DUAL-SUPPORT ONLINE POLICY OPTIMIZATION FOR FEW-STEP AUTOREGRESSIVE VIDEO MODELS

Anonymous authors
Paper under double-blind review

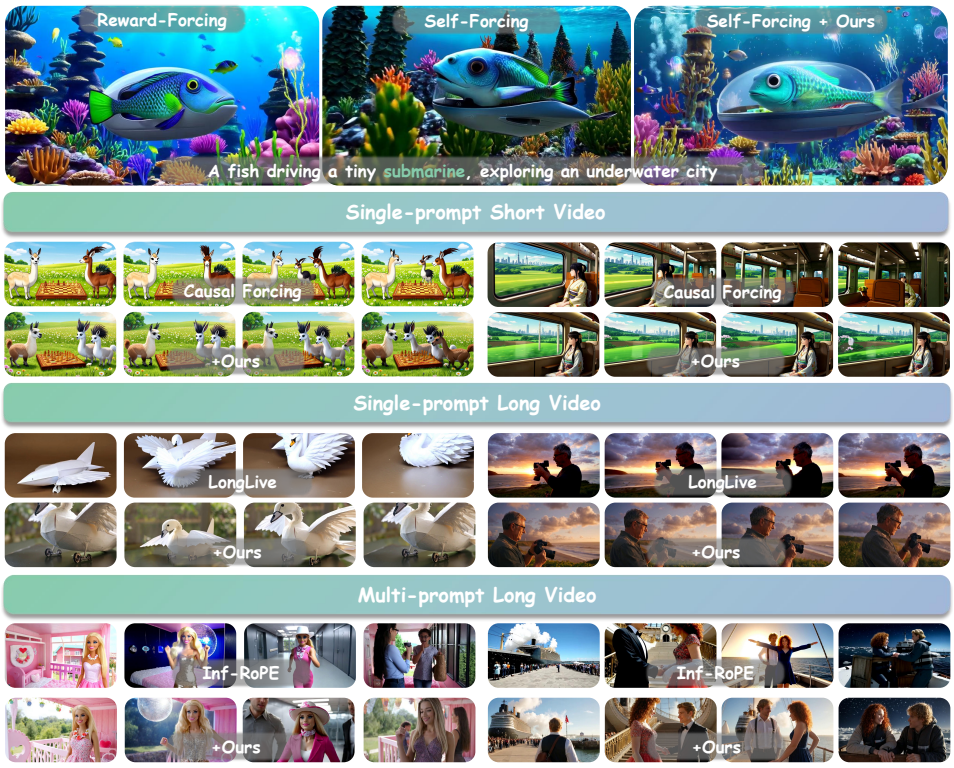


Figure 1: Our method efficiently aligns distilled streaming video models with human preferences without re-distillation, enhancing baselines (e.g., Causal Forcing Zhu et al. (2026), LongLive Yang et al. (2025) and Infinite-RoPE Yesiltepe et al. (2025)) by mitigating artifacts and improving temporal consistency. We demonstrate boosted perceptual quality across: (Top) single-prompt short, (Middle) single-prompt long, and (Bottom) multi-prompt long video generation.

ABSTRACT

Few-step causal autoregressive (AR) video models enable efficient streaming generation through timestep distillation and KV-cached causal inference, making online optimization directly on the final deployed model increasingly practical. However, post-distillation optimization introduces a *deployment-support mismatch*: the training updates may be estimated at timesteps and causal contexts that differ from those actually encountered by the few-step streaming policy at inference. We identify two complementary sources of this mismatch. The *distilled timestep support* is the discrete timestep set retained by the few-step sampler, while the *streaming context support* is the distribution of causal contexts induced by autoregressive rollout, approximated during training by a lagged behavior policy. Based on this observation, we propose a dual-support online policy optimization framework that performs relative forward-process updates only on distilled timesteps and optimizes candidate continuations under policy-generated long-horizon contexts. Historical KV states are detached so that optimization remains local while context exposure remains long-horizon. Across multiple distilled AR video backbones and generation horizons, our method con-

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.

sistently improves preference-oriented quality and motion dynamics while preserving the original few-step inference procedure. Our results show that effective post-distillation alignment depends not only on the optimization objective, but also on matching the state distribution encountered by the final generative policy at inference.

1 INTRODUCTION

Recent advances in video generation have produced increasingly capable diffusion models He et al. (2022); Ho et al. (2022); Blattmann et al. (2023b;a); Chen et al. (2023); Gupta et al. (2024); Wan et al. (2025); Zheng et al. (2024). However, conventional bidirectional video diffusion models remain costly for real-time and interactive generation because sampling requires repeated denoising over large spatiotemporal contexts. Few-step causal autoregressive (AR) models Huang et al. (2025a); Yang et al. (2025); Cui et al. (2025); Shin et al. (2025); Huang et al. (2025b); Hong et al. (2025) instead distill strong bidirectional teachers into efficient streaming policies with few-step sampling and KV-cached causal inference. Besides accelerating generation, their low rollout cost makes online policy optimization practical directly on the final causal model.

Distillation, however, changes not only model capacity but also the distribution of states visited at inference. Few-step students may exhibit over-saturated appearance, weakened motion, or degraded generation quality Lu et al. (2025), and autoregressive rollout can further propagate these errors through subsequent contexts. This motivates *post-distillation* alignment: directly improving the final causal student using reward feedback from its own rollouts.

Our key observation is that naively transferring online optimization to a distilled causal policy creates a *deployment-support mismatch*. The final policy is queried only at the discrete timesteps retained by the few-step sampler and under causal contexts induced by its own preceding outputs. In contrast, optimization over generic diffusion times or repeatedly reset contexts estimates updates under a different state distribution. For a fixed behavior policy, this mismatch can be traced to two sources: the timestep distribution used for optimization and the causal-context distribution induced by autoregressive rollout. We refer to them as the *distilled timestep support* and the *streaming context support*. Based on this formulation, we develop a **dual-support online policy optimization** framework that matches both distributions encountered at inference. We perform relative forward-process optimization only on the distilled timestep set, and estimate policy updates under long-horizon contexts generated by a lagged streaming policy. Historical KV states are detached so that optimization remains local while context exposure remains long-horizon. This yields a deployment-matched local optimization procedure without reconstructing the teacher's dense diffusion trajectory or backpropagating through the full autoregressive rollout. Our approach consistently improves preference-oriented quality and motion dynamics across multiple distilled AR video backbones Huang et al. (2025a); Yang et al. (2025); Zhu et al. (2026) while preserving the original few-step inference procedure.

Our contributions are threefold: (1) we identify a **deployment-support mismatch** in post-distillation online optimization and decompose it into timestep-support and streaming-context discrepancies; (2) we develop a dual-support optimization framework that matches both distributions using relative forward updates on distilled timesteps and policy-generated streaming contexts; and (3) we validate the formulation across multiple distilled AR backbones, generation horizons, and online RL baselines, consistently improving preference-oriented quality and motion dynamics while preserving few-step inference.

2 RELATED WORK

2.1 FEW-STEP CAUSAL VIDEO GENERATION

Diffusion models have achieved strong performance in video synthesis He et al. (2022); Ho et al. (2022); Singer et al. (2022); Blattmann et al. (2023b;a); Chen et al. (2023); Gupta et al. (2024); Zhao et al. (2024); Xing et al. (2024), while Diffusion Transformers (DiTs) Bao et al. (2023); Peebles & Xie (2023) further scale this paradigm to large video generators Yang et al. (2024); Bao et al. (2024); Kong et al. (2024); Wan et al. (2025). Their bidirectional computation and multi-step

sampling, however, remain costly for streaming generation. Causal and autoregressive models generate videos incrementally Wu et al. (2021); Hong et al. (2022); Wu et al. (2022); Yan et al. (2021); Kondratyuk et al. (2023); Deng et al. (2024). Recent approaches combine causal generation with timestep distillation to obtain efficient few-step streaming models. CausVid Yin et al. (2025) distills a bidirectional teacher into a block-causal student, while Self-Forcing and related methods Huang et al. (2025a); Yang et al. (2025); Yesiltepe et al. (2025); Cui et al. (2025); Guo et al. (2025) train on self-generated rollouts to better match autoregressive inference. Causal Forcing Zhu et al. (2026) further studies distillation strategies for long-horizon causal generation.

2.2 Adaptation of Distilled Generators

Post-training a timestep-distilled generator can be sensitive to how optimization interacts with its few-step inference process. PSO Miao et al. (2025) shows that naive fine-tuning may degrade distilled generation and introduces relative optimization to better preserve few-step behavior. D-OPSD Jiang et al. (2026) and OPSD-V Liu et al. (2026) instead optimize distilled models on states encountered during their own inference process, using compatible teacher supervision. Other work improves the distillation stage itself. Reward Forcing Lu et al. (2025) injects reward signals into distribution matching to preserve motion dynamics, while Data-Forcing Distillation Chen et al. (2026) targets diversity loss and over-saturation during distillation. Together, these works highlight the importance of preserving compatibility with the inference behavior of the distilled student.

2.3 Online Preference Optimization for Generative Models

DiffusionDPO Wallace et al. (2024) adapts diffusion models from offline preference pairs, while on-policy methods such as Dance-GRPO and Flow-GRPO Xue et al. (2025); Liu et al. (2025a) optimize rewards through the reverse denoising process. DiffusionNFT Zheng et al. (2025) instead performs relative reward optimization through forward corruption, avoiding reverse-process likelihood estimation and trajectory backpropagation. Recent work extends online optimization to autoregressive and streaming generators. WorldCompass Wang et al. (2026) applies forward-process preference optimization to autoregressive world models, while Vidu S2 Zhang et al. (2026) combines self-generated trajectory replay with streaming NFT. GARDO He et al. (2025) studies reward-hacking mitigation through disagreement-aware regularization. Our work builds on these directions and focuses on post-distillation online optimization of few-step causal video models, jointly accounting for their distilled timestep support and long-horizon streaming contexts.

3 Methodology

We study post-distillation online preference optimization for few-step causal video models. Our key observation is that the final policy operates on two inference-specific supports: the discrete timestep set retained after distillation and the causal contexts accumulated during streaming generation. Accordingly, we perform relative forward-process optimization on the *distilled timestep support* and optimize candidate continuations under the *streaming context support* generated by a lagged behavior policy. The following sections describe these two components and the practical stabilization used for online training.

3.1 Problem Setup: Dual-Support Policy Optimization

Let c denote the conditioning signal and $x_{1:N} = (x_1, \ldots, x_N)$ a sequence of generated video clips. A causal autoregressive policy factorizes as

$$P_\theta(x_{1:N} \mid c) = \prod_{n=1}^{N} p_\theta(x_n \mid C_n, c), \tag{1}$$

where C_n is the causal context accumulated from preceding clips and represented by the rolling KV cache. The distilled policy is sampled on a fixed timestep set $\mathcal{T}_{\text{distill}} = \{t_K, \ldots, t_1\}$, $K \ll T$, which we call the *distilled timestep support*. Meanwhile, autoregressive inference induces a distribution over causal contexts C_n through the model's own preceding generations; we call this the *streaming*

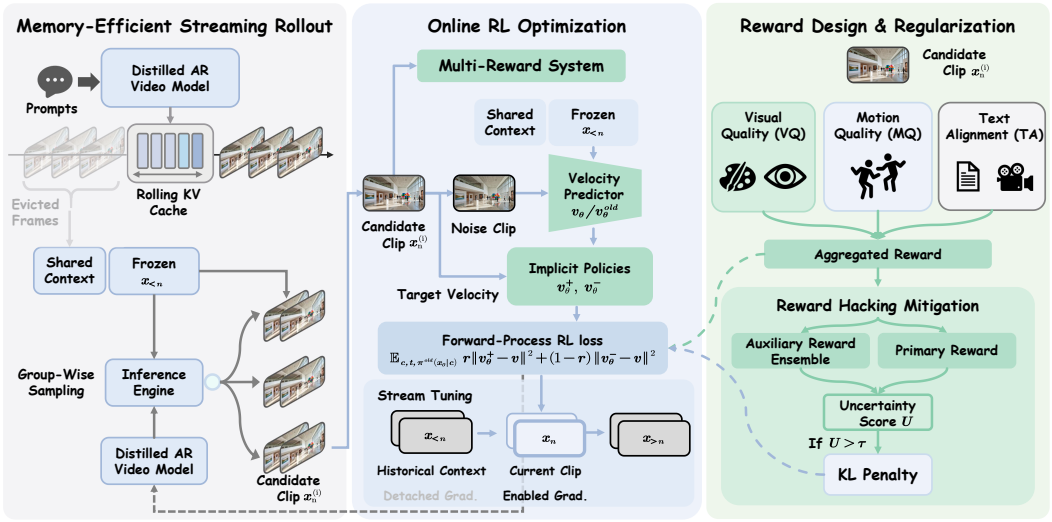


Figure 2: **Overview of our framework.** Given a distilled causal video model, we sample candidates from shared streaming contexts with a rolling KV cache (left). We optimize only the active clip at distilled timesteps using a relative forward-process objective with detached history (middle). Multi-reward scoring and disagreement-gated reference regularization stabilize training (right).

context support. Post-distillation online alignment therefore involves two complementary support-matching requirements: policy updates should be evaluated at the timesteps used by the few-step sampler and under the contexts encountered during streaming generation. We use a lagged behavior policy $\pi_{\theta_{\text{old}}}$ to sample both streaming contexts and candidate continuations, as detailed in Sec. 3.3.

3.2 Relative Forward Optimization on Distilled Timesteps

Direct reward optimization through the reverse diffusion trajectory typically requires storing intermediate denoising states and may depend on sampler-specific likelihoods. We instead build on relative forward-process optimization Zheng et al. (2025), which re-noises generated endpoints into training states and avoids backpropagation through the sampling trajectory.

Relative forward update. For a generated endpoint x_0, we sample $t \in \mathcal{T}_{\text{distill}}$ and construct

$$x_t = (1-t)x_0 + t\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \qquad u_t = \epsilon - x_0. \tag{2}$$

Let $v_{\text{old}} = v_{\theta_{\text{old}}}(x_t, t, C_n, c)$ and $v_\theta = v_\theta(x_t, t, C_n, c)$. Following DiffusionNFT Zheng et al. (2025), we define

$$v^+ = (1-\beta)v_{\text{old}} + \beta v_\theta, \qquad v^- = (1+\beta)v_{\text{old}} - \beta v_\theta, \tag{3}$$

and optimize

$$\mathcal{L}_{\text{policy}} = \tilde{r}\,\|v^+ - u_t\|_2^2 + (1-\tilde{r})\,\|v^- - u_t\|_2^2. \tag{4}$$

High-reward samples favor the positive relative direction and low-reward samples the opposite direction, while the update remains anchored to the current distilled policy.

Crucially, we sample t only from $\mathcal{T}_{\text{distill}}$. After timestep distillation, the final sampler queries only this small set of coarse transitions; arbitrary intermediate times do not directly participate in inference. Restricting the reward update to the same timestep set therefore concentrates optimization on the states exercised by the few-step sampler, without assuming that the student still represents the teacher's original dense probability flow.

Group-relative reward. For each causal context, we sample G candidate continuations and compare them within the group. For candidate i, we use

$$A^{(i)} = R^{(i)} - \frac{1}{G}\sum_{j=1}^{G} R^{(j)}, \qquad \tilde{r}_i = \frac{1}{2}\left[1 + \text{clip}\left(\frac{A^{(i)}}{A_{\max}}, -1, 1\right)\right], \tag{5}$$

where $A_{\max} = \max_j |A^{(j)}| + \epsilon$. Because candidates share the same prompt and causal prefix, this relative comparison reduces context-dependent reward offsets and emphasizes differences among alternative continuations.

Post-distillation interpretation. Forward regression remains well defined even when the policy has already undergone timestep distillation. For a non-negative endpoint weight $w(x_0, C_n, c)$, consider

$$\mathcal{L}_w(v) = \mathbb{E}\left[w \, \|v(x_t, t, C_n, c) - u_t\|_2^2\right], \qquad v_w^* = \frac{\mathbb{E}[w\, u_t \mid x_t, t, C_n, c]}{\mathbb{E}[w \mid x_t, t, C_n, c]}. \tag{6}$$

The optimum is exactly the conditional regression target induced by the reweighted endpoint distribution $p_w(x_0 \mid C_n, c) \propto w(x_0, C_n, c) p_{\text{old}}(x_0 \mid C_n, c)$. Thus, forward preference optimization can be defined directly on endpoints sampled from the distilled policy and does not require reconstructing the teacher's dense trajectory.

Distillation can nevertheless introduce mismatch between the current predictor and the corresponding forward-regression optimum. Let $a = 2\tilde{r} - 1$, $\mu = \mathbb{E}[u_t \mid x_t, t, C_n]$, and $b = v_{\text{old}} - \mu$. The population-optimal relative perturbation $\delta = v_\theta - v_{\text{old}}$ satisfies

$$\delta^* = \frac{1}{\beta} \operatorname{Cov}(a, u_t \mid x_t, t, C_n) - \frac{\mathbb{E}[a \mid x_t, t, C_n]}{\beta} b. \tag{7}$$

The first term captures the reward-directed update, while the second exposes the bias induced by distilled-flow mismatch. Group-relative centering does not eliminate this bias exactly, but reduces shared reward offsets and can suppress it in practice. Full derivations are provided in the Appendix.

3.3 Streaming Optimization under Self-Generated Contexts

Matching distilled timesteps alone is insufficient for causal generation, since each continuation is conditioned on preceding model outputs and errors can propagate through the rollout. We therefore optimize under contexts generated by a lagged streaming policy $\pi_{\theta_{\text{old}}}$ using the same rolling KV cache as inference. At step n, the context C_n retains S sink clips and the L most recent clips, and we sample G candidate continuations from the shared context:

$$x_n^{(i)} \sim \pi_{\theta_{\text{old}}}(\cdot \mid C_n, c), \qquad i = 1, \ldots, G. \tag{8}$$

The shared prefix makes rewards directly comparable while exposing optimization to long-horizon policy-generated contexts. During optimization, we detach the historical KV state, $\operatorname{stopgrad}(C_n)$, and backpropagate only through the active clip. This provides long-horizon state exposure with local credit assignment, avoiding backpropagation through the full autoregressive history. The lagged policy tracks the current model throughout training and provides a stable approximation to its evolving streaming context distribution.

3.4 Reward Design and Stabilization

Online optimization against learned rewards can exploit imperfections in the reward signal as the policy moves away from its initial distribution. We therefore combine complementary reward dimensions and selectively constrain policy drift when these signals disagree.

Multi-reward scoring. Each candidate receives $R = w_{\text{VQ}} R_{\text{VQ}} + w_{\text{MQ}} R_{\text{MQ}} + w_{\text{TA}} R_{\text{TA}}$, with $\sum_k w_k = 1$. We use HPSv3 Ma et al. (2025) for visual quality, grayscale VideoAlign Liu et al. (2025b) for motion quality, and RGB VideoAlign for text–video alignment. The three components are equally weighted in our experiments, and the resulting score is used to construct the group-relative preference signal in Eq. equation 5. Combining these signals discourages improvements in visual preference from being achieved at the expense of motion or semantic alignment.

Disagreement-gated reference regularization. Averaging multiple rewards does not fully prevent reward exploitation. Following GARDO He et al. (2025), we use disagreement among reward models to identify candidates for which the primary reward may be unreliable. Let $\Delta_{\text{rank}}^{(i)}$ denote the discrepancy between the ranking of candidate i under the primary reward and its mean ranking under auxiliary reward models, and define $\mathcal{M}^{(i)} = \mathbb{1}[\Delta_{\text{rank}}^{(i)} > \tau]$. We regularize only these

Table 1: Quantitative results on VBench benchmarks. Integrating our approach into existing distilled models improves aggregate quality and motion metrics; semantic scores are largely preserved with small trade-offs.

Method	Total↑	Quality↑	Semantic↑	HPSv3↑	MQ↑	Throughput↑
AR Models						
SkyReels-V2 Chen et al. (2025)	82.67	84.70	74.53	9.08	1.59	0.49
MAGI-1 Teng et al. (2025)	79.18	82.04	67.74	7.95	1.52	0.19
NOVA Deng et al. (2024)	80.12	80.39	79.05	8.21	1.63	0.88
PyramidFlow Jin et al. (2024)	81.72	84.74	69.62	8.76	1.50	6.70
Distilled AR Models						
CausVid Yin et al. (2025)	81.20	84.05	69.80	7.56	1.22	17.0
Reward Forcing Lu et al. (2025)	84.13	84.84	81.32	8.74	1.65	23.1
Self-Forcing Huang et al. (2025a)	83.74	84.48	80.77	9.36	1.65	17.0
+ Ours	83.79+.05	84.51+.03	80.92+.15	10.72+1.36	1.71+.06	17.0
LongLive Yang et al. (2025)	83.22	83.68	81.37	9.38	1.51	20.7
+ Ours	84.93+1.71	85.83+2.15	81.36-.01	11.03+1.65	1.64+.13	20.7
Causal Forcing Zhu et al. (2026)	84.04	84.59	81.84	9.48	1.69	17.0
+ Ours	84.46+.42	85.15+.56	81.72-.12	10.84+1.36	1.80+.11	17.0

Table 2: Quantitative results on VBench-Long benchmarks. Integrating our method improves aggregate long-video metrics, with occasional semantic trade-offs.

Method	Total↑	Quality↑	Semantic↑	HPSv3↑	MQ↑
SkyReels-V2 Chen et al. (2025)	75.29	80.77	53.37	8.72	1.54
FramePack Zhang et al. (2025)	81.95	83.61	75.32	8.94	1.58
Self-Forcing Huang et al. (2025a)	81.59	83.82	72.70	9.12	1.61
+ Ours	**82.03**	**84.36**	**72.71**	**10.38**	**1.72**
LongLive Yang et al. (2025)	83.52	85.44	75.82	9.21	1.48
+ Ours	**84.07**	**86.12**	**75.87**	**10.67**	**1.64**
Causal Forcing Zhu et al. (2026)	82.87	84.36	**76.91**	9.28	1.65
+ Ours	**84.24**	**86.18**	76.48	**10.52**	**1.74**

high-disagreement candidates:

$$\mathcal{L}_{\text{ref}} = \frac{\sum_i \mathcal{M}^{(i)} \left\| v_\theta^{(i)} - v_{\theta_{\text{ref}}}^{(i)} \right\|_2^2}{\max\left(1, \sum_i \mathcal{M}^{(i)}\right)}, \qquad \mathcal{L} = \mathcal{L}_{\text{policy}} + \lambda_{\text{ref}} \mathcal{L}_{\text{ref}}. \tag{9}$$

This selectively limits policy drift when reward signals conflict while leaving low-disagreement updates largely unconstrained. The velocity-space MSE is a practical regularization surrogate rather than an exact likelihood KL. The behavior policy is updated by exponential moving average, while the reference policy is periodically refreshed according to the criterion described in the implementation details.

4 Experiments

4.1 Experimental Setup

We evaluate our post-distillation alignment method on few-step causal autoregressive (AR) video models. The primary backbones are Self-Forcing Huang et al. (2025a), Causal-Forcing Zhu et al. (2026), and LongLive Yang et al. (2025). Training prompts are sampled from VidProM Wang & Yang (2024), using the filtered subset introduced by DanceGRPO Xue et al. (2025). We use LoRA with rank $r = 256$ and scaling factor $\alpha = 256$. VideoAlign Liu et al. (2025b) and HPSv3 Ma et al. (2025) provide the reward signals used during alignment. Implementation and evaluation details are provided in the Supplement.

4.2 Short-Horizon Post-Distillation Alignment

We first test whether post-distillation alignment improves different few-step AR backbones in the standard 5-second, single-prompt setting. Following the VBench protocol Huang et al. (2024),

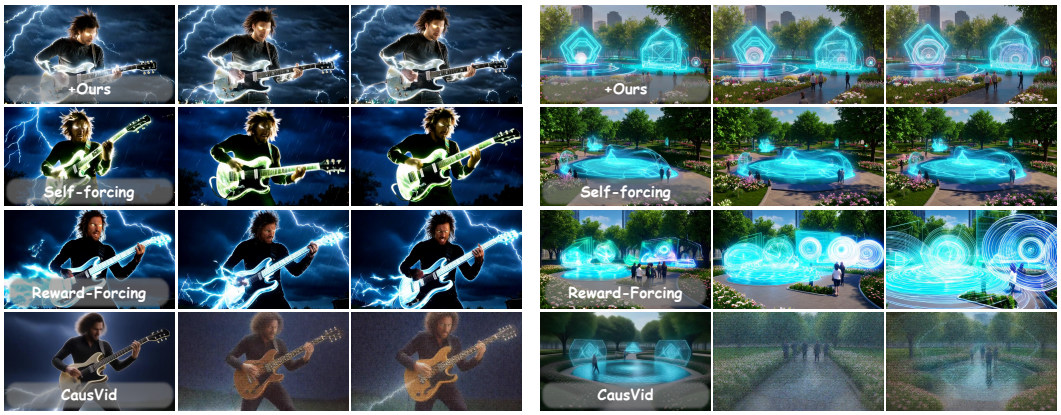


Figure 3: Qualitative comparison under the short-video, single-prompt setting. Our method produces sharper textures and more coherent motion than the compared baselines. Additional examples are provided in the supplementary material.



Figure 4: Qualitative results under the single-prompt long-video setting. Our framework (+Ours) effectively translates alignment optimizations from short videos to extended temporal horizons. Our approach delivers enhanced visual details and more stable throughout the sequence.

we evaluate 946 standard prompts. For the Self-Forcing comparison, we also use its augmented prompt test set, whose prompts are expanded with Qwen2.5-7B-Instruct Bai et al. (2023) and the Wan2.1 Wan et al. (2025) system prompts. We report both standard generation metrics and reward-oriented metrics. HPSv3 and VideoAlign-based scores are also used during training and therefore measure direct reward alignment, whereas VBench is an independent evaluation of general video quality. Across the Self-Forcing, Causal-Forcing, and LongLive variants, our method improves preference and motion scores while preserving the original few-step inference throughput; individual semantic metrics can trade off. Quantitative results are shown in Table 1 and qualitative comparisons



Figure 5: Qualitative comparison of multi-prompt long-video generation. We compare the LongLive Yang et al. (2025) with our method. The generated sequences exhibit noticeable improvements in visual aesthetics and fine details during complex narrative transitions.

Table 3: Quantitative evaluation on long video generation. We compare these overall metrics alongside CLIP Scores evaluated across 10-second intervals (0-60s).

Method	Quality Score↑	Consistency Score↑	Aesthetic Score↑	CLIP Score↑					
				0-10	10-20	20-30	30-40	40-50	50-60
SkyReels-V2 Chen et al. (2025)	81.55	94.72	56.83	25.31	23.40	22.50	21.62	21.67	20.91
FramePack Zhang et al. (2025)	84.40	96.77	59.44	26.51	22.60	22.18	21.53	21.98	21.62
Self-Forcing Huang et al. (2025a)	83.94	95.74	58.45	26.24	**24.87**	23.46	21.92	**22.05**	21.07
+ Ours	**84.72**	**95.98**	**59.62**	**26.42**	24.75	**23.95**	**22.40**	21.85	**21.50**
LongLive Yang et al. (2025)	84.28	96.05	59.89	26.63	25.77	**24.65**	23.99	**24.52**	24.11
+ Ours	**85.15**	**96.16**	**60.75**	**26.80**	**26.15**	24.45	**24.55**	24.30	**24.65**
Causal-Forcing Zhu et al. (2026)	84.12	**95.88**	59.15	26.45	**25.60**	**23.98**	22.85	22.48	22.45
+ Ours	**84.95**	95.63	**60.32**	**26.58**	25.12	23.85	**23.40**	**23.10**	**22.95**

in Figure 3. For an additional held-out reward-oriented check, we evaluate HPSv3 and Motion Quality on 100 diverse MovieGenBench prompts Polyak et al. (2024).

4.3 Long-Horizon Streaming Alignment

We next ask whether short-horizon improvements extend to streaming generation. Under the VBench-Long protocol, we generate a 30-second video for each official prompt and evaluate the clips produced by the standard script. LongLive natively supports long videos, while Self-Forcing and Causal-Forcing are extended with Infinity-RoPE Yesiltepe et al. (2025) for this evaluation. We compare each base model with and without our post-training alignment under the same inference procedure, and include SkyReels-V2 Chen et al. (2025) and FramePack Zhang et al. (2025) as contextual baselines. Table 2 reports VBench-Long and preference-oriented scores. The gains show that alignment learned on short clips transfers to the long-horizon streaming setting, although individual semantic or consistency metrics may trade off. Qualitative examples are shown in Figure 4.

4.4 Multi-Prompt Long-Horizon Generation

We further evaluate interactive multi-prompt generation. Following LongLive Yang et al. (2025), we use 100 groups of six successive 10-second prompts, producing 60-second videos. Short-context baselines are adapted by switching prompts during autoregressive rollout, while LongLive uses its native interactive generation procedure. We segment each video at prompt boundaries and compute CLIP scores at 10-second intervals. Table 3 shows improved aggregate quality and aesthetics with mixed changes across interval-level semantic scores. These results show that post-trained policies improve visual quality while preserving clip-wise instruction following across successive prompt changes; the clip-wise CLIP evaluation does not by itself establish long-term temporal consistency. Qualitative examples are provided in Figure 5.

Table 4: **Comparison with other online optimization on the same causal AR backbone.** DanceGRPO-AR and Flow-GRPO-AR denote our adaptations of the corresponding objectives to the same distilled causal student.

	Optimization		5s			30s		
Method	**Update**	**Streaming Ctx.**	**VBench↑**	**HPSv3↑**	**MQ↑**	**VBench-L↑**	**HPSv3↑**	**MQ↑**
Distilled Base	–	–	84.04	9.48	1.69	82.87	9.28	1.65
DanceGRPO-AR	Reverse GRPO	No	83.88	9.76	1.65	82.31	9.43	1.59
Flow-GRPO-AR	Reverse GRPO	No	83.94	9.71	1.66	82.45	9.40	1.61
DiffusionNFT	Forward	No	84.11	10.03	1.71	82.95	9.72	1.66
Streaming DanceGRPO-AR	Reverse GRPO	Yes	83.96	9.84	1.67	83.12	9.69	1.65
Streaming Flow-GRPO-AR	Reverse GRPO	Yes	84.00	9.79	1.68	83.24	9.65	1.67
Dual-Support (Ours)	**Forward**	**Yes**	**84.46**	**10.84**	**1.80**	**84.24**	**10.52**	**1.74**

Table 5: **Ablation studies.** We analyze the two support-matching components and the optimization strategy used for long-horizon streaming. Additional ablations on reward design and training stabilization are provided in the supplementary material.

Variant	**VBench-L↑**	**HPSv3↑**	**MQ↑**	**Peak Mem.↓**
Dual-Support Optimization				
Dense t + Reset Context	82.95	9.72	1.66	–
Distilled t + Reset Context	83.43	10.14	1.69	–
Dense t + Streaming Context	83.58	10.02	1.71	–
Distilled t + Streaming Context	**84.24**	**10.52**	**1.74**	–
Streaming Optimization Strategy				
Full-Sequence Backpropagation	–	–	–	OOM
Clip-Wise Backpropagation	84.05	10.39	1.72	96.3
Detached History	84.16	10.47	1.73	75.2
Shared Context + Detached History	**84.24**	**10.52**	**1.74**	**54.4**

4.5 Comparison with Existing Online RL Methods

We compare existing online RL objectives on the same distilled Causal-Forcing model under matched training and rollout budgets in Table 4. DanceGRPO-AR and Flow-GRPO-AR adapt the original objectives to the 4-step causal setting, while DiffusionNFT serves as the vanilla forward-process baseline. The 5-second results isolate the update rule, whereas the 30-second streaming results further evaluate the effect of policy-generated causal contexts.

4.6 Ablation Studies

Table 5 evaluates the two core design choices of our method. The dual-support study shows that matching distilled timesteps and streaming contexts each improves long-horizon generation, with the combination performing best. We further compare streaming optimization strategies: detaching historical context greatly reduces memory while preserving quality, and shared-context sampling provides the most efficient configuration. Additional reward and hyperparameter ablations are deferred to the supplementary material.

5 Conclusion

We presented a dual-support framework for post-distillation online preference optimization of few-step causal autoregressive video models. By aligning policy optimization with both the distilled timestep support and the streaming context support, our method consistently improves preference-oriented quality and motion dynamics across multiple AR backbones and generation horizons while preserving few-step inference. These results show that effective post-distillation alignment depends not only on the optimization objective itself, but also on matching the states and contexts actually encountered by the final policy at inference. We hope this perspective provides a useful basis for further improving efficient streaming generative models through online optimization

AI USE STATEMENT

We used generative AI tools for language polishing, brainstorming, and limited code assistance. We did not use generative AI tools to generate experimental results, fabricate data, or make final scientific claims. All AI-assisted text, code, and suggestions were manually reviewed, revised, and verified by the authors, who take full responsibility for the final content and artifacts.

ETHICS STATEMENT

This work studies post-distillation alignment and video generation for research purposes. Realistic synthetic videos may be misused to fabricate events or misrepresent model capabilities. We recommend clearly identifying generated content and avoiding deployment in safety-critical settings without human oversight and real-world validation. Our evaluations did not collect sensitive personal information. We will follow the licenses of all source datasets and document intended use, limitations, and potential misuse risks in any public release.

REPRODUCIBILITY STATEMENT

We provide a complete description of the model design, rollout construction, training and inference procedures, and benchmark setup in the main paper and appendix. Section 3 specifies the task formulation, dual-support optimization, and streaming rollout protocol, while the experimental section and supplementary material describe the baselines, evaluation settings, and implementation details. We will release code, data-processing scripts, and configuration files to support reproduction of the reported results.

REFERENCES

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22669–22679, 2023.

Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233*, 2024.

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

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22563–22575, 2023b.

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.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

Siyi Chen, Shaowei Liu, Yixuan Jia, Zian Wang, Huan Ling, Qing Qu, and Jun Gao. Data-forcing distillation: Restoring diversity and fidelity in few-step video generation. *arXiv preprint arXiv:2606.18478*, 2026.

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

Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. *arXiv preprint arXiv:2412.14169*, 2024.

Yuwei Guo, Ceyuan Yang, Hao He, Yang Zhao, Meng Wei, Zhenheng Yang, Weilin Huang, and Dahua Lin. End-to-end training for autoregressive video diffusion via self-resampling. *arXiv preprint arXiv:2512.15702*, 2025.

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

Haoran He, Yuxiao Ye, Jie Liu, Jiajun Liang, Zhiyong Wang, Ziyang Yuan, Xintao Wang, Hangyu Mao, Pengfei Wan, and Ling Pan. Gardo: Reinforcing diffusion models without reward hacking. *arXiv preprint arXiv:2512.24138*, 2025.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022.

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

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

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

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025a.

Yubo Huang, Hailong Guo, Fangtai Wu, Shifeng Zhang, Shijie Huang, Qijun Gan, Lin Liu, Sirui Zhao, Enhong Chen, Jiaming Liu, et al. Live avatar: Streaming real-time audio-driven avatar generation with infinite length. *arXiv preprint arXiv:2512.04677*, 2025b.

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.

Dengyang Jiang, Xin Jin, Dongyang Liu, Zanyi Wang, Mingzhe Zheng, Ruoyi Du, Xiangpeng Yang, Qilong Wu, Zhen Li, Peng Gao, Harry Yang, and Steven Hoi. D-opsd: On-policy self-distillation for continuously tuning step-distilled diffusion models. *arXiv preprint arXiv:2605.05204*, 2026.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024.

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. *arXiv preprint arXiv:2312.14125*, 2023.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

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.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025a.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, et al. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. *arXiv preprint arXiv:2512.04678*, 2025.

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*, pp. 15086–15095, 2025.

Zichen Miao, Zhengyuan Yang, Kevin Lin, Ze Wang, Zicheng Liu, Lijuan Wang, and Qiang Qiu. Tuning timestep-distilled diffusion model using pairwise sample optimization. In *International Conference on Learning Representations*, 2025.

Erwann Millon. Krea Realtime 14B: Real-time video generation, 2025. URL `https://github.com/krea-ai/realtime-video`.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

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. *arXiv preprint arXiv:2511.01266*, 2025.

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

Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

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

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

Zehan Wang, Tengfei Wang, Haiyu Zhang, Xuhui Zuo, Junta Wu, Haoyuan Wang, Wenqiang Sun, Zhenwei Wang, Chenjie Cao, Hengshuang Zhao, et al. Worldcompass: Reinforcement learning for long-horizon world models. *arXiv preprint arXiv:2602.09022*, 2026.

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.

Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *European conference on computer vision*, pp. 720–736. Springer, 2022.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pp. 399–417. Springer, 2024.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

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

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

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

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

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22963–22974, 2025.

Jintao Zhang, Kai Jiang, Jintao Chen, Xu Wang, Deyuan Liu, Jungang Li, Dechuang Chen, Ming Lin, Jingjiang Zhou, Haopeng Jin, Qi Jia, Xiaohang Wang, Yaole Wang, Zhanqiang Zhang, Ran Li, Zhengkun Huang, Shuyue Xiong, Yuji Wang, Zikun Dai, Hui He, Yang Luo, Mang Ning, Weiqi Feng, Chengyang Ye, Xinyue Lin, Min Zhao, Hongzhou Zhu, Hengkai Tan, Zeyuan Wang, Chendong Xiang, Kaiwen Zheng, Zhijie Deng, Fan Bao, Jianfei Chen, and Jun Zhu. Vidu s2: Real-time interactive, editable, and spatial video generation. *arXiv preprint arXiv:2609.11638*, 2026.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. *arXiv preprint arXiv:2504.12626*, 2025.

Min Zhao, Hongzhou Zhu, Chendong Xiang, Kaiwen Zheng, Chongxuan Li, and Jun Zhu. Identifying and solving conditional image leakage in image-to-video diffusion model. *Advances in Neural Information Processing Systems*, 37:30300–30326, 2024.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

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

A Implementation Details and Hyperparameters

We provide the implementation configuration in Table S1. The notation follows the main paper: K denotes the number of inference steps, while $\mathcal{T}_{\text{distill}}$ denotes the retained timestep set.

Table S1: Implementation and hyperparameter configuration for our method. We detail the model, optimization, forward-process, reference-regularization, and streaming-rollout settings.

Module	Hyperparameter	Value
Model & Video Specs	Base Architecture	Causal Wan 2.1
	Video Resolution ($H \times W$)	480×832
LoRA Fine-Tuning	Rank (r)	256
	Scaling Factor (α)	256
	Dropout Rate	0.0
	Gradient Checkpointing	True
Optimization	Hardware	48 $\times$ NVIDIA GPUs
	Precision Mode	`bf16` (Mixed Precision)
	Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999$)
	Learning Rate (η)	1e-5
	Weight Decay	1e-4
	Epsilon (ϵ)	1e-8
	Max Gradient Norm	1.0
Diffusion Process	Number of Inference Steps (K)	4
	Distilled Timestep Set	$\{1000, 750, 500, 250\}$
	Normalized timestep	$t_k = \text{index}_k/1000$ in the forward path
	Timestep Shift	5.0
	Forward Process Noise Level	0.7
Reference Regularization	Interpolation Strength (β)	1.0
	Reference Regularization Weight (λ_{ref})	1e-4
	Advantage Normalization	Group-relative, $A_{\max} = \max_i \lvert A^{(i)} \rvert + \epsilon$
	Reward Mapping	$\tilde{r}_i = \frac{1}{2}(1 + \text{clip}(A^{(i)}/A_{\max}, -1, 1))$
	EMA Decay Rate (γ)	0.9
	EMA Update Interval	1 step
Streaming Rollout	Window Selection Strategy	Random Choice
	Rolling Window Size (L)	21 (Self-Forcing) / 15 (LongLive)
	Frame Sink Size (S)	3

B Baseline Adaptation Details

The comparison in Sec. 4.5 uses the same distilled Causal-Forcing checkpoint, 4-step rollout protocol, LoRA parameterization, prompt groups, reward models, and candidate budget for all methods. **DanceGRPO-AR** applies the DanceGRPO group-relative objective to the student's 4-step process while keeping the causal checkpoint and timestep schedule unchanged. **Flow-GRPO-AR** similarly adapts the Flow-GRPO objective to the same few-step causal setting, including any sampler conversion required by the original method. **DiffusionNFT** serves as the vanilla forward-process baseline using the same generated endpoints and reward normalization, but without explicitly matching the distilled timestep and streaming-context supports used by our method.

C Additional Analysis

This section provides additional analysis for the dual-support formulation. Rather than establishing global convergence guarantees, our goal is to clarify the behavior of the local objectives used in our method: why forward regression remains meaningful after timestep distillation, how distilled-flow mismatch enters the relative update, and what is optimized when long streaming histories are detached.

C.1 Training Dynamics Across Backbones

Figure S1 summarizes the reward-oriented training outcomes across the three distilled backbones. The dashed lines mark each backbone's score before alignment, while the colored bars show the

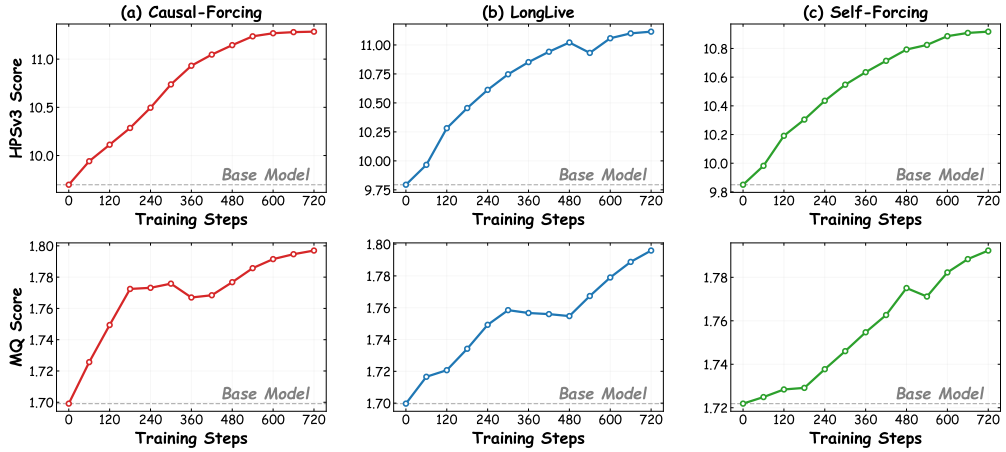


Figure S1: Cross-backbone reward-oriented training outcomes. We evaluate our method on three distilled models. Dashed gray lines indicate the scores of the respective base models before alignment; the colored bars show the post-training scores. Our method improves both HPSv3 and Motion Quality (MQ) across all three backbones.

corresponding post-training scores. Improvements in both HPSv3 and Motion Quality (MQ) across backbones indicate that the optimization signal transfers across model architectures rather than relying on a single backbone-specific failure mode. This comparison reports final training outcomes; it is not a per-step learning-curve analysis.

C.2 Endpoint Reweighting after Timestep Distillation

Fix a causal context C and conditioning signal c. Let $p_{\text{old}}(x_0 \mid C, c)$ denote any endpoint distribution, including one produced by a timestep-distilled policy. For $t \in \mathcal{T}_{\text{distill}}$, we use

$$x_t = (1-t)x_0 + t\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \qquad u_t = \epsilon - x_0. \tag{10}$$

For a non-negative endpoint weight $w(x_0, C, c)$, consider

$$\mathcal{L}_w(v) = \mathbb{E}\left[w(x_0, C, c)\, \|v(x_t, t, C, c) - u_t\|_2^2\right]. \tag{11}$$

Conditioned on $z = (x_t, t, C, c)$, define $m(z) = \mathbb{E}[w \mid z]$ and $q(z) = \mathbb{E}[w u_t \mid z]$. The conditional objective becomes

$$\mathcal{L}_w(v \mid z) = \mathbb{E}\left[w\left(\|v\|_2^2 - 2v^\top u_t + \|u_t\|_2^2\right) \middle| z\right] \tag{12}$$

$$= m(z)\|v\|_2^2 - 2v^\top q(z) + \mathbb{E}[w\|u_t\|_2^2 \mid z]. \tag{13}$$

The last term is independent of v. Differentiating the remaining terms gives

$$\nabla_v \mathcal{L}_w(v \mid z) = 2m(z)v - 2q(z). \tag{14}$$

Therefore, whenever $m(z) > 0$, the unique pointwise optimum is

$$v_w^*(x_t, t, C, c) = \frac{\mathbb{E}[w\, u_t \mid x_t, t, C, c]}{\mathbb{E}[w \mid x_t, t, C, c]}. \tag{15}$$

The Hessian is $2m(z)I \succ 0$, so the solution is unique.

This optimum admits a direct endpoint-reweighting interpretation. Define

$$p_w(x_0 \mid C, c) = \frac{w(x_0, C, c)p_{\text{old}}(x_0 \mid C, c)}{Z_w(C, c)}, \qquad Z_w(C, c) = \mathbb{E}_{p_{\text{old}}}[w \mid C, c]. \tag{16}$$

Under this reweighted endpoint law and the same forward corruption kernel, the conditional expectation of u_t is exactly Eq. equation 15. Hence reward-weighted forward regression remains well defined even when the endpoint distribution comes from a timestep-distilled student.

Importantly, this argument does not require the distilled model to recover the teacher's original dense probability-flow trajectory. It only requires forward-corrupted states constructed from endpoints sampled by the current policy. Restricting t to $\mathcal{T}_{\text{distill}}$ is a separate inference-matching choice that concentrates these updates on timesteps used by the final sampler.

C.3 Relative Update under Distilled-Flow Mismatch

We next analyze the exact relative objective used in the main method. Fix (x_t, t, C, c) and denote

$$v_0 = v_{\text{old}}(x_t, t, C, c), \qquad \delta = v_\theta - v_0, \qquad a = 2\tilde{r} - 1.$$

Since $v^+ = v_0 + \beta\delta$ and $v^- = v_0 - \beta\delta$, the conditional objective is

$$\mathcal{L}(\delta) = \mathbb{E}\left[\frac{1+a}{2}\|v_0 + \beta\delta - u_t\|_2^2 + \frac{1-a}{2}\|v_0 - \beta\delta - u_t\|_2^2 \middle| x_t, t, C, c\right]. \tag{17}$$

Derivation. Let $e = u_t - v_0$. Then

$$v_0 + \beta\delta - u_t = \beta\delta - e, \qquad v_0 - \beta\delta - u_t = -\beta\delta - e.$$

Expanding the two squared terms gives

$$\|\beta\delta - e\|_2^2 = \beta^2\|\delta\|_2^2 + \|e\|_2^2 - 2\beta\delta^\top e, \tag{18}$$

$$\|-\beta\delta - e\|_2^2 = \beta^2\|\delta\|_2^2 + \|e\|_2^2 + 2\beta\delta^\top e. \tag{19}$$

Substituting these expressions into Eq. equation 17, the coefficients of the common terms sum to one, while the two cross terms combine according to

$$-\frac{1+a}{2}(2\beta\delta^\top e) + \frac{1-a}{2}(2\beta\delta^\top e) = -2\beta a\,\delta^\top e.$$

Therefore,

$$\mathcal{L}(\delta) = \mathbb{E}\left[\|e\|_2^2 + \beta^2\|\delta\|_2^2 - 2\beta a\,\delta^\top e \mid x_t, t, C, c\right]. \tag{20}$$

Differentiating with respect to δ yields

$$\nabla_\delta \mathcal{L} = 2\beta^2\delta - 2\beta\mathbb{E}[ae \mid x_t, t, C, c]. \tag{21}$$

Since the Hessian is $2\beta^2 I \succ 0$ for $\beta > 0$, the unique population-optimal perturbation is

$$\delta^* = \frac{1}{\beta}\mathbb{E}\left[a(u_t - v_0) \mid x_t, t, C, c\right]. \tag{22}$$

To make the effect of timestep distillation explicit, define

$$\mu = \mathbb{E}[u_t \mid x_t, t, C, c], \qquad b = v_0 - \mu. \tag{23}$$

Then

$$\mathbb{E}[a(u_t - v_0) \mid x_t, t, C, c] = \mathbb{E}[a(u_t - \mu - b) \mid x_t, t, C, c] \tag{24}$$

$$= \mathbb{E}[a(u_t - \mu) \mid x_t, t, C, c] - \mathbb{E}[a \mid x_t, t, C, c]\, b. \tag{25}$$

Because μ is the conditional mean of u_t,

$$\mathbb{E}[a(u_t - \mu) \mid x_t, t, C, c] = \mathbb{E}[au_t \mid x_t, t, C, c] - \mathbb{E}[a \mid x_t, t, C, c]\mathbb{E}[u_t \mid x_t, t, C, c] \tag{26}$$

$$= \mathrm{Cov}\left(a, u_t \mid x_t, t, C, c\right). \tag{27}$$

Hence

$$\delta^* = \frac{1}{\beta}\,\mathrm{Cov}\left(a, u_t \mid x_t, t, C, c\right) - \frac{\mathbb{E}[a \mid x_t, t, C, c]}{\beta} b. \tag{28}$$

The first term captures reward-dependent variation in the forward target. The second term explicitly shows how mismatch $b = v_0 - \mu$ between the distilled predictor and the corresponding forward-regression optimum enters the update. The derivation does not assume that v_0 remains an exact dense-time probability-flow field.

C.4 Effect of Group-Relative Centering

For a group of G candidates sharing the same prompt and causal context, we define

$$A^{(i)} = R^{(i)} - \frac{1}{G}\sum_{j=1}^{G} R^{(j)}. \tag{29}$$

By construction,

$$\sum_{i=1}^{G} A^{(i)} = \sum_{i} R^{(i)} - G\left(\frac{1}{G}\sum_{j} R^{(j)}\right) = 0. \tag{30}$$

The same construction is invariant to any additive reward bias shared by the group. If $\widehat{R}^{(i)} = R^{(i)} + b(C, c)$, then

$$\widehat{A}^{(i)} = R^{(i)} + b(C,c) - \frac{1}{G}\sum_{j}\Big(R^{(j)} + b(C,c)\Big) \tag{31}$$

$$= R^{(i)} - \frac{1}{G}\sum_{j} R^{(j)} = A^{(i)}. \tag{32}$$

Under the normalization in the main paper, $A_{\max} = \max_j |A^{(j)}| + \epsilon$, so

$$a_i = 2\tilde{r}_i - 1 = \frac{A^{(i)}}{A_{\max}},$$

and therefore $\sum_i a_i = 0$ within each sampled group.

However, this sample-wise identity does not imply exact conditional centering. In general, $\mathbb{E}[a \mid x_t, t, C, c]$ need not vanish because conditioning on the corrupted state changes the distribution over candidates. From Eq. equation 28, if

$$|\mathbb{E}[a \mid x_t, t, C, c]| \le \varepsilon_a, \tag{33}$$

then the mismatch component satisfies

$$\left\| \frac{\mathbb{E}[a \mid x_t, t, C, c]}{\beta} b \right\|_2 \le \frac{\varepsilon_a}{\beta}\|b\|_2. \tag{34}$$

Thus group-relative centering exactly removes group-shared additive reward offsets and empirically suppresses, but does not guarantee cancellation of, the distilled-flow mismatch term.

C.5 Detached Streaming Optimization

We next clarify the objective optimized when long streaming histories are generated by a lagged behavior policy and detached before the active-window update.

Let Ξ_n collect all random variables sampled for the local update at generation step n, including the streaming context, candidate continuations, rewards, sampled timesteps, and forward noise. Let $q^{(n)}_{\theta_{\text{old}}}(\Xi_n \mid c)$ denote their joint distribution under the fixed behavior policy. Define

$$\mathcal{J}_{\text{local}}(\theta; \theta_{\text{old}}) = \mathbb{E}_{\Xi_n \sim q^{(n)}_{\theta_{\text{old}}}}\left[\ell_\theta(\Xi_n)\right]. \tag{35}$$

Gradient of the detached surrogate. During one local update, θ_{old} and the sampled Ξ_n are treated as fixed. Therefore the sampling distribution does not depend on the trainable parameter θ:

$$\nabla_\theta q^{(n)}_{\theta_{\text{old}}}(\Xi_n \mid c) = 0. \tag{36}$$

Under the usual regularity conditions that permit exchanging differentiation and expectation,

$$\nabla_\theta \mathcal{J}_{\text{local}} = \nabla_\theta \int q^{(n)}_{\theta_{\text{old}}}(\Xi_n \mid c)\ell_\theta(\Xi_n)\, d\Xi_n \tag{37}$$

$$= \int q^{(n)}_{\theta_{\text{old}}}(\Xi_n \mid c)\nabla_\theta \ell_\theta(\Xi_n)\, d\Xi_n \tag{38}$$

$$= \mathbb{E}_{\Xi_n \sim q^{(n)}_{\theta_{\text{old}}}}\left[\nabla_\theta \ell_\theta(\Xi_n)\right]. \tag{39}$$

Thus stochastic gradients computed from detached rollouts are unbiased estimators of the gradient of this fixed-behavior local surrogate.

This is different from a fully on-policy long-horizon objective. If the context distribution itself were generated by the current policy,

$$\mathcal{J}_{\mathrm{on}}(\theta) = \mathbb{E}_{\Xi_n \sim q_\theta^{(n)}}[\ell_\theta(\Xi_n)],$$

then its gradient would contain both a local loss derivative and a derivative through the sampling distribution:

$$\nabla_\theta \mathcal{J}_{\mathrm{on}} = \mathbb{E}_{q_\theta^{(n)}}\left[\nabla_\theta \ell_\theta(\Xi_n) + \ell_\theta(\Xi_n)\nabla_\theta \log q_\theta^{(n)}(\Xi_n)\right]. \tag{40}$$

Our detached-history update intentionally omits the second term. It therefore provides *long-horizon state exposure* while retaining *local credit assignment*, rather than performing full sequence-level policy optimization.

C.6 Why Policy-Generated Histories Matter

For two autoregressive trajectory distributions,

$$P_\theta(x_{1:N} \mid c) = \prod_{n=1}^{N} p_\theta(x_n \mid x_{<n}, c), \qquad P_0(x_{1:N} \mid c) = \prod_{n=1}^{N} p_0(x_n \mid x_{<n}, c), \tag{41}$$

the trajectory-level KL is

$$D_{\mathrm{KL}}(P_\theta \| P_0) = \mathbb{E}_{P_\theta}\left[\log \frac{P_\theta(x_{1:N} \mid c)}{P_0(x_{1:N} \mid c)}\right]. \tag{42}$$

Using the autoregressive factorization,

$$\log \frac{P_\theta(x_{1:N} \mid c)}{P_0(x_{1:N} \mid c)} = \log \prod_{n=1}^{N} \frac{p_\theta(x_n \mid x_{<n}, c)}{p_0(x_n \mid x_{<n}, c)} \tag{43}$$

$$= \sum_{n=1}^{N} \log \frac{p_\theta(x_n \mid x_{<n}, c)}{p_0(x_n \mid x_{<n}, c)}. \tag{44}$$

Taking expectation under P_θ gives

$$D_{\mathrm{KL}}(P_\theta \| P_0) = \sum_{n=1}^{N} \mathbb{E}_{x_{1:n} \sim P_\theta}\left[\log \frac{p_\theta(x_n \mid x_{<n}, c)}{p_0(x_n \mid x_{<n}, c)}\right] \tag{45}$$

$$= \sum_{n=1}^{N} \mathbb{E}_{x_{<n} \sim P_\theta}\left[D_{\mathrm{KL}}\left(p_\theta(\cdot \mid x_{<n}, c) \| p_0(\cdot \mid x_{<n}, c)\right)\right]. \tag{46}$$

The second equality follows from the tower property by first conditioning on $x_{<n}$.

Equation equation 46 is not our optimization objective. Its role is interpretive: conditional deviations between causal policies are naturally averaged over histories visited by the policy. This motivates exposing post-training to policy-generated streaming contexts rather than only to repeatedly reset prefixes.

D Efficiency and Memory Analysis

D.1 Memory Scaling with Detached Rolling Contexts

Suppose the active streaming cache retains S sink clips and L recent clips, and each retained clip requires at most M_{KV} memory. The historical cache therefore satisfies

$$M_{\mathrm{hist}} \le (S + L) M_{\mathrm{KV}}. \tag{47}$$

Because S and L are fixed, this term is independent of the total number of generated clips N.

Let M_{fixed} denote model parameters and optimizer states, and let M_{active} denote gradient-bearing activations for the current optimization window and candidate group. Peak memory can then be written as

$$M_{\text{peak}} = O\left(M_{\text{fixed}} + (S+L)M_{\text{KV}} + M_{\text{active}}\right). \tag{48}$$

This bounded-history statement relies on two distinct operations. Detaching historical states prevents the backward graph from growing with rollout length, while rolling eviction prevents stale KV tensors from accumulating in memory. Detachment alone would not bound stored cache memory if all historical KV tensors were retained.

D.2 Limitation of Post-Distillation Reweighting

Consider a distilled endpoint distribution p_D and its exponentially reweighted form

$$p_\beta(x) = \frac{p_D(x)e^{\beta R(x)}}{Z_\beta}, \qquad Z_\beta = \int p_D(x)e^{\beta R(x)}dx, \tag{49}$$

where $R(x)$ is finite and $0 < Z_\beta < \infty$.

For any measurable region A such that $p_D(A) = 0$,

$$p_\beta(A) = \int_A \frac{p_D(x)e^{\beta R(x)}}{Z_\beta}dx \tag{50}$$

$$= 0. \tag{51}$$

Conversely, because $e^{\beta R(x)}/Z_\beta > 0$ wherever $p_D(x) > 0$, a region with positive p_D mass also retains positive mass after finite reweighting. Thus positive reward reweighting preserves the zero-probability regions of the original distilled distribution.

This result should be interpreted narrowly. It does not imply that online optimization necessarily reduces diversity, nor does it prevent large changes in the relative probability of modes already represented by the student. It only shows that purely on-policy reweighting cannot directly recover a region assigned exactly zero probability by the distilled model.

D.3 Scope of Reference Regularization

The disagreement-gated reference penalty is a practical stabilizer rather than a theoretical KL constraint. Our regularizer operates directly in velocity space,

$$\mathcal{L}_{\text{ref}} \propto \|v_\theta - v_{\theta_{\text{ref}}}\|_2^2, \tag{52}$$

whereas a distributional KL is defined on the endpoint or trajectory distribution induced by the complete generative process. Without additional assumptions relating perturbations of the velocity field to changes in the induced probability law, these quantities are not generally equivalent.

We therefore interpret the reference term only as a selective surrogate for limiting policy drift on high-disagreement samples. It does not constitute an exact trust-region constraint and does not imply monotonic reward improvement.

E Additional Ablation Studies

The following studies quantify implementation and stabilization choices that are intentionally kept out of the main ablation, whose purpose is to test the dual-support hypothesis.

E.1 Streaming Optimization Scheme

The streaming optimization strategy is summarized in the main ablation table 5. The detached clip-level scheme provides the practical memory/quality trade-off used in our long-horizon experiments.

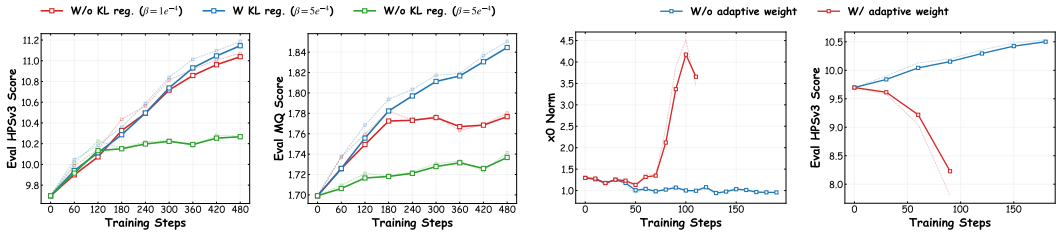
E.2 Reward Design

The multi-reward variants quantify the trade-off between visual quality, motion quality, and text–video alignment.



(a) Multi-reward design (b) Different β values

Figure S2: **Ablation studies on reward formulation and interpolation strength. (a)** Optimizing a single objective induces reward hacking and degrades other quality dimensions, whereas our aggregated formulation balances visual aesthetics and motion consistency. **(b)** The parameter β controls the implicit contrast between positive and negative samples, with $\beta = 1.0$ yielding the optimal trade-off for quality.



(a) Ablation on selective reference regularization. (b) Ablation on adaptive weight.

Figure S3: Ablation study of stabilization techniques in our method. (a) Effect of the selective reference regularization penalty. (b) Impact of the dynamic adaptive weight mechanism.

E.3 Reference Regularization

The reference-regularization variants in Figure S3 quantify stabilization effects. We refer to the velocity-space penalty as reference regularization rather than an exact KL constraint; its scope is discussed in Sec. D.3.

E.4 Adaptive Weighting and Effect of β

Figure S3b reports the effect of removing the adaptive x_0 weighting used by DiffusionNFT. Figure S2 reports the sensitivity to the interpolation strength β.

F Evaluation Details

F.1 Prompt Sets and Augmentation

The short-video evaluation uses the 946-prompt VBench set and the augmented Self-Forcing prompt set described in Sec. 4.2. The held-out reward-oriented check uses 100 diverse MovieGenBench prompts Polyak et al. (2024).

F.2 VBench and VBench-Long

We use the standard VBench protocol for 5-second videos and VBench-Long for 30-second single-prompt videos. Self-Forcing and Causal-Forcing use Infinity-RoPE for the long-horizon protocol, while LongLive uses its native streaming support.

Table S2: Hyperparameter sensitivity analysis on group size G, window size L, KL weight λ, and interpolation strength β.

G	HPS	MQ	L	HPS	MQ	λ_{ref}	HPS	MQ	β	HPS	MQ
8	10.42	1.73	16	10.51	1.72	0.001	10.71	1.77	0.3	10.35	1.74
16	10.68	1.77	32	10.84	1.80	0.01	10.84	1.80	0.5	10.84	1.80
24	10.84	1.80	48	10.87	1.81	0.05	10.62	1.74	0.7	10.71	1.76
32	10.86	1.79	64	10.88	1.81	0.1	10.28	1.68	0.9	10.52	1.71

Table S3: **Illustrative human preference evaluation (provisional placeholders).** Each row reports majority-vote outcomes over 100 held-out prompts, with three independent raters per prompt. "Win" denotes a prompt for which the majority preferred Ours, and "Loss" denotes a majority preference for the baseline. The win rate is computed after excluding ties; the interval is a prompt-level bootstrap 95% confidence interval. These values are predicted placeholders and must be replaced by measured results before submission.

Criterion	Win	Tie	Loss	Win Rate	95% CI
Visual quality	67	18	15	81.7%	[72.8, 89.4]
Motion quality	72	14	14	83.7%	[75.2, 90.9]
Text alignment	58	24	18	76.3%	[65.6, 84.8]
Long-term consistency	70	16	14	83.3%	[74.1, 90.2]
Overall preference	73	13	14	83.9%	[74.7, 90.1]

F.3 Multi-Prompt Evaluation

Multi-prompt evaluation uses 100 groups of six 10-second prompts and reports clip-wise scores at the prompt boundaries. Clip-wise CLIP measures instruction following for each segment; it does not by itself establish long-term temporal consistency.

F.4 Human Preference Evaluation

To complement reward-model scores, we plan a blinded pairwise evaluation on 100 held-out prompts that are not used for training or hyperparameter tuning. For each prompt, we compare Ours with the strongest baseline using the same generated videos and randomly swap the left/right presentation order. Three independent raters judge visual quality, motion quality, text alignment, long-term consistency, and overall preference without seeing method names. We aggregate the three judgments by majority vote per prompt and report win/tie/loss counts. The win rate excludes ties, and confidence intervals are obtained by bootstrapping prompts rather than individual ratings. Table S3 contains provisional predicted values to define the reporting format; all entries must be replaced with measured human-study results before submission.

G Training Algorithms

We provide detailed pseudocode for the training procedures. Algorithm 1 outlines the core forward-process reinforcement learning pipeline with selective reference regularization. It details the multi-reward evaluation, uncertainty quantification via rank disagreement, and the selective application of reference regularization to mitigate reward hacking during short-clip optimization. To scale this alignment to long-sequence generation without exceeding memory constraints, Algorithm 2 presents the streaming training procedure for long video. This algorithm leverages a rolling KV cache and detaches historical context, applying gradient updates exclusively to local active windows while maintaining long-range temporal coherence.

H Additional Qualitative Results

We present additional qualitative comparisons on 30-second long video generation under the single-prompt setting in Figures **??**–S28. Qualitative comparisons showing improvements over baseline

Algorithm 1 NFT Training with Selective Reference Regularization

Require: Policy π_θ, behavior policy $\pi_{\theta_{\text{old}}}$, reference policy $\pi_{\theta_{\text{ref}}}$; reward functions $\{R_m\}_{m=1}^{M}$; prompts $\mathcal{D}$; hyperparameters β, λ_{ref}, $K_{\max}$, τ_{ref}
Ensure: Optimized policy π_θ

Initialize $\theta_{\text{old}} \leftarrow \theta$, $\theta_{\text{ref}} \leftarrow \theta$
Initialize risk buffer $\mathcal{B} \leftarrow \emptyset$, risk ratio $\rho \leftarrow \rho_0$, $k_{\text{last}} \leftarrow 0$
for each epoch k **do**
  // Phase 1: Rollout and multi-reward evaluation
  Sample prompts $\{c_i\}_{i=1}^{B} \sim \mathcal{D}$
  Generate samples $\{x_0^{(i)}\}$ using $\pi_{\theta_{\text{old}}}$
  Compute rewards $\{r_m^{(i)} = R_m(x_0^{(i)}, c_i)\}$ for each reward model m
  Compute advantages $\{A^{(i)}\}$ from primary reward
  // Compute reward uncertainty via rank disagreement
  for each reward model m **do**
    $\text{rank}_m^{(i)} \leftarrow \text{Rank}(\{r_m^{(j)}\}_{j=1}^{B})$ ▷ Rank samples by reward m
  end for
  $\Delta_{\text{rank}}^{(i)} \leftarrow \text{rank}_{\text{primary}}^{(i)} - \frac{1}{M-1}\sum_{m \neq \text{primary}} \text{rank}_m^{(i)}$
  // Adaptive threshold via risk compensation
  $\rho \leftarrow \textsc{UpdateRiskRatio}(\mathcal{B}, \{\Delta_{\text{rank}}^{(i)}\}, \rho)$
  $\tau \leftarrow \text{Percentile}(\{\Delta_{\text{rank}}^{(i)} | \Delta_{\text{rank}}^{(i)} \geq 0\}, 100 - 100\rho)$
  $\mathcal{M}^{(i)} \leftarrow \mathbb{1}[\Delta_{\text{rank}}^{(i)} > \tau]$ ▷ High-uncertainty mask
  // Phase 2: Forward process optimization
  for each mini-batch $(x_0, A, \mathcal{M})$ **do**
    Sample $t \sim \mathcal{U}(\mathcal{T}_{\text{distill}})$
    $\epsilon \sim \mathcal{N}(0, I)$; $x_t \leftarrow (1-t)x_0 + t\epsilon$; $u_t \leftarrow \epsilon - x_0$
    $v_\theta, v_{\theta_{\text{old}}}, v_{\theta_{\text{ref}}} \leftarrow \pi_\theta(x_t, t), \pi_{\theta_{\text{old}}}(x_t, t), \pi_{\theta_{\text{ref}}}(x_t, t)$
    // Policy loss
    $v^+ \leftarrow \beta \cdot v_\theta + (1-\beta) \cdot v_{\theta_{\text{old}}}$
    $v^- \leftarrow (1+\beta) \cdot v_{\theta_{\text{old}}} - \beta \cdot v_\theta$
    $\tilde{r} \leftarrow \text{clip}(A/A_{\max})/2 + 0.5$
    $\mathcal{L}_{\text{policy}} \leftarrow \tilde{r}\|v^+ - u_t\|^2 + (1-\tilde{r})\|v^- - u_t\|^2$
    // Selective reference regularization: only on high-uncertainty samples
    $\mathcal{L}_{\text{ref}} \leftarrow \frac{1}{|\mathcal{M}|}\sum_{i:\mathcal{M}^{(i)}=1} \|v_\theta^{(i)} - v_{\theta_{\text{ref}}}^{(i)}\|^2$
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{policy}} + \lambda_{\text{ref}} \cdot \mathcal{L}_{\text{ref}}$
    Update: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$
  end for
  // Adaptive reference-policy update
  if $\mathcal{L}_{\text{ref}} > \tau_{\text{ref}}$ **or** $k - k_{\text{last}} > K_{\max}$ **then**
    $\theta_{\text{ref}} \leftarrow \theta$ ▷ Reset reference policy
    $k_{\text{last}} \leftarrow k$
  end if
  $\theta_{\text{old}} \leftarrow \gamma \cdot \theta_{\text{old}} + (1-\gamma) \cdot \theta$
end for

distilled models and competing methods. As shown in Figure S4, our method consistently produces higher-quality and more temporally coherent videos compared to the Krea 14B baseline Millon (2025), demonstrating that our approach yields notable improvements even when applied on top of large-scale 14B models.

I Limitations and Failure Cases

Finite post-distillation reweighting cannot assign mass to regions that have zero probability under the distilled student, so the method cannot recover missing modes without additional exploration or data. In practice, moving too far from the reference policy can also expose reward-model weaknesses;

Algorithm 2 Streaming Training for Long Video

Require: Policy π_θ, behavior policy $\pi_{\theta_{\text{old}}}$, reference policy $\pi_{\theta_{\text{ref}}}$; rewards $\{R_m\}$; prompts $\mathcal{D}$; window size W, total frames F

Ensure: Optimized policy π_θ

Initialize $\theta_{\text{old}} \leftarrow \theta$, $\theta_{\text{ref}} \leftarrow \theta$, risk buffer $\mathcal{B} \leftarrow \emptyset$, $\rho \leftarrow \rho_0$
for each epoch k **do**
  // Select training window (synced across GPUs)
  $s \leftarrow \textsc{SelectWindow}(F, W), \quad F_{\text{req}} \leftarrow s + W$
  // Phase 1: Efficient rollout with multi-reward
  Sample prompts $\{c_i\}_{i=1}^{B} \sim \mathcal{D}$
  Generate partial videos $\{x_0^{(i)}[0 : F_{\text{req}}]\}$ using $\pi_{\theta_{\text{old}}}$
  Extract window: $x_{0,W}^{(i)} \leftarrow x_0^{(i)}[s : s + W]$
  Compute rewards $\{r_m^{(i)} = R_m(x_{0,W}^{(i)}, c_i)\}$ for each model m
  Compute advantages $\{A^{(i)}\}$
  // Reward uncertainty via rank disagreement
  for each reward model m **do**
    $\text{rank}_m^{(i)} \leftarrow \text{Rank}(\{r_m^{(j)}\}_{j=1}^{B})$
  end for
  $\Delta_{\text{rank}}^{(i)} \leftarrow \text{rank}_{\text{primary}}^{(i)} - \frac{1}{M-1} \sum_{m \neq \text{primary}} \text{rank}_m^{(i)}$
  $\rho \leftarrow \textsc{UpdateRiskRatio}(\mathcal{B}, \{\Delta_{\text{rank}}^{(i)}\}, \rho)$
  $\tau \leftarrow \text{Percentile}(\{\Delta_{\text{rank}}^{(i)} \geq 0\}, 100 - 100\rho)$
  $\mathcal{M}^{(i)} \leftarrow \mathbb{1}[\Delta_{\text{rank}}^{(i)} > \tau]$
  // Phase 2: Window-focused optimization with selective reference regularization
  for each mini-batch $(x_0, A, \mathcal{M})$ **do**
    Sample $t \sim \mathcal{U}(\mathcal{T}_{\text{distill}})$ and $\epsilon \sim \mathcal{N}(0, I)$; $x_t \leftarrow (1-t)x_0 + t\epsilon$; $u_t \leftarrow \epsilon - x_0$
    $v_\theta, v_{\theta_{\text{old}}}, v_{\theta_{\text{ref}}} \leftarrow \textsc{StreamingNFT}(x_t, t, s, W)$
    Extract: $v_{\theta,W}, v_{\theta_{\text{old}},W}, v_{\theta_{\text{ref}},W}, u_{t,W}$
    // Policy loss on window
    $v^+ \leftarrow \beta \cdot v_{\theta,W} + (1-\beta) \cdot v_{\theta_{\text{old}},W}$
    $v^- \leftarrow (1+\beta) \cdot v_{\theta_{\text{old}},W} - \beta \cdot v_{\theta,W}$
    $\tilde{r} \leftarrow \text{clip}(A/A_{\max})/2 + 0.5$
    $\mathcal{L}_{\text{policy}} \leftarrow \tilde{r}\|v^+ - u_{t,W}\|^2 + (1-\tilde{r})\|v^- - u_{t,W}\|^2$
    // Selective reference regularization on high-uncertainty samples
    $\mathcal{L}_{\text{ref}} \leftarrow \frac{1}{|\mathcal{M}|} \sum_{i:\mathcal{M}^{(i)}=1} \|v_{\theta,W}^{(i)} - v_{\theta_{\text{ref}},W}^{(i)}\|^2$
    Update: $\theta \leftarrow \theta - \eta \nabla_\theta (\mathcal{L}_{\text{policy}} + \lambda_{\text{ref}} \mathcal{L}_{\text{ref}})$
  end for
  // Adaptive reference update
  if $\mathcal{L}_{\text{ref}} > \tau_{\text{ref}}$ **or** $k - k_{\text{last}} > K_{\max}$ **then**
    $\theta_{\text{ref}} \leftarrow \theta, \quad k_{\text{last}} \leftarrow k$
  end if
  $\theta_{\text{old}} \leftarrow \gamma \cdot \theta_{\text{old}} + (1-\gamma) \cdot \theta$
end for

the multi-reward score and disagreement-gated reference regularization are stabilizers rather than guarantees against reward hacking. Finally, clip-wise evaluation across prompt boundaries does not establish global temporal consistency for 60-second videos.

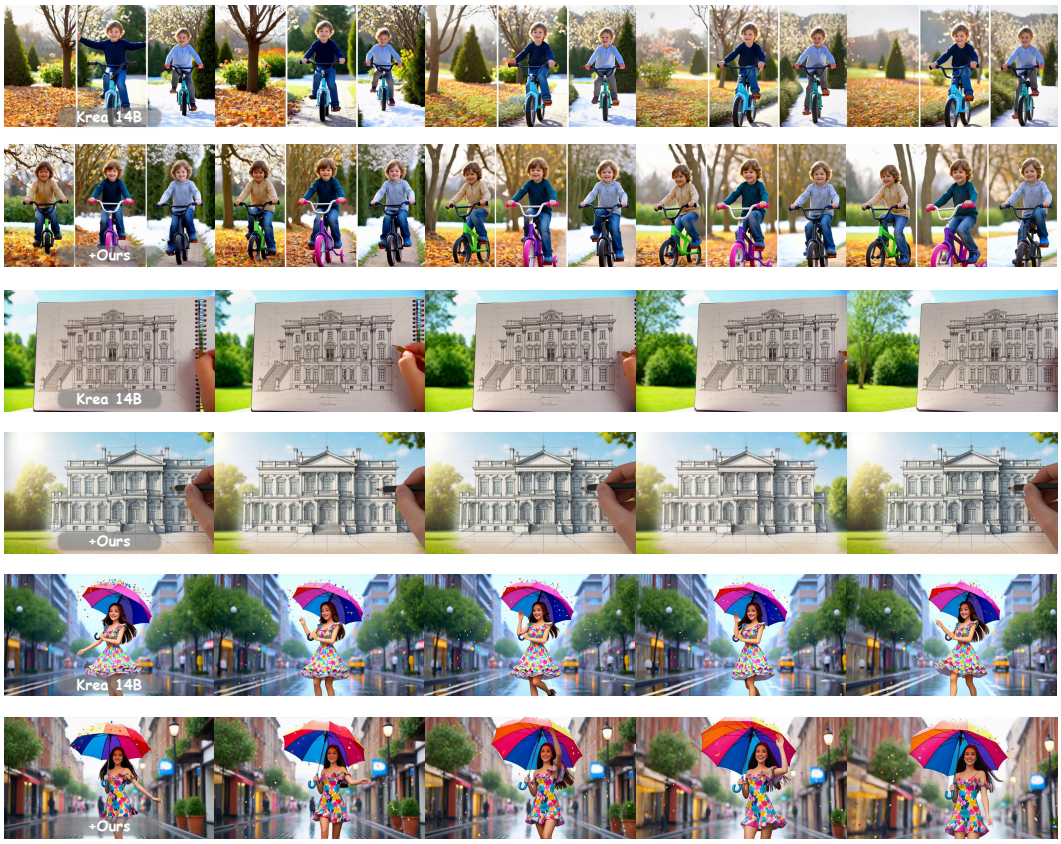


Figure S4: Qualitative comparison between Krea 14B (odd rows) and with our method (+Ours, even rows)

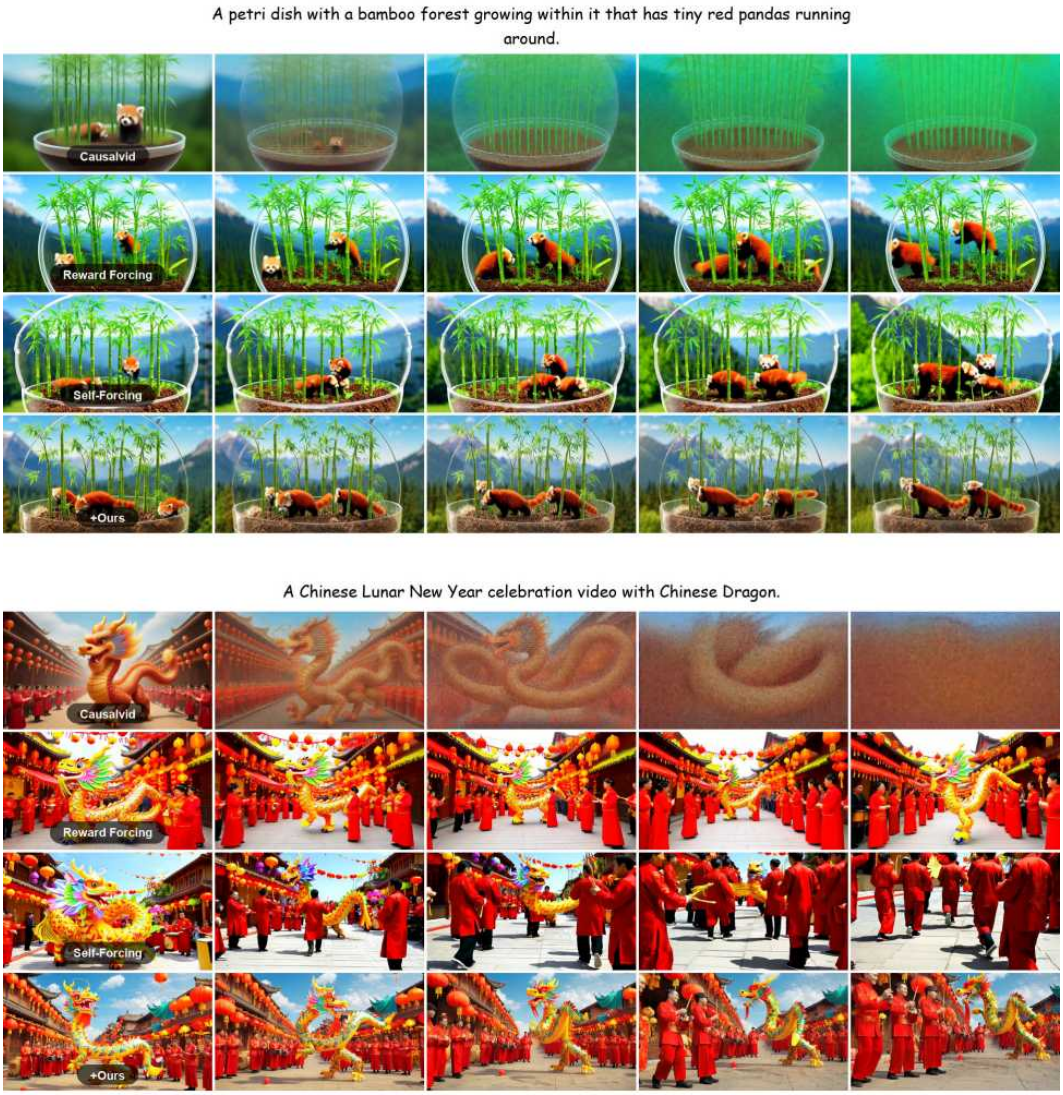


Figure S5: Additional qualitative comparison on long video generation (30s) under the single-prompt setting (Set 1).

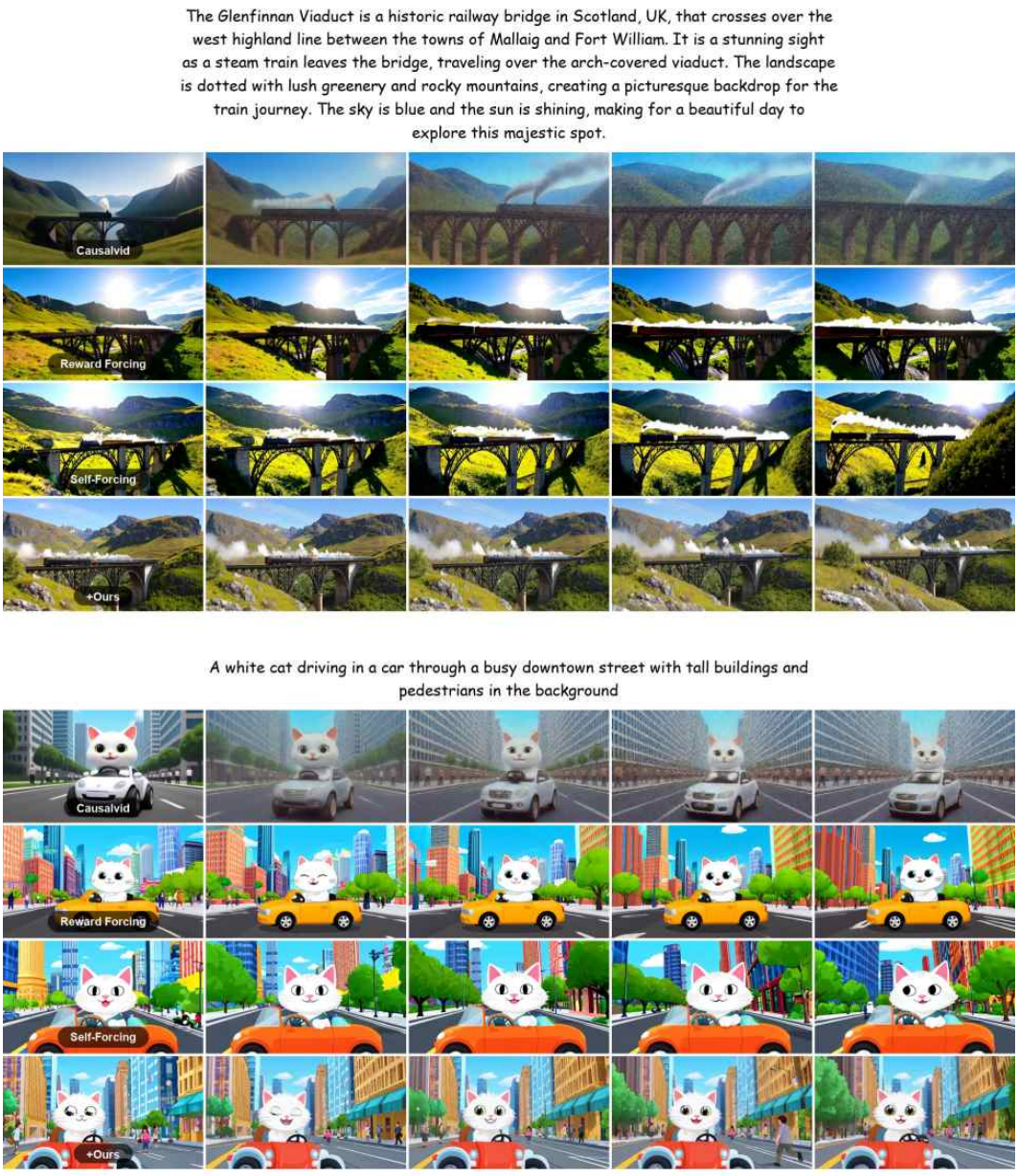


Figure S6: Additional qualitative comparison on long video generation (30s) under the single-prompt setting (Set 2).

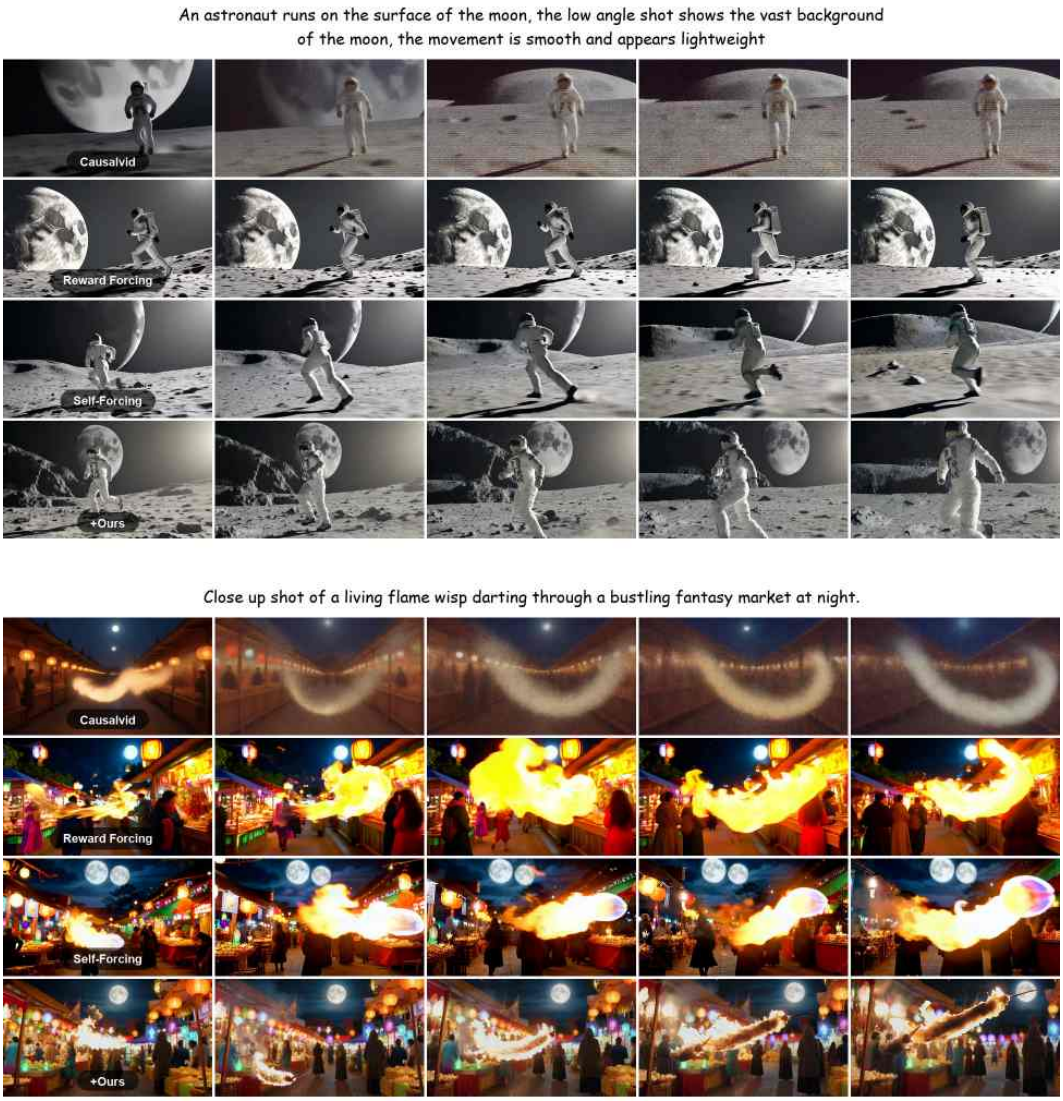


Figure S7: Additional qualitative comparison on long video generation (30s) under the single-prompt setting (Set 3).

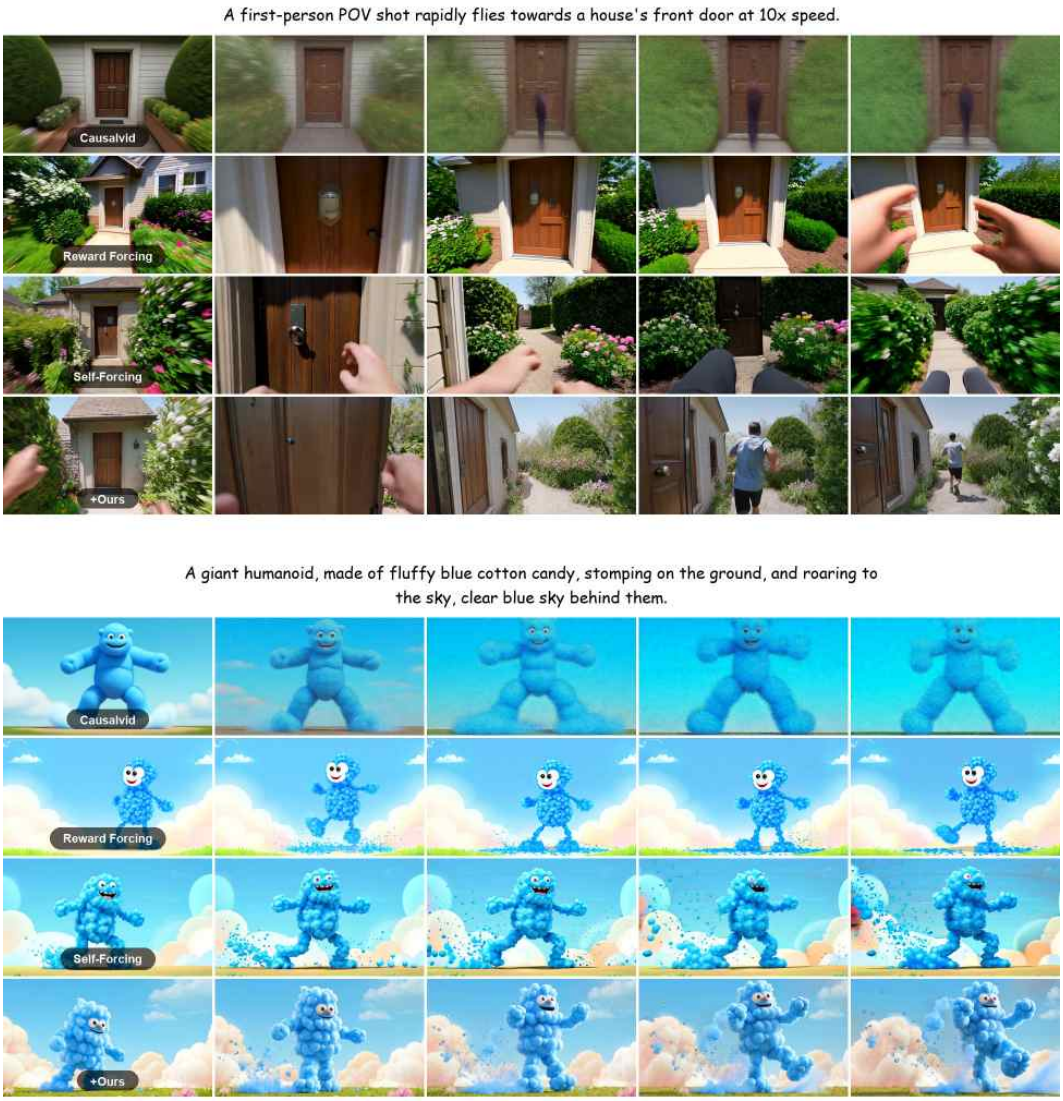


Figure S8: Additional qualitative comparison on long video generation (30s) under the single-prompt setting (Set 4).

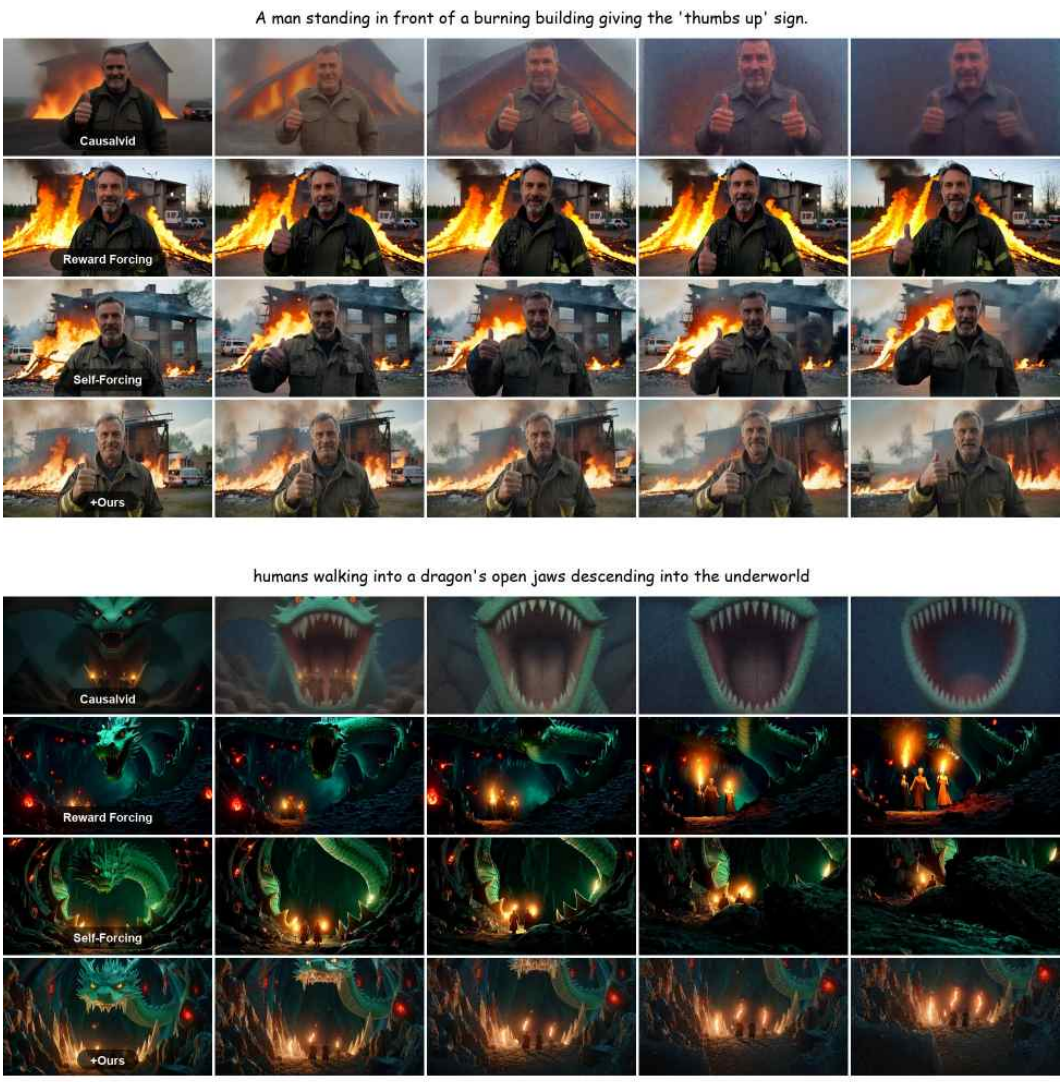


Figure S9: Additional qualitative comparison on long video generation (30s) under the single-prompt setting (Set 5).

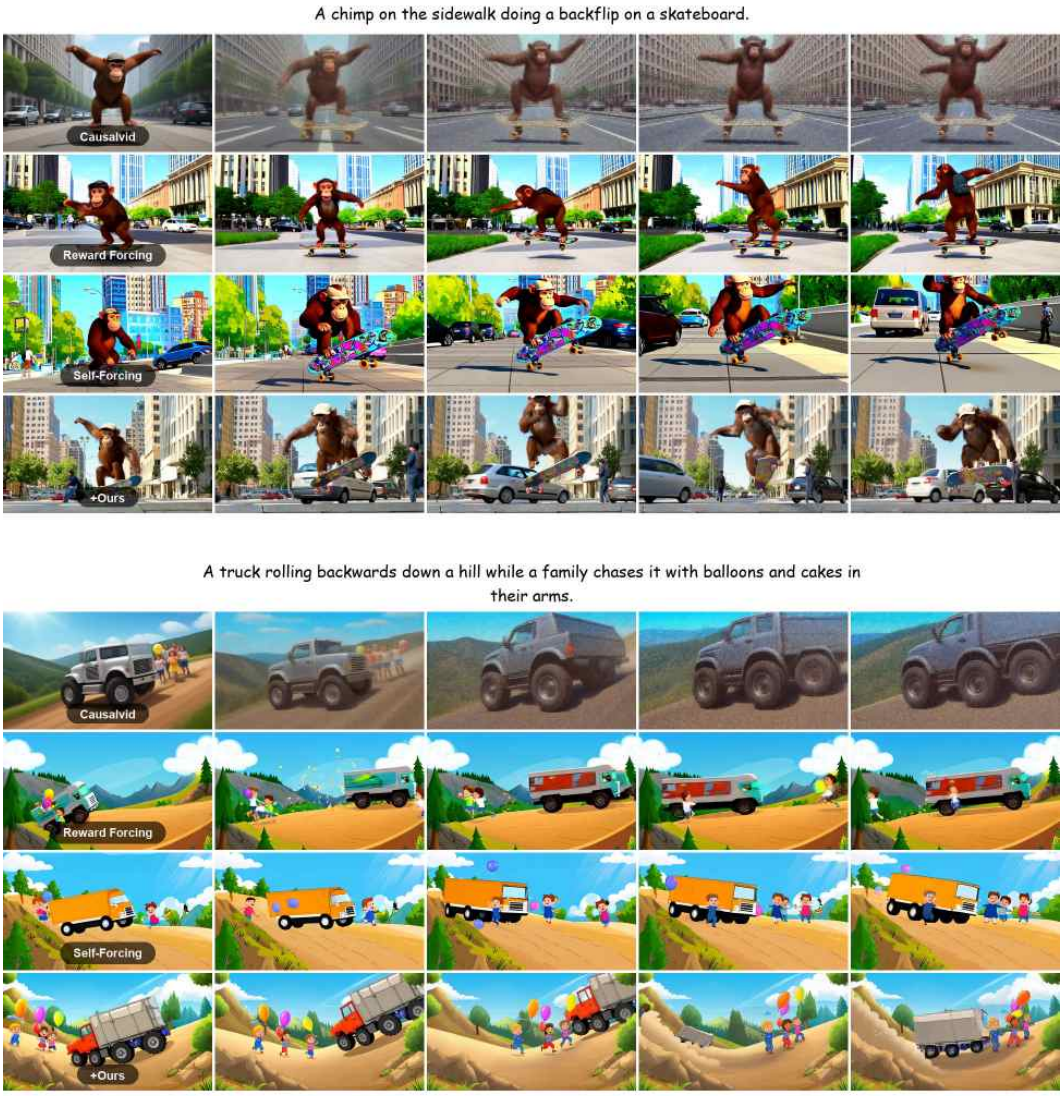


Figure S10: Additional qualitative comparison on long video generation (30s) under the single-prompt setting (Set 6).

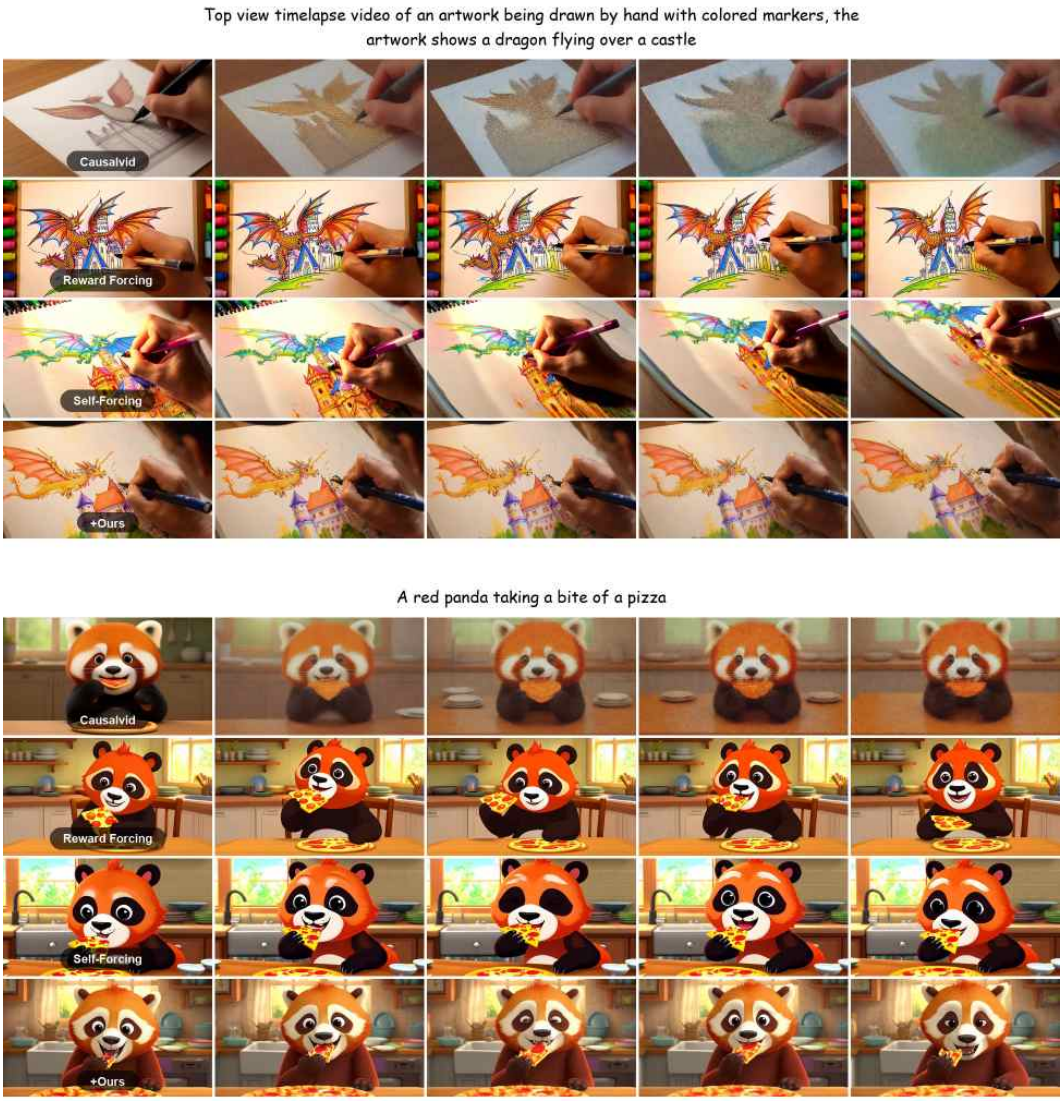


Figure S11: Additional qualitative comparison on long video generation (30s) under the single-prompt setting (Set 7).

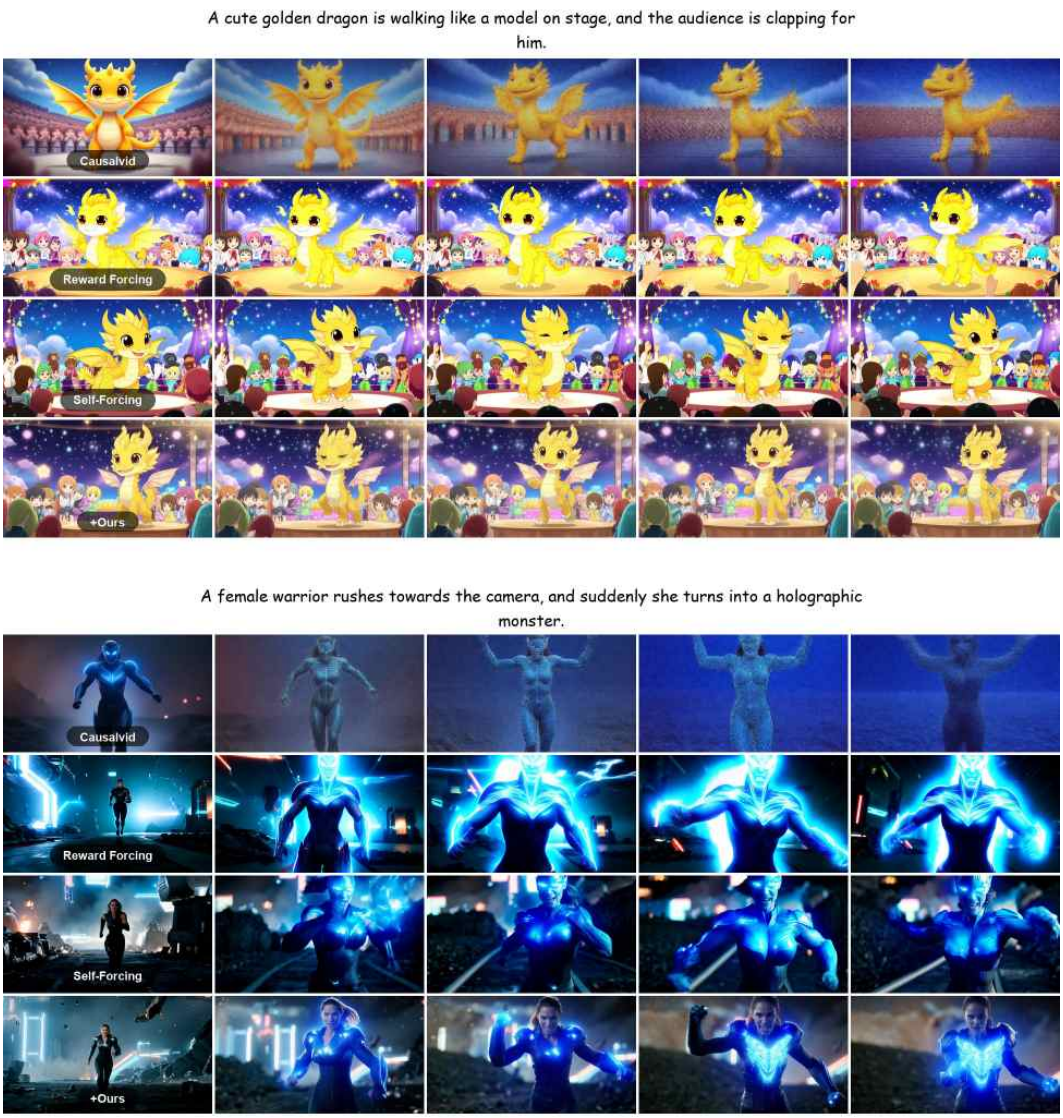


Figure S12: Additional qualitative comparison on long video generation (30s) under the single-prompt setting (Set 8).

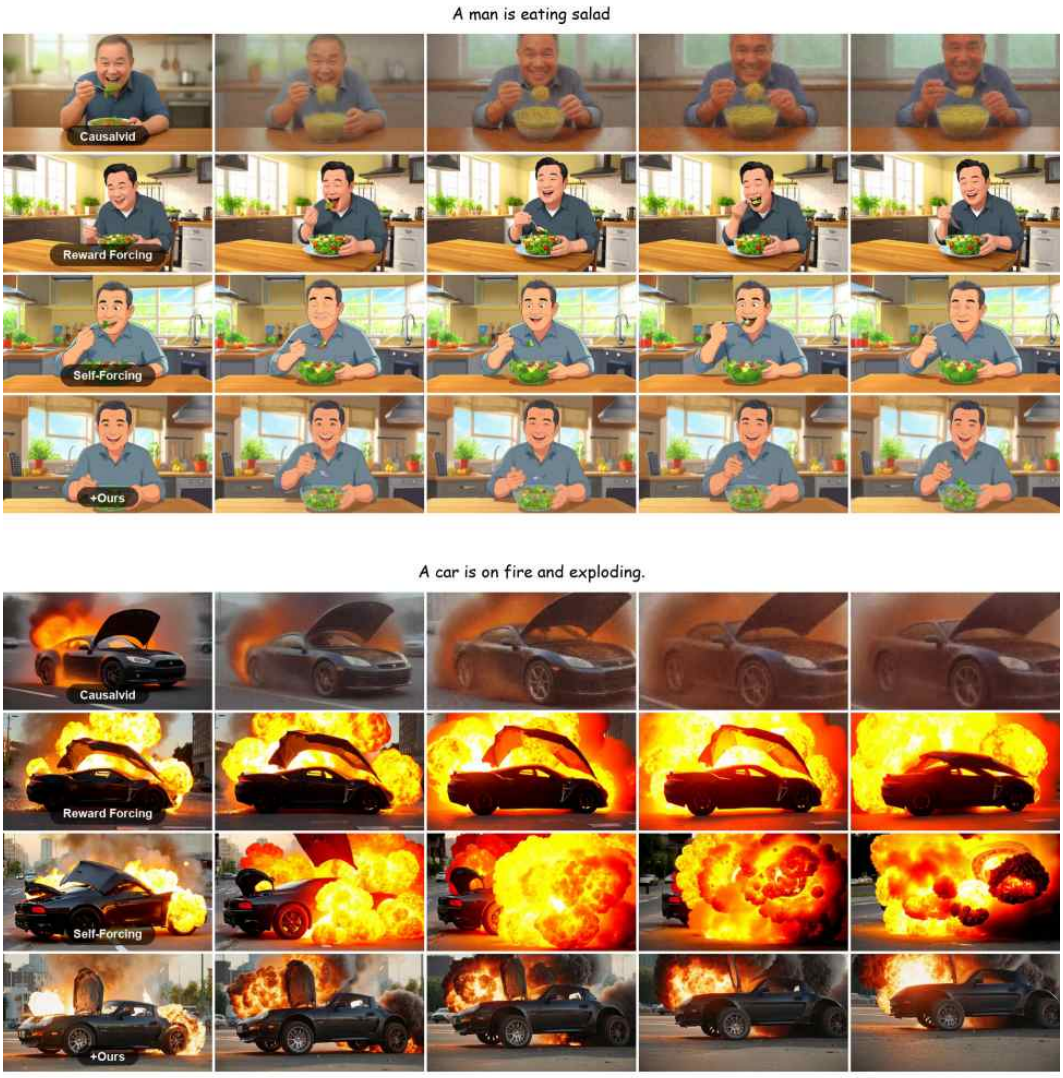


Figure S13: Additional qualitative comparison on long video generation (30s) under the single-prompt setting (Set 9).

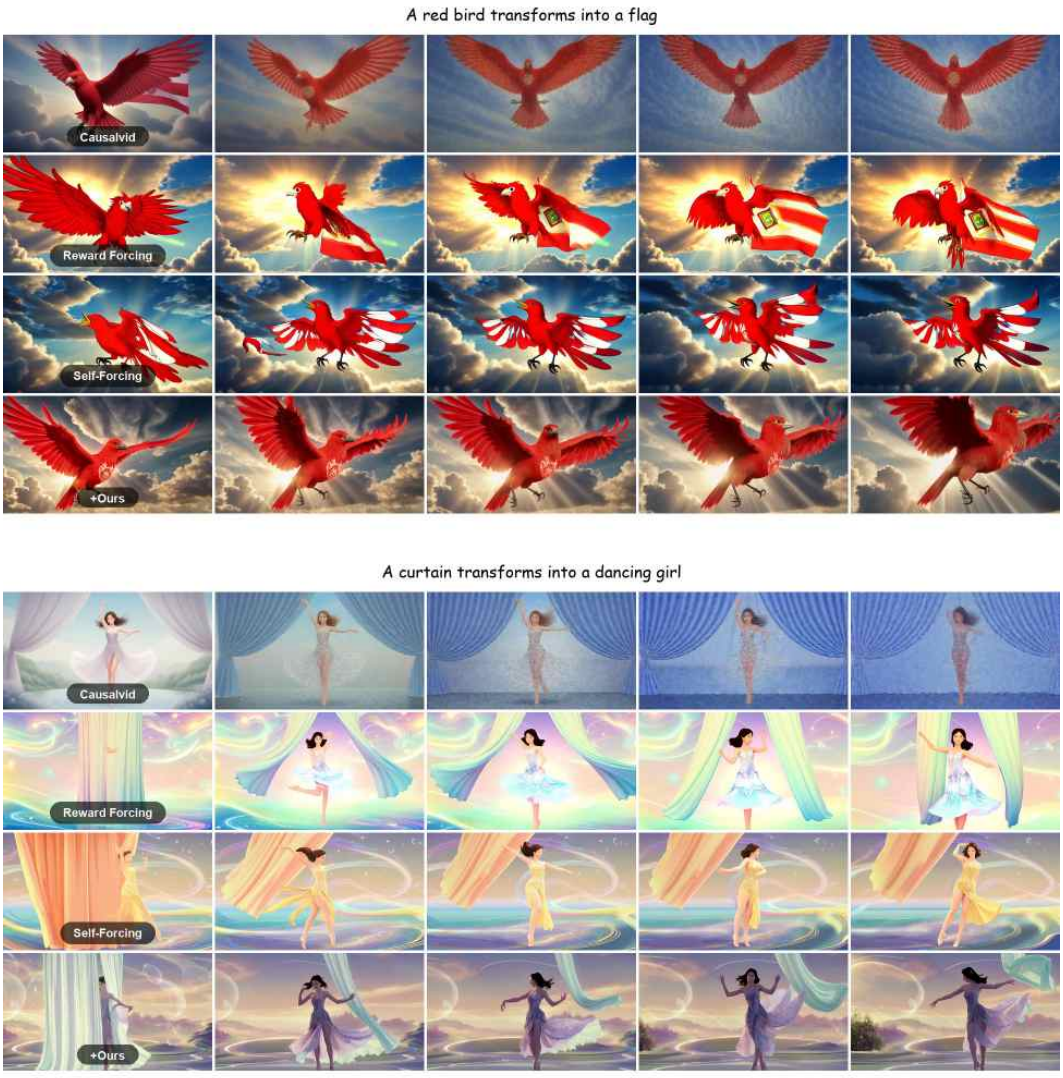


Figure S14: Additional qualitative comparison on long video generation (30s) under the single-prompt setting (Set 10).

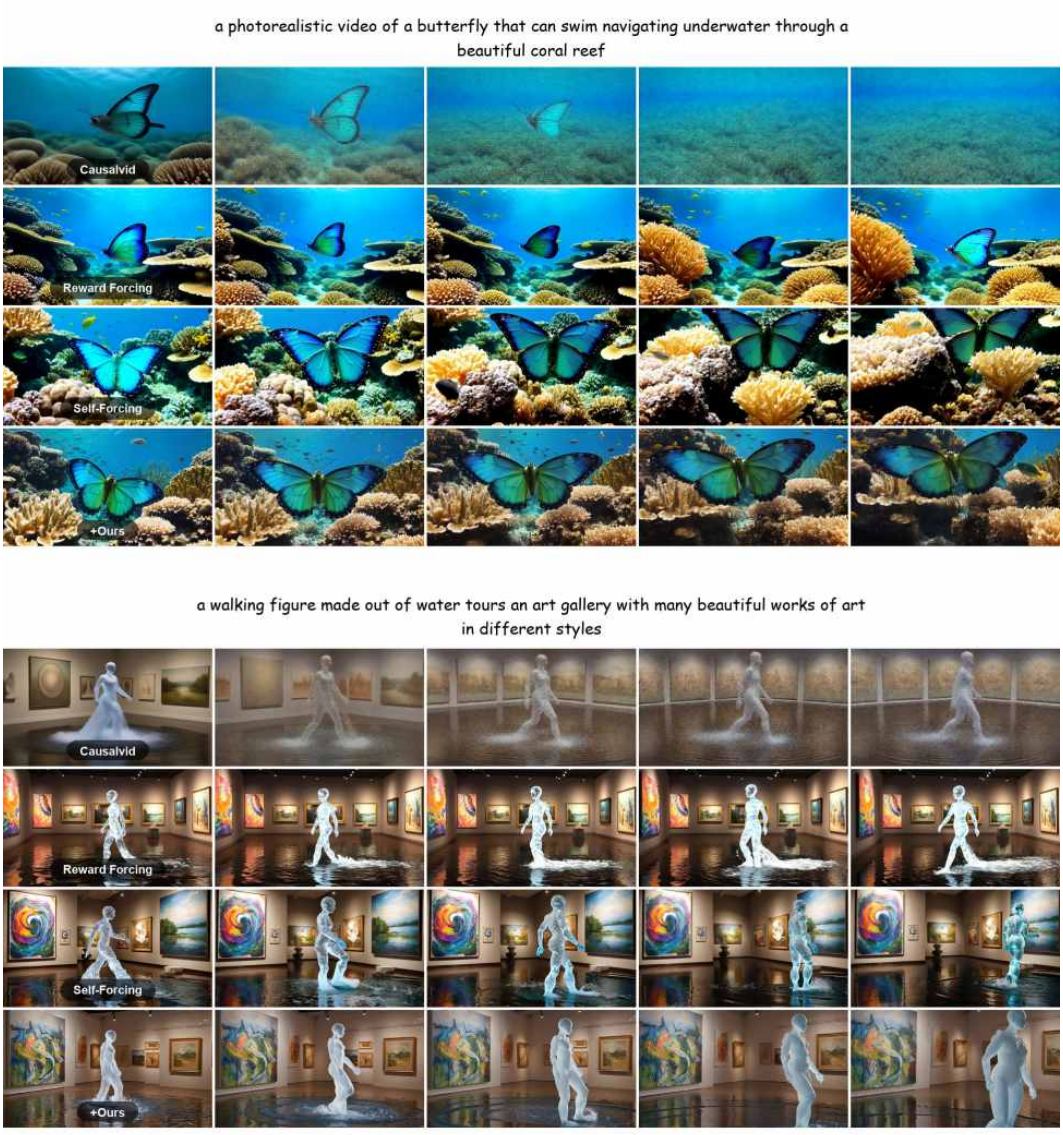


Figure S15: Additional qualitative comparison on long video generation (30s) under the single-prompt setting (Set 11).

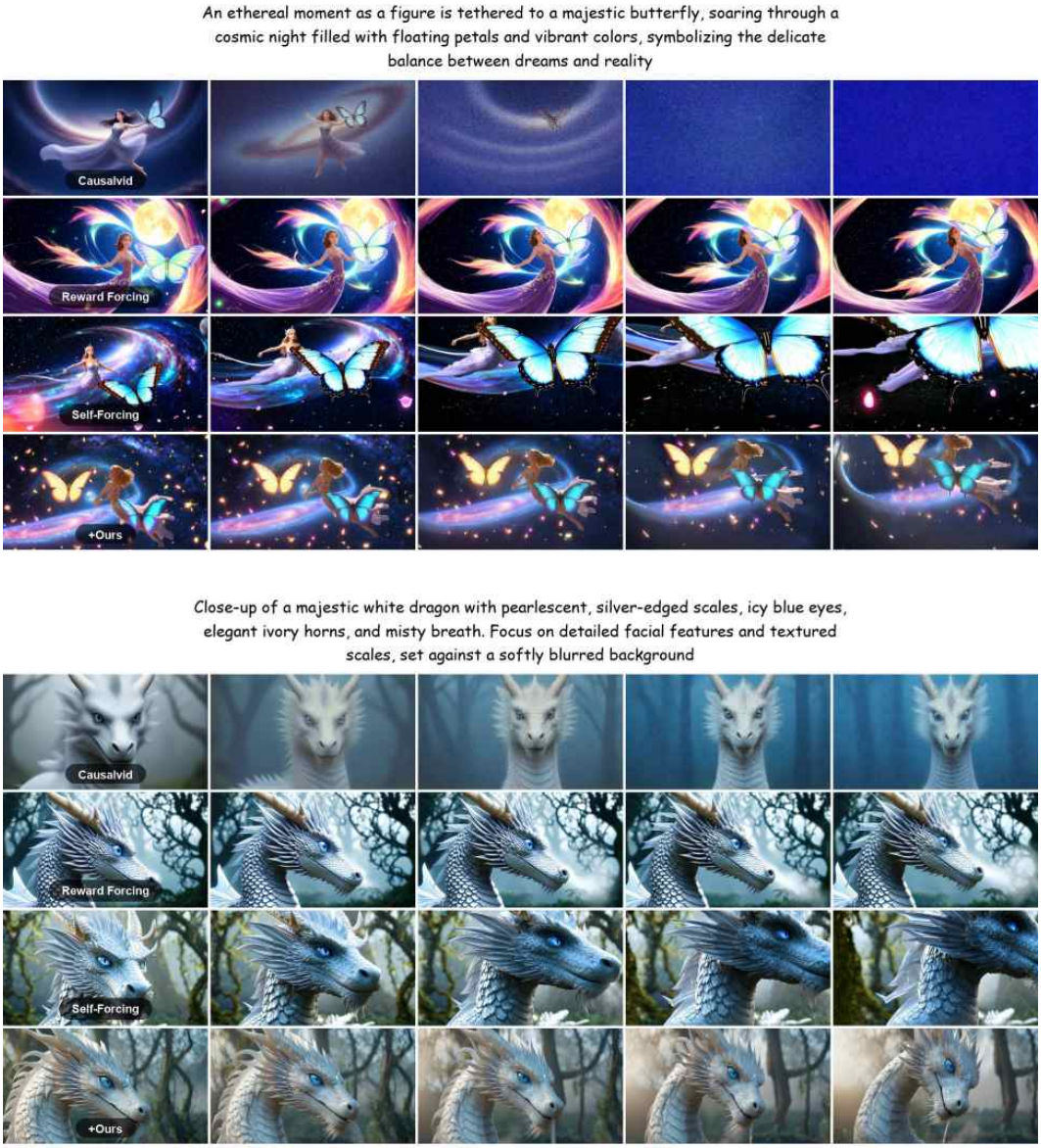


Figure S16: Additional qualitative comparison on long video generation (30s) under the single-prompt setting (Set 12).



Figure S17: Additional qualitative comparison on long video generation (30s) under the single-prompt setting (Set 13).

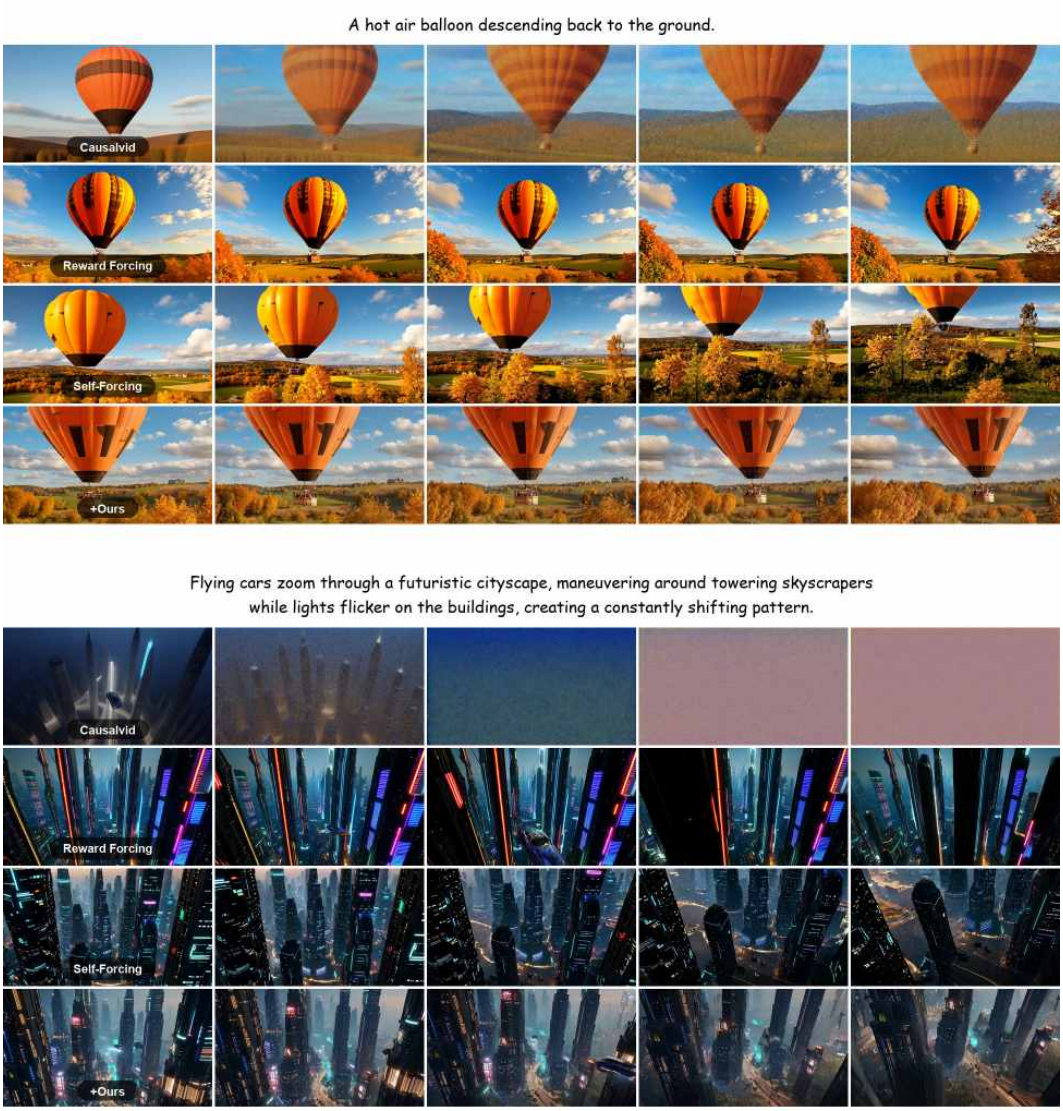


Figure S18: Additional qualitative comparison on long video generation (30s) under the single-prompt setting (Set 14).



Figure S19: Additional ualitative comparison on long video generation (30s) under the single-prompt setting (Set 15).



Figure S20: Additional ualitative comparison on long video generation (30s) under the single-prompt setting (Set 16).

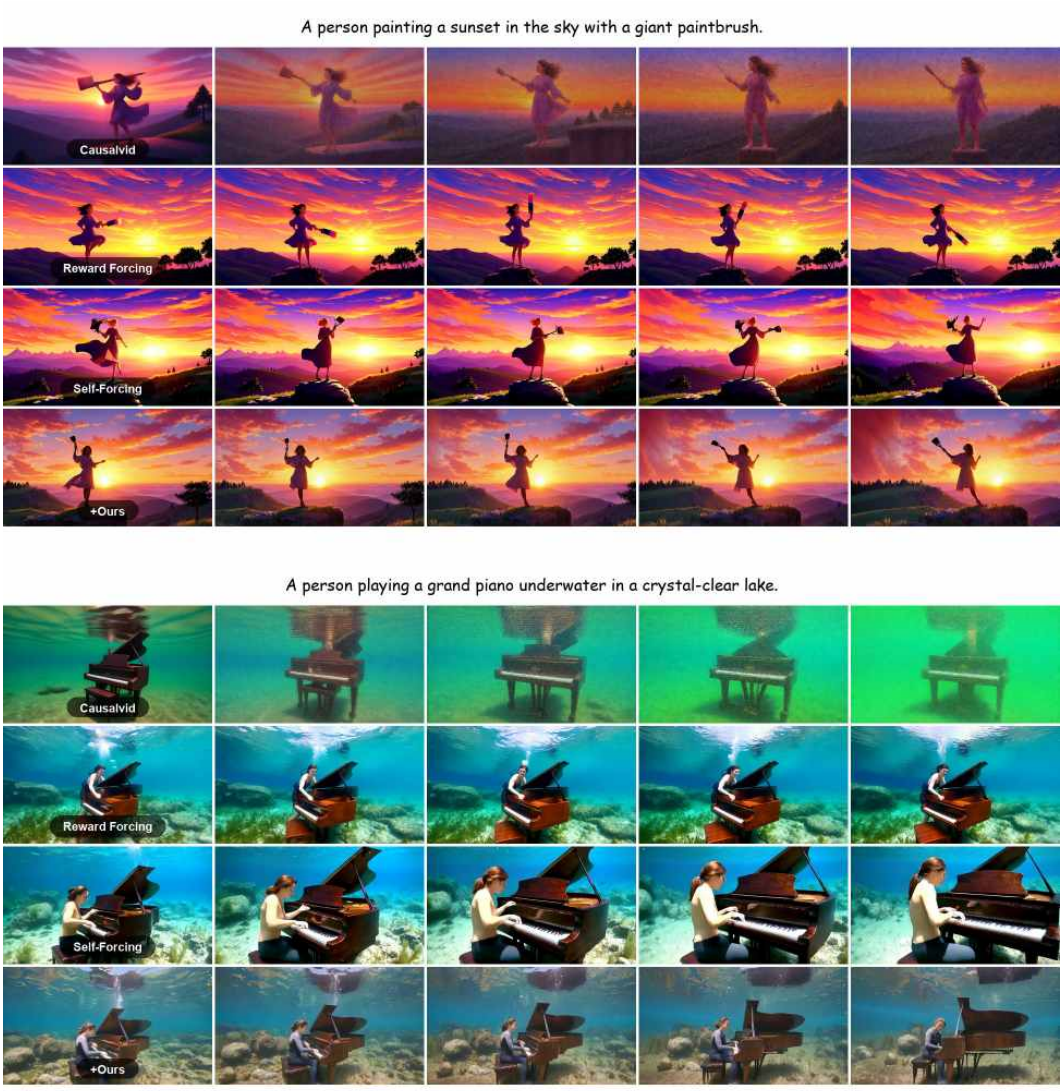


Figure S21: Additional qualitative comparison on long video generation (30s) under the single-prompt setting (Set 17).

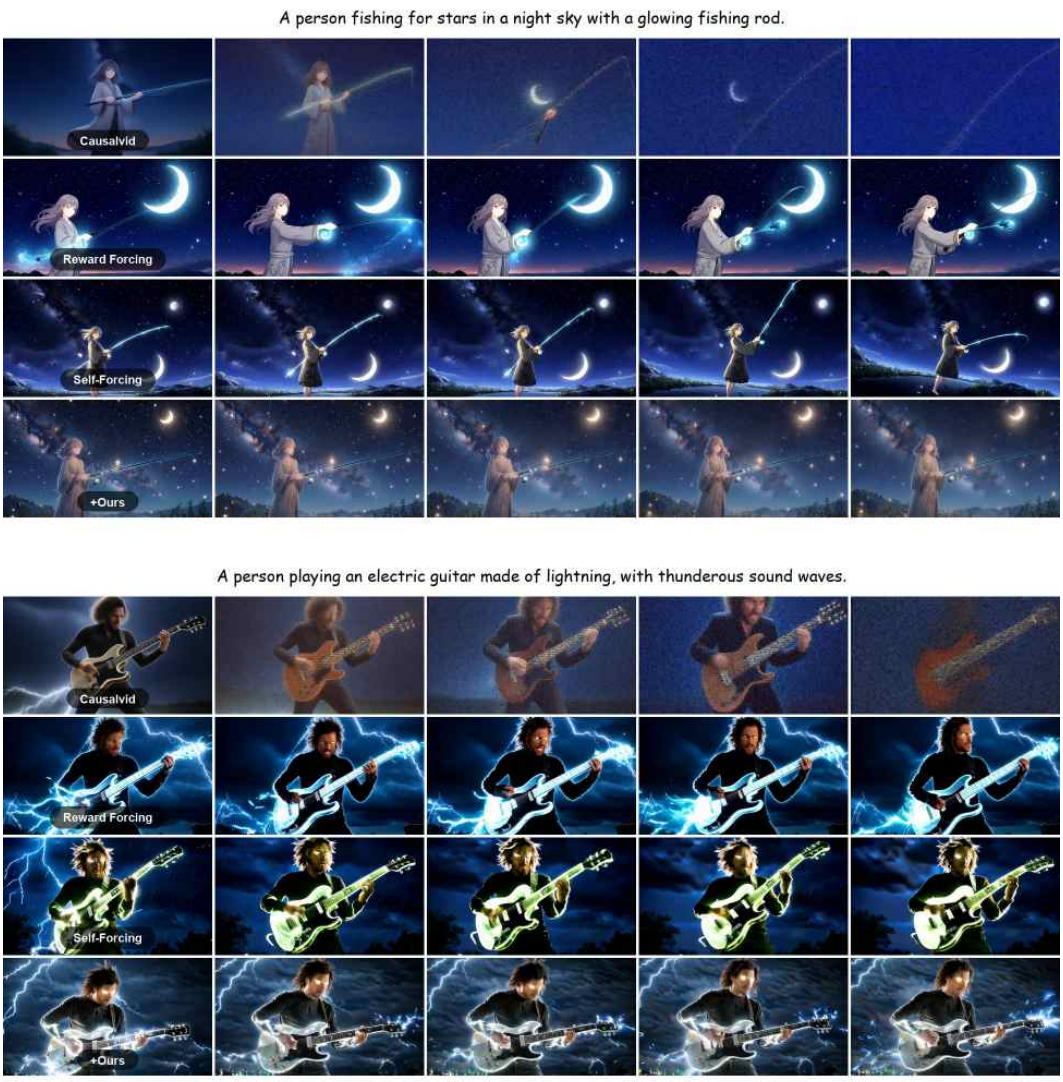


Figure S22: Additional qualitative comparison on long video generation (30s) under the single-prompt setting (Set 18).



Figure S23: Additional qualitative comparison on long video generation (30s) under the single-prompt setting (Set 19).

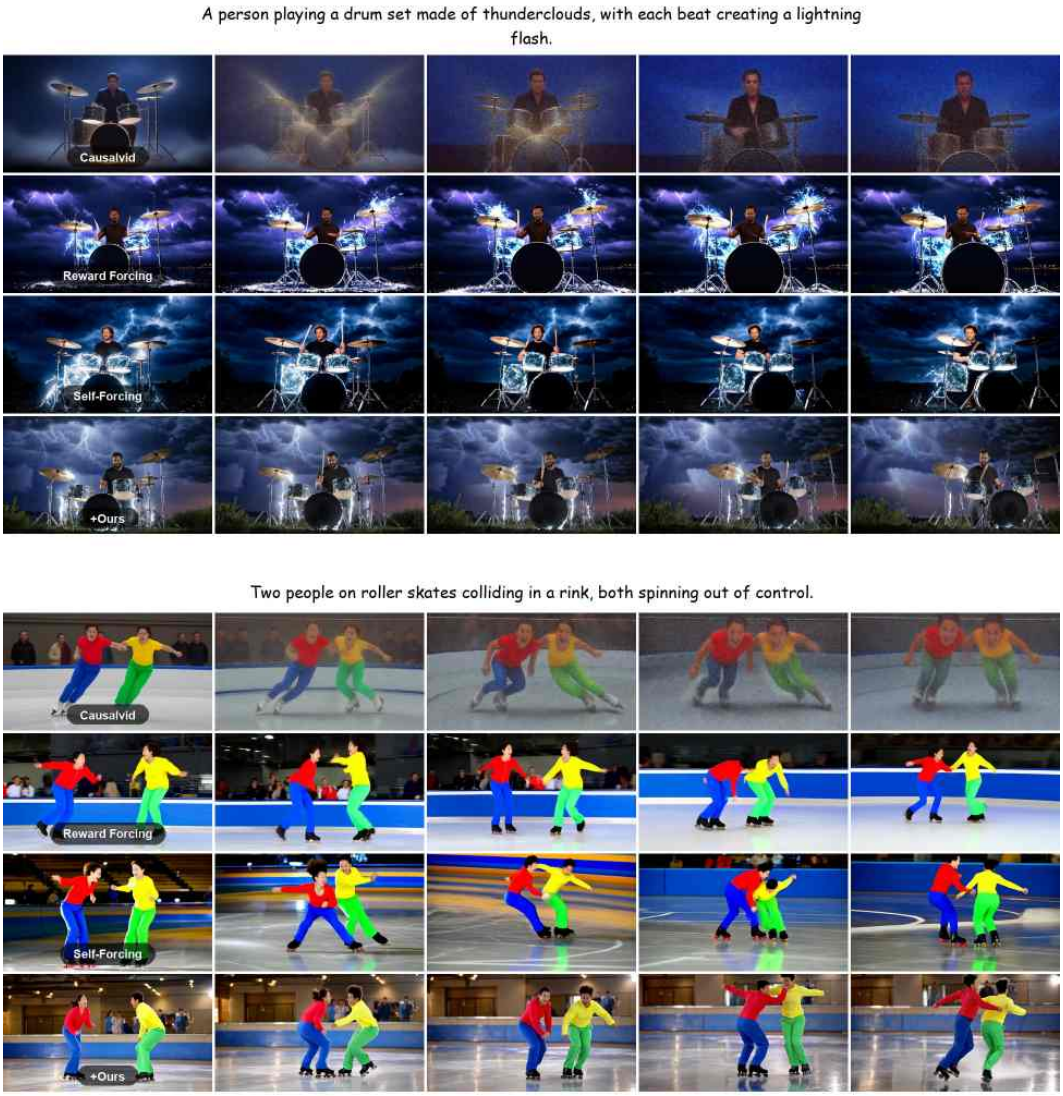


Figure S24: Additional qualitative comparison on long video generation (30s) under the single-prompt setting (Set 20).

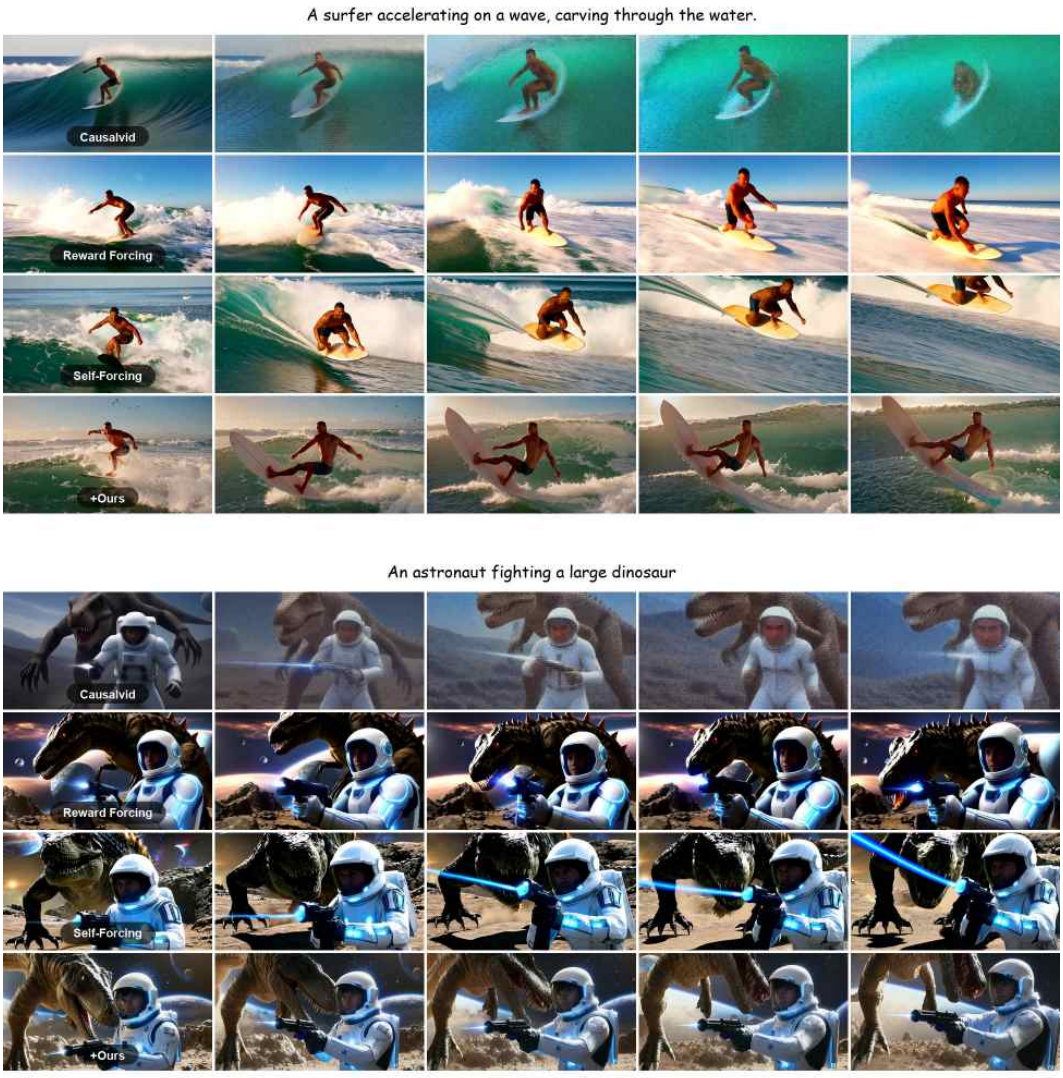


Figure S25: Additional qualitative comparison on long video generation (30s) under the single-prompt setting (Set 21).

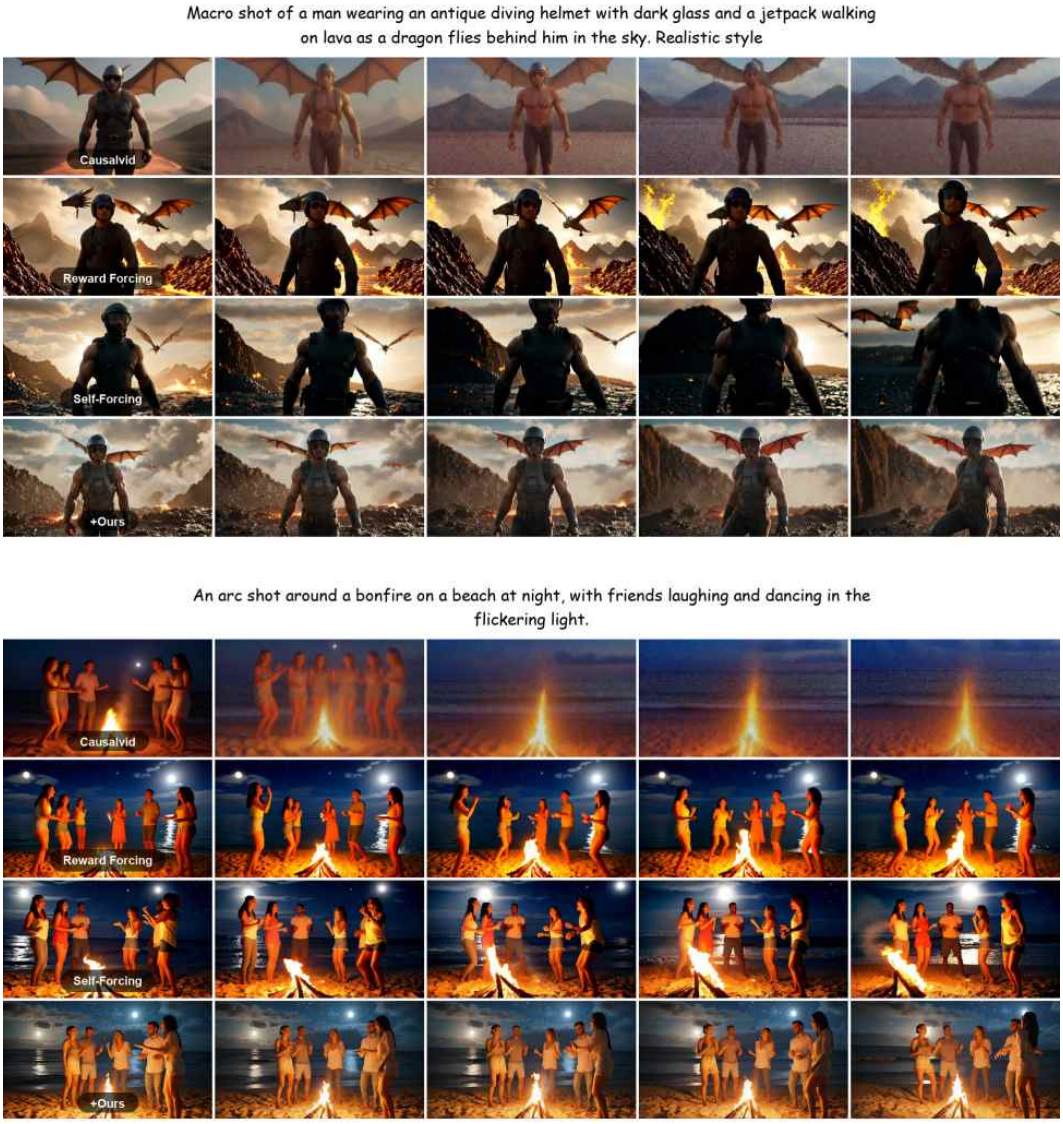


Figure S26: Additional qualitative comparison on long video generation (30s) under the single-prompt setting (Set 22).

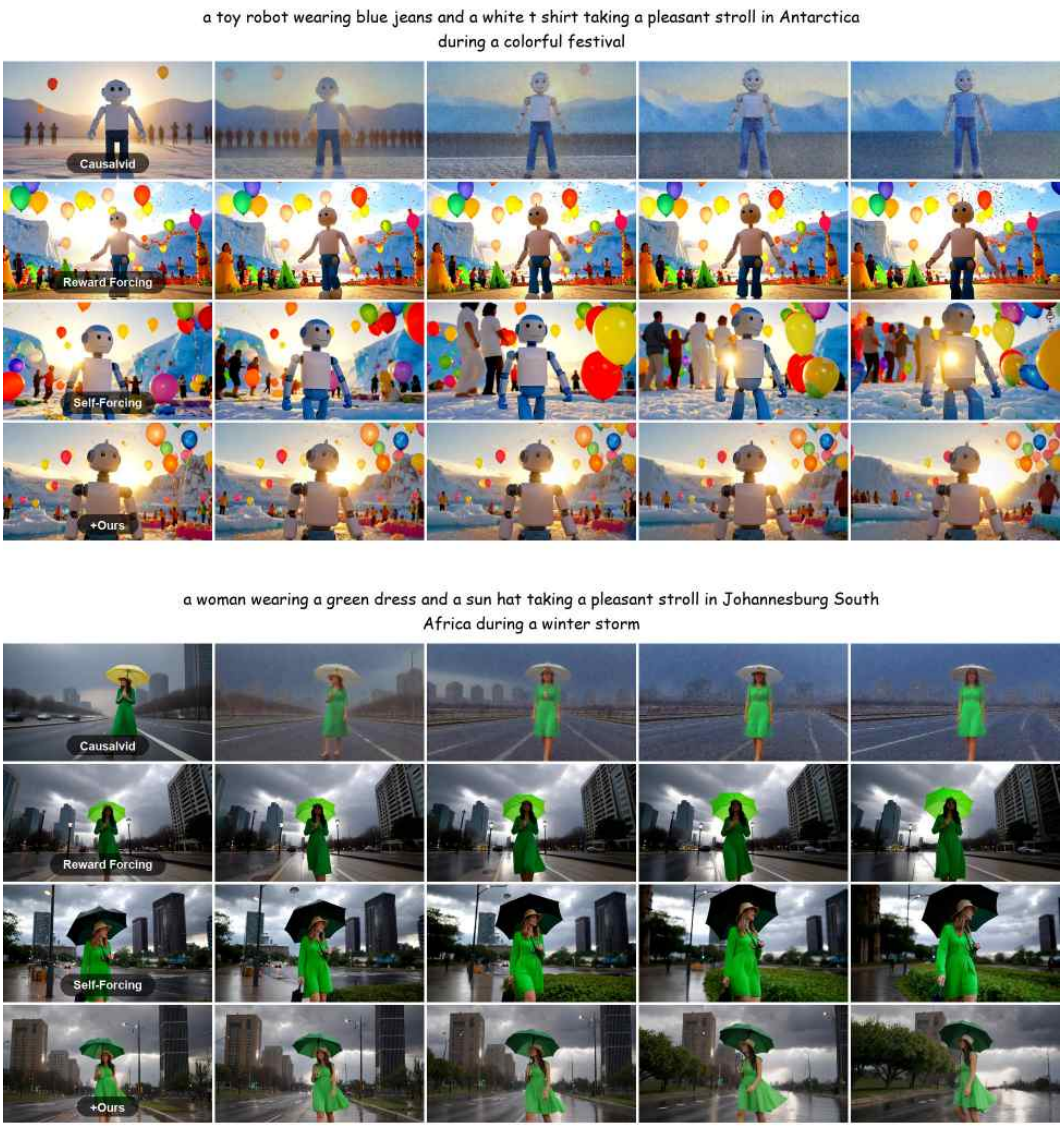


Figure S27: Additional qualitative comparison on long video generation (30s) under the single-prompt setting (Set 23).

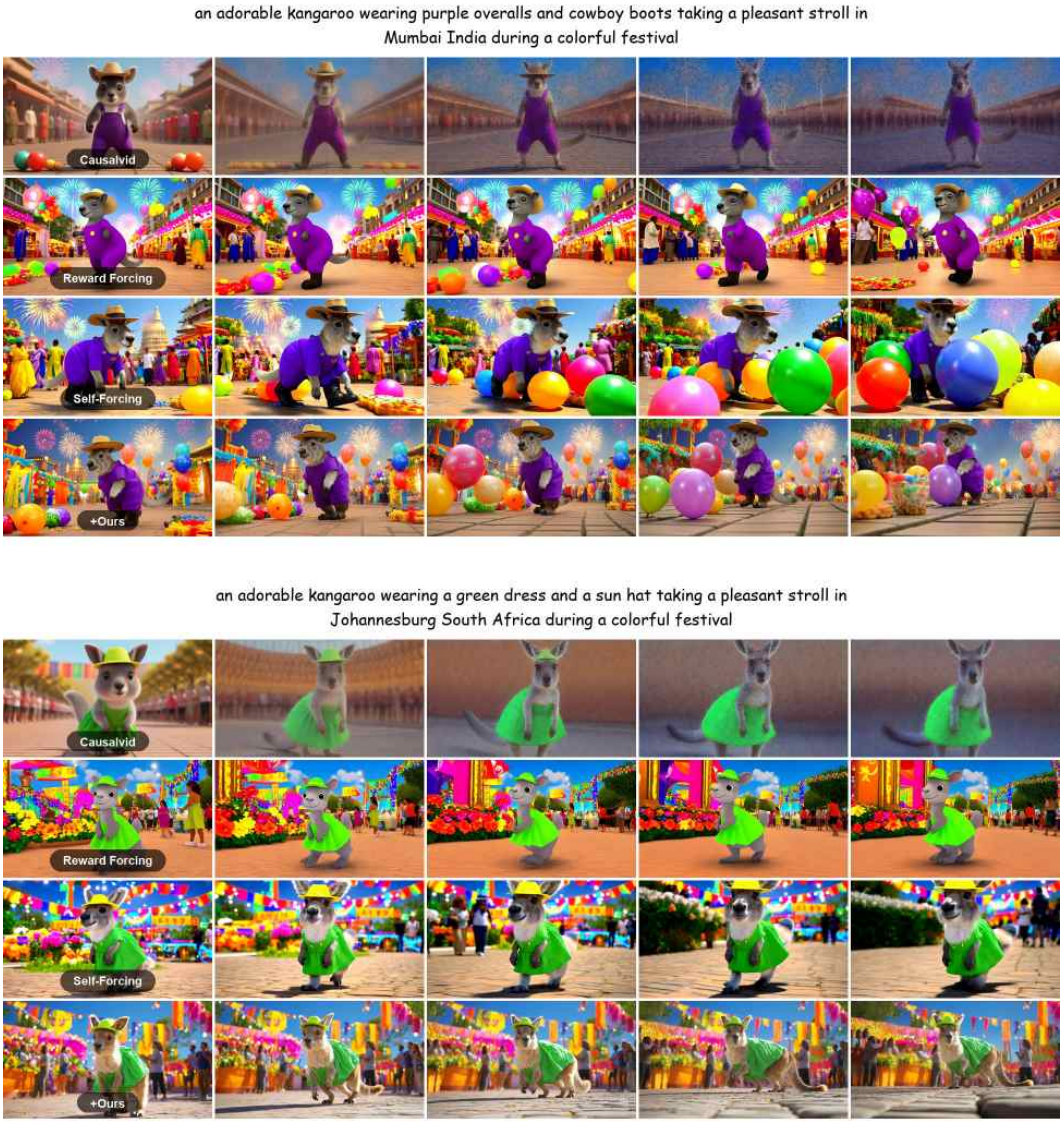


Figure S28: Additional qualitative comparison on long video generation (30s) under the single-prompt setting (Set 24).