ROLLOUT-MARGINAL DISTILLATION FOR LONG-HORIZON AUTOREGRESSIVE VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive (AR) video diffusion enables low-latency, streamable video generation, but prediction errors often accumulate over long rollouts. Training the generator on its own rollouts exposes it to these imperfect histories. However, existing video-level distribution matching distillation (DMD) scores the whole rollout jointly. Because a chunk is evaluated together with its past and future, its correction can favor matching artifacts in the surrounding context merely to preserve temporal consistency. To provide a clearer visual-quality signal, we introduce Rollout-Marginal Distillation (RMD). RMD retains the generated history for AR prediction but scores each chunk independently against a chunk teacher, ensuring its quality correction is not compromised by an imperfect temporal context. To compensate for the lack of temporal context in independent chunk scoring, RMD subsequently applies video-level DMD to restore temporal coherence. Extensive experiments demonstrate that RMD maintains high visual quality far beyond its training horizon and outperforms video-level DMD baselines.

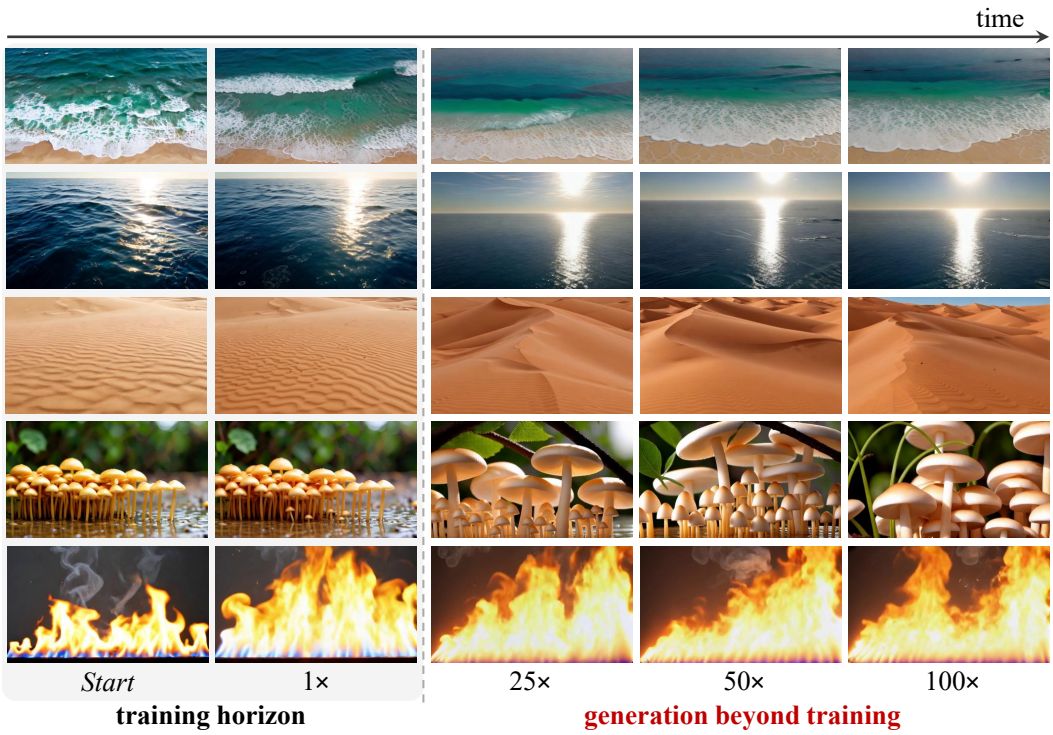
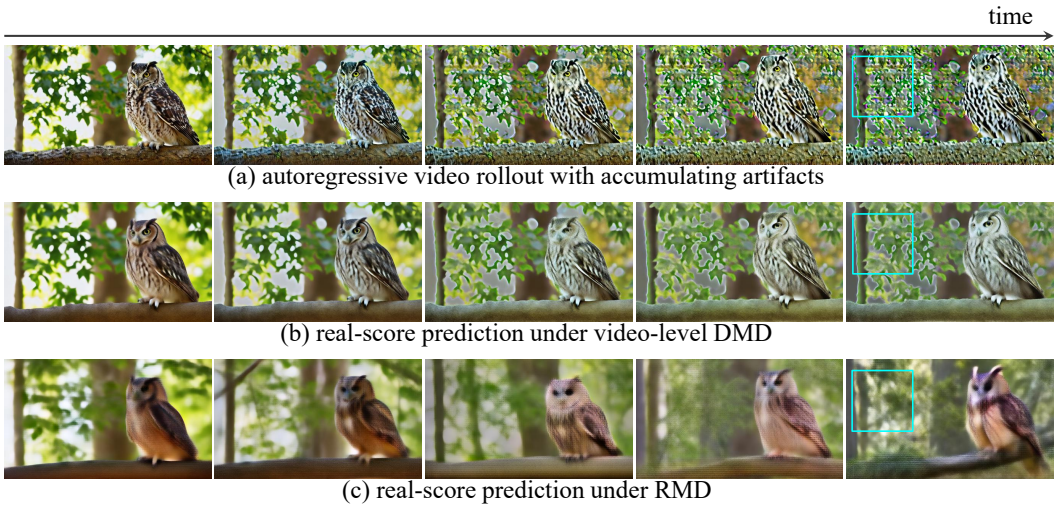


Figure 1: **Long-horizon autoregressive video generation.** Trained on 5-second rollouts, RMD maintains high visual quality over 500-second rollouts (100× the training horizon). Notably, generation relies solely on a fully sliding context window, without fixed anchors, additional explicit memory mechanisms, or modifications to the generator architecture.

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.



(a) autoregressive video rollout with accumulating artifacts

(b) real-score prediction under video-level DMD

(c) real-score prediction under RMD

Figure 2: **Comparison of real-score predictions under video-level DMD and RMD.** (a) The autoregressive rollout accumulates visual artifacts. (b) Video-level DMD scores the sequence jointly, causing the real-score prediction to retain these historical artifacts. (c) RMD scores the current chunk independently, providing a clean target regardless of the degraded context.

1 INTRODUCTION

Due to high computational and memory costs, video diffusion models are typically trained on short clips (Wan et al., 2025; Huang et al., 2026a). However, many real-world applications demand much longer sequences, including autonomous-driving simulation, interactive and embodied world models, and physics-grounded visual simulation (Zhang et al., 2025; Bruce et al., 2024; Liu et al., 2024). Autoregressive (AR) models naturally support this by extending video generation chunk by chunk (Yin et al., 2025; Huang et al., 2026a). Specifically, the diffusion process for each new chunk conditions on previously generated frames. As the video grows, a sliding context window ensures the model only attends to the most recent history, keeping the generation cost bounded. However, this sequential process introduces an inherent challenge: each generated chunk inevitably contains minor prediction errors, which then become the context for future steps. Consequently, errors barely noticeable in a short sequence can compound and severely degrade visual quality over a long rollout.

To mitigate error accumulation, self-rollout training exposes the generator to its own imperfect predictions (Huang et al., 2026a; Zhu et al., 2026). Since a generated video rarely matches real videos perfectly, no paired ground-truth continuation exists to supervise the next steps. To overcome this lack of paired data, distribution matching distillation (DMD) is often employed (Yin et al., 2024b; 2025). Instead of requiring exact continuations, a bidirectional diffusion teacher evaluates the entire sequence jointly. It provides corrective feedback by matching a fake score tracking the student's output against a real score representing the teacher's high-quality prior.

However, DMD-trained models still struggle to generate very long sequences. We attribute this to a bias introduced by the video teacher's joint supervision. As visual artifacts accumulate over a long autoregressive rollout (e.g., the severe texture and color corruption in the background foliage in Figure 2(a)), the teacher must evaluate the new chunk alongside this flawed history. Because the teacher cannot modify past frames, enforcing a clean, natural background in the current chunk would cause an abrupt visual shift. Consequently, to preserve temporal consistency, the video-level teacher often retains these accumulated errors, which manifest as non-physical ghosting artifacts, in its real-score prediction (Figure 2(b)). The temporal consistency constraint thus fundamentally limits the teacher's ability to correct local errors.

We break this coupling with Rollout-Marginal Distillation (RMD). The generator still uses its history to produce continuations, but the teacher evaluates the generated chunk independently, without checking the history or future chunks. As shown in Figure 2(c), context-free scoring ensures the teacher provides a clean, high-fidelity real score despite the degraded history. RMD achieves this by

matching the student's chunk marginal distribution to a teacher-defined chunk distribution. To provide this supervision, we prepare the RMD teacher by adapting a video diffusion model on individual chunks extracted from videos generated by the original bidirectional model. However, because independent chunk supervision does not enforce cross-chunk relationships, generated videos might exhibit motion jitter between chunks. To resolve this, we subsequently refine the generator with a video-level DMD stage, explicitly smoothing these transitions to promote temporal coherence.

Extensive experiments demonstrate that RMD achieves significantly higher visual quality than video-level DMD baselines beyond the training horizon. Our quantitative evaluation covers rollouts of approximately 60 seconds ($12\times$ the 5-second training horizon). In addition, Figure 1 provides a qualitative example extending to 500 seconds ($100\times$ the training horizon). Generation relies entirely on a fully sliding context window, without fixed anchors such as frame sinks (Yang et al., 2026; Li et al., 2026), additional explicit memory mechanisms, or changes to the generator architecture.

2 Related Work

Few-step autoregressive video generation. Causal video diffusion generates frames or blocks sequentially with KV caching for low-latency streaming. CausVid converts a pretrained bidirectional video diffusion model into a causal generator and extends DMD to few-step video generation (Yin et al., 2025). Self-Forcing further performs autoregressive rollout during training, feeding generated outputs back as context and supervising the sequence with a holistic video objective (Huang et al., 2026a). Causal Forcing instead studies initialization and uses an autoregressive teacher for ODE distillation before applying Self-Forcing DMD (Zhu et al., 2026). Causal-rCM similarly combines teacher-forced consistency training with self-forced DMD refinement (Zheng et al., 2026a). Solaris introduces *Checkpointed Self Forcing*, which decouples serial self-rollout from parallel differentiable replay to reduce training memory (Savva et al., 2026). Together, these methods establish causal students and on-policy training pipelines. We adopt the Solaris replay mechanism directly and focus RMD on the distribution matched during on-policy DMD.

Distribution matching distillation. DMD trains a few-step generator by minimizing a reverse KL divergence to the distribution of a pretrained diffusion model, using the difference between real and fake scores as generator supervision (Yin et al., 2024b). DMD2 removes the costly regression objective and improves stability through a two-time-scale update rule, adversarial learning, and multi-step training (Yin et al., 2024a). Subsequent work has combined distribution matching with adversarial learning (Lu et al., 2025), replaced KL-based matching with an optimal-transport objective (Wang et al., 2026), or coupled reverse-KL refinement with score-regularized consistency training (Zheng et al., 2026b). These developments primarily modify the divergence, regularization, or optimization of distribution matching. By contrast, RMD retains the DMD score-difference update but changes its output space from a joint rollout to the chunk marginal induced by autoregressive sampling.

Long rollouts and causal supervision. LongLive improves extended AR generation through streaming long-video tuning, a finite attention window, and a persistent frame sink (Yang et al., 2026). Rolling Sink studies cache maintenance beyond the training horizon and introduces a training-free cache-maintenance strategy (Li et al., 2026). More directly related to training-time supervision, OPSD-V post-trains few-step AR generators on their inference-time trajectories while supplying the teacher with a cleaner cache that can incorporate real long-video context (Liu et al., 2026). Most closely related to our analysis, Context-Matched Distillation (CMD) observes that a bidirectional full-clip teacher can use future information unavailable to a causal student (Bandyopadhyay et al., 2026). CMD addresses this issue with a causal teacher and prefix scoring under the student's realized rollout context. RMD takes a different route. It does not construct a teacher conditional for each self-rollout history. Instead, it uses a bidirectional diffusion prior as a chunk-distribution target, scores the rollout chunks independently, and subsequently restores joint temporal structure through video-level refinement. It therefore requires neither paired continuations for generated histories nor real long-video context.

3 ROLLOUT-MARGINAL DISTILLATION

In this section, we present Rollout-Marginal Distillation (RMD), an approach designed to disentangle visual quality supervision from temporal context in autoregressive video generation. We first review video-level distribution matching distillation (DMD) and formalize how its joint scoring mechanism inherently couples the current chunk's gradient to past and future frames (Section 3.1). To break this coupling, we introduce our rollout-marginal objective, which independently penalizes unrealistic chunks without conditioning on the imperfect context (Section 3.2). We then detail the training procedure in practice (Section 3.3). Finally, because independent scoring ignores cross-chunk relationships, we describe a subsequent video-level refinement stage that explicitly restores temporal coherence (Section 3.4).

3.1 REVISITING VIDEO-LEVEL DISTRIBUTION MATCHING

To understand the necessity of rollout-marginal supervision, we first examine how video-level DMD improves a causal video generator and why its joint scoring mechanism introduces a critical flaw.

Given a text prompt c, let the causal generator G_θ produce a latent rollout $\mathbf{x} = (x_1, \ldots, x_T)$, where each x_i is a chunk. The distribution of the generated video $\mathbf{x}$ can be naturally factorized as:

$$q_\theta(\mathbf{x} \mid c) = \prod_{i=1}^{T} q_\theta(x_i \mid x_{<i}, c). \tag{1}$$

Following the self-rollout framework (Huang et al., 2026a; Zhu et al., 2026), the generator conditions each continuation on its own generated history $x_{<i}$, exposing it to the prediction errors it will encounter at inference. Because there are no ground-truth continuations paired with these imperfect generated histories, video-level distribution matching distillation (DMD) is employed to supervise the entire student rollout jointly (Yin et al., 2024b;a; 2025).

At noise level τ, let $\mathbf{x}_\tau = \alpha_\tau \mathbf{x} + \sigma_\tau \boldsymbol{\epsilon}$ denote a noised rollout. The video-level objective minimizes the reverse KL divergence between the noised student video distribution $q_{\theta,\tau}$ and a pretrained teacher video distribution p_τ:

$$\mathcal{L}_{\text{video}}(\theta) = \mathbb{E}_{c,\tau}\Big[D_{\text{KL}}\big(q_{\theta,\tau} \,\|\, p_\tau\big)\Big]. \tag{2}$$

The generator is updated using the difference between two diffusion scores:

$$\nabla_\theta \mathcal{L}_{\text{video}} \approx \mathbb{E}\Big[w(\tau) J_\theta(\mathbf{x})^\top \big(s_{\text{fake}}(\mathbf{x}_\tau, \tau, c) - s_{\text{real}}(\mathbf{x}_\tau, \tau, c)\big)\Big], \tag{3}$$

where $J_\theta(\mathbf{x}) = \partial \mathbf{x} / \partial \theta$ is the generator Jacobian, $w(\tau)$ is a noise-level weight, s_{real} is a frozen teacher score representing the target distribution, and s_{fake} is an online score model trained to track the student's distribution.

The Coupling Dilemma. Equation 3 reveals a fundamental issue in applying video-level DMD to autoregressive rollouts. Both the real and fake score networks (s_{real} and s_{fake}) take the entire sequence $\mathbf{x}_\tau$ as input. Consequently, the correction signal assigned to any specific chunk x_i is coupled with its generated past ($x_{<i}$) and future ($x_{>i}$). If an artifact appears in the generated history, a joint score evaluation will likely penalize a sudden correction to natural colors in the current chunk, as such a transition breaks the temporal consistency of the sequence. The model is thereby forced to compromise local visual realism to accommodate an imperfect temporal context. To resolve this, we must disentangle the quality supervision of x_i from the rest of the rollout.

3.2 ROLLOUT-MARGINAL OBJECTIVE

Video-level DMD scores the entire sequence, tying each chunk's quality correction to its compatibility with the surrounding content. RMD breaks this dependency by matching chunks from the causal rollouts independently, while the generator still uses its history for prediction.

We model the initial and continuation chunks separately to account for their distinct latent statistics. Specifically, under the video VAE, the first latent frame encodes only the initial image without

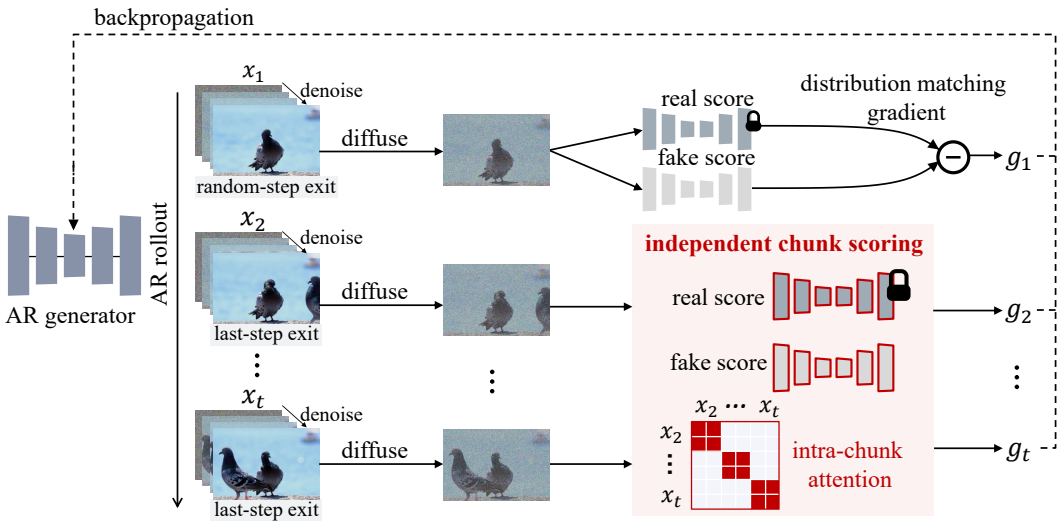


Figure 3: **Overview of Rollout-Marginal Training.** The AR generator produces a causal rollout, sampling the initial chunk (x_1) at a random-step exit and continuation chunks ($x_2, \ldots, x_T$) at the last-step exit. Adapted score networks evaluate them independently without temporal context. The resulting gradients ($g_1, \ldots, g_T$) are backpropagated via differentiable causal replay.

temporal compression (Wan et al., 2025; Huang et al., 2026a). This gives it completely different statistical properties compared to subsequent frames. Let $q_{\theta,1}(x_1 \mid c)$ denote the initial chunk distribution. To define the continuation *rollout marginal*, we pool the chunks $x_2, \ldots, x_T$ across the student's causal rollouts for prompt c:

$$q^{(T)}_{\theta,\text{chunk}}(x \mid c) = \frac{1}{T-1} \sum_{i=2}^{T} \mathbb{E}_{x_{<i} \sim q_\theta(x_{<i}|c)} [q_\theta(x \mid x_{<i}, c)] \,. \tag{4}$$

This distribution averages over both the generated histories and the continuation positions. Sampling from it requires the causal rollout, but evaluating a chunk's realism requires neither its history nor its rollout position.

At noise level τ, let $p_{1,\tau}$ and $p_{\text{chunk},\tau}$ denote the teacher-defined target distributions for the initial frame and continuation chunks, respectively. Weighting these two groups by their frequency in the rollout yields the RMD objective:

$$\mathcal{L}_{\text{RMD}}(\theta) = \mathbb{E}_{c,\tau} \left[\frac{1}{T} D_{\text{KL}}\big(q_{\theta,1,\tau} \,\|\, p_{1,\tau}\big) + \frac{T-1}{T} D_{\text{KL}}\big(q^{(T)}_{\theta,\text{chunk},\tau} \,\|\, p_{\text{chunk},\tau}\big) \right] . \tag{5}$$

Crucially, the generation process remains causal, meaning each sample x_i is still produced conditionally from its actual on-policy history $x_{<i}$. However, the supervision applied to each chunk is entirely context-free. This objective simply penalizes generated chunks that are unlikely under the teacher's chunk distribution. By doing so, it explicitly isolates visual quality supervision from the imperfect generated history. Instead of forcing the model to repeat past artifacts to maintain temporal consistency, this marginal matching ensures the teacher guides the current chunk toward high-quality targets, regardless of the flaws in the preceding frames.

3.3 Rollout-Marginal Training

To realize this decoupled supervision in practice, our training pipeline (Figure 3) optimizes the rollout-marginal objective by first generating the causal rollout, then evaluating the initial and continuation chunks against their distinct target distributions, and finally backpropagating the isolated gradients to the generator.

We initialize the AR generators from Causal Forcing's causal ODE checkpoints (Zhu et al., 2026). Following the self-forcing framework (Huang et al., 2026a), at each training iteration, we generate a

causal rollout of chunks $x_1, \ldots, x_T$, sampling each chunk in 4 denoising steps. To retain the temporal dynamics prior in the pretrained model, we carefully choose which denoising step to supervise for each chunk (Wu et al., 2026). Early steps in the sampling trajectory rely heavily on historical context to establish global motion, and applying independent supervision at these timesteps would disrupt the temporal consistency. Therefore, we adopt an asymmetric supervision strategy: for the initial chunk x_1, we compute the loss at a randomly selected intermediate exit step along the trajectory, whereas for the continuation chunks $x_2, \ldots, x_T$, we apply the loss only at the final, lowest-noise step of the rollout where the network refines local visual details.

To score the sampled chunks, we use separate teacher targets for the initial frame and continuations. As discussed in Section 3.2, Wan's causal VAE gives them different latent statistics. An unmodified video teacher would treat an isolated continuation chunk as the start of a video and expect an uncompressed first frame. We therefore adapt the Wan 14B video teacher (Wan et al., 2025) with LoRA (Hu et al., 2022) on continuation chunks extracted from videos generated by the original model. The adapted teacher supplies the continuation target $p_{\mathrm{chunk},\tau}$ in Equation 5, allowing each continuation to be scored without its generated history.

For each x_i, we add scoring noise and obtain clean predictions from its teacher and fake-score networks on the same noisy input. Let $D_{\mathrm{real},i}$ and $D_{\mathrm{fake},i}$ denote these two predictions. Their normalized difference gives a gradient estimate g_i for the chunk:

$$g_i = \frac{D_{\mathrm{fake},i} - D_{\mathrm{real},i}}{\mathrm{mean}\,|x_i - D_{\mathrm{real},i}|}. \tag{6}$$

To propagate g_i to the generator parameters, we follow checkpointed self-forcing (Savva et al., 2026). The causal rollout is generated without retaining its computation graph. A differentiable replay then recomputes each supervised prediction from its sampled noisy input and generated history, allowing g_i to be backpropagated through the generator.

The first latent frame has different statistics from later latents. Once it leaves the sliding window, the generator must predict without that distinctive frame, a condition absent from short training rollouts that can cause flicker. Following Self-Forcing (Huang et al., 2026a), we mask attention to the initial block when replaying predictions in the latter half of the training window.

3.4 Refinement with Video-Level DMD

Matching chunk marginals improves their appearance but does not directly constrain the transitions between them. Individually realistic chunks can still form a video with inconsistent appearance or motion. We therefore refine the AR generator with the video-level DMD objective in Equation 2.

The generator still produces causal self-rollouts, but the video teacher and online fake-score network now evaluate each complete rollout jointly. Their score difference updates the generator through the same differentiable replay, supervising motion and consistency across chunks. This stage begins from a generator already trained using the rollout-marginal objective. We use smaller learning rates to limit the changes made during refinement.

4 Experiments

4.1 Experimental Setup

Implementation Details. We distill the pretrained Wan 14B video diffusion model into a causal Wan 1.3B generator, initialized from Causal Forcing's causal ODE checkpoints (Zhu et al., 2026). RMD is trained in two stages using a metadata set comprising 70,000 unique prompts sourced from Self Forcing (Huang et al., 2026a). Our models are trained under a strictly fixed 81-frame horizon, without performing longer rollouts during the training phase.

All trainable networks are optimized using AdamW (Loshchilov & Hutter, 2019) ($\beta_1 = 0$, $\beta_2 = 0.999$, weight decay 0.01) with a global batch size of 8. During the rollout-marginal training stage, we alternate 5 fake-score updates for every 1 generator update. When the chunk size is 1, both the generator and fake-score networks use a learning rate of 2×10^{-5}, and we train for 900 steps. When the chunk size is 3, the generator learning rate remains 2×10^{-5}, while the fake-score learning rates

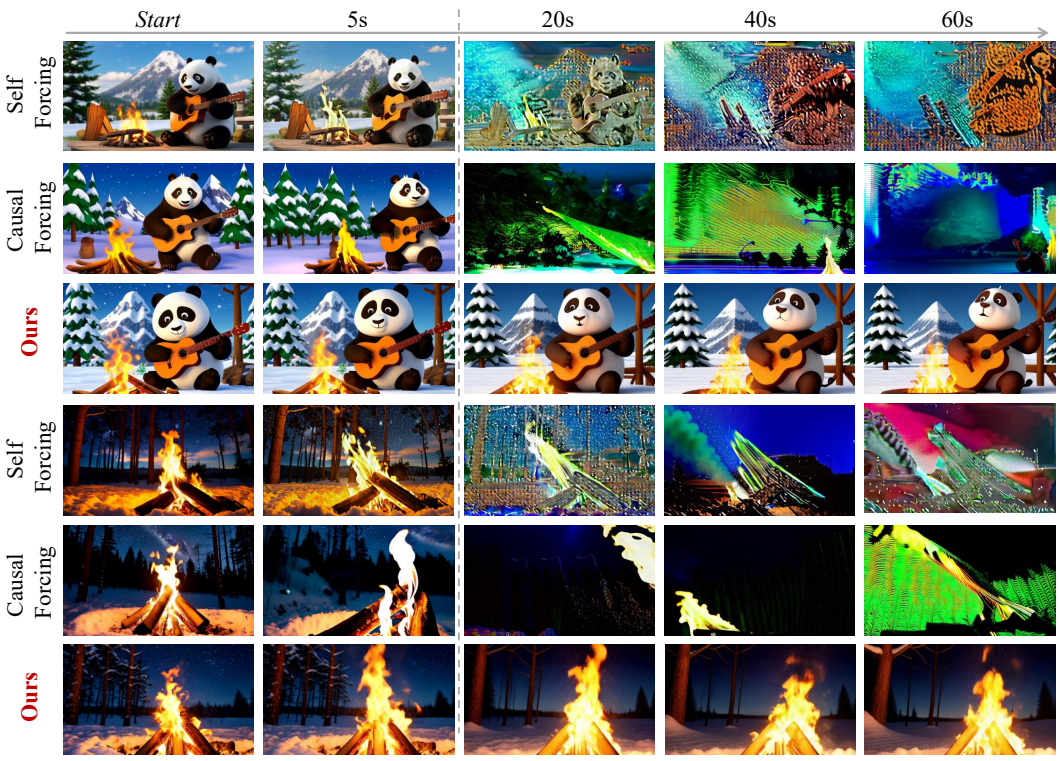


Figure 4: **Qualitative comparison of long-horizon generation.** While baseline methods suffer from severe color saturation and structural collapse beyond their training horizon, RMD consistently preserves sharp subject details and stable scene structures.

are reduced to 2×10^{-7}, and we train for 750 steps. In the subsequent video-level refinement stage, the generator and fake-score learning rates are reduced to 2×10^{-6} and 4×10^{-7}, respectively, and both models are trained for 800 steps.

Evaluation Protocol. We adopt the VBench-Long suite (Huang et al., 2024; 2026b) for our primary evaluation, using the complete standard set of 944 unique prompts with five generations per prompt. All videos are generated at 832×480 resolution and 16 FPS, and evaluated at a duration of approximately 60 seconds for both chunk sizes. The 500-second rollout in Figure 1 is a qualitative example and is not included in these aggregate metrics.

4.2 Comparison with Baselines

We compare RMD with Self Forcing (Huang et al., 2026a) and Causal Forcing (Zhu et al., 2026). Both baselines share our standard sliding-window generation protocol, operating entirely without fixed frame anchors or modifications to the underlying autoregressive logic. This isolates the distillation objectives and ensures RMD introduces zero additional computational overhead during inference. We report results separately using chunk sizes of 1 and 3.

Figure 4 presents corresponding qualitative comparisons over extended rollouts. Within the 5-second training horizon, all methods generate reasonable frames. However, as generation progresses far beyond this boundary (20s, 40s, 60s), the baselines exhibit visual degradation including color saturation and structural collapse. In contrast, RMD consistently preserves clear subject details and stable scene structures throughout the 1-minute trajectory.

Table 1 presents the quantitative comparison on 1-minute generated videos. RMD establishes a clear advantage across both configurations. With a chunk size of 1, our method improves the total score by over 10 absolute points compared with the baselines (81.26 vs. 70.94 and 69.88), while attaining quality and semantic scores of 84.52 and 68.24, respectively. With a chunk size of 3, RMD achieves

Table 1: **Quantitative comparison on long-horizon generation.** Evaluated on VBench-Long, RMD consistently outperforms video-level distillation baselines across all metrics.

Chunk size	Method	Total ↑	Quality ↑	Semantic ↑
1	Self Forcing	70.94	77.89	43.17
	Causal Forcing	69.88	78.34	36.03
	RMD (ours)	**81.26**	**84.52**	**68.24**
3	Self Forcing	77.44	81.99	59.22
	Causal Forcing	74.88	80.53	52.27
	RMD (ours)	**81.48**	**85.03**	**67.24**

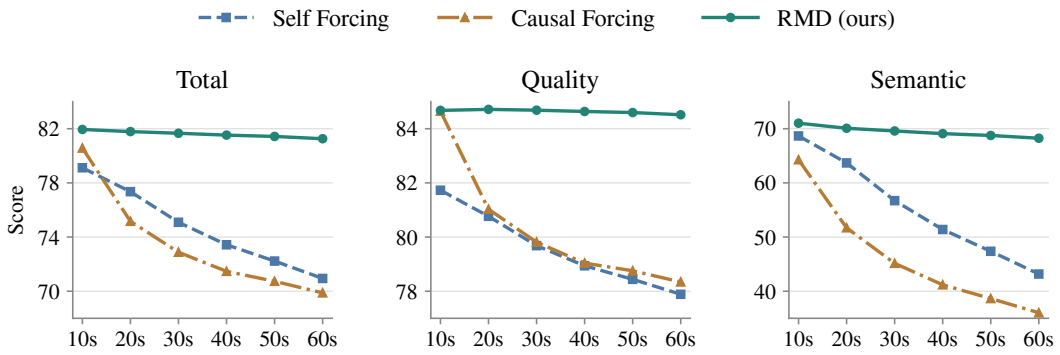


Figure 5: **Performance over increasing rollout lengths.** VBench-Long scores are evaluated at durations ranging from 10 to 60 seconds. As generation extends, RMD maintains stable visual quality and significantly mitigates the performance degradation suffered by baseline methods.

a total score of 81.48, with corresponding quality and semantic scores of 85.03 and 67.24. Together, these quantitative and qualitative results demonstrate that explicitly decoupling chunk supervision from historical context effectively prevents the degradation typical of joint video-level distillation.

4.3 Long-Horizon Extrapolation

To assess performance degradation over extended rollouts, we evaluate the same videos generated with a chunk size of 1 at durations ranging from 10 to 60 seconds. Figure 5 plots the performance dynamics over these trajectories. RMD maintains stable visual quality as the rollout extends: its quality score changes only from 84.68 at 10 seconds to 84.52 at 60 seconds, where it retains a semantic score of 68.24. In contrast, the baselines degrade steeply as prediction errors compound; for example, the total score of Self Forcing drops from 79.12 to 70.94 over the same interval. This robust extrapolation confirms that independent chunk supervision effectively mitigates the accumulation of visual artifacts.

4.4 Ablation Studies

Effectiveness of the Rollout-Marginal Objective. Table 2 evaluates the components of RMD using a chunk size of 1. Applying video-level DMD alone yields a total score of 72.44, and qualitatively, the generation rapidly collapses into severe color artifacts (Figure 6). Optimizing exclusively for the rollout-marginal objective (*Marginal only*) raises the quality score to 84.90, confirming that context-free scoring effectively prevents the visual degradation.

Effectiveness of Refinement with Video-Level DMD. While the *Marginal only* variant achieves high visual realism, it lacks cross-chunk constraints and falls short on temporal coherence, which is visibly manifested as jagged flickering in its spatio-temporal slice (Figure 6). The full RMD pipeline resolves this via subsequent refinement with video-level DMD, which improves subject consistency, background consistency, and motion smoothness (Table 2(b)). As shown in Figure 6, this combined

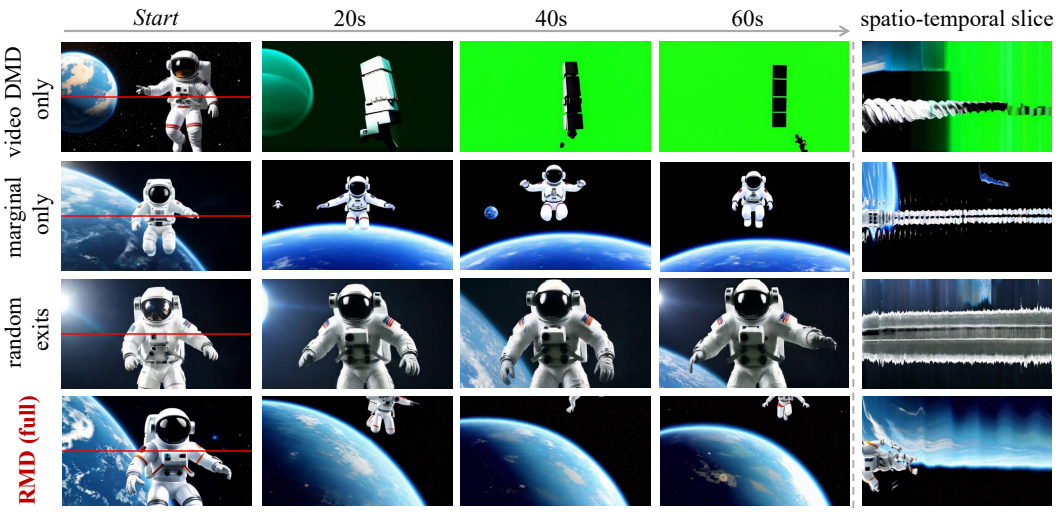


Figure 6: **Qualitative ablation results.** The spatio-temporal slice is extracted along the red line. Removing the rollout-marginal objective (*video DMD only*) causes severe color artifacts. Removing video-level refinement (*marginal only*) introduces temporal flickering (jagged slice artifacts). Removing asymmetric denoising exits (*random exits*) destroys the temporal prior, resulting in severe temporal jitter. Full *RMD* achieves both high visual quality and temporal coherence.

Table 2: **Quantitative ablation results.** (a) Removing refinement with video DMD (*marginal only*) or asymmetric exits (*random exits*) degrades overall performance. (b) Temporal metrics confirm that this refinement is essential for temporal coherence.

(a) Aggregate scores

Variant	Total ↑	Quality ↑	Semantic ↑
video DMD only	72.44	78.68	47.50
marginal only	80.64	**84.90**	63.59
random exits	78.29	82.48	61.56
RMD (full)	**81.26**	84.52	**68.24**

(b) Temporal metrics

Metric	Marginal only	RMD (full)
Subject Cons. ↑	97.29	**98.05**
Background Cons. ↑	96.26	**96.89**
Temporal Flickering ↑	99.06	**99.41**
Motion Smoothness ↑	98.79	**98.96**

approach achieves the best total and semantic scores (81.26 and 68.24, respectively) while restoring smoother, more continuous trajectories in the spatio-temporal slice.

Effectiveness of Asymmetric Denoising Exits. A variant that applies random-step exits uniformly across all chunks suffers from severe temporal jitter (Figure 6) and performs worse across all metrics, with total and quality scores of 78.29 and 82.48, respectively (Table 2(a)). This supports our asymmetric design: supervising the initial chunk across random steps preserves the pretrained temporal prior, while restricting continuation chunks to the final exit strictly refines local details.

5 Conclusion

Autoregressive video generation inherently suffers from error accumulation because joint temporal scoring forces the teacher to compromise local visual quality for sequence-level consistency. To resolve this coupling, we introduce Rollout-Marginal Distillation (RMD), which disentangles spatial realism from temporal coherence. By independently matching chunk marginals and subsequently refining global dynamics, RMD provides a clean, context-free supervision signal. Quantitative evaluations over 60-second rollouts demonstrate that RMD effectively prevents artifact accumulation beyond the training horizon, while a 500-second qualitative example illustrates its potential at substantially longer durations. We believe this decoupled paradigm offers a robust foundation for future long-horizon, streamable video generation.

AI USE STATEMENT

In accordance with ICLR guidelines, we declare the use of Large Language Models (LLMs) during the preparation of this manuscript. Specifically, AI tools were utilized to aid in text polishing and grammar correction to improve the readability of the manuscript. Additionally, AI-assisted search tools were employed for literature retrieval and discovery during the initial scoping of related work. The authors have meticulously reviewed and verified all AI-generated suggestions, and take full responsibility for the final content, originality, and scientific integrity of this paper.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the full reproducibility of the Rollout-Marginal Distillation (RMD) framework. All quantitative evaluations are conducted on standard public benchmarks (e.g., VBench-Long). Upon acceptance, the full source code, pre-trained model weights, and inference scripts will be made publicly open-source to facilitate future research.

ETHICS STATEMENT

The advancement of high-fidelity, long-horizon video generation presents significant benefits for creative industries, physical simulation, and virtual reality. However, we acknowledge the potential dual-use risks associated with our technology, particularly concerning the generation of deceptive content, deepfakes, or misinformation. While our current research focuses on the foundational algorithmic improvements of autoregressive video generation, we strongly advocate for the concurrent development and application of robust generative detection strategies, such as invisible watermarking and cryptographic provenance tracking, prior to any wide-scale deployment. Furthermore, all training data utilized in this study were derived from standard, publicly available academic datasets, adhering strictly to established data usage and privacy protocols.

REFERENCES

Hmrishav Bandyopadhyay, Xuanchi Ren, Zijian Huang, Jay Zhangjie Wu, Tianshi Cao, Ruilong Li, Bryan Chu, Sanja Fidler, Yi-Zhe Song, and Zian Wang. Context-matched distillation: Teacher causality for autoregressive video distillation. *arXiv preprint arXiv:2608.13391*, 2026.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C.Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=bJbSbJskOS.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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

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

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang,

Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and Versatile Benchmark Suite for Video Generative Models . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 48(03):3268–3285, March 2026b. ISSN 1939-3539. doi: 10.1109/TPAMI.2025.3633890. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2025.3633890.

Haodong Li, Shaoteng Liu, Zhe Lin, and Manmohan Chandraker. Rolling sink: Bridging limited-horizon training and open-ended testing in autoregressive video diffusion. *arXiv preprint arXiv:2602.07775*, 2026.

Hongyu Liu, Chun Wang, Feng Gao, Xuanhua He, Yue Ma, Ziyu Wan, Yong Zhang, Xiaoming Wei, and Qifeng Chen. Opsd-v: On-policy self-distillation for post-training few-step autoregressive video generators. *arXiv preprint arXiv:2607.08766*, 2026.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pp. 360–378. Springer, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Yanzuo Lu, Yuxi Ren, Xin Xia, Shanchuan Lin, Xing Wang, Xuefeng Xiao, Andy J Ma, Xiaohua Xie, and Jian-Huang Lai. Adversarial distribution matching for diffusion distillation towards efficient image and video synthesis. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16818–16829. IEEE, 2025.

Georgy Savva, Oscar Michel, Daohan Lu, Suppakit Waiwitlikhit, Timothy Meehan, Dhairya Mishra, Srivats Poddar, Jack Lu, and Saining Xie. Solaris: Building a multiplayer video world model in minecraft. *arXiv preprint arXiv:2602.22208*, 2026.

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.

Yutong Wang, Haiyu Zhang, Tianfan Xue, Yu Qiao, Yaohui Wang, Chang Xu, and Xinyuan Chen. Vdot: Efficient unified video creation via optimal transport distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9273–9283, 2026.

Tianhe Wu, Ruibin Li, Lei Zhang, and Kede Ma. Diversity-preserved distribution matching distillation for fast visual synthesis. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=R3JzLU2qZj.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Ying-Cong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=nCAODkpsPJ.

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

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

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

Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, Xun Cao, and Wei Yin. Epona: Autoregressive diffusion world model for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Kaiwen Zheng, Guande He, Min Zhao, Jintao Zhang, Huayu Chen, Jianfei Chen, Chen-Hsuan Lin, Ming-Yu Liu, Jun Zhu, and Qianli Ma. Causal-rcm: A unified teacher-forcing and self-forcing open recipe for autoregressive diffusion distillation in streaming video generation and interactive world models. *arXiv preprint arXiv:2606.25473*, 2026a.

Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=2uNlM353RI.

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. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=BYInOck3gr.