LONGFORCING: SCALING TEMPORAL MANIFOLD MATCHING FOR LONG-HORIZON WORLD ROLLOUT

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive video diffusion models have emerged as a foundation for long-horizon world rollout. However, when extended far beyond their training windows, these models suffer from progressive drift: visual quality deteriorates, motion becomes inconsistent, and object identity gradually collapses as generated frames are recursively reused as context. We study a *supervision-horizon gap*: short-horizon distillation leaves later rollout states without direct supervision. We show that matching short-window marginals need not determine long-horizon behavior, while longer on-policy trajectories expose delayed failures to training. This motivates *temporal manifold matching at scale* for long-horizon world models. Building on this insight, we introduce **LongForcing**, a distillation approach that scales teacher supervision to long student self-rollouts. LongForcing preserves the short-horizon teacher-forcing and causal ODE-distillation stages, while replacing the real-score teacher in the final distribution-matching stage with an extended-horizon bidirectional teacher. The long-horizon teacher supervises extended student self-rollouts using the standard distribution-matching formulation, without modifying the student architecture or increasing inference cost. We evaluate LongForcing on text-to-video generation and interactive action-conditioned world models, achieving a VBench Quality score of 85.52 and improving WorldRoamBench 60s action score by 23.9 points in first-person ablations. Qualitative stress tests further sample interactive rollouts up to **24 hours**, far beyond the training horizon. These results suggest that scaling supervision along the temporal axis is a practical step toward long-horizon world models.

1 INTRODUCTION

"We can only see a short distance ahead, but we can see plenty there that needs to be done."
— ALAN TURING

Autoregressive (AR) video diffusion models are becoming a foundational paradigm for long-horizon world rollout. An AR model generates video chunk by chunk, with each new chunk conditioned on previously generated frames. Repeating this process enables rollouts far longer than individual training clips. This capability is essential for both interactive world models (Bruce et al., 2024; Google DeepMind, 2025; Valevski et al., 2025; Decart et al., 2024; Zhang et al., 2025; Sun et al., 2026), where a world must remain coherent and responsive as the user acts, and long-form text-to-video systems (Kim et al., 2024; Huang et al., 2025a), where visual quality, motion, and identity must persist far beyond the initial clip. However, the same recursive mechanism that enables long rollouts also makes AR models vulnerable to *rollout drift*. Because every generated chunk becomes context for the next, small errors compound, the conditioning distribution moves away from the data seen during training, and the model gradually loses visual fidelity.

Modern real-time AR video generators are often obtained through a three-stage distillation pipeline. Teacher forcing first adapts a bidirectional video model to causal conditioning, ODE distillation compresses its denoising trajectory into a few-step generator, and distribution matching distillation (DMD) aligns the student's self-generated distribution with a teacher (Yin et al., 2025; Huang et al., 2025a; Zhu et al., 2026a). This recipe is highly effective within the few-second horizon on which it is trained, but deployment requires closed-loop generation over minutes. Teacher forcing learns from

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

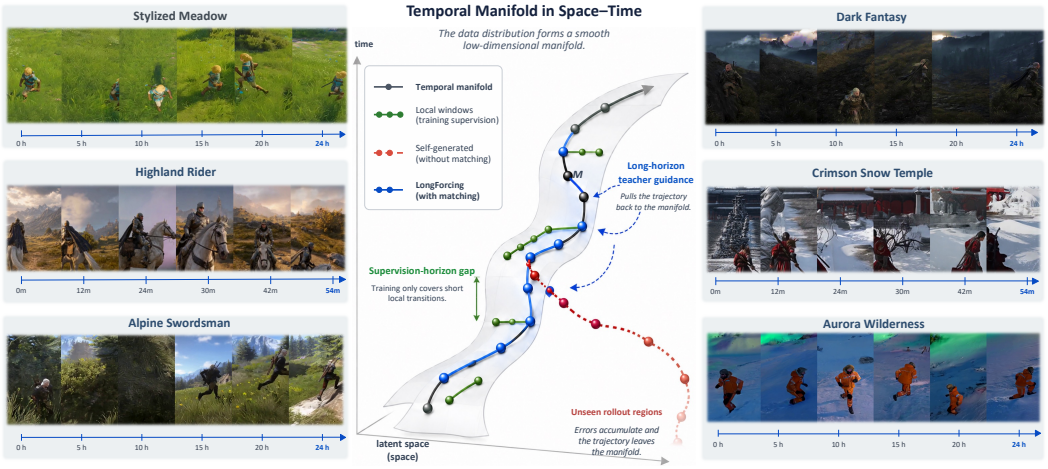

Figure 1: LongForcing extends teacher guidance over student rollouts. Qualitative snapshots from the 5B interactive model span stress-test rollouts sampled up to **24 hours**.

short clean histories, causal ODE distillation uses short teacher-generated trajectories, and short-horizon DMD evaluates only short student rollouts; later rollout states receive no direct supervision from this objective.

Recent approaches mitigate long-horizon drift through sparse local matching (Cui et al., 2026a), inference-time memory and positional modifications (Li et al., 2026; Yi et al., 2026; Cui et al., 2026b), or context perturbation (Zhu et al., 2026b). Local matching can supervise late rollout states, but its jointly scored windows remain short; inference-time modifications instead stabilize the conditioning context. Neither mechanism by itself matches the longer joint path, leaving cross-window dependencies that can affect visual fidelity, identity, dynamics, and action consistency.

These failures reveal a *supervision-horizon gap* between short-window training and long-horizon deployment. Our principle is that **temporal manifold matching should scale with the trajectories visited on policy**, where temporal manifold informally describes high-probability regions of the teacher's path distribution. Section 4 formalizes this view using path distributions, KL divergence, and total variation. Matching short windows does not determine the distribution of a long rollout, whereas matching clean trajectories controls drift within the supervised horizon. Beyond this horizon, stability also depends on teacher leakage, student–teacher mismatch along visited states, and recovery from drifted states. This motivates extending distribution matching over longer student rollouts while keeping teacher forcing and ODE distillation short.

Guided by this principle, we introduce **LongForcing**, which replaces Causal Forcing's DMD real-score teacher with an extended-horizon teacher for longer self-generated trajectories. The student rolls out on policy, while an extended-horizon teacher supplies distribution-level supervision across a long, contiguous portion of each self-generated training trajectory, requiring no architectural changes. Empirically, LongForcing improves text-to-video VBench Quality by 0.93 points over Causal Forcing and raises WorldRoamBench 60s action score by 23.9 points (37% relative) in first-person ablations. Qualitative stress tests sample interactive rollouts up to **24 hours**. These results position temporal manifold matching at scale as a practical step toward long-horizon world models.

Our contributions are threefold:

- We formalize the *supervision-horizon gap*: longer student rollouts expose delayed failures, and joint path matching constrains dependencies left undetermined by local-window marginals.
- We introduce **LongForcing**, which extends the teacher's supervision horizon in distribution-matching distillation to guide longer student-generated trajectories.
- We evaluate LongForcing across interactive world models and text-to-video generation, measuring improvements in action following and video quality alongside their dependence on supervision length.

2 Related Work

2.1 Autoregressive Long-Video Generation

Autoregressive video generation has evolved from discrete-token transformers (Wu et al., 2021; Hong et al., 2023; Kondratyuk et al., 2024; Yan et al., 2021) to diffusion and flow models that stream through temporally varying noise levels, causal attention, or sliding-window denoising (Chen et al., 2024; Xie et al., 2025; Sun et al., 2025; Sand.ai et al., 2025; Kodaira et al., 2026). Few-step variants further use video DMD, student-generated contexts, progressive-window denoising, corrected causal ODE initialization, or diagonal distillation (Yin et al., 2025; Huang et al., 2025a; Liu et al., 2026b; Zhu et al., 2026a; Liu et al., 2026a). To extend generation beyond the training horizon, training-free methods modify noise schedules, denoising windows, attention, positional encodings, or temporal memory (Qiu et al., 2024; Kim et al., 2024; Yi et al., 2026; Cui et al., 2026b; Zhao et al., 2026; Kim et al., 2026), whereas training-based approaches use self-generated segments, long-context tuning, hierarchical denoising, and drift simulation (Cui et al., 2026a; Yang et al., 2026; Chen et al., 2026; Zou et al., 2026; Yuan et al., 2026). Context Forcing (Chen et al., 2026) also uses a long-context teacher, combining continuation-conditioned DMD with Slow-Fast Memory. LongForcing retains Causal Forcing's student architecture and inference procedure, replacing its DMD teacher to score a longer contiguous, jointly noised student path.

2.2 Interactive World Models

Interactive world models use video generation as a controllable simulator for prediction, exploration, and closed-loop interaction. Recent systems demonstrate playable game simulation (Valevski et al., 2025; Decart et al., 2024; Alonso et al., 2024), camera and action-controlled world exploration (Zhang et al., 2025; Li et al., 2025; Sun et al., 2026), and real-time interactive generation (Google DeepMind, 2025; Tencent Hunyuan, 2025; Shin et al., 2026). Unlike prompt-driven video generation, interactive rollout must preserve both world state and the consequences of continuously changing user actions. Over extended sessions, accumulated errors can therefore degrade not only visual fidelity and identity, but also action responsiveness and physical consistency. WorldRoamBench (Xu et al., 2026) explicitly evaluates these long-horizon properties across diverse interactive environments. LongForcing scales teacher supervision of on-policy rollouts along the temporal axis in both interactive and text-conditioned generation. Qualitative stress tests sample interactive rollouts up to **24 hours**, complementing the benchmark evaluation.

3 Drift in Autoregressive Video Rollouts

We first review autoregressive video diffusion and its standard distillation pipeline, then connect the empirical manifestations of long-video drift to the supervision coverage of closed-loop rollout.

3.1 Autoregressive Video Diffusion and Distillation

Video diffusion. Following standard practice, video generation takes place in the latent space of a pretrained spatiotemporal VAE. Let $\mathbf{z}$ denote a clean video latent. At noise level $s \in [0, 1]$, the forward process constructs (Lipman et al., 2023; Zhu et al., 2026a) $\mathbf{z}^{(s)} = \alpha_s \mathbf{z} + \sigma_s \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the superscript (s) denotes the diffusion noise level. A score or velocity network learns the reverse dynamics, whose probability-flow ODE maps noise to a clean latent.

Autoregressive rollout. Partition a latent video into K causal chunks $\mathbf{z}_{1:K}$, where each chunk is generated from the preceding chunks and condition $\mathbf{c}$. The resulting path distribution factorizes as $p_\theta(\mathbf{z}_{1:K} \mid \mathbf{c}) = \prod_{k=1}^{K} p_\theta(\mathbf{z}_k \mid \mathbf{z}_{<k}, \mathbf{c})$, where $\mathbf{c}$ denotes a text prompt or a prompt augmented with per-chunk actions. At inference, each prediction is appended to the history, enabling streaming while accumulating error.

Few-step distillation and distribution matching. An efficient causal student is typically obtained through teacher forcing on short clean histories, causal ODE distillation into a few-step generator, and DMD refinement (Zhu et al., 2026a; Yin et al., 2024b;a). DMD compares bidirectional real and

fake scores on student-generated closed-loop paths; Self Forcing (Huang et al., 2025a) applies this refinement on policy, but only within a short rollout window.

3.2 Why Autoregressive Rollouts Drift

Long-video drift can appear as position or color shift, amplified blur and noise, and action inconsistency (Yuan et al., 2026). These mechanisms share a temporal structure: generated chunks recursively change the conditioning distribution. The relevant train–test mismatch is therefore between the short prefixes evaluated during training and the evolving on-policy path traversed during long deployment.

Fix a prefix-measurable drift detector $g_t(\mathbf{z}_{1:t}) \in [0, 1]$ and threshold γ, and define the first drift time and within-horizon event

$$\tau_{\text{drift}} = \inf\{t \geq 1 : g_t > \gamma\}, \qquad \mathcal{B}_H(\gamma) = \{\tau_{\text{drift}} \leq H\}. \tag{1}$$

For a fixed rollout distribution Q and $H_1 < H_2$, lengthening the path exposes exactly the delayed failure mass

$$\Pr_Q(\tau_{\text{drift}} \leq H_2) - \Pr_Q(\tau_{\text{drift}} \leq H_1) = \Pr_Q(H_1 < \tau_{\text{drift}} \leq H_2). \tag{2}$$

Thus an objective computed only from H_1-step prefixes cannot evaluate failures whose first hitting time exceeds H_1. A longer rollout changes which delayed states can receive a teacher signal; whether optimization corrects them is addressed in Section 4.

Figure 2 reveals structured late-rollout shift in the shared causal-VAE latent space: the channel-mean deviation grows beyond the short supervision horizon, while the signed heat map shows a consistent channel-wise direction. Appendix C.2 gives the exact reference window and aggregation procedure.

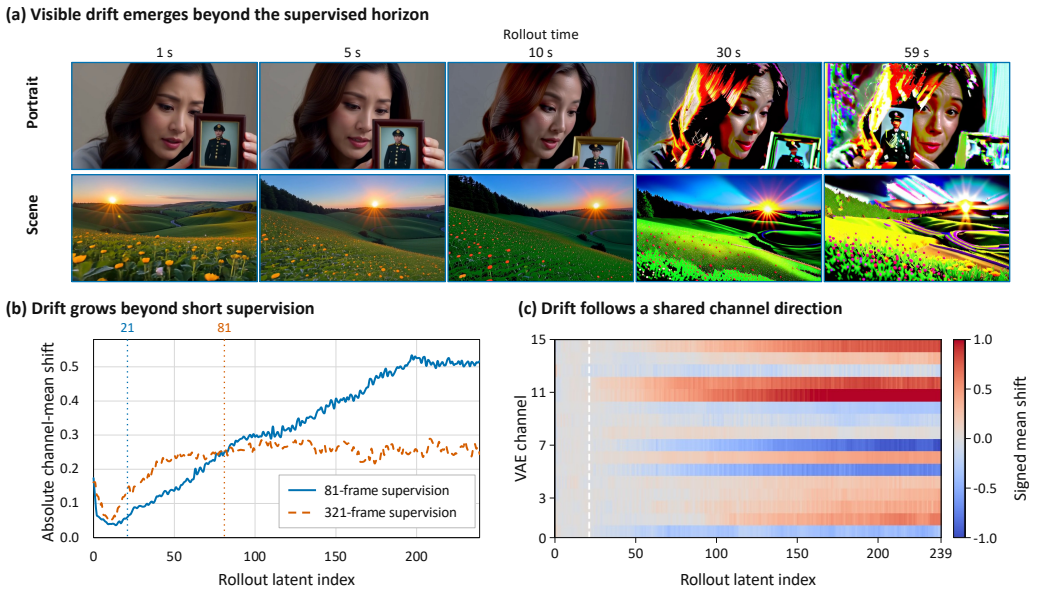

Figure 2: Visual and feature-space signatures of rollout drift. **(a)** Shared timestamps reveal late over-saturation and high-frequency artifacts. **(b)** Median absolute channel-mean shift from the early reference; dotted lines mark supervision horizons. **(c)** Median signed per-channel shift for the 81-frame model; persistent bands indicate a channel-wise direction shared across 20 videos.

4 Finite-Horizon Drift Control

We separate three questions behind LongForcing: which failures can be detected within a finite training window, what matching trajectories within that window guarantees, and under what conditions those guarantees extend to later blocks. Throughout, H denotes a horizon of H autoregressive

chunks; experiments report the corresponding number of video frames. Appendix B gives the complete setup, proofs, and scope conditions.

4.1 The Supervision-Horizon Gap

Let S_t denote the student's causal state after chunk t, and let $\mathcal{C}_H = (S_0, \mathbf{a}_{1:H}, \mathbf{c})$ collect the initial state, an action sequence fixed over the rollout, and persistent conditions such as the text prompt. For this shared condition, $Q_{\theta,H}$ denotes the causal student's H-chunk path distribution and $Q^{*,\mathrm{long}}_{\phi,H}$ the extended-horizon teacher's joint reference. The distinction between matching local windows and matching this joint path is fundamental.

Proposition 4.1 (Short windows do not identify the long path)**.** *For every $W < T$, two causal path distributions can agree on every contiguous W-step marginal while having disjoint T-step supports. Hence even perfect local-window matching need not determine the long trajectory distribution.*

Proof in Appendix B.2. Proposition 4.1 is an identifiability statement: local-window marginals leave cross-window coupling undetermined. Equation (2) gives a separate exposure statement. For a fixed rollout law, an H-chunk prefix objective cannot evaluate failures first reached after H; extending that rollout exposes exactly the delayed-failure mass in the added suffix. This exposure makes those states available for supervision, but does not imply monotonic improvement across models.

4.2 What Path Matching Guarantees

For a shared condition distribution μ_H, let $\overline{Q}_{\theta,H} = \mu_H Q_{\theta,H}$ and $\overline{Q}^{*}_{\phi,H} = \mu_H Q^{*,\mathrm{long}}_{\phi,H}$. The ideal clean path-matching objective is

$$\mathcal{L}^{\mathrm{ideal}}_H = \mathbb{E}_{\mathcal{C}_H \sim \mu_H} D_{\mathrm{KL}}\Big(Q_{\theta,H}(\cdot \mid \mathcal{C}_H) \,\Big\|\, Q^{*,\mathrm{long}}_{\phi,H}(\cdot \mid \mathcal{C}_H) \Big). \tag{3}$$

Theorem 4.2 (Within-horizon drift control)**.** *For every measurable path event $\mathcal{A}_H$,*

$$\Pr_{\overline{Q}_{\theta,H}}(\mathcal{A}_H) \le \Pr_{\overline{Q}^{*}_{\phi,H}}(\mathcal{A}_H) + \sqrt{\mathcal{L}^{\mathrm{ideal}}_H/2}. \tag{4}$$

Proof in Appendix B.3. Taking $\mathcal{A}_H = \mathcal{B}_H(\gamma)$ bounds student drift by teacher drift plus the path-matching slack. Increasing H therefore enlarges the delayed events covered by the guarantee, but does not automatically shrink the slack or control failures after H.

4.3 Teacher-Guided Correction over Finite Depth

For a fixed evaluation stride B, let $\mathcal{D}$ be a measurable degraded region and $\mathcal{G} = \mathcal{D}^c$; when the drift detector is state-measurable, $\mathcal{D}$ can be its thresholded region from Eq. (1). Define $b_j = \Pr(S_{jB} \in \mathcal{D})$, $a_j = \Pr(S_{(j+1)B} \in \mathcal{D} \mid S_{jB} \in \mathcal{G})$, and $\eta_j = \Pr(S_{(j+1)B} \in \mathcal{G} \mid S_{jB} \in \mathcal{D})$. Total probability gives

$$b_{j+1} = (1 - b_j)a_j + (1 - \eta_j)b_j. \tag{5}$$

Let λ^*_j be teacher leakage on the student's on-policy good-start distribution, η^*_j the corresponding teacher return from degraded starts, and $\epsilon^{\mathcal{G}}_j, \epsilon^{\mathcal{D}}_j$ the clean B-step student–teacher path KLs on those starts. Conditional Pinsker yields

$$a_j \le \lambda^*_j + \sqrt{\epsilon^{\mathcal{G}}_j/2}, \qquad \eta_j \ge \Big[\eta^*_j - \sqrt{\epsilon^{\mathcal{D}}_j/2}\Big]_+. \tag{6}$$

Proposition 4.3 (Conditional finite-depth drift control)**.** *Suppose for $0 \le j < J$ that $\lambda^*_j \le \bar{\lambda}$, $\eta^*_j \ge \bar{\eta}^*$, $\epsilon^{\mathcal{G}}_j \le \bar{\epsilon}_{\mathcal{G}}$, and $\epsilon^{\mathcal{D}}_j \le \bar{\epsilon}_{\mathcal{D}}$. Define $A = \min\{1, \bar{\lambda} + \sqrt{\bar{\epsilon}_{\mathcal{G}}/2}\}$ and $\underline{\eta} = [\bar{\eta}^* - \sqrt{\bar{\epsilon}_{\mathcal{D}}/2}]_+ > 0$. Then*

$$b_j \le \min\left\{1, (1-\underline{\eta})^j b_0 + \frac{A}{\underline{\eta}}\left[1 - (1-\underline{\eta})^j\right]\right\}, \quad 0 \le j \le J. \tag{7}$$

Proof in Appendix B.4. Here A controls newly injected degradation, while $\underline{\eta}$ lower-bounds student recovery using the teacher's return probability and the degraded-start path mismatch. Equation (7) controls degraded-state occupancy at these block boundaries and is nontrivial when $A < \underline{\eta}$. These are sufficient conditions, not guarantees supplied automatically by LongForcing; feedback actions and a stronger conditional long-run extension are treated in Appendix B.4–B.5.

5 LONGFORCING

LongForcing builds on Causal Forcing (Zhu et al., 2026a). We retain its short-horizon teacher-forcing and causal ODE-distillation stages, and replace the real-score teacher in the final asymmetric DMD stage with an extended-horizon bidirectional teacher. This teacher supervises longer student self-rollouts using the standard DMD formulation, motivated by the analysis in Section 4.

5.1 LONG-HORIZON TRAJECTORY MATCHING

Objective. With the long-horizon teacher as the reference, we apply the standard noisy DMD objective to an H-chunk student path, extending the short supervision horizon W:

$$\mathcal{J}_H^{\text{noisy}} = \mathbb{E}_{\mathcal{C}_H \sim \mu_H,\, s}\left[D_{\text{KL}}\Big(Q^{(s)}_{\theta,H}(\cdot \mid \mathcal{C}_H) \,\Big\|\, Q^{*,(s)}_{\phi,H}(\cdot \mid \mathcal{C}_H) \Big) \right], \tag{8}$$

where $Q^{(s)}_{\theta,H}$ and $Q^{*,(s)}_{\phi,H}$ are the student and teacher path distributions at shared noise level s. As in asymmetric DMD, the student generates causally and both score networks evaluate its noised path bidirectionally. The longer scoring horizon covers later prefixes; Appendix A.2–A.4 gives the objective's scope and score estimator.

Extended-horizon teacher. The frozen real-score model is trained or fine-tuned on long sequences so that it can score the complete H-step path; the online fake-score model uses the same long bidirectional context to track the current student distribution. For interactive training, a pre-specified action plan is supplied over the same horizon. The extended horizon increases supervision coverage; the guarantees remain conditional on teacher quality and student–teacher mismatch.

Algorithm. Algorithm 1 summarizes the pipeline. In Stage 3, the long-horizon real-score teacher supervises the longer self-generated path.

Algorithm 1 LongForcing training

Require: short horizon W; long horizon H; extended teacher s_{real}

1: Train a causal model by teacher forcing on W-step real clips.
2: Use this causal model as the ODE teacher to initialize G_θ on W-step clips.
3: Initialize the online model s_{fake} from s_{real}.
4: **for** each LongForcing iteration **do**
5: Sample $\mathcal{C}_H$ and roll out $\mathbf{z}_{1:H} \sim Q_{\theta,H}(\cdot \mid \mathcal{C}_H)$ on policy.
6: Sample s and noise the full path to obtain $\mathbf{z}^{(s)}_{1:H}$.
7: Evaluate s_{real} and s_{fake} bidirectionally on $\mathbf{z}^{(s)}_{1:H}$.
8: Update G_θ using the DMD estimator and s_{fake} on their respective schedules (Appendix A.1).
9: **end for**
10: **return** few-step causal student G_θ

6 EXPERIMENTS

6.1 SETUP

Interactive world model. We train action-conditioned models on Wan2.2-5B (Team Wan, 2025b) and Wan2.1-1.3B (Team Wan, 2025a) and evaluate 60-second WorldRoamBench rollouts: third-person for the 5B comparison and first-person for the 1.3B horizon ablation. Teacher forcing and ODE distillation use 121 and 81 frames for the 5B and 1.3B models, respectively, while DMD extends both to 321 frames. DMD training uses AdamW on eight GPUs with a global batch size of 8 for 1000 steps. The generator and fake-score network use constant learning rates of 2×10^{-6} and 4×10^{-7}, respectively, with $(\beta_1, \beta_2) = (0, 0.999)$ and weight decay 0.01.

Text-to-video model. For text-to-video, we follow Causal Forcing (Zhu et al., 2026a) on Wan2.1-1.3B (Team Wan, 2025a) at 832×480 resolution. We retain its teacher-forcing and causal ODE

configurations, replace the DMD real-score teacher, and use 321-frame DMD rollouts. DMD uses AdamW on 8 GPUs with a global batch size of 8 and learning rates of 2×10^{-6} and 4×10^{-7} for the generator and fake-score network. We evaluate with the official 5-second VBench protocol (Huang et al., 2024) and VBench-Long (Huang et al., 2025b) on 30-second generations. See Appendix A.1 for further details.

6.2 Main Results

Interactive world model. Table 1 reports control-rate-weighted scores for 60-second third-person rollouts. Scores incorporate control rate as defined in the table caption. LongForcing-5B achieves weighted action and visual scores of 76.81 and 57.89, the highest among the open-source models shown. Its control rate is 87.88% (87/99), compared with 40.40% for LingBot-World v2.

Table 1: **WorldRoamBench results.** Except Ctrl., higher-is-better scores are multiplied by $p = \mathrm{Ctrl.}/100$; Aesth. Drift is $100[1 - p(1 - d)]$ for normalized drift $d \in [0, 1]$. Best and second-best open-source results are bold and underlined.

Model	Size	WASD Action Interaction				Visual Quality			
		Action ↑	Ctrl. ↑	Partial Acc. ↑	Traj. ↑	Visual ↑	Aesthetic ↑	Aesth. Drift ↓	Imaging ↑
Proprietary Models									
Genie 3 (Google DeepMind, 2025)	–	69.73	79.80	65.75	73.70	56.26	48.11	32.08	45.30
HappyOyster (Alibaba Cloud, 2026)	–	84.02	96.97	80.48	87.57	65.63	59.10	20.61	47.44
Open-Source Models									
LingBot-World (Robbyant Team et al., 2026)	14B	18.81	23.23	18.50	19.12	15.99	14.32	81.70	13.84
HY-World 1.5 (Tencent Hunyuan, 2025)	8B	24.92	29.29	24.57	25.28	19.11	17.52	79.48	16.52
LingBot-World v2 (Gao et al., 2026)	14B	<u>36.09</u>	<u>40.40</u>	<u>35.35</u>	<u>36.84</u>	<u>28.42</u>	<u>25.61</u>	<u>68.75</u>	<u>25.04</u>
LongForcing (ours)	5B	**76.81**	**87.88**	**73.74**	**79.89**	**57.89**	**51.69**	**31.93**	**48.26**

Text-to-video: 5-second VBench. Table 2 reports 5-second VBench results. At matched inference cost, LongForcing leads distilled autoregressive models with 84.56 Total and 85.52 Quality, improving over Causal Forcing by 0.52 and 0.93 points, respectively. These results suggest that longer-horizon supervision also benefits short-clip visual quality, extending its utility beyond long-rollout stability.

Text-to-video: 30-second VBench-Long. Table 3 reports 30-second VBench-Long results, where LongForcing achieves the highest Total and Quality scores among the compared methods, at 82.83 and 83.99, improving over Causal Forcing by 1.78 and 0.55 points, respectively. The full results in Appendix C.1 show the highest background consistency score of 96.58. Semantic alignment also improves from 71.52 to 78.19, complementing the gains in visual quality and background consistency.

Table 2: **5-second VBench.**

Model	Total ↑	Quality ↑	Semantic ↑
CausVid (Yin et al., 2025)	81.33	83.98	70.72
Self Forcing (Huang et al., 2025a)	83.74	84.48	<u>80.77</u>
Causal Forcing (Zhu et al., 2026a)	<u>84.04</u>	<u>84.59</u>	**81.84**
LongForcing (ours)	**84.56**	**85.52**	80.72

Table 3: **30-second VBench-Long.**

Model	Total ↑	Quality ↑	Semantic ↑
LongLive (Yang et al., 2026)	<u>82.80</u>	83.21	**81.17**
Self Forcing (Huang et al., 2025a)	81.49	82.67	76.77
Causal Forcing (Zhu et al., 2026a)	81.05	<u>83.44</u>	71.52
LongForcing (ours)	**82.83**	**83.99**	<u>78.19</u>

6.3 Ablation Studies

Text-to-video supervision horizon. Table 4 compares supervision-horizon configurations from 81 to 321 frames (approximately 5 to 20 seconds). Each DMD rollout length uses its corresponding separately trained teacher checkpoint. The 321-frame setting achieves the highest 5-second total score and 30-second total, quality, and semantic scores.

Table 4: **Supervision-horizon configurations for text-to-video generation.** Horizons are rollout lengths in frames. Best and second-best scores are **bold** and underlined.

Frames	Approx. span	VBench (5s) Total ↑	VBench-Long (30s) ↑			
			Total	Quality	Semantic	Dynamic
81	∼5s	84.36	81.87	83.71	74.54	58.15
161	∼10s	83.76	78.02	81.59	63.77	56.67
241	∼15s	84.38	79.18	83.75	60.87	**94.91**
321	∼20s	**84.56**	**82.83**	**83.99**	**78.19**	60.93

Interactive world-model supervision horizon. In Table 5, extending the DMD supervision horizon from 81 to 321 frames raises action score from 64.39 to 88.26 and strict accuracy from 42.89 to 85.16 on 60-second rollouts, while reducing aesthetic drift from 56.58 to 42.37. The 321-frame setting achieves the best results across all action metrics and most visual metrics, supporting the benefit of longer supervision in exposing accumulated control and consistency errors.

Table 5: **Effect of the DMD supervision horizon on first-person interactive world modeling.** The comparison varies the DMD supervision horizon with a fixed 1.3B backbone under the first-person WorldRoamBench 60s protocol; scores are unweighted. Higher is better except for Aesthetic Drift. The best and second-best results are shown in **bold** and underlined, respectively.

Frames	Approx. span	WASD Action Interaction				Visual Quality			
		Action ↑	Strict Acc. ↑	Partial Acc. ↑	Traj. ↑	Visual ↑	Aesthetic ↑	Aesth. Drift ↓	Imaging ↑
81	∼5s	64.39	42.89	63.01	87.27	45.05	39.81	56.58	46.40
161	∼10s	56.16	43.21	50.00	75.28	47.01	41.06	60.18	**52.29**
241	∼15s	83.39	78.24	85.14	86.78	55.19	49.93	43.30	50.25
321	∼20s	**88.26**	**85.16**	**90.56**	**89.05**	**56.30**	**51.61**	**42.37**	50.84

Time-resolved rollout drift. Figure 3 shows how text-to-video generation deteriorates as rollouts grow longer. Extending supervision from 81 to 321 frames reduces the mean drift score at 59 seconds from 67.46 to 31.91, a relative reduction of 52.7%. The gap between the curves grows over time, indicating that longer supervision slows error accumulation and helps maintain quality during extended generation, although some drift remains. Appendix C.3 provides the metric definition, evaluation protocol, and corresponding interactive results.

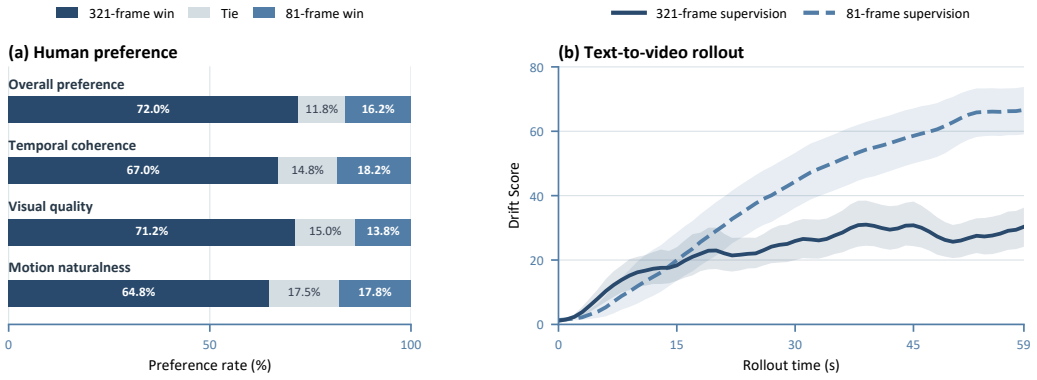

Figure 3: **Rollout drift and human preference across supervision horizons. (a)** Blinded pairwise preferences. **(b)** Mean Drift Score.

Human study. Across 400 judgments per criterion from 20 evaluators on 20 shared cases, 321-frame LongForcing receives 72.0% overall preference, compared with 16.2% for the 81-frame model and 11.8% ties (Figure 3). Preferences also favor longer supervision in visual quality by 71.2% to 13.8%, temporal coherence by 67.0% to 18.2%, and motion naturalness by 64.8% to 17.8%. Appendix C.4 gives the evaluation protocol.

6.4 Qualitative Results

Figure 4A–D shows recognizable scenes and characters across snapshots spanning up to 54 minutes. In the 600-second comparisons (E–F), LingBot-World v2 (14B) (Gao et al., 2026) exhibits over-saturation and repetitive textures (E), and scene and character appearance shifts (F), while LongForcing-5B retains more of the initial scene structure and visual style. These comparisons assess visual consistency at the displayed timestamps.

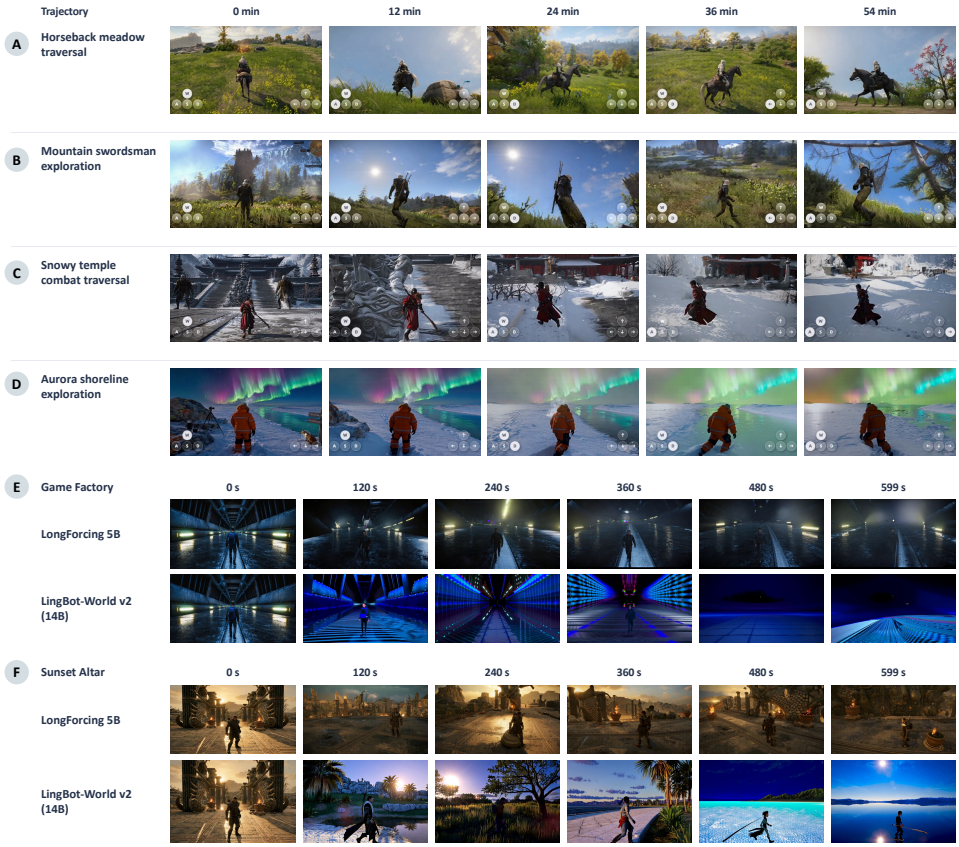

Figure 4: A–D: LongForcing rollouts over up to 54 minutes. E–F: LongForcing-5B (top) and LingBot-World v2 (14B; bottom) at matched timestamps from 0 to 599 seconds.

7 Conclusion

We identify a *supervision-horizon gap* in autoregressive video generation: short-horizon supervision leaves later rollout behavior underdetermined, while longer on-policy trajectories expose delayed failures to teacher guidance. LongForcing addresses this gap by replacing Causal Forcing's DMD real-score teacher with an extended-horizon teacher that supervises longer student self-rollouts, while retaining its short teacher-forcing and causal ODE-distillation stages. Across text-to-video generation and interactive world models, this improves VBench Quality by 0.93 points and first-person action score by 23.9 points over the respective baselines, without changing the architecture or increasing inference cost. Horizon ablations and time-resolved evaluations show reduced rollout drift, complemented by qualitative interactive stress tests sampled up to 24 hours. These findings position temporal supervision as a scaling dimension for video world models. Integrating coherent visual rollouts with persistent memory and hierarchical planning offers a path toward persistent interactive worlds, goal-directed behavior, and long-term reasoning.

AI USE STATEMENT

AI tools assisted with revising prose and theoretical exposition, LaTeX formatting, and scripts used to assemble and visualize analysis outputs. The authors designed the method and experiments, ran the experiments, verified the reported results and mathematical arguments, and take full responsibility for the paper.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, Appendix A describes the training pipeline, teacher initialization, training data, optimization settings, and computational resources. Appendix B provides the formal assumptions and proofs of our theoretical results. Section 6.1 and Appendix C document the evaluation setup, drift metrics, aggregation procedures, and human preference study protocol.

REFERENCES

Alibaba Cloud. Alibaba launches HappyOyster, a world model product for real-time immersive creation and interaction. Alibaba Cloud Blog, 2026.

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in Atari. In *Advances in Neural Information Processing Systems*, volume 37, pp. 58757–58791, 2024. doi: 10.52202/079017-1873.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4603–4623. PMLR, 2024.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion Forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0759.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. SkyReels-V2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

Shuo Chen, Cong Wei, Sun Sun, Tiancheng Shen, Ping Nie, Kai Zou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context Forcing: Consistent autoregressive video generation with long context. In *International Conference on Machine Learning*, 2026.

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

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. LoL: Longer than longer, scaling video generation to hour. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38132–38142, 2026b.

Decart, Julian Quevedo, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. Decart and Etched, 2024.

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

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, et al. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026.

Google DeepMind. Genie 3. Google DeepMind, 2025.

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.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. CogVideo: Large-scale pretraining for text-to-video generation via transformers. In *International Conference on Learning Representations*, 2023.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. doi: 10.52202/085713-5576.

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

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. doi: 10.1109/TPAMI.2025.3633890.

Fan Jiang, Zhaoxu Sun, Mengchao Wang, Ziyu Zhu, Chiyu Wang, Yunpeng Zhang, et al. ABot-World-0: Infinite interactive world rollout on a single desktop GPU. *arXiv preprint arXiv:2607.19191*, 2026.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, et al. Pyramidal flow matching for efficient video generative modeling. In *International Conference on Learning Representations*, 2025.

Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. FIFO-Diffusion: Generating infinite videos from text without training. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2853.

Youngrae Kim, Qixin Hu, C.-C. Jay Kuo, and Peter A. Beerel. MemRoPE: Training-free infinite video generation via evolving memory tokens. In *European Conference on Computer Vision*, 2026.

Akio Kodaira, Tingbo Hou, Ji Hou, Markos Georgopoulos, Felix Juefei-Xu, Masayoshi Tomizuka, and Yue Zhao. StreamDiT: Real-time streaming text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29200–29210, 2026.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. VideoPoet: A large language model for zero-shot video generation. In *International Conference on Machine Learning*, 2024.

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.

Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-GameCraft: High-dynamic interactive game video generation with hybrid history condition. *arXiv preprint arXiv:2506.17201*, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Jinxiu Liu, Xuanming Liu, Kangfu Mei, Yandong Wen, Ming-Hsuan Yang, and Weiyang Liu. Streaming autoregressive video generation via diagonal distillation. In *International Conference on Learning Representations*, 2026a.

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*, 2026b.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15086–15095, 2025.

Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. FreeNoise: Tuning-free longer video diffusion via noise rescheduling. In *International Conference on Learning Representations*, 2024.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

Gareth O. Roberts and Jeffrey S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004. doi: 10.1214/154957804100000024.

Sand.ai, Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, W. Q. Zhang, et al. MAGI-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

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

Mingzhen Sun, Weining Wang, Gen Li, Jiawei Liu, Jiahui Sun, Wanquan Feng, Shanshan Lao, Siyu Zhou, Qian He, and Jing Liu. AR-Diffusion: Asynchronous video generation with auto-regressive diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7364–7373, 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. In *International Conference on Machine Learning*, 2026.

Team Wan. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025a.

Team Wan. Wan2.2: Open and advanced large-scale video generative models. GitHub repository, 2025b.

Tencent Hunyuan. HY-World 1.5: A systematic framework for interactive world modeling with real-time latency and geometric consistency. Technical report, Tencent Hunyuan, 2025.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations*, 2025.

Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. GODIVA: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021.

Desai Xie, Zhan Xu, Yicong Hong, Hao Tan, Difan Liu, Feng Liu, Arie Kaufman, and Yang Zhou. Progressive autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 6376–6386, 2025.

Ting-Bing Xu, Jiacheng Sui, Zhe Gao, Kewei Shi, Wenjin Yang, Zhicheng Liu, Zhaoxu Sun, Mingchao Sun, Hongyu Pan, Fan Jiang, et al. WorldRoamBench: An open-world benchmark for long-horizon stability of interactive world models. *arXiv preprint arXiv:2606.31672*, 2026.

Zhucun Xue, Jiangning Zhang, Teng Hu, Haoyang He, Yinan Chen, Yuxuan Cai, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, et al. UltraVideo: High-quality UHD video dataset with comprehensive captions. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5445.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. VideoGPT: Video generation using VQ-VAE and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. LongLive: Real-time interactive long video generation. In *International Conference on Learning Representations*, 2026.

Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep Forcing: Training-free long video generation with deep sink and participative compression. In *International Conference on Machine Learning*, 2026.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems*, volume 37, pp. 47455–47487, 2024a. doi: 10.52202/079017-1505.

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

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

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

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*, 2025.

Zengqun Zhao, Yanzuo Lu, Ziquan Liu, Jifei Song, Jiankang Deng, and Ioannis Patras. Relax Forcing: Relaxed KV-memory for consistent long video generation. In *British Machine Vision Conference*, 2026. Accepted for publication.

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 *International Conference on Machine Learning*, 2026a. arXiv:2602.02214.

Yixuan Zhu, Jiaqi Feng, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Jiwen Lu, and Jie Zhou. Astra: General interactive world model with autoregressive denoising. In *International Conference on Learning Representations*, 2026b.

Kai Zou, Dian Zheng, Hongbo Liu, Tiankai Hang, Bin Liu, and Nenghai Yu. HiAR: Efficient autoregressive long video generation via hierarchical denoising. In *European Conference on Computer Vision*, volume 17008 of *Lecture Notes in Computer Science*, pp. 225–242. Springer, 2026. doi: 10.1007/978-3-032-37026-6_13.

A Implementation Details

A.1 Training Pipeline

LongForcing retains the teacher-forcing and causal ODE-distillation stages of Causal Forcing and replaces the real-score teacher in its final DMD stage with an extended-horizon bidirectional teacher. Teacher forcing uses short clean histories and causal ODE distillation uses short trajectories from the causal teacher; the resulting student generates a contiguous H-chunk closed-loop path, scored jointly by the replacement teacher and the online fake-score network. The student architecture, sampler, and inference procedure remain unchanged. Under the causal VAE mapping, 81, 161, 241, and 321 decoded RGB frames correspond to 21, 41, 61, and 81 temporal latent positions, respectively; these positions are rounded and grouped into autoregressive chunks of three latent frames.

Interactive teacher: initialization and data. The interactive teachers use the pretrained Wan2.2-5B and Wan2.1-1.3B backbones, adapted to action-conditioned video generation on the ABot-World training data (Jiang et al., 2026). We train the interactive teachers for 20,000 steps with a learning rate of 10^{-5} and a global batch size of 256 video clips. We use action-annotated video sequences, with the action sequence aligned to the video over the teacher's temporal window. The teacher retains bidirectional temporal attention so that it can evaluate the complete student-generated path with its corresponding actions. The data collection and annotation pipeline is described in ABot-World (Jiang et al., 2026); the long teacher provides the frozen real-score network for our 321-frame DMD stage.

Text-to-video teacher: long-video adaptation. We initialize the bidirectional teacher from the pretrained Wan2.1-T2V-1.3B checkpoint (Team Wan, 2025a) and fine-tune it on UltraVideo-Long (Xue et al., 2025) for three epochs, with each training example visited once per epoch. Videos are sampled at 16 FPS and processed at 832×480 resolution, using 321-frame clips for the long teacher. The horizon study uses separate teacher checkpoints for the 81-, 161-, 241-, and 321-frame DMD configurations, pairing each rollout length with its corresponding teacher. The pretrained Wan VAE and UMT5 text encoder remain frozen; video latents and text embeddings are cached before optimization. We optimize a supervised flow-matching loss on the noisy video latents, sampling a diffusion timestep and Gaussian noise anew for each example and applying the scheduler's timestep-dependent loss weight. Adaptation uses rank-64 LoRA with scaling parameter $\alpha = 64$ in the DiT attention projections (`q,k,v,o`) and feed-forward projections (`ffn.0,ffn.2`); the underlying pretrained weights remain frozen. AdamW uses a base learning rate of 10^{-4}, weight decay 0.01, and its default momentum parameters $(0.9, 0.999)$. The implementation uses BF16 model parameters, evaluates the mean-squared-error loss in FP32, enables gradient checkpointing, and processes one clip per GPU without gradient accumulation, giving a global batch size of 32 clips. After adaptation, the resulting teacher is frozen during DMD; only the causal generator and the online fake-score network are updated in that stage.

Resources and DMD configuration. We follow Causal Forcing (Zhu et al., 2026a) for DMD timestep sampling and loss weighting. LongForcing extends the real-score teacher's supervision horizon and the student rollout length. Interactive teacher training uses 256 GPUs, while text-to-video teacher training uses 32 NVIDIA H20 GPUs. The subsequent DMD stage runs for 1,000 global steps on eight GPUs with a global batch size of eight trajectories in both text-to-video and interactive experiments. The generator and fake-score learning rates are 2×10^{-6} and 4×10^{-7}, respectively, with AdamW momentum parameters $(0, 0.999)$ and weight decay 0.01. For text-to-video DMD training, the fake-score network is updated at every global step, whereas the generator is updated once every five global steps, giving a $5{:}1$ ratio of fake-score to generator updates. For both the 1.3B and 5B interactive world models, the fake-score network is updated at every global step and the generator once every three global steps during DMD training.

A.2 Ideal Objective and Practical Scope

The ideal clean objective assigns weight only to prefixes reached inside the H-step training rollout. Let $q_{\theta,t}$ and $q^*_{\phi,t}$ denote the student and teacher chain-rule conditionals induced by their respective

joint path distributions. Then

$$\mathcal{L}_H^{\text{ideal}} = \sum_{t=1}^{H} \mathbb{E}_{\substack{\mathcal{C}_H \sim \mu_H, \\ \mathbf{z}_{<t} \sim Q_{\theta,H}}} D_{\text{KL}}\big(q_{\theta,t}(\cdot \mid \mathbf{z}_{<t}, \mathcal{C}_H) \,\big\|\, q^*_{\phi,t}(\cdot \mid \mathbf{z}_{<t}, \mathcal{C}_H)\big)\,. \tag{A1}$$

The expectation is under the student's on-policy prefix distribution; the teacher conditional is derived from its joint distribution and need not be implemented as a causal model. Increasing H gives delayed prefixes in Eq. (2) nonzero weight, whereas prefixes first reached after H remain outside the objective.

Equation (3) describes clean-path matching, while Eq. (8) defines its noisy-path counterpart used to motivate DMD. Appendix A.4 states the conditions under which the score-difference estimator is unbiased for the noisy objective. The clean-path bounds in Section 4 apply when their stated KL conditions hold; extensions to later blocks use the conditions in Appendix B.4. For the interactive world model, training uses an action plan specified before the matched path and supplied to the teacher over the same horizon; extending the theory to state-feedback actions requires a joint state–action reference distribution.

Long-horizon DMD increases training memory and compute because the student unrolls farther and both score networks process a longer path. At fixed spatial resolution, dense bidirectional attention has quadratic attention-score computation in temporal sequence length; total memory and runtime also depend on the implementation. For a fixed output length and inference configuration, Long-Forcing retains the student architecture and sampling procedure, adding no inference operations.

A.3 Comparison of Long-Horizon Rollout Strategies

Figure A1 contrasts the supervision coverage and inference mechanisms of representative long-horizon rollout strategies. The comparison distinguishes where teacher supervision is applied from mechanisms that modify inference-time conditioning.

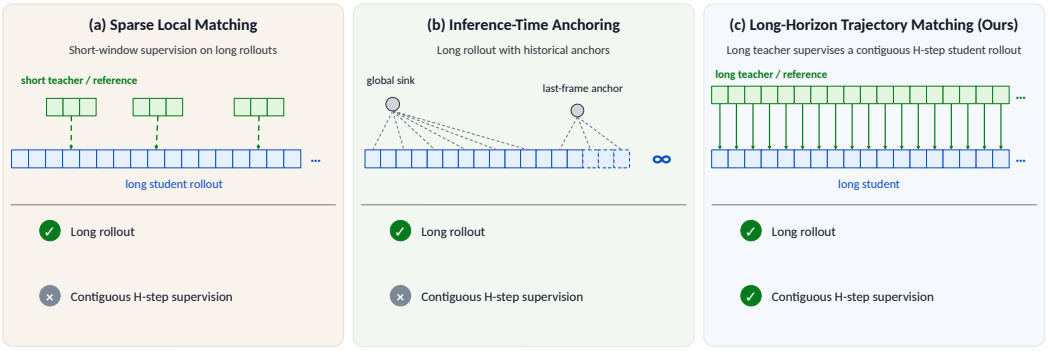

Figure A1: Mechanistic comparison of sparse local matching, inference-time anchoring, and Long-Forcing's contiguous H-step on-policy trajectory matching under an extended-horizon teacher.

A.4 DMD Score-Difference Estimator

For a shared condition $\mathcal{C}_H$, let $\mathbf{z} = G_\theta(\boldsymbol{\xi}; \mathcal{C}_H)$ be a clean student path and $\mathbf{y} = \alpha_s \mathbf{z} + \sigma_s \boldsymbol{\epsilon}$ its noised version. With s_{real} and s_{fake} denoting estimates of the teacher and student noisy scores, the unweighted, full-Jacobian score-difference estimator is

$$\widehat{\mathbf{g}}_{\text{DMD}} = \alpha_s \left(\frac{\partial \mathbf{z}}{\partial \theta}\right)^{\top} \text{sg}\big[s_{\text{fake}}(\mathbf{y}, s) - s_{\text{real}}(\mathbf{y}, s)\big]\,. \tag{A2}$$

Equation (A2) presents the ideal, unweighted estimator; our implementation follows the sampling and weighting scheme of Causal Forcing (Zhu et al., 2026a). Here sg treats the score difference as fixed during the generator update; condition arguments are suppressed. The factor α_s follows from differentiating the forward-noising map. For a condition and noise-level sampling law independent of θ, exact scores, and full rollout differentiation, the expectation equals $\nabla_\theta \mathcal{J}_H^{\text{noisy}}$, assuming

differentiable positive noisy densities and valid interchange of differentiation and integration. In particular, the expected explicit derivative of the student log-density vanishes by normalization, leaving the pathwise score difference. Stop-gradient implements this expression without differentiating the score networks; score estimation, reweighting, normalization, or rollout-gradient truncation can introduce approximations to this ideal update. Both score networks attend bidirectionally over the sampled path: the real-score network is frozen during DMD, whereas the fake-score network is updated online to track the student's rollout distribution.

B Theoretical Details and Proofs

B.1 Formal Path Setup

All state, action, and latent spaces are taken to be standard Borel spaces, so the conditional path laws below exist. We use natural logarithms in KL and $d_{\mathrm{TV}}(P,Q) = \sup_A |P(A) - Q(A)|$. Let $S_t = \Gamma(S_{t-1}, \mathbf{z}_t)$ be the student's causal state after chunk t, including any finite context or cache required for the next prediction; fixed conditions and model parameters are suppressed in Γ. All event comparisons use the same measurable path-to-state map and detector for student and teacher paths, with the model parameters held fixed. For matched horizon H, define $\mathcal{C}_H = (S_0, \mathbf{a}_{1:H}, \mathbf{c})$, where $\mathbf{a}_{1:H}$ is fixed before the window and $\mathbf{c}$ contains persistent conditions. The causal student induces

$$Q_{\theta,H}(\mathbf{z}_{1:H} \mid \mathcal{C}_H) = \prod_{t=1}^{H} p_\theta(\mathbf{z}_t \mid S_{t-1}, \mathbf{a}_{1:t}, \mathbf{c}), \tag{B1}$$

whereas the bidirectional teacher defines the joint reference $Q^{*,\mathrm{long}}_{\phi,H}(\mathbf{z}_{1:H} \mid \mathcal{C}_H)$ without requiring a causal implementation. Fixing the actions open-loop prevents conditioning on actions generated from the same path; feedback policies instead require a joint state–action reference law.

For $\mathcal{C}_H \sim \mu_H$, the joint measures are $\overline{Q}_{\theta,H} = \mu_H Q_{\theta,H}$ and $\overline{Q}^*_{\phi,H} = \mu_H Q^{*,\mathrm{long}}_{\phi,H}$. Because they share μ_H, Eq. (3) is equivalently $D_{\mathrm{KL}}(\overline{Q}_{\theta,H} \| \overline{Q}^*_{\phi,H})$. For the exposure comparison, both events are evaluated on the same rollout law with the same initial-condition distribution and fixed detector. Identifying a shorter event with its extension to the longer path space gives $\mathcal{B}_{H_1}(\gamma) \subseteq \mathcal{B}_{H_2}(\gamma)$ for $H_1 < H_2$. This construction does not require independently trained teachers at different horizons to have consistent marginals.

B.2 Non-Identifiability and Delayed Exposure

Proof of Proposition 4.1. For $b \in \{0,1\}$, define the uniform binary path law

$$Q_T^{(b)} = \mathrm{Unif}\left\{ x_{1:T} \in \{0,1\}^T : \sum_{t=1}^{T} x_t \equiv b \pmod 2 \right\}. \tag{B2}$$

Thus $Q_T^{(0)}$ assigns mass to even-parity paths and $Q_T^{(1)}$ to odd-parity paths. Both admit causal realizations: sample the first $T-1$ bits independently from $\mathrm{Bernoulli}(1/2)$, then set the last bit to enforce the prescribed parity. For any coordinate set I with $|I| = k < T$, fixing x_I leaves exactly 2^{T-k-1} completions of either parity. Each full-path support has size 2^{T-1}, hence

$$Q_T^{(0)}(x_I) = Q_T^{(1)}(x_I) = 2^{-k}. \tag{B3}$$

In particular, every contiguous $W < T$ window has the same uniform marginal under both laws. Their full-path supports are disjoint, giving $d_{\mathrm{TV}}(Q_T^{(0)}, Q_T^{(1)}) = 1$. □

Concrete example: three steps and two-step windows. For $T = 3$ and $W = 2$, the two laws are

$$\begin{aligned} Q_3^{(0)} &= \mathrm{Unif}\{000, 011, 101, 110\}, \\ Q_3^{(1)} &= \mathrm{Unif}\{001, 010, 100, 111\}. \end{aligned} \tag{B4}$$

Under either law, both windows (x_1, x_2) and (x_2, x_3) assign probability $1/4$ to each of $00, 01, 10, 11$. A discriminator restricted to either window therefore cannot distinguish the laws,

while the full-path event $\{x_1 + x_2 + x_3 \text{ is even}\}$ has probability one under $Q_3^{(0)}$ and zero under $Q_3^{(1)}$. The missing information is the coupling across windows: local agreement alone leaves a global path constraint undetermined.

The proposition concerns objectives defined by local-window marginals; exact matching of every causal conditional on its full support would determine the joint path. For a measurable region $\mathcal{R}$, let $\tau_{\mathcal{R}} = \inf\{t : S_t \in \mathcal{R}\}$ and $h_{\mathcal{R}}(H) = \Pr(\tau_{\mathcal{R}} \le H)$. With N independent horizon-H rollouts,

$$\Pr(\mathcal{R} \text{ is never exposed}) = (1 - h_{\mathcal{R}}(H))^N. \tag{B5}$$

If $h_{\mathcal{R}}(H_1) = 0$, more horizon-H_1 samples cannot expose $\mathcal{R}$; sequential depth cannot be replaced by sample count. This is an exposure statement for a fixed rollout distribution, not a monotonic improvement claim across separately trained students.

B.3 Within-Horizon Path Control

Proof of Theorem 4.2. Conditionally on $\mathcal{C}_H$, Pinsker's inequality gives

$$d_{\mathrm{TV}}\Big(Q_{\theta,H}, Q^{*,\mathrm{long}}_{\phi,H}\Big) \le \sqrt{\tfrac{1}{2} D_{\mathrm{KL}}\Big(Q_{\theta,H} \| Q^{*,\mathrm{long}}_{\phi,H}\Big)}. \tag{B6}$$

Average over μ_H and apply Jensen's inequality to obtain $d_{\mathrm{TV}}(\overline{Q}_{\theta,H}, \overline{Q}^*_{\phi,H}) \le \sqrt{\mathcal{L}_H^{\mathrm{ideal}}/2}$. The event bound follows from the definition of total variation. □

The same argument controls any measurable $F_H \in [0,1]$:

$$\left|\mathbb{E}_{\overline{Q}_{\theta,H}} F_H - \mathbb{E}_{\overline{Q}^*_{\phi,H}} F_H\right| \le \sqrt{\mathcal{L}_H^{\mathrm{ideal}}/2}. \tag{B7}$$

Taking $F_H = H^{-1} \sum_{t=1}^{H} \mathbf{1}\{S_t \in \mathcal{R}\}$ yields the finite-horizon occupancy bound without assuming ergodicity. Indeed, $F_H = \int_0^1 \mathbf{1}\{F_H > u\}\, du$; integrating the event bound for $\{F_H > u\}$ proves the bounded-functional inequality above. The next subsection extends the analysis to later block boundaries under explicit leakage and recovery conditions.

B.4 Finite-Depth Drift Control

For an open-loop action block, define the complete condition

$$X_j = (S_0, \mathbf{z}_{1:jB}, \mathbf{a}_{1:(j+1)B}, \mathbf{c}), \qquad \nu_j = \mathrm{Law}_{Q_\theta}(X_j), \tag{B8}$$

where the future suffix $\mathbf{a}_{jB+1:(j+1)B}$ is fixed before the block. If the block is a conditional of a matched H-step joint path, use $X_j = (\mathcal{C}_H, \mathbf{z}_{1:jB})$ instead, retaining the entire teacher condition, including actions after the block. The full prefix may be replaced only by an augmented state sufficient for both student and teacher path laws. Let $\rho_j^{\mathcal{G}} = \nu_j(\cdot \mid S_{jB} \in \mathcal{G})$ and $\rho_j^{\mathcal{D}} = \nu_j(\cdot \mid S_{jB} \in \mathcal{D})$ when the corresponding events have positive mass; otherwise their conditional values are immaterial in Eq. (5).

For each start x, the teacher reference is a specified measurable continuation kernel conditioned on that same prefix and action block. Within a matched joint path, it is the corresponding chain-rule conditional; a continuation outside that horizon requires a separately specified reference. Let λ_j^* be the teacher's average terminal probability of $\mathcal{D}$ from $\rho_j^{\mathcal{G}}$ and η_j^* its average terminal probability of $\mathcal{G}$ from $\rho_j^{\mathcal{D}}$. Define

$$\epsilon_j^{\mathcal{G}} = \mathbb{E}_{x \sim \rho_j^{\mathcal{G}}} D_{\mathrm{KL}}\Big(Q_{\theta,B}(\cdot \mid x) \,\Big\|\, Q^{*,\mathrm{long}}_{\phi,B}(\cdot \mid x)\Big), \tag{B9}$$

and define $\epsilon_j^{\mathcal{D}}$ analogously under $\rho_j^{\mathcal{D}}$. The finite-KL assumptions entail student-to-teacher absolute continuity almost everywhere under the respective start distribution. Write P_j^x and R_j^x for these student and teacher block laws, and $K_j(x) = D_{\mathrm{KL}}(P_j^x \| R_j^x)$. For either terminal region $E \in \{\mathcal{G}, \mathcal{D}\}$, its probability difference is at most $d_{\mathrm{TV}}(P_j^x, R_j^x) \le \sqrt{K_j(x)/2}$, since the terminal state is

a measurable function of the start and block path. Averaging under the appropriate conditional start law and applying Jensen's inequality gives

$$\begin{aligned} a_j - \lambda_j^* &\le \mathbb{E}_{\rho_j^{\mathcal{G}}} \sqrt{K_j/2} \le \sqrt{\epsilon_j^{\mathcal{G}}/2}, \\ \eta_j^* - \eta_j &\le \mathbb{E}_{\rho_j^{\mathcal{D}}} \sqrt{K_j/2} \le \sqrt{\epsilon_j^{\mathcal{D}}/2}. \end{aligned} \tag{B10}$$

Combining the second inequality with $\eta_j \ge 0$ proves Eq. (6).

Relation to the path objective. Let $\mathcal{L}_{B,j} = \mathbb{E}_{\nu_j} K_j$ denote the block-specific on-policy KL. Splitting its expectation by the start region gives

$$\mathcal{L}_{B,j} = (1 - b_j)\epsilon_j^{\mathcal{G}} + b_j \epsilon_j^{\mathcal{D}}. \tag{B11}$$

For $0 < b_j < 1$, nonnegativity therefore implies $\epsilon_j^{\mathcal{G}} \le \mathcal{L}_{B,j}/(1 - b_j)$ and $\epsilon_j^{\mathcal{D}} \le \mathcal{L}_{B,j}/b_j$; a zero-mass term contributes zero. When all blocks use the chain-rule conditionals of the same matched joint laws and $H = JB$, the KL chain rule gives $\mathcal{L}_H^{\text{ideal}} = \sum_{j=0}^{J-1} \mathcal{L}_{B,j}$. For $JB < H$, the remaining suffix contributes an additional nonnegative conditional KL. If instead a block loss is averaged under a different training start law $\widetilde{\mu}_j$, transfer to ρ_j^E requires a measure comparison: whenever $\rho_j^E \ll \widetilde{\mu}_j$ and $d\rho_j^E/d\widetilde{\mu}_j \le M_j^E$, integration gives $\epsilon_j^E \le M_j^E \mathbb{E}_{\widetilde{\mu}_j} K_j$. These relations control conditional averages; the finite-depth proposition explicitly assumes the required good- and degraded-start bounds at each evaluated block.

Proof of Proposition 4.3. Under the proposition's assumptions, Eq. (6) gives $a_j \le A$ and $\eta_j \ge \underline{\eta}$. Together with Eq. (5), this implies $b_{j+1} \le A(1 - b_j) + (1 - \underline{\eta})b_j \le A + (1 - \underline{\eta})b_j$. Probability bounds are taken in $[0, 1]$, so $r = 1 - \underline{\eta} \in [0, 1)$. Induction yields $b_j \le r^j b_0 + A \sum_{k=0}^{j-1} r^k = r^j b_0 + A(1 - r^j)/\underline{\eta}$ for $1 \le j \le J$. Together with the case $j = 0$ and $b_j \le 1$, this proves Eq. (7). The monotone relaxation avoids any sign requirement on $1 - A - \underline{\eta}$. □

Partial start coverage. Coverage is measured under the relevant conditional start law. If the student's per-start leakage is at most A and its per-start recovery is at least $\underline{\eta}$ only on subsets with masses at least $1 - \delta_j^{\mathcal{G}}$ under $\rho_j^{\mathcal{G}}$ and $1 - \delta_j^{\mathcal{D}}$ under $\rho_j^{\mathcal{D}}$, respectively, the averaged bounds become

$$a_j \le A + (1 - A)\delta_j^{\mathcal{G}}, \qquad \eta_j \ge (1 - \delta_j^{\mathcal{D}})\underline{\eta}. \tag{B12}$$

To see this, bound leakage on uncovered good starts by one and recovery on uncovered degraded starts by zero, then average over the two subsets. These adjusted quantities enter the recursion at each block. With complete degraded-start coverage and $\delta_j^{\mathcal{G}} \le \delta$, each affine update acquires at most $(1-A)\delta$; summing its geometrically weighted contributions gives additional envelope slack at block j of at most $(1 - A)\delta[1 - (1 - \underline{\eta})^j]/\underline{\eta}$.

Feedback actions. The total-probability identity in Eq. (5) also holds for feedback policies, without Markov or stationarity assumptions. For its KL-based bounds, specify a policy $\pi_t(d\mathbf{a}_t \mid h_{t-1})$ and a nonanticipating teacher continuation $q_t^*(d\mathbf{z}_t \mid h_{t-1}, \mathbf{a}_t)$, where h_{t-1} contains the available history and policy state. From the same block start, define joint action–latent laws

$$\begin{aligned} P_j^\pi &= \prod_{t=jB+1}^{(j+1)B} \pi_t(d\mathbf{a}_t \mid h_{t-1})\, p_\theta(d\mathbf{z}_t \mid h_{t-1}, \mathbf{a}_t), \\ R_j^\pi &= \prod_{t=jB+1}^{(j+1)B} \pi_t(d\mathbf{a}_t \mid h_{t-1})\, q_t^*(d\mathbf{z}_t \mid h_{t-1}, \mathbf{a}_t). \end{aligned} \tag{B13}$$

Under finite KL, the shared policy factors cancel in the likelihood ratio, and the chain rule expresses $D_{\mathrm{KL}}(P_j^\pi \| R_j^\pi)$ as the sum of conditional latent KLs averaged over student histories and actions. Applying the preceding Pinsker argument to these joint laws proves the feedback version with its own KL, leakage, and recovery quantities. This construction requires a specified nonanticipating reference; conditioning a bidirectional open-loop teacher on actions realized through feedback does not supply that reference automatically. Throughout, b_j measures occupancy at boundaries spaced by the fixed stride B; within-block visits are covered by path events or the bounded functionals in Appendix B.3.

B.5 Conditional Long-Run Extension

Fix a persistent condition $\mathbf{c}$ and a time-homogeneous stationary policy π, and let Y_j include the generator cache and policy state sufficient for a time-homogeneous Markov boundary process. Suppose its initial law is supported on a measurable class $\mathcal{R}_{c,\pi}$, with $P^{c,\pi}_{\theta,B}(y, \mathcal{R}_{c,\pi}) = 1$ for every y in that class, and its block kernel satisfies the Doeblin minorization $P^{c,\pi}_{\theta,B}(y, \cdot) \geq \beta_{c,\pi} m_{c,\pi}(\cdot)$ there, with $\beta_{c,\pi} \in (0, 1]$ and $m_{c,\pi}$ a probability measure supported on the class. The regions $\mathcal{G}$ and $\mathcal{D}$ are lifted to Y_j through its visual-state component. Then the process for this fixed condition and policy has a unique stationary distribution $\nu^{c,\pi}_\infty$, and its empirical degraded-boundary fraction converges almost surely to $\nu^{c,\pi}_\infty(\mathcal{D})$. If, pointwise on $\mathcal{R}_{c,\pi}$, good states leak with probability at most $A_{c,\pi}$ and degraded states return with probability at least $\eta_{c,\pi} > 0$, then

$$\nu^{c,\pi}_\infty(\mathcal{D}) \leq \frac{A_{c,\pi}}{A_{c,\pi} + \eta_{c,\pi}}. \tag{B14}$$

Proof. Suppress c, π and write $P = P^{c,\pi}_{\theta,B}$. For $0 < \beta < 1$, the residual kernel $R = (P - \beta m)/(1 - \beta)$ is a probability kernel on the closed class. Thus $P = \beta m + (1 - \beta)R$, and

$$\nu_\infty = \beta \sum_{k=0}^{\infty} (1 - \beta)^k m R^k \tag{B15}$$

is a probability measure satisfying $\nu_\infty P = \nu_\infty$ by shifting the series. Since Markov kernels contract total variation,

$$d_{\mathrm{TV}}(\nu P^k, \nu' P^k) \leq (1 - \beta)^k d_{\mathrm{TV}}(\nu, \nu'). \tag{B16}$$

This proves uniqueness and uniform convergence to ν_∞. The decomposition also constructs regeneration times: at each transition, independently refresh from m with probability β, otherwise use R. After the first refresh, cycles are independent and identically distributed, with geometric lengths of mean $1/\beta$ and bounded per-step rewards. The renewal-reward strong law gives almost-sure convergence of the empirical average of $\mathbf{1}_\mathcal{D}$; its limit equals $\nu_\infty(\mathcal{D})$ by the displayed uniform convergence and boundedness. For $\beta = 1$, the process is i.i.d. from m after one transition, giving the same conclusions. Finally, stationarity and the pointwise leakage/return bounds give $b_\infty \leq A_{c,\pi}(1 - b_\infty) + (1 - \eta_{c,\pi})b_\infty$; rearranging proves the stated fraction bound. □

This is the standard minorization and regeneration argument for general-state-space chains (Roberts & Rosenthal, 2004). This extension assumes the stated uniform minorization, leakage, and recovery conditions for the fixed prompt and policy. It bounds the long-run fraction of degraded block boundaries, rather than the probability of ever visiting a degraded state.

C Evaluation Details and Additional Results

C.1 Full VBench Results

Table 6: **Full 5-second VBench comparison.**

Model	#Params	Resolution	Throughput (FPS) ↑	Latency (s) ↓	VBench Total ↑	Quality ↑	Semantic ↑
Bidirectional video diffusion models							
LTX-Video (HaCohen et al., 2024)	1.9B	768×512	8.98	13.5	79.83	81.88	71.62
Wan2.1 (Team Wan, 2025a)	1.3B	832×480	0.78	103	83.37	84.30	79.65
Autoregressive video diffusion models							
NOVA (Deng et al., 2025)	0.6B	768×480	0.88	4.1	80.31	80.66	78.92
Pyramid Flow (Jin et al., 2025)	2B	640×384	6.70	2.5	80.75	83.41	70.11
SkyReels-V2 (Chen et al., 2025)	1.3B	960×540	0.49	112	81.97	83.96	74.01
MAGI-1 (Sand.ai et al., 2025)	4.5B	832×480	0.19	282	78.88	81.67	67.72
Distilled autoregressive video models							
CausVid (Yin et al., 2025)	1.3B	832×480	17.0	0.69	81.33	83.98	70.72
Self Forcing (Huang et al., 2025a)	1.3B	832×480	17.0	0.69	83.74	84.48	<u>80.77</u>
Causal Forcing (Zhu et al., 2026a)	1.3B	832×480	17.0	0.69	<u>84.04</u>	<u>84.59</u>	**81.84**
LongForcing (ours)	1.3B	832×480	17.0†	0.69†	**84.56**	**85.52**	80.72

† Reference throughput and latency quoted from Causal Forcing (Zhu et al., 2026a) for its inference configuration, which LongForcing retains.

Table 7: **Full 30-second VBench-Long comparison.**

Model	Aggregate			Long-horizon consistency		Visual quality		Temporal evolution	
	Total	Quality	Semantic	Subject Cons.	Background Cons.	Aesthetic	Imaging	Dynamic	Temporal Style
LongLive (Yang et al., 2026)	<u>82.80</u>	83.21	**81.17**	**98.36**	<u>96.25</u>	**63.77**	**68.35**	38.89	**24.15**
Self Forcing (Huang et al., 2025a)	81.49	82.67	76.77	<u>97.76</u>	95.00	61.72	<u>67.70</u>	41.93	23.56
Causal Forcing (Zhu et al., 2026a)	81.05	<u>83.44</u>	71.52	96.24	95.59	58.85	66.92	**75.46**	22.61
LongForcing (ours)	**82.83**	**83.99**	<u>78.19</u>	97.40	**96.58**	<u>62.87</u>	64.19	<u>60.93</u>	<u>23.61</u>

Best and second-best results are shown in **bold** and <u>underlined</u>, respectively. All scores are percentages.

C.2 Causal-VAE Drift Statistics

We re-encode 20 videos per model with the shared causal VAE. Let $m_{v,c,t}$ be the spatial mean of channel c at zero-based latent index t in video v, and define its early reference as

$$\bar{m}_{v,c} = \frac{1}{20} \sum_{r=1}^{20} m_{v,c,r}. \tag{C1}$$

This reference excludes the first latent ($t = 0$) and averages latent indices 1–20, corresponding to zero-based RGB frame indices 1–80 (80 frames). Panel (b) of Figure 2 plots $\mathrm{median}_{v,c} |m_{v,c,t} - \bar{m}_{v,c}|$ at each t. For the 81-frame model, panel (c) plots the signed channel-wise statistic $\mathrm{median}_v(m_{v,c,t} - \bar{m}_{v,c})$. These statistics summarize the magnitude and channel-wise direction of latent shifts relative to the early reference window.

C.3 Time-Resolved Drift Score

We sample frames at 1 Hz from 16-FPS videos, using indices $0, 16, \ldots, 944$ (60 observations per video). High saturation and RGB clipping are measured at decoded resolution. With OpenCV HSV saturation and value normalized to $[0, 1]$, the high-saturation ratio is the fraction of all pixels satisfying $S > 0.85$ and $V > 0.15$. The RGB-clipping ratio is the fraction with any 8-bit RGB channel at most 5 or at least 250. For chroma edges, frames are resized to a 256-pixel short side with aspect ratio preserved, integer-rounded dimensions, and OpenCV area interpolation. After RGB-to-YCrCb conversion, let g_Y, g_{Cr}, and g_{Cb} be the Euclidean magnitudes of horizontal and vertical 3×3 Sobel responses, computed in floating point on the unnormalized 8-bit channel scale. With $g_C = \sqrt{g_{Cr}^2 + g_{Cb}^2}$, the chroma-edge ratio is the fraction of resized pixels satisfying $g_C > 18$ and $g_C > 1.5(g_Y + 1)$. For text-to-video HPSv3 (Ma et al., 2025), each matched pair uses identical filename-derived prompt prefixes, held fixed across models and timestamps. These scores measure temporal changes relative to each video's early reference. Interactive HPSv3 uses the prompt associated with each scene ID.

For video i and metric m, let b_{im} be the mean over seconds 0–4. We orient deterioration as $d_{im}(t) = b_{im} - x_{im}(t)$ for HPSv3 and $d_{im}(t) = x_{im}(t) - b_{im}$ for three artifact ratios: high saturation, RGB clipping, and chroma edges. The normalized deterioration is

$$z_{im}(t) = \mathrm{clip}\left(\frac{[d_{im}(t)]_+}{q_m}, 0, 1 \right). \tag{C2}$$

The frozen scales are the pooled 95th percentiles of $[d_{im}(t)]_+$, including zeros, over both the 81- and 321-frame text-to-video groups, all 20 matched videos per group, all 60 sampled seconds, and each metric independently: $q_m = (12.5961, 0.2977, 0.4786, 0.3202)$ for HPSv3, high saturation, RGB clipping, and chroma edges, respectively. The per-video score is

$$D_i(t) = 100 \left[\tfrac{1}{2} z_{i,\mathrm{HPS}}(t) + \tfrac{1}{6} \sum_{m \in \{\mathrm{sat,clip,chroma}\}} z_{im}(t) \right]. \tag{C3}$$

Thus HPSv3 and the correlated artifact family receive equal total weight. The score measures deterioration relative to each video's early reference: lower values indicate smaller increases in the selected degradation indicators. Absolute perceptual quality and motion are assessed separately by the benchmark and human evaluations.

Figure C1 applies the same frozen scales to both regimes. Curves use a centered 5-second moving average, truncated to the available observations at each endpoint; numerical endpoint comparisons below use the unsmoothed 59-second observations. The pointwise 95% percentile bootstrap intervals resample complete prompt or scene trajectories 2,000 times and apply the same smoothing within each replicate, keeping the normalization scales fixed. At each timestamp, the interval endpoints are the 2.5th and 97.5th percentiles of the bootstrap means. The text-to-video and interactive panels contain 20 prompt-matched and 23 scene-matched videos per model, respectively.

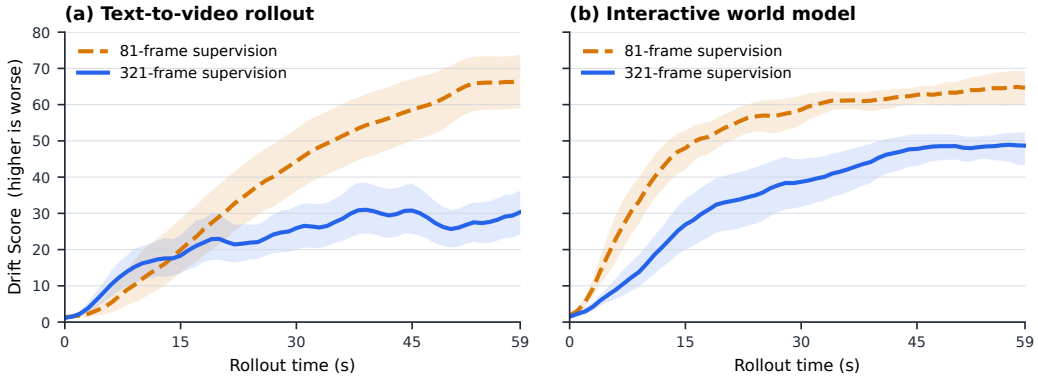

Figure C1: **Full time-resolved drift comparison.** Curves show mean Drift Score with 95% bootstrap confidence intervals for text-to-video and interactive rollouts.

At 59 seconds, 321-frame supervision reduces the mean Drift Score from 67.46 to 31.91 for text-to-video generation and from 64.76 to 48.95 for interactive world modeling, corresponding to relative reductions of 52.7% and 24.4%, respectively.

C.4 Human Preference Evaluation

Our analysis covers 20 shared cases evaluated by all 20 participants: 10 text-to-video prompts and 10 interactive world-model scenes. Each pair compares 81- and 321-frame supervision with model identities hidden and left/right placement randomized. Evaluators select A better, Tie, or B better for overall preference, temporal coherence, visual quality, and motion naturalness. Each participant provides one response per case and criterion, yielding $20 \times 20 = 400$ judgments per criterion and 1,600 judgments across the four criteria. Figure 3 reports descriptive preference proportions, assigning equal weight to each evaluator–case response and retaining ties in the denominator.

D Additional Qualitative Results

D.1 Text-to-Video Rollouts

Figure D1 follows four 60-second text-to-video rollouts trained with 321-frame supervision. Across portrait, human-motion, animated-aerial, and deep-space prompts, the displayed frames show recognizable subjects and scene style as content evolves beyond the supervised horizon.

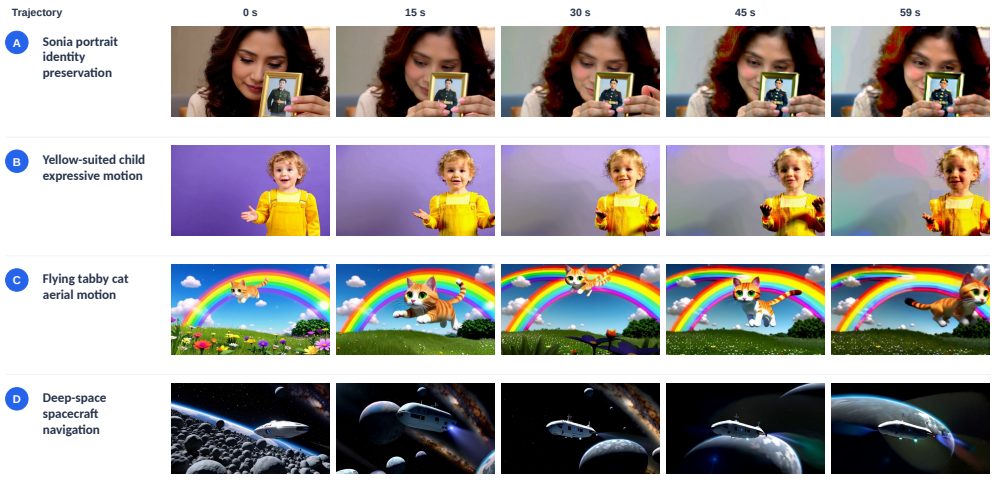

Figure D1: Qualitative 60-second text-to-video rollouts. Each row follows one LongForcing trajectory trained with 321-frame supervision; columns show frames at fixed video timestamps.

D.2 Supervision-Horizon Comparison

Figure D2 compares matched prompts or initial scenes. The 81-frame text-to-video examples exhibit progressive saturation and structural distortion, while the interactive examples develop horizontal banding and geometry collapse. With 321-frame supervision, the displayed snapshots from all four trajectories preserve more scene structure and do not show the same late-stage artifacts.

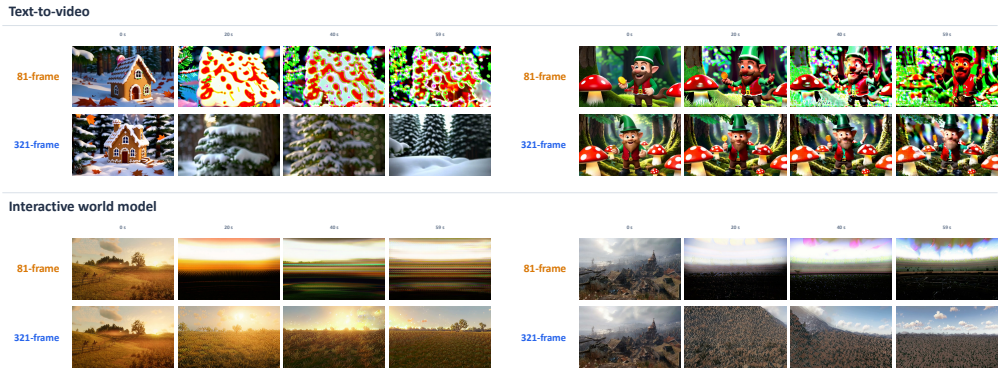

Figure D2: Qualitative comparison across supervision horizons. Upper and lower rows use 81- and 321-frame supervision, respectively, at matched rollout times from 0 to 59 seconds.

E Discussion

When does LongForcing help? Horizon scaling targets failures first reached beyond the shorter training window. Extending the on-policy rollout makes these later prefixes available to the teacher, while the horizon ablations quantify the resulting empirical changes in generation quality. The benefit depends on both supervision coverage and the quality of the teacher signal on those prefixes.

Cost. The additional computation occurs during training: the student generates longer paths and both score networks process longer sequences. At fixed spatial resolution, dense bidirectional attention has quadratic attention-score computation in temporal length; practical memory and runtime

depend on the attention and gradient implementations. For a fixed inference configuration and output length, the student uses the same architecture and sampling operations as the short-horizon baseline.

Scope. The finite-horizon results quantify teacher-relative path agreement, while the blockwise extension additionally assumes controlled leakage, mismatch, and recovery on the student's start distribution. The long-run extension states its uniform dynamical assumptions separately. These conditions concern the specified model, domain, and action protocol. Visual quality, motion, and action following are evaluated separately, complementing the drift analysis.

How long is enough? A useful supervision horizon depends on when task-relevant dependencies and failure modes become observable, rather than on a universal duration. For scene roaming, delayed effects may involve revisiting a location or maintaining an object after a long absence; for goal-directed interaction, an action's consequence may appear much later. Our horizon sweep measures the benefit of extending supervision within the tested range; determining a sufficient horizon for a new task calls for evaluating those dependencies over the intended deployment duration. A promising direction is to allocate longer training rollouts when delayed errors remain frequent, and to compare this adaptive allocation with fixed horizons under matched compute budgets. Teacher quality and persistent memory offer complementary ways to make additional temporal coverage useful.

From visual rollout to long-term planning. Maintaining a coherent visual trajectory and deciding which events should occur over a long story or task are complementary problems. LongForcing strengthens the former by supervising longer self-generated paths; long-term planning additionally requires representing goals, organizing events, and retaining consequences across distant steps. This suggests a hierarchical approach in which a planner maintains an abstract state and selects subgoals, persistent memory records relevant past events, and a video model renders the resulting interactions. Long-horizon supervision could then assess whether the rendered trajectory respects those plans and remembered consequences. Evaluating delayed action effects, goal completion, and narrative consistency alongside visual quality would help determine whether longer rollouts also support more coherent long-term behavior.