta, t_{\text{denoise}})$, a perturbation-based, causal measure of context-frame influence that reveals a two-axis structure jointly across temporal distance and denoising step.
- **Method.** We propose Recency Forcing with Temporal Response Bias (TRB), which scaffolds attention using this dependency structure, improving few-step AR generation quality and, as a further effect, letting short-clip-trained models generate substantially longer videos.
- **Implementation.** We introduce Biased Attention Reparameterization (BAR), an exact FlashAttention-compatible implementation of TRB with negligible inference overhead.

2 Related Work

Video generation models have evolved from U-Net diffusion models Blattmann et al. (2023); Guo et al. (2023) to scalable Diffusion Transformers (DiT) Peebles & Xie (2022), with recent models such as CogVideoX Yang et al. (2024), HunyuanVideo Kong et al. (2024), Wan Wan et al. (2025), Cosmos Agarwal et al. (2025), and LTX-Video HaCohen et al. (2024) achieving high-quality synthesis via transformer architectures over spatiotemporal tokens. These models generate all frames jointly using bidirectional attention, serving as the backbone for many AR models. Among them, Wan Wan et al. (2025) has emerged as the dominant open-source base for AR video generation Yin et al. (2025); Huang et al. (2025a); Zhu et al. (2026); we likewise build on Wan2.1.

Autoregressive video generation. A growing line of work extends bidirectional video diffusion to autoregressive generation Henschel et al. (2025); Li et al. (2024); Liu et al. (2024); Ren et al. (2025); Weng et al. (2024); Yuan et al. (2025); Zhang et al. (2025b). Early AR methods such as NOVA Deng et al. (2024), SkyReels Chen et al. (2025), and PyramidFlow Lei et al. (2023) generate chunk-by-chunk but require many sampling steps. More recent distillation-based methods achieve real-time generation: CausVid Yin et al. (2025) distills a bidirectional DiT into a few-step causal generator; Self-Forcing Huang et al. (2025a) reduces exposure bias by conditioning on the model's own outputs; Causal Forcing Zhu et al. (2026) and Causal Forcing++ Zhao et al. (2026a) further improve the ODE pretraining stage. All of these are trained on short clips ($\sim$5 seconds) and degrade over longer horizons – the regime we target.

Autoregressive long video generation. The autoregressive paradigm is a natural fit for long-horizon generation, but distilled AR video diffusion models trained on short clips often suffer from drift when rolled out far beyond their training horizon Huang et al. (2025a); Zhu et al. (2026). Recent work improves long-horizon generation by introducing an *attention sink* Xiao et al. (2023) as a persistent anchor and incorporating additional mechanisms. Existing approaches broadly fall into three categories. **(1) KV-cache compression and pruning** reduces memory growth by selectively discarding or compressing historical context. Representative examples include DeepForcing Yi et al. (2025), PackForcing Mao et al. (2026), RelaxForcing Zhao et al. (2026b), and MALT Yu et al. (2025b). **(2) RoPE-based methods** modify positional encoding to improve length extrapolation, including ∞-RoPE Yesiltepe et al. (2025), Rolling Forcing Liu et al. (2026), LoL Cui et al. (2026), and Anchor variants Yang et al. (2026); Li et al. (2026). **(3) Correction-based methods** expose models to their own degraded outputs or use feedback signals to stabilize long rollouts, such as Self-Resampling Guo et al. (2026), BAgger Po et al. (2025), Pathwise Xiang et al. (2026), Self-Forcing++ Cui et al. (2025), and Reward Forcing Lu et al. (2025).

These approaches manage memory, positional encoding, or training-time correction, but none directly measures how a generated frame depends on individual past frames. In contrast, we introduce a perturbation-based measure of this dependency and adapt attention weighting according to the measured response, which also mitigates KV eviction discrepancies during long-horizon inference.

3 Method

3.1 Preliminaries

Flow matching. Flow matching Lipman et al. (2022); Liu et al. (2022); Albergo et al. (2025) trains a velocity field $\boldsymbol{v}_\theta(\boldsymbol{x}_t, t)$ to transport a Gaussian prior to the data distribution. Following Wan2.1 Wan et al. (2025) and its distillations Yin et al. (2025); Huang et al. (2025a); Zhu et al. (2026), we use $t = 1$ for pure noise and $t = 0$ for clean data, with linear path $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{\epsilon}$ and training loss

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}} \big\| \boldsymbol{v}_\theta(\boldsymbol{x}_t, t) - (\boldsymbol{\epsilon} - \boldsymbol{x}_0) \big\|_2^2. \tag{1}$$

Autoregressive video generation. AR video models factorize the joint distribution as

$$p_\theta(\boldsymbol{x}_0^{1:N}) = \prod_{i=1}^{N} p_\theta(\boldsymbol{x}_0^i \mid \boldsymbol{x}_0^{<i}), \tag{2}$$

predicting each frame via a diffusion process conditioned on past generated frames $\boldsymbol{x}_0^{<i}$ (*context frames*). Generating a single frame requires many denoising steps (e.g., 50 in Wan2.1), and naively

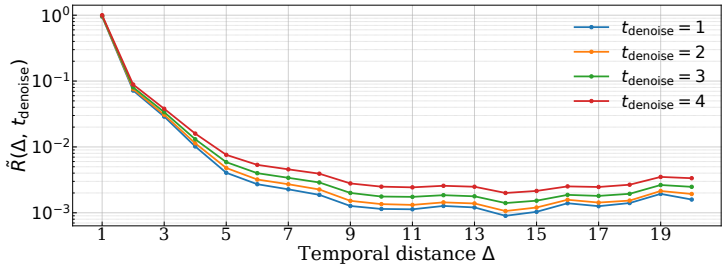


Figure 2: Normalized positional response $\tilde{R}(\Delta,\ t_{\text{denoise}})$ showing that context influence decays sharply with temporal distance and varies across denoising steps.

training with teacher-forced ground-truth context $\boldsymbol{x}_0^{<i}$ induces exposure bias, since the model never learns to condition on its own imperfect rollouts. Both issues make vanilla AR generation prohibitively slow and prone to drift for streaming use.

Distilled AR models. Distillation methods Huang et al. (2025a); Zhu et al. (2026) address both issues by training a few-step student generator $\boldsymbol{v}_\theta$ with distribution matching distillation (DMD), mapping noise directly to a clean frame in K steps (e.g., $K = 4$). They close the exposure-bias gap by training on the same self-rollouts, $\hat{\boldsymbol{x}}_0 = [\hat{\boldsymbol{x}}_0^1, \ldots, \hat{\boldsymbol{x}}_0^N]$, encountered at inference. The student is optimized via a distribution-matching gradient against a teacher velocity field $\boldsymbol{v}^{\text{real}}$, using an auxiliary fake field $\boldsymbol{v}_\phi^{\text{fake}}$ trained on the student's own outputs:

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[\left(\boldsymbol{v}_\phi^{\text{fake}}(\hat{\boldsymbol{x}}_t, t) - \boldsymbol{v}^{\text{real}}(\hat{\boldsymbol{x}}_t, t)\right)\frac{\partial \hat{\boldsymbol{x}}_0^i}{\partial \theta}\right], \tag{3}$$

yielding a student capable of fast, few-step causal generation robust to self-rollout drift.

Building on the distilled AR models, we now turn to the central question raised in Sec. 1: how does a generated frame actually depend on individual context frames? In few-step distilled AR models, generation is causal and denoising is compressed into a handful of steps, each of which may rely on past frames differently, and this dependency structure has not been directly measured. To answer this question, we introduce a perturbation-based sensitivity measure that directly quantifies this dependency, which we term the **positional response**.

3.2 Analysis: How Context Frames Influence Current Predictions

Positional response. Considering a 4-step AR model, e.g., Causal Forcing Zhu et al. (2026) based on Wan2.1-1.3B, we define the positional response $R(\Delta, t_{\text{denoise}})$ to quantify how strongly predictions depend on a context frame at temporal distance Δ during denoising step $t_{\text{denoise}} \in \{1, 2, 3, 4\}$, where $t_{\text{denoise}} = 1$ and $t_{\text{denoise}} = 4$ correspond to the highest- and lowest-noise steps, respectively.

$$R(\Delta,\ t_{\text{denoise}}) = \mathbb{E}\left\|\boldsymbol{v}_\theta\big(\boldsymbol{x}^i_{t_{\text{denoise}}} \mid \boldsymbol{x}_0^{<i}\big) - \boldsymbol{v}_\theta\big(\boldsymbol{x}^i_{t_{\text{denoise}}} \mid \delta_\Delta(\boldsymbol{x}_0^{<i})\big)\right\|^2, \tag{4}$$

where $\delta_\Delta(\boldsymbol{x}_0^{<i})$ replaces the frame at position Δ with one from an alternative rollout of the same prompt generated using a different random seed. We use replacement rather than additive noise to avoid introducing out-of-distribution perturbations. For cross-step comparison, we normalize by the response at $\Delta = 1$: $\tilde{R}(\Delta,\ t_{\text{denoise}}) = R(\Delta,\ t_{\text{denoise}})/R(1,\ t_{\text{denoise}})$.

Observations. Fig. 2 shows $\tilde{R}$ averaged over 200 MovieGen prompts Polyak et al. (2024) sampled from the same prompt pool used for VBench-Long evaluation, together with 20 alternative contexts, measured on Causal Forcing Zhu et al. (2026). The analysis is conducted on a large and diverse prompt set spanning a broad range of scenes, motions, camera trajectories, and subject configurations, as detailed in Sec. D, allowing us to characterize positional-response patterns over a representative portion of the evaluation distribution.

Three patterns emerge consistently: **(1) monotonic ordering**, where curves remain ordered by t_{denoise} across all distances; **(2) steep decay**, where $\tilde{R}$ decreases by orders of magnitude within a few frames; and **(3) step-dependent long-range reach**, where responses are similar at short distances but diverge at large Δ. Together, these observations indicate that denoising steps differ primarily in how they use distant context rather than nearby context.

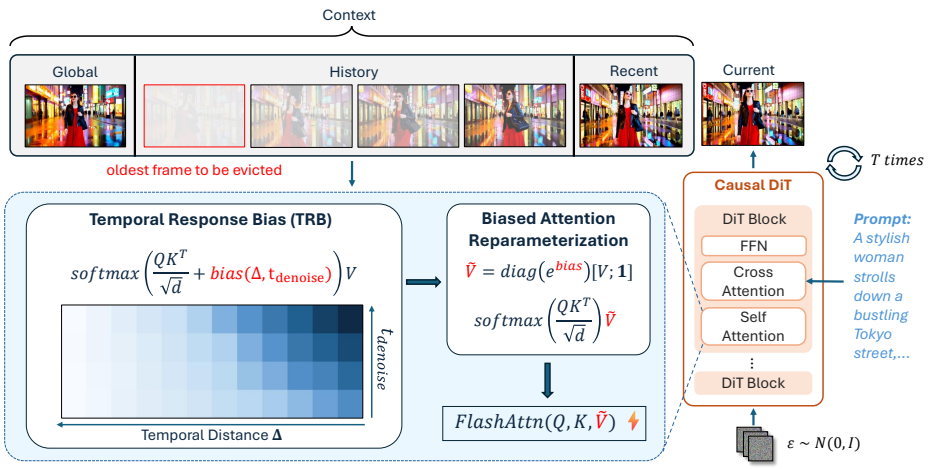


Figure 3: **Overview of Recency Forcing.** Context is partitioned into global sink frames (L_{global}), history frames (L_{history}), and recent frames (L_{recent}). Standard AR methods use unrestricted content-based attention over this window and evict distant frames at inference, creating a KV eviction mismatch. Recency Forcing addresses this by injecting a **Temporal Response Bias (TRB)** into attention logits, decaying history-frame influence as a function of temporal distance Δ and denoising timestep t_{denoise}. While a direct implementation requires flexible kernels (e.g., FlexAttention), **Biased Attention Reparameterization (BAR)** gives an exact reformulation that runs on FlashAttention at negligible overhead.

Implications. The measured positional response suggests that temporal context usage in these models is neither uniform across distance nor across denoising steps. In particular, distant context has progressively less influence on model predictions, while later denoising steps retain access to a broader temporal window than earlier steps. This observation motivates an attention mechanism whose treatment of historical context varies with both temporal distance and denoising step.

Our approach. Motivated by this analysis, we keep the full context window and instead modulate *how strongly* distant context is attended to. We instantiate this idea as **Recency Forcing**, whose core component is **Temporal Response Bias (TRB)**, a non-positive, timestep-dependent bias on pre-softmax attention logits derived from $\tilde{R}$. TRB encourages attention patterns that better match the measured positional response while preserving the full context window. To make this practical, we introduce **Biased Attention Reparameterization (BAR)**, an exact reformulation of TRB compatible with FlashAttention Dao (2023) under our query-grouping setting and with negligible overhead. Details follow in Secs. 3.3 and 3.4.

3.3 Temporal Response Bias (TRB)

Context partition. We partition the context into three regions (Fig. 3): **global** (L_{global} initial frames serving as an attention sink), **recent** (L_{recent} nearest frames supporting short-term motion), and **history** (the intermediate frames). Including the current denoising frames L_{current}, the attention window size is defined as

$$L_{\text{attention}} = L_{\text{global}} + L_{\text{history}} + L_{\text{recent}} + L_{\text{current}}.$$

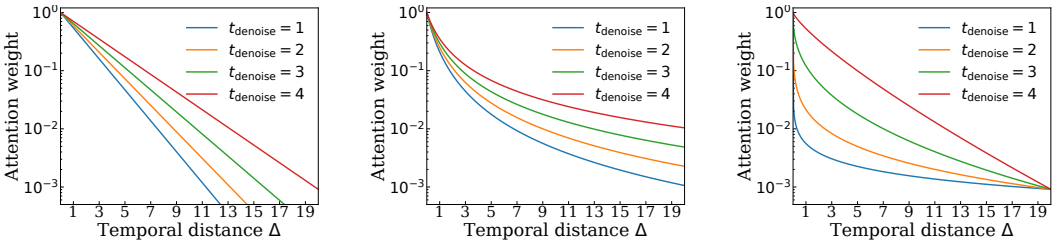
Since $R_{\Delta=1}$ remains maximal throughout denoising, we leave recent context unattenuated, while preserving the global context as an attention sink Liu et al. (2026); TRB acts only on the history.

Temporal response bias. To continuously attenuate historical context, we add a bias term to the scaled dot-product attention (SDPA) logits before the softmax. The bias depends on both temporal distance Δ and denoising timestep t_{denoise}:

$$\text{SDPA}(Q, K, V;\, B) = \text{Softmax}\Big(QK^\top/\sqrt{d} + B\Big)\, V, \tag{5}$$

where $B \in \mathbb{R}^{N\times N}$, $B_{i,j} = \text{bias}(\Delta_{i,j}, t_{\text{denoise}})$, and $\Delta_{i,j} = i - j$.

The positional-response analysis in Sec. 3.2 suggests three desirable properties for the bias: **(1) denoising-timestep-dependent strength**, stronger early and weaker late; **(2) rapid decay with** Δ,



(a) Linear decay. (b) Logarithmic decay. (c) Power-law decay.

Figure 4: Candidate decay functions designed to follow the observed empirical pattern.

matching the observed drop in $\tilde{R}$; and **(3) non-positive values** ($B_{i,j} \leq 0$), attenuating distant context rather than amplifying it.

Relation to context truncation. Context truncation (Yang et al., 2025; Lu et al., 2025) is a limiting case of temporal attenuation, retaining only the global sink and recent frames, equivalent to assigning a bias of $-\infty$ to the history region. This discards temporal information the model still relies on, degrading motion coherence (Fig. 6). TRB instead preserves the full context window and continuously modulates attention to historical frames.

Instantiating the bias. The principles above specify the desired behavior of TRB but do not uniquely determine its functional form. We therefore consider two instantiations: a fully learned bias and a structured parametric bias that encodes these principles by construction.

(1) Learnable bias. The most flexible instantiation imposes no predefined structure. We model the bias as a learnable lookup table $\text{bias}_{\text{learn}}[\Delta, t_{\text{denoise}}]$ over temporal distance and denoising step, initialize it from $\log(\tilde{R})$, register it as a model parameter, and optimize it jointly with the generator.

(2) Parametric bias. Alternatively, we consider three structured families: linear and logarithmic decay use $\alpha(t_{\text{denoise}}) > 0$ (decreasing with t_{denoise}), while power-law decay uses $\beta > 0$ and $\gamma(t_{\text{denoise}}) > 0$ (increasing with t_{denoise}):

$$\text{bias}_{\text{lin}} = -\alpha(t_{\text{denoise}}) \cdot (\Delta - L_{\text{recent}}), \tag{6}$$

$$\text{bias}_{\text{log}} = -\alpha(t_{\text{denoise}}) \cdot \log(1 + \Delta - L_{\text{recent}}), \tag{7}$$

$$\text{bias}_{\text{pow}} = -\beta \cdot \tilde{\Delta}^{\,\gamma(t_{\text{denoise}})}, \quad \tilde{\Delta} = \frac{\Delta - L_{\text{recent}}}{L_{\text{history}}}, \tag{8}$$

$$\alpha(t_{\text{denoise}}) = \alpha_{\text{base}} - \frac{t_{\text{denoise}}}{4}, \qquad \gamma(t_{\text{denoise}}) = \gamma_{\text{base}} \cdot 2^{\,t_{\text{denoise}}}, \tag{9}$$

where $\beta > 0$ is a scale constant controlling the decay range. All three functions are illustrated in Figs. 4a to 4c. Linear decay suppresses context at a constant rate, logarithmic decay under-suppresses distant frames relative to $\tilde{R}$, and power-law decay produces a convex profile that better matches the observed steep drop-off in positional response, concentrating suppression on the most distant frames.

3.4 Biased Attention Reparameterization (BAR)

TRB requires injecting a bias into the pre-softmax attention logits. While this can be implemented directly using flexible attention kernels such as FlexAttention Dong et al. (2024), doing so incurs additional overhead and is incompatible with FlashAttention Dao (2023), which does not expose intermediate logits. We address this with **Biased Attention Reparameterization (BAR)**, an exact reformulation that absorbs the bias into the value branch, enabling TRB to be implemented using a FlashAttention call with only a lightweight value augmentation. For a query frame q_i, since $B_{i,j}$ depends only on $\Delta_{i,j}$, it can be factored as $w_j = \exp(B_{i,j}) \in (0, 1]$, giving:

$$o_i = \text{Softmax}\left(\frac{q_i K^\top}{\sqrt{d}} + B_{i,:}\right) V = \frac{\sum_j \exp\left(\frac{q_i k_j^\top}{\sqrt{d}} + B_{i,j}\right) v_j}{\sum_j \exp\left(\frac{q_i k_j^\top}{\sqrt{d}} + B_{i,j}\right)} = \frac{\sum_j \exp\left(\frac{q_i k_j^\top}{\sqrt{d}}\right) w_j v_j}{\sum_j \exp\left(\frac{q_i k_j^\top}{\sqrt{d}}\right) w_j}. \tag{10}$$

Absorbing w_j into an augmented value $\tilde{v}_j = w_j \cdot [v_j; 1] \in \mathbb{R}^{d+1}$ where $[\cdot;\cdot]$ is concatenation and stacking into $\tilde{V} \in \mathbb{R}^{N\times(d+1)}$, this is equivalent to standard attention followed by a scalar normalization:

$$\tilde{o}_i = \mathrm{Softmax}\left(\frac{q_i K^\top}{\sqrt{d}}\right)\tilde{V}, \qquad o_i = \tilde{o}_i[:d] \,/\, \tilde{o}_i[\,d\,]. \tag{11}$$

This is algebraically identical to Eq. (5) with no approximation.

Exactness. The reparameterization in Eqs. (10) and (11) is derived for a fixed query temporal index i. BAR is therefore exact whenever all query tokens in an attention call share the same temporal index, which holds during autoregressive generation where frames (or chunks) are generated one at a time.

Efficiency. BAR requires only three lightweight modifications to a standard attention call: (i) rescaling each value vector v_j by its decay weight $w_j = \exp(B_{i,j})$; (ii) appending w_j to form the augmented value $\tilde{v}_j$; and (iii) normalizing the output by its final channel. Steps (i)–(ii) incur $O(Nd)$ cost and step (iii) incurs $O(N)$ cost, both negligible compared to the $O(N^2 d)$ attention computation. The attention operation itself remains an unmodified FlashAttention call, retaining its memory efficiency and throughput. Compared with direct logit modification in FlexAttention, BAR achieves more than $11\times$ speedup with identical outputs. Detailed runtimes are reported in Tab. 1, while end-to-end generation costs for 5-second video on H100 are summarized in Tab. 2.

Table 1: Runtime of bias injection and BAR.

Method	Time (ms)↓	Speedup↑
FlexAttn + bias	70.34	1.00×
Ours (FlashAttn)	6.06	**11.61**×

Table 2: End-to-end generation latency.

Method	Runtime (s)
Self Forcing	4.912
Recency Forcing + BAR	5.076

4 Experiments

Implementation details. We build Recency Forcing on Wan2.1-T2V-1.3B Wan et al. (2025), replacing standard SDPA with BAR-based TRB in all self-attention layers. The *training-free* variant is applied directly to chunk-wise Self-Forcing Huang et al. (2025a), while the *training-based* variant intialized from Causal Forcing ODE checkpoint Zhu et al. (2026). We use $L_{\text{attention}} = 21$ and $L_{\text{global}} = L_{\text{recent}} = 3$. Additional implementation details are provided in Appendix- Sec. C.

Evaluation benchmarks. We evaluate on VBench Huang et al. (2023) for short-video quality and VBench-Long Huang et al. (2025b) for long-horizon consistency following Rolling Forcing Liu et al. (2026), with a disjoint 50-prompt set for ablations. All videos are 60 seconds at 16 fps, 832×480. We additionally conduct a user study with 45 participants across four aspects: color consistency, motion dynamics, subject consistency, and overall quality. Details are in Appendix-Sec. B.

4.1 Comparison with prior approaches

Quantitative results. Tab. 3 reports VBench results for short video quality, and Tab. 4 reports VBench-Long results for long-horizon generation. We highlight three takeaways. **First, short video quality is preserved.** Our method achieves the best overall score on VBench, demonstrating that integrating TRB introduces no degradation in short-video quality – the bias has negligible effect when all context frames fit within the training window. **Second, long-horizon stability is substantially improved.** On VBench-Long, our method achieves the highest dynamic degree (75.48) while maintaining strong imaging quality (70.50) and subject consistency (97.68). This indicates that TRB effectively mitigates long-horizon drift without sacrificing visual fidelity or motion naturalness. **Third, the training-free variant is competitive.** Among training-free methods, our approach achieves the best overall VBench-Long score, confirming that the positional response curves $\tilde{R}$ provide a reliable prior for setting the decay schedule without any additional training.

Qualitative results. Fig. 5 shows 60-second video generations across methods. Competing methods exhibit three main failure modes. **Self-Forcing** shows progressive drift in color, texture, and sharpness, leading to quality degradation. **LongLive** reduces drift but introduces repetition due to its global sink mechanism (red boxes in Fig. 5). **Deep-Forcing** reduces drift but introduces compression artifacts such as motion blur. **Infinity-RoPE** mitigates drift but suffers from gradual subject inconsistency over time. In contrast, our method maintains consistent visual quality and stable motion coherence over the full 60 seconds without noticeable artifacts or inconsistency.

Table 3: **Comparison on VBench on generated short (5-second) videos**. Higher is better.

Model	Total Score	Quality Score	Semantic Score
Autoregressive Video Diffusion Models			
SkyReels-V2 Chen et al. (2025)	82.67	84.70	74.53
MAGI-1 ai et al. (2025)	79.18	82.04	67.74
NOVA Deng et al. (2024)	80.12	80.39	79.05
Pyramid Flow Jin et al. (2025)	81.72	84.74	69.62
Distilled Autoregressive Video Models			
CausVid Yin et al. (2025)*	81.20	84.05	69.80
Self Forcing Huang et al. (2025a)	84.31	85.07	<u>81.28</u>
Causal Forcing Zhu et al. (2026)	84.04	84.59	**81.84**
Rolling Forcing Liu et al. (2026)	83.19	83.97	80.06
Long Live Yang et al. (2025)	<u>84.87</u>	**86.97**	76.47
Reward Forcing Lu et al. (2025)	84.20	84.06	81.38
Recency Forcing (Ours)	**85.08**	<u>86.20</u>	80.59

Table 4: **Results on VBench-Long Huang et al. (2025b) with 60-second videos.** Higher is better.

Model	Quality	Dynamic	Motion Smoothness	Temporal Flickering	Imaging	Aesthetic	Subject Consistency	Background Consistency
Training-free								
Infinity RoPE Yesiltepe et al. (2025)	82.30	56.22	98.53	97.41	67.14	59.20	97.31	96.15
Deep Forcing Yi et al. (2025)	82.14	56.01	98.10	96.88	68.22	59.85	97.35	96.30
Recency Forcing (Ours)	82.63	56.27	98.23	96.98	68.94	60.48	**97.81**	**96.64**
Training-based								
Self Forcing Huang et al. (2025a)	80.11	34.50	98.48	**97.77**	65.93	56.55	96.97	96.24
Causal Forcing Zhu et al. (2026)	78.98	<u>65.19</u>	96.94	95.21	63.44	51.35	95.39	95.51
Rolling Forcing Liu et al. (2026)	81.56	34.68	<u>98.71</u>	<u>97.62</u>	<u>70.18</u>	59.93	<u>97.74</u>	<u>96.54</u>
Long Live Yang et al. (2025)	81.98	42.85	**98.75**	97.61	68.71	**61.48**	97.09	95.99
Reward Forcing Lu et al. (2025)	<u>82.73</u>	64.63	98.22	96.54	69.17	59.43	97.07	96.16
Recency Forcing (Ours)	**84.02**	**75.48**	98.00	96.32	**70.50**	<u>61.31</u>	97.68	96.49

4.2 Ablation Studies

We conduct ablation studies on a held-out set of 50 prompts as the validation set, evaluating on VBench-Long with 60-second videos. All metrics are **higher-is-better**.

Choice of bias function and context length. We compare full context usage without TRB ($L_{\text{attention}} = 21$), *context truncation* ($L_{\text{attention}} = 9$), a learnable bias, three parametric decay functions (Eqs. (6) to (8)). As shown in Tab. 5, retaining the full context without TRB results in extreme color and subject drift, while truncation removes useful temporal context outright, causing unnatural motion and unwanted scene changes that inflate dynamic degree (88.58), as shown in Fig. 6. Within TRB variants, the learnable bias underperforms, suggesting that unconstrained learning does not reliably recover the temporal attenuation pattern motivated in Sec. 3.3; among the parametric forms, power-law decay performs best, consistent with its convex profile closely matching $\tilde{R}$, while linear and logarithmic decay under-suppress distant context and lead to residual drift. We therefore adopt power-law decay for all subsequent experiments.

Effect of decay strength γ_{base}. We vary γ_{base} at inference time to study decay aggressiveness. As shown in Tab. 6, small values insufficiently suppress distant frames, while large values over-suppress history and harm motion coherence. We find $\gamma_{\text{base}} = 0.1$ provides the best trade-off.

Effect of recent context length. We vary $L_{\text{recent}} \in \{0, 3, 6\}$ to study recent context length. As shown in Tab. 7, increasing L_{recent} from 0 to 3 substantially improves motion quality, whereas increasing it to 6 degrades performance. We therefore use $L_{\text{recent}} = 3$.

5 Conclusion

We introduce the **positional response** $R(\Delta, t_{\text{denoise}})$, a perturbation-based measure that reveals a consistent two-axis structure in how distilled AR video diffusion models depend on temporal context. Guided by this analysis, we propose **Recency Forcing**, using a timestep-dependent **Temporal Response Bias (TRB)** to scaffold attention toward the measured response; a further effect of this scaffolding is that it suppresses distant-frame influence before KV eviction, letting short-clip-

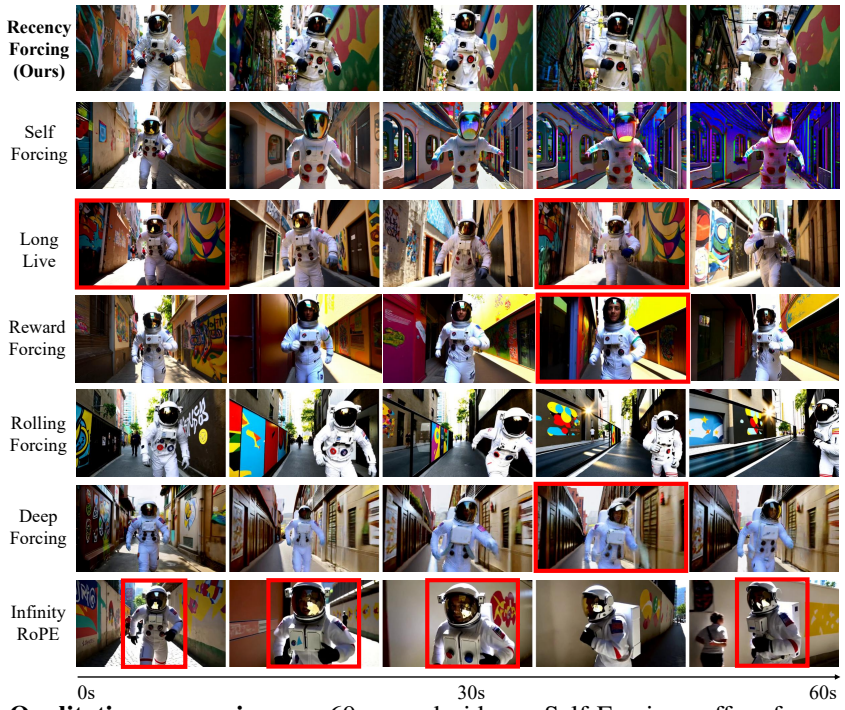


Figure 5: **Qualitative comparison** on 60-second videos. Self-Forcing suffers from color drift, LongLive produces scene repetition (red boxes), DeepForcing introduces blur, and Reward Forcing, Infinity-RoPE loses subject consistency. Recency Forcing preserves visual fidelity and motion throughout. For better visualization, we refer the reader to Appendix- Sec. A.



Figure 6: Qualitative comparison of context truncation ($L_{\text{attention}} = 9$) and Recency Forcing over a 60-second rollout. Truncation produces unstable motion and scene changes (middle row), while Recency Forcing maintains consistent subject identity and visual quality throughout (bottom row).

Table 5: Bias decay and context study.

Context Type	Quality	Imaging	Dynamic
Standard ($L_{\text{attention}} = 21$)	80.56	65.09	78.32
Truncated ($L_{\text{attention}} = 9$)	81.10	66.40	**88.58**
Learnable decay	82.42	68.96	62.40
Linear decay	83.26	68.90	66.06
Log decay	83.16	**69.42**	72.06
Power-law decay	**84.17**	69.05	81.55

Table 6: Effect of decay strength γ_{base}.

γ_{base}	Quality	Imaging	Dynamic
0.01	**84.36**	**69.63**	79.42
0.1	84.17	69.05	81.55
0.5	82.91	67.05	81.94
1.0	82.61	65.98	**83.93**

Table 7: Effect of recent context size.

L_{recent}	Quality	Imaging	Dynamic
0	82.80	**69.86**	40.90
3	**84.17**	69.05	**81.55**
6	82.14	69.51	49.03

trained models generate substantially longer videos. **Biased Attention Reparameterization (BAR)** implements TRB exactly within FlashAttention at negligible overhead. On 60-second generations, both training-free and training-based variants improve long-horizon quality and consistency over their base models, with the training-based variant achieving state-of-the-art VBench-Long performance.

REFERENCES

Nvidia Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, JiaoJiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Samuel Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Kl'ar, Grace Lam, Shiyi Lan, Laura Leal-Taixé, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum A. Reda, Xiao-Shuai Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne P. Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Rajan Varghese, Hao Wang, Haoxiang Wang, Hengyi Wang, Tingwei Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yuan Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai. *ArXiv*, abs/2501.03575, 2025. URL https://api.semanticscholar.org/CorpusID:275342594.

Sand. ai, Han Sack Teng, Hong Jia, Lei Sun, Lingzhi Li, Maolin Li, Min Tang, Shu Han, Tianning Zhang, W. Q. Zhang, Weifeng Luo, Xiaoyang Kang, Yuchen Sun, Yue Cao, Yunpeng Huang, Yutong Lin, Yuxin Fang, Zewei Tao, Zheng Zhang, Zhongshu Wang, Zixun Liu, Daiqi Shi, Guoli Su, Hanwen Sun, Hong-Fei Pan, Jie Wang, Jie Sheng, Mingyan Cui, Min Hu, Ming Yan, Shucheng Yin, Si-Hui Zhang, Tingting Liu, Xi Yin, Xiaoyu Yang, Xin Song, Xuan Hu, Yankai Zhang, and Yu-Qian Li. Magi-1: Autoregressive video generation at scale. *ArXiv*, abs/2505.13211, 2025. URL https://api.semanticscholar.org/CorpusID:278768832.

Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 37:58757–58791, 2024.

A. Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, and Dominik Lorenz. Stable video diffusion: Scaling latent video diffusion models to large datasets. *ArXiv*, abs/2311.15127, 2023. URL https://api.semanticscholar.org/CorpusID:265312551.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Juncheng Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zhenghao Chen, Chengcheng Ma, Weiming Xiong, Wei Wang, Nuo Pang, Kang Kang, Zhi-Xin Xu, Yuzhe Jin, Yu Liang, Yu-Ning Song, Peng Zhao, Bo Xu, Di Qiu, Debang Li, Zhengcong Fei, Yang Li, and Yahui Zhou. Skyreels-v2: Infinite-length film generative model. *ArXiv*, abs/2504.13074, 2025. URL https://api.semanticscholar.org/CorpusID:277856899.

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

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

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

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. *ArXiv*, abs/2412.14169, 2024. URL https://api.semanticscholar.org/CorpusID:274822761.

Juechu Dong, Boyuan Feng, Driss Guessous, Yanbo Liang, and Horace He. Flex attention: A programming model for generating optimized attention kernels. *arXiv preprint arXiv:2412.05496*, 2(3):4, 2024.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Y. Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *ArXiv*, abs/2307.04725, 2023. URL https://api.semanticscholar.org/CorpusID:259501509.

Yuwei Guo, Ceyuan Yang, Hao He, Yang Zhao, Meng Wei, Zhenheng Yang, Weilin Huang, and Dahua Lin. End-to-end training for autoregressive video diffusion via self-resampling. In *European Conference on Computer Vision*, pp. 324–344. Springer, 2026.

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

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

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

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *ArXiv*, abs/2506.08009, 2025a. URL https://api.semanticscholar.org/CorpusID:279251392.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qin Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818, 2023. URL https://api.semanticscholar.org/CorpusID:265506207.

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

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *International Conference on Learning Representations*, volume 2025, pp. 23378–23402, 2025.

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.

Jiarui Lei, Xiaobo Hu, Yue Wang, and Dong Liu. Pyramidflow: High-resolution defect contrastive localization using pyramid normalizing flow. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14143–14152, 2023. URL https://api.semanticscholar.org/CorpusID:257365925.

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.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025.

Zongyi Li, Shujie Hu, Shujie Liu, Long Zhou, Jeongsoo Choi, Lingwei Meng, Xun Guo, Jinyu Li, Hefei Ling, and Furu Wei. Arlon: Boosting diffusion transformers with autoregressive models for long video generation. *arXiv preprint arXiv:2410.20502*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *ArXiv*, abs/2210.02747, 2022. URL https://api.semanticscholar.org/CorpusID:252734897.

Haozhe Liu, Shikun Liu, Zijian Zhou, Mengmeng Xu, Yanping Xie, Xiao Han, Juan C Pérez, Ding Liu, Kumara Kahatapitiya, Menglin Jia, et al. Mardini: Masked autoregressive diffusion for video generation at scale. *arXiv preprint arXiv:2410.20280*, 2024.

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.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

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. *arXiv preprint arXiv:2512.04678*, 2025.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, Tong He, Jiangmiao Pang, Yu Qiao, and Kaipeng Zhang. Yume-1.5: A text-controlled interactive world generation model. *arXiv preprint arXiv:2512.22096*, 2025.

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. *arXiv preprint arXiv:2603.25730*, 2026.

William S. Peebles and Saining Xie. Scalable diffusion models with transformers. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2022. URL https://api.semanticscholar.org/CorpusID:254854389.

Ryan Po, Eric Ryan Chan, Changan Chen, and Gordon Wetzstein. Bagger: Backwards aggregation for mitigating drift in autoregressive video diffusion models. *ArXiv*, abs/2512.12080, 2025. URL https://api.semanticscholar.org/CorpusID:283895805.

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.

Sucheng Ren, Chen Chen, Zhenbang Wang, Liangchen Song, Xiangxin Zhu, Alan Yuille, Yinfei Yang, and Jiasen Lu. Autoregressive video generation beyond next frames prediction. *arXiv preprint arXiv:2509.24081*, 2025.

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

Junshu Tang, Jiacheng Liu, Jiaqi Li, Longhuang Wu, Haoyu Yang, Penghao Zhao, Siruis Gong, Xiang Yuan, Shuai Shao, Linfeng Zhang, et al. Hunyuan-gamecraft-2: Instruction-following interactive game world model. *arXiv preprint arXiv:2511.23429*, 2025.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

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

Wenming Weng, Ruoyu Feng, Yanhui Wang, Qi Dai, Chunyu Wang, Dacheng Yin, Zhiyuan Zhao, Kai Qiu, Jianmin Bao, Yuhui Yuan, et al. Art-v: Auto-regressive text-to-video generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7395–7405, 2024.

Xunzhi Xiang, Zixuan Duan, Guiyu Zhang, Haiyu Zhang, Zhe Gao, Junta Wu, Shaofeng Zhang, Tengfei Wang, Qi Fan, and Chunchao Guo. Pathwise test-time correction for autoregressive long video generation. *ArXiv*, abs/2602.05871, 2026. URL https://api.semanticscholar.org/CorpusID:285304326.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks, 2024. *URL https://arxiv. org/abs/2309.17453*, pp. 1, 2023.

Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 1(2):6, 2023.

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*, 2025.

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. *arXiv preprint arXiv:2408.06072*, 2024.

Hidir Yesiltepe, Tuna Han Salih Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. *arXiv preprint arXiv:2511.20649*, 2025.

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, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, 2025.

Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11590–11599, 2025a.

Sihyun Yu, Meera Hahn, Dan Kondratyuk, Jinwoo Shin, Agrim Gupta, José Lezama, Irfan Essa, David Ross, and Jonathan Huang. Malt diffusion: Memory-augmented latent transformers for any-length video generation. *arXiv preprint arXiv:2502.12632*, 2025b.

Hangjie Yuan, Weihua Chen, Jun Cen, Hu Yu, Jingyun Liang, Shuning Chang, Zhihui Lin, Tao Feng, Pengwei Liu, Jiazheng Xing, et al. Lumos-1: On autoregressive video generation from a unified model perspective. *arXiv preprint arXiv:2507.08801*, 2025.

Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, et al. Matrix-game: Interactive world foundation model. *arXiv preprint arXiv:2506.18701*, 2025a.

Yuan Zhang, Jiacheng Jiang, Guoqing Ma, Zhiying Lu, Haoyang Huang, Jian min Yuan, and Nan Duan. Generative pre-trained autoregressive diffusion transformer. *ArXiv*, abs/2505.07344, 2025b. URL https://api.semanticscholar.org/CorpusID:278501597.

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

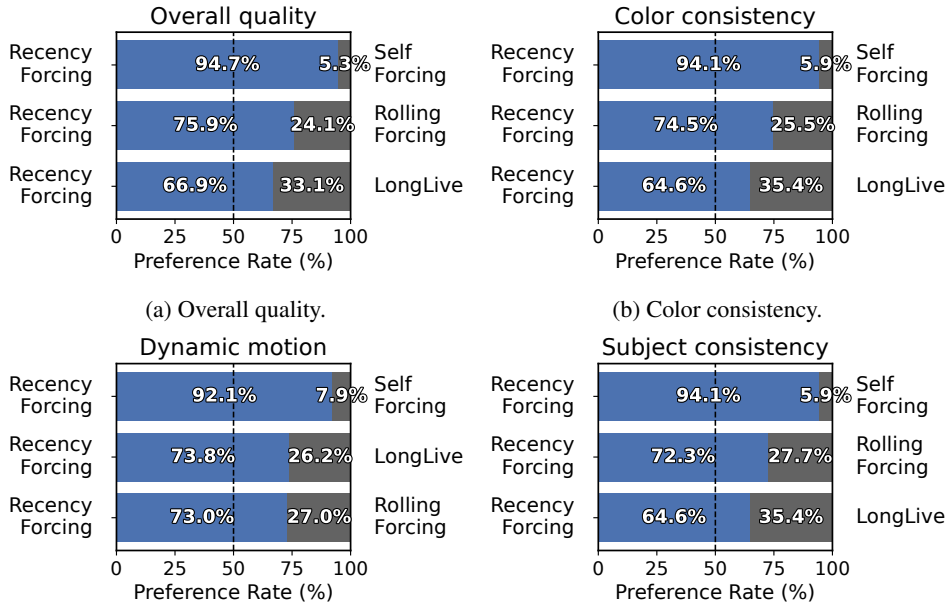
Zengqun Zhao, Yanzuo Lu, Ziquan Liu, Jifei Song, Jiankang Deng, and Ioannis Patras. Relax forcing: Relaxed kv-memory for consistent long video generation. *arXiv preprint arXiv:2603.21366*, 2026b.

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

A Additional qualitative results

We provide additional qualitative results for both the main comparison and the ablation study in the *Supplementary Video* (accessible via `index.html`). We also include comments in each video to indicate when errors or artifacts appear throughout the 1-minute generation.

B User Study



(a) Overall quality.

(b) Color consistency.

(c) Dynamic motion.

(d) Subject consistency.

Figure 7: **User preference study.** Preference rates of **Recency Forcing** against baseline methods across four evaluation aspects: **(a)** overall quality, **(b)** color consistency, **(c)** dynamic motion, and **(d)** subject consistency. Recency Forcing is consistently preferred in all settings. The largest margin is observed in *dynamic motion*, highlighting improved long-range identity preservation beyond what is reflected by automatic metrics.

We conduct a human evaluation using a one-versus-other protocol.

Setup. We conduct a user study with 45 participants and evaluate on randomly sampled 50 prompts. For each prompt, we generate videos from four methods: Self Forcing Huang et al. (2025a), Rolling Forcing Liu et al. (2026), Longlive Yang et al. (2025), and our Recency Forcing. Each participant is assigned up to 20 pairwise comparisons randomly sampled from four methods.

Protocol. We adopt a Two-Alternative Forced Choice (2AFC) protocol. Participants are shown two videos at a time and asked to select the better one according to the following criteria:

- **Color consistency**: which video has more consistent colors across frames?
- **Dynamic motion**: which video show more natural and convincing motion dynamics?
- **Subject consistency**: which video keep the subject more consistent over time?
- **Overall quality**: Which video has better overall quality?

The evaluation is conducted via a web-based interface with randomized ordering to avoid bias.

Statistics. In total, we collected 3384 responses. Results are reported as preference percentages.

Results. As shown in Fig. 7, participants consistently prefer Recency Forcing across all four evaluation aspects. The largest margin appears in motion dynamics, suggesting that TRB's suppression

of distant-frame drift improves perceptual identity preservation — an effect not fully captured by automatic metrics. These results further support our quantitative and qualitative findings.

C Implementation Details

We implement Recency Forcing on top of a 1.3B autoregressive video diffusion model following Causal Forcing Zhu et al. (2026), built on the Wan2.1-T2V-1.3B backbone Wan et al. (2025). We replace standard SDPA with BAR-based TRB in all self-attention layers. The model generates videos at a resolution of 832×480, $L_{\text{attention}} = 21$ latent frames, $L_{\text{global}} = L_{\text{recent}} = L_{\text{current}} = 3$ latent frames, with TRB applied to the remaining history frames.

Analysis. The positional response $R(\Delta, t_{\text{denoise}})$ analysis is conducted on both chunk-wise and frame-wise Causal Forcing Zhu et al. (2026); only the frame-wise version is reported in Sec. 3.2 for conceptual clarity.

Training. Our method can be directly applied to pretrained distilled autoregressive video models Huang et al. (2025a); Zhu et al. (2026). For the training-based variant, we initialize from a pretrained causal ODE checkpoint of Causal Forcing Zhu et al. (2026) and fine-tune only the DMD stage with TRB following the Self-Forcing Huang et al. (2025a) training protocol. Training is conducted on short clips (5 seconds) sampled from a filtered and LLM-augmented version of VidProM Wang & Yang (2024) for 3,500 steps with batch size 4 on H100 GPUs. During training, we sample $\alpha_{\text{base}} \sim \mathcal{U}(1, 2)$, $\gamma_{\text{base}} \sim \mathcal{U}(0, 1)$, and set $\beta = L_{\text{history}}$.

Inference. During inference, KV-cache eviction is applied as in prior work, while TRB adaptively reduces reliance on evicted frames. All experiments use chunk-wise Causal Forcing Zhu et al. (2026) as the base model, with a chunk size of 3 latent frames, temporal distance Δ is calculated between chunks. For the training-free setting, we apply our method to Self-Forcing Huang et al. (2025a) to ensure a fair comparison with other training-free methods (e.g., Deep Forcing Yi et al. (2025) and ∞-RoPE Yesiltepe et al. (2025)). For long video generation, we adopt relative RoPE Liu et al. (2026); Yi et al. (2025); Zhao et al. (2026b), where the RoPE indices of the global context are shifted to align with the immediately preceding history frames. For fair comparison, we also incorporate the Causal Forcing baseline with both global sink tokens and relative RoPE.

D Prompt Analysis

We report the distribution of the 200 MovieGen Video Bench prompts (Polyak et al., 2024) used for the diagnostic $\tilde{R}$ analysis. This analysis uses no benchmark scores; all bias-family and hyperparameter selection uses the disjoint 50-prompt validation set in Sec. 4.2.

Concept coverage. Fig. 8 reports the distribution of the prompts following MovieGen Video Bench official concept split. Human-centered content is most frequent (70), followed by scene (42), physics (37), unusual activity (37), camera motion (32), unusual subject (17), and animal (16), covering both ordinary and deliberately atypical content.

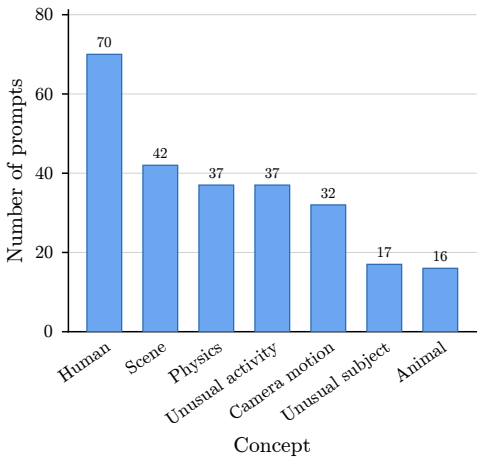
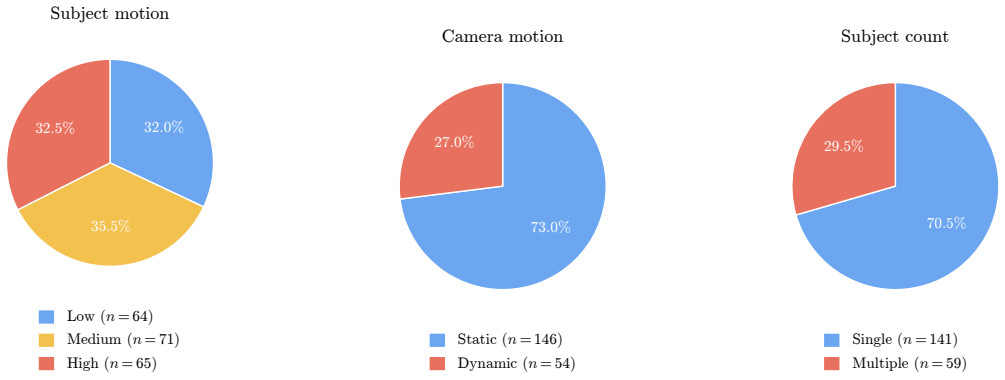


Figure 8: Concept distribution of the $R(\Delta, t_{\text{denoise}})$ analysis prompt set, following the concept taxonomy of MovieGen Video Bench.

Motion and subject configuration. We further manually annotate all 200 prompts along three content axes, each partitioning the set exactly: *subject motion* (low: the subject is near-static / medium: moderate locomotion or object manipulation / high: fast, large-amplitude motion), *camera motion* (static: locked-off framing / dynamic: any deliberate rig movement, from a slow push-in to a fast fly-through), and *subject count* (single / multiple salient foreground subjects). As shown in Fig. 9, the resulting set spans all categories along every axis and reflects the composition of the MovieGen prompt benchmark.



(a) Subject motion. (b) Camera motion. (c) Subject count.

Figure 9: Distribution of the $R(\Delta, t_{\text{denoise}})$ analysis prompt set, categorized by subject motion, camera motion, and subject count.

E GENERALIZATION TO FRAME-WISE SETTING

We further provide results in the frame-wise autoregressive setting, complementing the chunk-wise evaluation in the main paper.

Setup. Among the baselines considered in Sec. 4, Causal Forcing Zhu et al. (2026) is the only method with a frame-wise variant, so we adopt it as the point of comparison. We train Recency Forcing in frame-wise mode, reusing the same TRB schedule from chunkwise setting, and evaluate it on long-video generation under the same protocol as Sec. 4. Results are reported in Tab. 8.

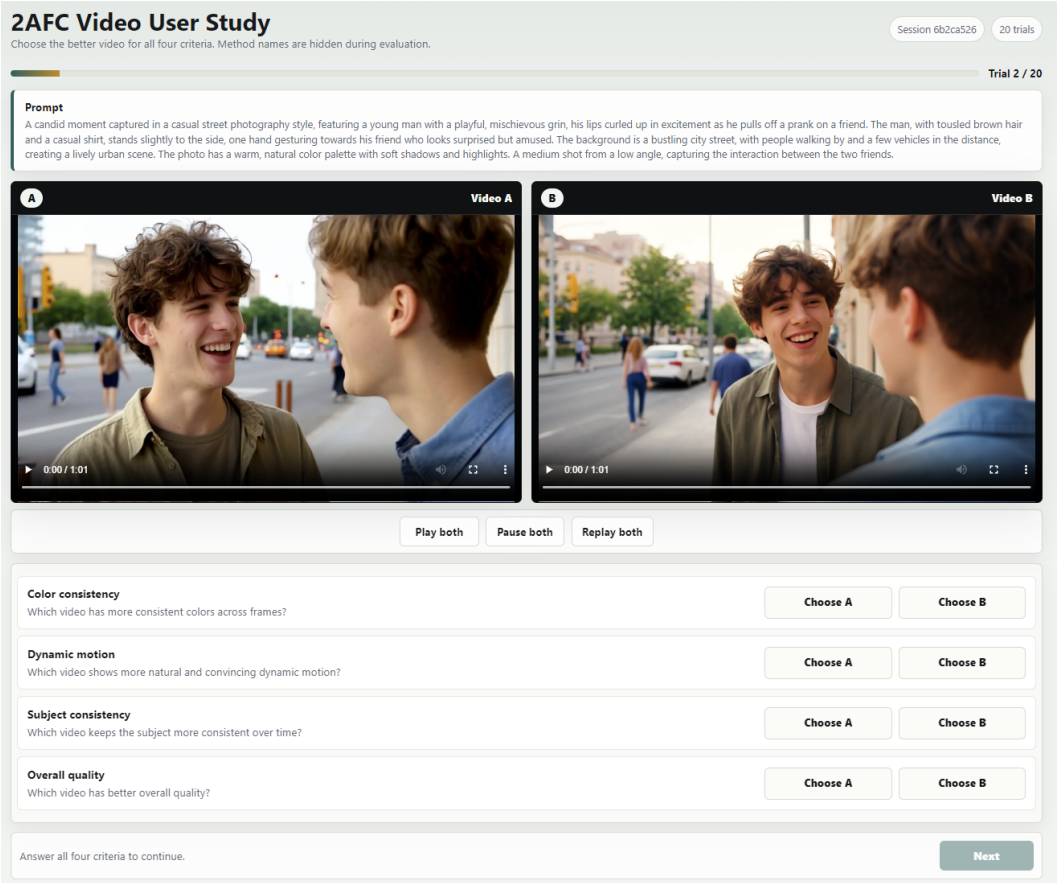


Figure 10: An example interface for our user preference study, where videos generated by different methods are displayed in a randomized left/right arrangement.

Table 8: **Generalization to the frame-wise setting.** Results on VBench-Long Huang et al. (2025b) evaluated on 60-second videos, under the same protocol as Tab. 4. Higher is better.

Model	Quality	Dynamic	Motion Smoothness	Temporal Flickering	Imaging	Aesthetic	Subject Consistency	Background Consistency
Frame-wise								
Causal Forcing Zhu et al. (2026)	79.65	**79.71**	96.75	93.96	65.54	51.96	94.72	94.95
Recency Forcing (Ours)	**82.62**	60.71	**98.18**	**96.56**	**71.60**	**57.89**	**97.47**	**96.18**

Results. Recency Forcing transfers cleanly to the frame-wise setting, improving Quality by $+2.97$ and improving on other fidelity and consistency metrics.

F Details of Biased Attention Reparameterization

We provide the full derivation of **Biased Attention Reparameterization (BAR)**.

For a single query frame q_i, biased attention is defined as:

$$o_i = \mathrm{Softmax}\left(\frac{q_i K^\top}{\sqrt{d}} + B_{i,:}\right) V. \tag{12}$$

Expanding the softmax:

$$o_i = \frac{\sum_j \exp\left(\frac{q_i k_j^\top}{\sqrt{d}} + B_{i,j}\right) v_j}{\sum_j \exp\left(\frac{q_i k_j^\top}{\sqrt{d}} + B_{i,j}\right)}. \tag{13}$$

Since $B_{i,j}$ depends only on position, define:

$$w_j = \exp(B_{i,j}) \in (0, 1]. \tag{14}$$

Substituting:

$$o_i = \frac{\sum_j \exp\left(\frac{q_i k_j^\top}{\sqrt{d}}\right) w_j v_j}{\sum_j \exp\left(\frac{q_i k_j^\top}{\sqrt{d}}\right) w_j}. \tag{15}$$

We absorb w_j into an augmented value:

$$\tilde{v}_j = w_j \cdot [v_j; 1] \in \mathbb{R}^{d+1}, \tag{16}$$

where $[;]$ denotes concatenation. Stacking all augmented values gives $\tilde{V} \in \mathbb{R}^{N\times(d+1)}$.

The attention can then be rewritten as standard attention:

$$\tilde{o}_i = \mathrm{Softmax}\left(\frac{q_i K^\top}{\sqrt{d}}\right) \tilde{V}. \tag{17}$$

Finally, the original output is recovered by:

$$o_i = \tilde{o}_i[:d] \,/\, \tilde{o}_i[d]. \tag{18}$$

This reformulation is algebraically identical to the original biased attention with no approximation, enabling efficient implementation using FlashAttention kernels.

G LIMITATIONS

While our method extends autoregressive generation far beyond the training horizon, it is not unbounded: we observe noticeable quality degradation once the generated duration exceeds roughly four to five minutes (Fig. 11). Our experiments are also confined to a single backbone. We adopt Wan2.1 1.3B Wan et al. (2025), the most common testbed used by the baselines in this line of work, and do not evaluate generalization to other video diffusion models such as CogVideoX Yang et al. (2024) or LTX HaCohen et al. (2024), nor to larger variants within the Wan family. Finally, we do not assess how the method transfers to downstream settings that place different demands on long-horizon modeling, such as action-conditioned world models or long-context retrieval tasks. We leave these directions to future work.

H AI USE STATEMENT

In this work, we used generative AI tools to assist in writing the mathematical claims, to refine hypotheses, to provide feedback on our research methodology and experiments, and to assist in implementing methods. We have not used generative AI tools to generate synthetic data sets, develop theoretical models, provide critical ingredients for proofs, clean or reformat datasets, or interpret results; qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools for translation and general writing assistance. We have reviewed all AI-assisted work: AI-suggested hypotheses and methodological feedback were evaluated by the authors and checked against a manual literature survey, all derivations were verified by the authors, and LLM-assisted code was tested for correctness. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.



Figure 11: Failure cases of our method at extremely long horizons (5-minute videos). In the first-row example, the house's column gradually drifts and changes color after approximately 2 minutes. In the second-row example, the bird disappears as the video approaches the 4-minute mark, then reappears with inconsistent visual appearance compared to earlier frames.