

Pathwise Test-Time Correction for Autoregressive Long Video Generation

Xunzhi Xiang^{1,2*}, Zixuan Duan^{1*}, Guiyu Zhang³, Haiyu Zhang², Zhe Gao¹, Junta Wu², Shaofeng Zhang⁴, Tengfei Wang^{2†}, Qi Fan^{1†}, and Chunchao Guo²

¹ Nanjing University, Nanjing, China

fanqi@nju.edu.cn

² Tencent Hunyuan, Shanghai, China

³ Chinese University of Hong Kong, Shenzhen, China

⁴ University of Science and Technology of China, Hefei, China

*Equal Contribution †Corresponding Authors

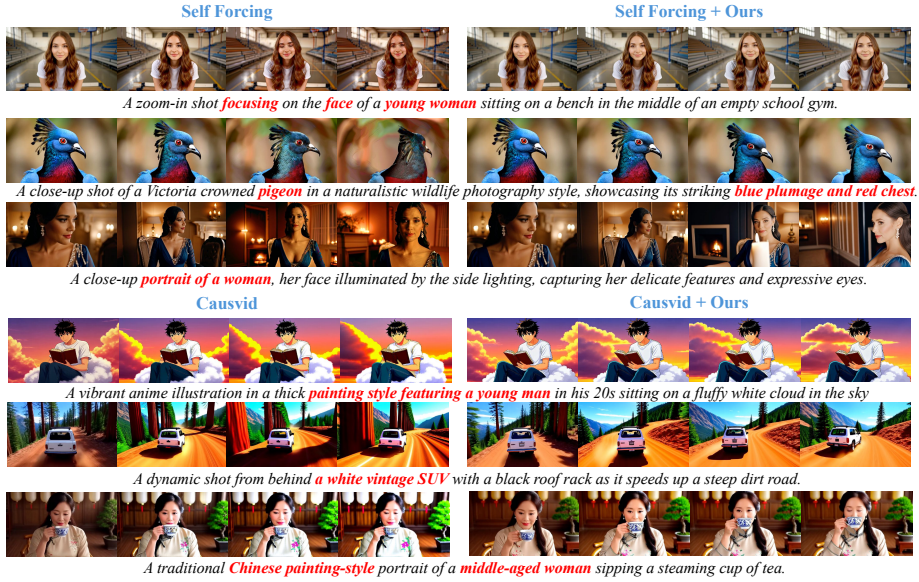


Fig. 1: 30-second video generation examples. Our method reduces error accumulation in CausVid and Self-Forcing, enabling longer and more stable videos with improved visual consistency. All samples are generated with the **same random seed** for fair comparison.

Abstract. Distilled autoregressive diffusion models facilitate real-time short video synthesis but suffer from severe error accumulation during long-sequence generation. While existing Test-Time Optimization (TTO) methods prove effective for images or short clips, we identify that they fail to mitigate drift in extended sequences due to unstable reward landscapes and the hypersensitivity of distilled parameters. To overcome these limitations, we introduce Test-Time Correction (TTC), a training-free alternative. Specifically, TTC utilizes the initial frame as a stable reference

anchor to calibrate intermediate stochastic states along the sampling trajectory. Extensive experiments demonstrate that our method seamlessly integrates with various distilled models, extending generation lengths with a slight overhead while matching the quality of resource-intensive training-based methods on 30-second benchmarks.

Keywords: Video Generation · Training-Free · Test-Time Correction

1 Introduction

Video generation [14, 24, 30–32, 48] has advanced rapidly with the development of diffusion-based generative models [14, 26, 33, 36, 39, 43], which now enable the high-quality synthesis of complex motion [17, 46, 58] and visual appearance [12, 54]. However, scaling these diffusion priors to extended video sequences remains a formidable challenge. Beyond the escalating computational costs associated with longer contexts, maintaining temporal coherence over extended horizons is difficult without incurring excessive latency, thereby limiting their deployment in real-time applications.

To overcome these limitations, recent studies [20, 51] have shifted from bidirectional modeling to step-distilled autoregressive generation, enabling true real-time video synthesis. However, these methods remain constrained by cascading error accumulation: since each frame is conditioned on prior outputs, initial inaccuracies compound over time, resulting in temporal drift and long-horizon degradation. While recent extensions [8, 49] like Rolling Forcing [29], LongLive [47], Self-Forcing++ [7], and WorldPlay [40] have achieved minute-level consistency through sink mechanisms and windowed DMD retraining, they necessitate substantial computational overhead for model fine-tuning. Consequently, a pivotal question arises: *Can we improve the stability of autoregressive video generation purely at inference time, bypassing the need for retraining the base model?*

TTO [41, 44] has emerged as a compelling alternative for enhancing video quality without the need for retraining. However, while effective for short-video synthesis [10, 53], our toy experiments reveal that scaling TTO to long-horizon

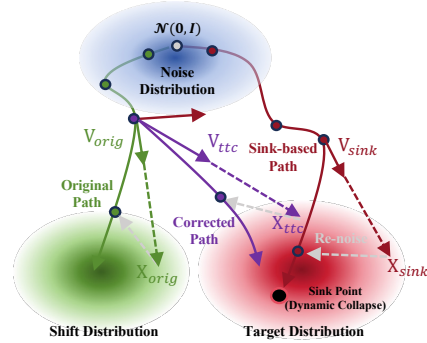


Fig. 2: Comparison of sampling strategies. The **Original Path** suffers from error accumulation, while the **Sink-based Path** collapses into a *Sink Point*. Our **TTC strategy** avoids this via reference-conditioned denoising and explicit *Re-noising*.

autoregressive generation faces a dual bottleneck consisting of the inherent difficulty in defining reward functions for long-range consistency and the extreme optimization sensitivity of distilled models. We observe that in these distilled models, even infinitesimal test-time gradients often trigger reward collapse and fail to mitigate cumulative error. Therefore, we propose TTC, which is a training-free framework that shifts the paradigm from parameter-space optimization to sampling-space stochastic intervention. TTC is grounded in the insight that few-step distilled samplers are inherently stochastic as they perturb intermediate states with injected noise. This property implies that intermediate predictions are not fixed outcomes but rather malleable latent states that can be rectified by subsequent diffusion steps to align with the global initial context while preserving the underlying sampling distribution.

Specifically, as shown in Figure 2, TTC applies a small number of correction steps along the stochastic sampling path, **only after the global structure has stabilized**. This delay prevents the generation from falling into sink-collapse [8], a phenomenon where newly generated frames repeatedly regress toward the sink frames instead of evolving naturally.

At chosen steps in the sampling path, TTC performs reference-conditioned denoising by utilizing the initial frame context to anchor a corrected clean prediction. Then, this corrected state is re-noised back to the variance level corresponding to the current timestep, which ensures that the intervention remains compatible with the expected noise distribution. By integrating correction into the stochastic sampling path of the autoregressive diffusion process rather than directly replacing the denoised prediction, this mechanism suppresses long-term error accumulation and temporal drift without retraining, while preserving high-fidelity temporal coherence over extended durations.

In this work, we show that long-horizon stability in autoregressive video generation can be achieved through test-time intervention alone. Our method suppresses error accumulation with a slight computational overhead, without requiring any retraining. As a result, it extends the stable generation length of distilled autoregressive models from a few seconds to over 30 seconds, while achieving visual quality comparable to state-of-the-art training-based methods. Consistent improvements across multiple model architectures demonstrate that TTC is a robust and general solution for stabilizing distilled autoregressive diffusion models.

2 Related Work

Bidirectional Models for Video Generation. Diffusion models have been widely adopted for video generation, with recent works typically formulating video synthesis as a sequence-level joint denoising problem. Under this bidirectional diffusion paradigm, all frames are denoised simultaneously via spatiotemporal attention, enabling the model to leverage global temporal context and produce temporally coherent, high-fidelity videos [2, 19, 25, 34, 50, 56]. Large-scale systems such as Hunyuan [26] and Wan [43] further demonstrate the effective-

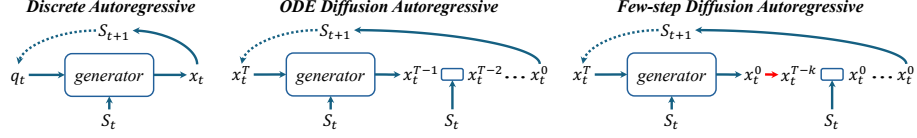


Fig. 3: Variants of autoregressive video generation. Discrete AR uses single-step deterministic prediction, multi-step diffusion follows a deterministic ODE trajectory, while few-step distilled diffusion performs stochastic sampling with intermediate noise injection.

ness of this joint denoising formulation at scale. However, because the entire sequence must be processed as a whole during inference, this paradigm inherently precludes streaming or incremental generation, limiting its applicability in real-time and interactive scenarios.

Autoregressive Models for Video Generation. Autoregressive video diffusion models generate videos sequentially in a strict causal manner, conditioning each new frame or segment on the historical context of previously generated content [4, 5, 9, 11, 20, 23, 29, 42, 47, 51]. While this formulation naturally supports streaming generation with low initial latency, it is inherently susceptible to error accumulation, where minor deviations propagate and amplify across steps, leading to severe temporal drift and degraded coherence in long videos. To mitigate this issue, recent works introduce planning methods [45, 55] or explicit memory mechanisms [3, 6, 18, 22, 40, 52], strategies that typically necessitate complex architectural modifications and extensive re-training.

Test-time Image/Video Generation. Test-time generation methods [27, 41, 44] aim to enhance the performance of pre-trained models directly during the inference phase. Test-time scaling improves quality by iteratively searching over multiple candidates, as seen in Video-T1 [28] and EvoSearch [13], yet this comes at the price of prohibitive computational cost. Similarly, Test-time optimization refines generation via auxiliary parameter updates that necessitate instance-specific training, such as HyperNoise [10], AutoRefiner [53], and SLOWFAST-VGEN [15]. In contrast, our approach distinguishes itself from both paradigms by being fully training-free, avoiding both the overhead of candidate search and the complexity of parameter optimization.

3 Test-time Optimization for Distilled Models

3.1 Background: Few-step Distilled Sampling

In this section, we formulate autoregressive video generation as next-chunk prediction under a context-conditional generative model. Given a video sequence $\{x_1, \dots, x_N\}$, the joint distribution factorizes as

$$p(x_{1:N}) = \prod_{t=1}^N p_\theta(x_t | S_t), \quad S_t = \{x_1, \dots, x_{t-1}\}, \quad (1)$$

where S_t denotes the context at step t , consisting of all previously generated frames or chunks. As illustrated in Figure 3, existing autoregressive video generation methods typically fall into three categories. Discrete autoregressive models generate each chunk through a single deterministic prediction conditioned on past outputs, while multi-step autoregressive diffusion models approximate the same conditional distribution via a deterministic ODE-based sampling trajectory. In contrast, few-step distilled diffusion models replace deterministic ODE solvers with a *stochastic sampling process* that explicitly injects noise at intermediate steps.

Under this formulation, conditional distribution $p_\theta(x_t | S_t)$ is no longer realized as a single deterministic mapping, but through a *stochastic diffusion sampling trajectory* defined over a sparse set of diffusion steps $\{T_0 = 0, T_1, \dots, T_K = T_{\max}\}$. Specifically, distilled video generation begins from Gaussian noise $x_t^{T_{\max}} \sim \mathcal{N}(0, I)$ and evolves progressively along this trajectory via a sequence of denoise–re-noise transitions. Specifically, at each denoising step T_j , the generation process starts from a noisy latent state and applies the denoising network to produce an estimate of the underlying clean latent representation,

$$x_{t,0}^{T_j} = G_\theta\left(\Psi(x_{t,0}^{T_{j+1}}, \epsilon_t^{T_j}, T_j); S_t, T_j\right), \quad (2)$$

where $G_\theta(\cdot)$ denotes the parameterized denoising network, and S_t represents the autoregressive context at step t .

After each denoising step, distilled diffusion models proceed by re-injecting noise according to the predefined schedule, mapping the clean estimate back onto the diffusion trajectory. Concretely, the estimated clean latent is re-noised to obtain the latent state at the next diffusion step,

$$x_t^{T_{j-1}} = \Psi\left(x_{t,0}^{T_j}, \epsilon_t^{T_{j-1}}, T_{j-1}\right), \quad \epsilon_t^{T_{j-1}} \sim \mathcal{N}(0, I), \quad (3)$$

thereby yielding a stochastic transition that advances the generation process to the next noise level.

The forward diffusion process $\Psi(\cdot)$ is defined as

$$\Psi(x, \epsilon_T, T) = \alpha_T x + \sigma_T \epsilon_T, \quad (4)$$

where α_T and σ_T are predefined diffusion coefficients corresponding to step T . Repeating this denoise–re-noise procedure across diffusion steps forms the complete stochastic sampling trajectory of distilled diffusion models.

3.2 Toy Experiment: Applying Test-time Optimization to Long Video Generation

Existing TTO methods improve generation quality by aligning the model distribution with a predefined reward function. Given a pre-trained generative model with output distribution p^{base} , TTO typically defines a reward-weighted target distribution

$$p^*(x) \propto p^{\text{base}}(x) \exp(r(x)), \quad (5)$$



Fig. 4: Comparison of two toy test-time optimization variants based on LoRA fine-tuning, illustrating distinct failure modes.

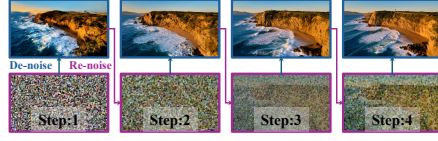


Fig. 5: Intermediate predictions along the stochastic sampling path. High-noise steps determine global structure, while low-noise steps refine appearance details under a fixed layout.

where $r(x)$ encodes preferences of samples x . The optimization objective can be formulated as minimizing the KL divergence between a parameterized distribution p^ϕ and the target distribution p^* ,

$$\min_{\phi} D_{\text{KL}}(p^\phi \| p^*) = \min_{\phi} D_{\text{KL}}(p^\phi \| p^{\text{base}}) - \mathbb{E}_{x \sim p^\phi} [r(x)], \quad (6)$$

which trades off reward maximization against deviation from the original model distribution. However, for long video generation, it remains challenging to design an explicit reward that effectively suppresses error accumulation. Temporal drift arises from coupled inconsistencies in semantics, appearance, and motion, which are difficult to characterize with a single hand-crafted objective. A naive alternative is to constrain each subsequent chunk’s predictive distribution to remain close to that of the initial frames, effectively anchoring generation to early content.

To assess this idea, we conduct two toy experiments using direct LoRA fine-tuning [16] at test time, following HyperNoise [10] and AutoRefiner [53]. Both variants use the same backbone and identical LoRA adapters, and differ only in their optimization objectives. The first variant fine-tunes LoRA with a standard denoising reconstruction loss on early frames across noise levels. The second variant replaces pixel-level reconstruction with a semantic consistency objective, enforcing similarity to early frames in pretrained feature spaces [35, 38].

Figure 4 shows that these two objectives lead to distinct failure modes. The reconstruction-based variant quickly collapses to a trivial solution, where later frames become near-duplicates of the initial frame, resulting in severe motion loss. In contrast, the semantic objective fails to effectively reduce long-horizon error accumulation, and the generated videos still exhibit temporal drift similar to the baseline. These results indicate that naive TTO, whether based on low-level reconstruction or high-level semantics, is insufficient for stable long-horizon generation.

4 From Test-Time Optimization to Test-Time Correction

Based on the above analysis, we identify two key limitations of TTO for long video generation.

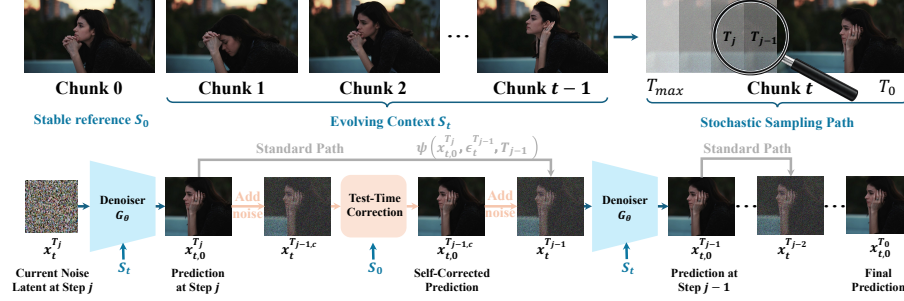


Fig. 6: Overall pipeline of our method. A sparse set of correction steps is inserted into the stochastic sampling path until the global structure stabilizes. At selected steps, TTC performs reference-conditioned denoising using the initial frame to obtain a corrected prediction, which is then re-noised to the current timestep to remain consistent with the expected noise distribution. This on-path, training-free correction suppresses long-term error accumulation and stabilizes long-horizon generation.

Reward design for error accumulation. Temporal drift stems from coupled errors in semantics, appearance, and motion, which are hard to capture with a single reward: low-level reconstruction suppresses motion, while high-level semantic objectives lack frame-wise correction signals.

Optimization Challenges and Collapse. Performing test-time optimization on distilled models presents significant training difficulties. The models tend to overfit rapidly to the auxiliary reward, causing the optimization trajectory to collapse into specific, degenerate solutions that violate the pre-trained generative prior.

Together, these limitations motivate a shift from parameter-updating test-time optimization to *test-time correction*, which avoids model updates and instead performs trajectory-aware interventions during sampling.

4.1 Correctability along the Stochastic Sampling Path

Distilled few-step diffusion models maintain a stochastic sampling trajectory through iterative noise re-injection, which prevents premature convergence and preserves flexibility in intermediate states. As shown in Figure 5, this trajectory exhibits a clear functional phase transition. At high noise levels, the denoising process primarily determines global structure, such as scene layout and spatial relationships. As the noise level decreases, the generation progressively shifts to an appearance refinement stage, where local textures and fine visual details are synthesized while the global structure remains largely fixed. This phase-wise behavior naturally suggests a principled test-time correction strategy. Rather than intervening throughout the sampling process, we apply alternative conditioning only during the appearance refinement stage, after the global structure has stabilized. At this stage, the model is less sensitive to structural changes, allowing

visual attributes to be adjusted without affecting layout or geometry. As a result, targeted test-time intervention can modulate appearance while preserving structural consistency.

Motivated by planning-based video generation models [45, 55], which relax strictly unidirectional prediction via cross-frame context, we consider applying test-time intervention at a *single sampling step* after structural stabilization. Specifically, at a designated step j^* , we restrict the visible context state S_t to include only the earliest frame, forcing the model to rely exclusively on the initial frame for subsequent appearance refinement and texture generation. This **single-point correction** process can be formalized as:

$$x_{t,0}^{T_{j-1}} = G_\theta\left(\Psi(x_{t,0}^{T_j}, \epsilon_t^{T_{j-1}}, T_{j-1}); S_t^{(j \rightarrow j^*)}, T_{j-1}\right), \quad (7)$$

where $S_t^{(j \rightarrow j^*)}$ denotes the modified context state. Specifically, at the designated sampling step j^* , the original autoregressive context S_t is replaced by the earliest-frame context S_0 , while all other sampling steps remain unchanged. Here, j^* corresponds to the stage at which the global layout and object structure have stabilized.

4.2 Pathwise Test-time Correction

Despite its conceptual simplicity, single-point latent correction frequently leads to visible artifacts in distilled autoregressive diffusion models, such as flickering, abrupt appearance changes, and temporal inconsistency. To overcome this, we propose a path-wise self-correction strategy as shown in Figure 6, which leverages the model’s stochastic nature to ensure smooth state transitions.

Instead of performing a hard correction at a single denoising step, our method first applies the intervention to the current prediction and then re-noises it back to the current noise level. By restarting the sampling process from this re-noised state, the correction is naturally integrated into the stochastic path. This avoids the abrupt state transitions typical of direct prediction replacement, allowing the model to smoothly assimilate the update while maintaining generation stability.

Formally, consider generating chunk t under a denoising schedule $T_{\max} > \dots > T_j > T_{j-1} > \dots > T_0$. At step j , given the current noisy latent $x_t^{T_j}$ and the evolving context S_t , the denoiser produces a clean prediction as:

$$x_{t,0}^{T_j} = G_\theta\left(x_t^{T_j}; S_t, j\right). \quad (8)$$

Instead of directly executing the next denoising update following the standard path in Figure 6, we first apply forward diffusion noise injection to the current prediction and explicitly map it to the next noise level T_{j-1} . Then, we replace the evolving context S_t with a stable reference context S_0 for denoising. This produces a reference-aligned corrected clean prediction as:

$$x_{t,0}^{T_{j-1},c} = G_\theta\left(\Psi\left(x_{t,0}^{T_j}, \epsilon_t^{j-1}, T_{j-1}\right); S_0, j-1\right), \quad (9)$$

Algorithm 1 Pathwise Test-time Correction

```

1: Input: Noise schedule  $\{T_J > \dots > T_0 = 0\}$ ; Generator  $G_\theta$ ; Evolving context  $S_t$ ;
   Ref context  $S_0$ ; Correction indices  $\mathcal{J}^*$ ; Diffusion forward process  $\Psi$ 
2: Output: Final prediction  $x_{t,0}^0$ 
3: Sample  $x_t^{T_J} \sim \mathcal{N}(0, I)$  # Initial gaussian noise
4: for  $j = J$  down to 1 do
5:    $x_{t,0}^{T_j} \leftarrow G_\theta(x_t^{T_j}; S_t, j)$  # Initial prediction with  $S_t$ 
6:   Sample  $\epsilon_t^{T_{j-1}} \sim \mathcal{N}(0, I)$ 
7:   if  $j - 1 \in \mathcal{J}^*$  then
8:     — Phase A: Reference-guided Correction —
9:      $x_t^{T_{j-1},c} \leftarrow \Psi(x_{t,0}^{T_j}, \epsilon_t^{T_{j-1}}, T_{j-1})$ 
10:     $x_{t,0}^{T_{j-1},c} \leftarrow G_\theta(x_t^{T_{j-1},c}; S_0, j - 1)$  # Correct trajectory using  $S_0$ 
11:    — Phase B: Re-noising & Re-denoising —
12:    Sample  $\tilde{\epsilon}_t^{T_{j-1}} \sim \mathcal{N}(0, I)$ 
13:     $x_t^{T_{j-1}} \leftarrow \Psi(x_{t,0}^{T_{j-1},c}, \tilde{\epsilon}_t^{T_{j-1}}, T_{j-1})$  # Inject new gaussian noise
14:     $x_{t,0}^{T_{j-1}} \leftarrow G_\theta(x_t^{T_{j-1}}; S_t, j - 1)$  # Finalize step with  $S_t$ 
15:   else
16:     $x_t^{T_{j-1}} \leftarrow \Psi(x_{t,0}^{T_j}, \epsilon_t^{T_{j-1}}, T_{j-1})$ 
17:     $x_{t,0}^{T_{j-1}} \leftarrow G_\theta(x_t^{T_{j-1}}; S_t, j - 1)$ 
18:   end if
19: end for
20: return  $x_{t,0}^{T_0}$ 

```

The corrected prediction is then mapped back to the same noise level T_{j-1} through noise injection. Denoising is subsequently resumed under the true evolving context S_t

$$x_{t,0}^{T_{j-1}} = G_\theta\left(\Psi\left(x_{t,0}^{T_{j-1},c}, \tilde{\epsilon}_t^{T_{j-1}}, T_{j-1}\right); S_t, j - 1\right), \quad (10)$$

This sequence of operations integrates test-time correction directly into the stochastic sampling process. All modified intermediate states are produced through valid diffusion transitions, as summarized in Algorithm 1. This effectively suppresses chunk-boundary flickering, mitigates long-horizon error accumulation, and preserves temporal coherence in autoregressive video generation.

5 Experiments

5.1 Settings

Baseline. We evaluate our test-time correction method on two baseline models, CausVid [51] and Self-Forcing [20]. Both baselines are built upon the Wan2.1-T2V-1.3B model [43] and generate 5-second video clips at 16 FPS with a resolution of 832×480 . More implementation details for fair comparison are included in the supplementary material.

Table 1: Comprehensive comparison with SOTA methods on prompt-conditioned 30-second video generation (Speed & VBench).

Method	Training Speed Free	VBench Metrics ↑						
		Total fps	Subject Consistency	Background Consistency	Dynamic Degree	Motion Smoothness	Imaging Quality	Aesthetic Quality
Rolling Forcing	✗	15.38	95.8	95.1	35.9	98.9	72.5	63.6
LongLive	✗	-	95.5	95.4	44.5	98.8	71.7	65.0
CausVid	✗	15.79	91.2	91.4	50.8	98.1	70.2	63.5
CV + Ours	✓	10.53	93.2	93.3	69.5	97.6	70.1	63.5
Self-Forcing	✗	15.79	92.5	93.2	62.5	98.0	72.5	63.4
SF + Ours	✓	10.53	94.0	94.2	60.2	98.3	72.7	63.8

Table 2: Comprehensive comparison with SOTA methods on prompt-conditioned 30-second video generation (Color-shift & JEPA). We report Color-shift metrics (L1, Correlation) and JEPA consistency (Variance, Difference).

Method	Training Speed Free	Color-shift		JEPA Consistency	
		L1 ↓	Correlation ↑	Variance ↓	Difference ↓
Rolling Forcing	✗	0.436	0.858	0.0162	0.201
LongLive	✗	0.701	0.724	0.0151	0.101
CausVid	✗	1.047	0.451	0.0199	0.313
CV + Ours	✓	0.607	0.778	0.0157	0.164
Self-Forcing	✗	1.028	0.479	0.0145	0.191
SF + Ours	✓	0.644	0.710	0.0108	0.170

Table 3: Comprehensive comparison with test-time scaling methods, including Best-of-N (BoN) and Search-over-Path (SoP), on prompt-conditioned 30-second video generation. We report Throughput (fps) and VBench metrics.

Method	Train. Speed Free	VBench Metrics ↑						
		Total fps	Subject Consistency	Background Consistency	Dynamic Degree	Motion Smoothness	Imaging Quality	Aesthetic Quality
Self-Forcing	✗	15.79	92.5	93.2	62.5	98.0	72.5	63.4
SF + BoN	✓	3.16	92.4	93.2	62.5	98.4	72.7	63.3
SF + SoP	✓	3.16	92.7	93.4	60.2	98.6	72.7	63.1
SF + Ours	✓	10.53	94.0	94.2	60.2	98.3	72.7	63.8

Standard Evaluation. We benchmark our method against representative autoregressive video diffusion baselines, including CausVid [51], Self-Forcing [20], Rolling Forcing [29], and LongLive [47]. Following standard protocols, evaluations are conducted using VBench [21] on 128 prompts randomly sampled from MovieGen [37]. Unless otherwise specified, we conduct all experiments in the 30-second video generation setting, using Self-Forcing as the default baseline. More experimental settings and implementation details for fair comparison are included in the supplementary material.

Additional Evaluation. To rigorously evaluate temporal coherence, we complement standard VBench quality metrics with temporal color histograms and

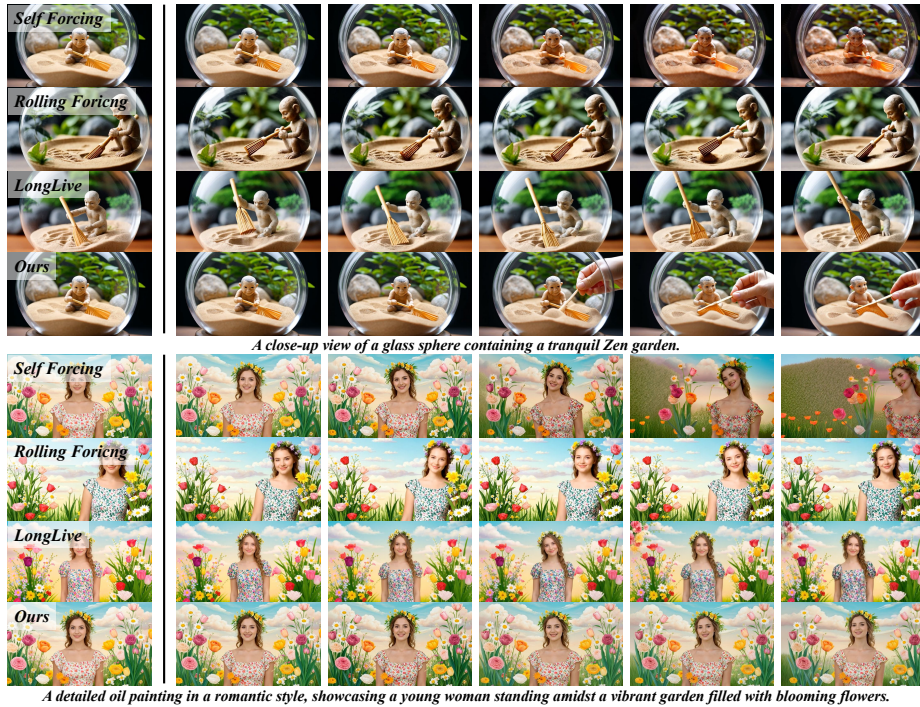


Fig. 7: Qualitative comparison of 30-second long-horizon video generation with Self-Forcing, Rolling Forcing, and LongLive. Our method significantly outperforms Self-Forcing and achieves temporal coherence and visual quality comparable to training-based methods.

JEPA scores [1] to assess long-term temporal drift. Since temporal consistency scores can be artificially improved by suppressing motion, allowing models to effectively cheat the evaluation, we conduct comparisons under matched dynamic degrees to ensure a fair and meaningful assessment. In addition, we use t-LPIPS [57] to explicitly measure visual discontinuities, serving as a direct proxy for flickering artifacts at autoregressive chunk boundaries in our ablation studies. Comprehensive experimental settings and implementation details for fair comparison are included in the supplementary material.

5.2 Results

Qualitative Results. As shown in Figure 7, integrating our method with Self-Forcing substantially reduces error accumulation in long-horizon video generation. While the original baselines exhibit temporal drift and visual degradation over time, our integration maintains stable temporal coherence and visual fidelity over 30-second sequences, especially in videos with complex motion and appearance changes. Under the 30-second setting, our training-free approach achieves

Table 4: Ablation study on noise-correction steps. We evaluate quality using VBench metrics alongside the Boundary metric.

Total NFE	Timesteps			VBench Metrics $\uparrow$						Boundary $\downarrow$
	750	500	250	Subject Consistency	Background Consistency	Dynamic Degree	Motion Smoothness	Imaging Quality	Aesthetic Quality	t-LPIPS
4	$\times$	$\times$	$\times$	92.5	93.2	62.5	98.0	72.5	63.4	0.178
5	$\checkmark$	$\times$	$\times$	93.6	94.3	60.2	98.6	72.6	63.2	0.161
5	$\times$	$\checkmark$	$\times$	93.2	93.9	60.9	98.5	72.8	63.1	0.182
5	$\times$	$\times$	$\checkmark$	93.6	94.1	57.0	98.5	72.9	63.4	0.183
6	$\times$	$\checkmark$	$\checkmark$	94.0	94.2	60.2	98.3	72.7	63.8	0.176
6	$\checkmark$	$\checkmark$	$\times$	93.1	93.9	61.7	98.4	73.0	63.1	0.170
7	$\checkmark$	$\checkmark$	$\checkmark$	93.4	94.2	62.5	98.5	72.4	63.8	0.169

visual quality comparable to, and in some cases better than, Rolling Forcing and LongLive, which rely on additional training or specialized mechanisms. These results demonstrate that our method provides an effective and general test-time solution for improving long-term temporal consistency in autoregressive video generation. More experimental settings and results are included in the supplementary material.

Quantitative Results. All quantitative results are summarized in Table 1 and Table 2. Under the 30-second generation setting on VBench, integrating our method into standard autoregressive baselines, Self-Forcing and CausVid, consistently improves long-horizon video generation quality across diverse prompts and scenes. In particular, our method substantially reduces error accumulation and temporal drift, leading to improved subject and background consistency while notably enhancing dynamic degree without sacrificing motion smoothness or imaging quality. Moreover, the proposed path-wise correction effectively stabilizes appearance evolution over time, as evidenced by lower color-shift L1 distances and higher histogram correlations between the first and last frames. At the semantic level, our method also improves JEPA consistency by reducing both the standard deviation and first-last score difference across the entire sequence, indicating more coherent long-term representations. Compared with training-based methods such as Rolling Forcing and LongLive, our approach achieves comparable long-horizon consistency and visual quality while preserving stronger motion dynamics and requiring no additional training or parameter updates at test time.

Comparison with Test-Time Scaling. We benchmark our approach against test-time scaling strategies, including Best-of-N (BoN) and Search-over-Path (SoP), as shown in Table 3. While these methods attempt to mitigate errors through redundant candidate generation or iterative search, they incur prohibitive computational overhead and inference latency. In contrast, our method embeds correction directly into a single stochastic sampling trajectory. This design drastically reduces inference costs compared to multi-sample scaling, and by *actively* rectifying structural deviations rather than *passively* selecting from

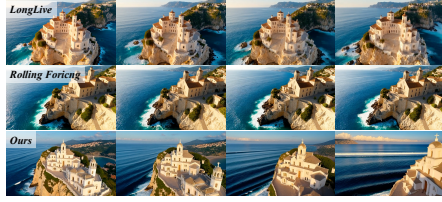


Fig.8: Comparison between the sink-based method and path-wise correction. The sink-based method overly constrains intermediate states, leading to degraded motion dynamics and reduced temporal variation.



Fig.9: Comparison of single-point and path-wise correction. Single-point correction causes temporal discontinuities, while on-path re-noising improves temporal stability and reduces flickering.

drifting candidates, it achieves superior suppression of long-term error accumulation with minimal overhead. More experimental settings and implementation details for are included in the supplementary material.

Ablation Study on Pathwise Correction. We compare *single-point* and *path-wise* correction to evaluate the role of the stochastic sampling trajectory in practice. Single-point correction directly replaces the latent at a fixed denoising step, whereas path-wise correction re-noises the corrected prediction to the same noise level and resumes denoising along the original trajectory. As shown in Figure 9 and Table 5, single-point correction frequently introduces flickering and temporal instability, leading to degraded consistency metrics and higher t-LPIPS scores. In contrast, path-wise correction achieves consistently higher temporal consistency and substantially lower t-LPIPS, resulting in more stable videos with improved temporal coherence. These results demonstrate that effective test-time intervention requires integrating corrections along the sampling path rather than directly replacing latents.

Ablation Study on Noise-correction Steps. After establishing the necessity of path-wise correction, we evaluate how different numbers and placements of correction steps along the sampling path affect long-horizon generation quality as shown in Table 4. Enabling correction at noise levels 750, 500, and 250, either individually or in combination, consistently outperforms the baseline without correction, demonstrating the robustness and effectiveness of our method across different configurations. Balancing generation quality and inference overhead, we choose correction at noise levels 500 and 250 as our default, since this configuration yields consistent improvements in both quantitative metrics and qualitative visual fidelity.

Comparison with the Sink-based Method. We compare our proposed path-wise correction with the Sink-based method. The Sink-based approach keeps the Sink frame as visible context throughout the entire denoising process, effectively imposing persistent conditioning. In contrast, path-wise correction is applied only at later stages after structural stabilization, where corrected predictions

Table 5: Comparison of correction strategies, including single-point correction and pathwise correction. We evaluate quality using VBench metrics alongside the Boundary metric.

Method	VBench Metrics $\uparrow$						Boundary $\downarrow$
	Subject Consistency	Background Consistency	Dynamic Degree	Motion Smoothness	Imaging Quality	Aesthetic Quality	t-LPIPS
Single-point	93.4	94.0	57.0	98.3	71.6	62.8	0.205
Pathwise	94.0	94.2	60.2	98.3	72.7	63.8	0.176

Table 6: Comparison on prompt-conditioned 5-second video generation. We evaluate quality using standard VBench metrics.

Method	VBench Metrics $\uparrow$					
	Subject Consistency	Background Consistency	Dynamic Degree	Motion Smoothness	Imaging Quality	Aesthetic Quality
CausVid	96.2	94.9	54.7	98.2	70.5	63.8
CausVid + Ours	96.6	95.2	68.0	97.8	70.5	64.2
Self-Forcing	97.0	96.2	62.5	98.7	72.9	64.5
Self-Forcing + Ours	97.0	96.3	62.5	98.7	73.0	64.6

are re-noised and integrated along the stochastic sampling trajectory. Because the Sink frame continuously participates in all denoising steps, the model becomes overly conditioned on it, causing generated content to remain visually and structurally close to the Sink frame. This static conditioning restricts motion and scene variation, suppressing temporal dynamics, as shown in Table 1 and Figure 8. By contrast, path-wise correction preserves structural flexibility in early stages and introduces correction only during appearance refinement, maintaining temporal coherence while retaining meaningful video dynamics.

Comparison on Short Video Generation. As shown in Table 6, we evaluate our method on short video generation. Although error accumulation is less pronounced under short temporal horizons, our method still consistently outperforms the baseline across most metrics. This indicates that the proposed correction strategy is not specialized to long-horizon generation, but also remains effective in short video settings. Together with the significant improvements observed for long video generation, these results demonstrate the robustness and general applicability of our approach.

User Study. We further conduct an anonymized human evaluation to assess the perceptual quality of different methods. The study includes 14 participants, and each method is evaluated on 20 generated videos. Participants rate each video on a 0–10 scale from five aspects: text alignment, temporal consistency, color consistency, dynamic quality, and overall human preference, where higher scores indicate better quality. As shown in Table 7, our method achieves the best scores on temporal consistency and dynamic quality, and obtains competitive results on text alignment, color consistency, and overall preference. These results show

Table 7: User Study on Short Video Generation. We report the average human evaluation scores from 14 participants. The evaluation covers text alignment, temporal consistency, color consistency, dynamic quality, and overall human preference. Higher scores indicate better quality.

Method	Human Evaluation Scores $\uparrow$				
	Text Alignment	Temporal Consistency	Color Consistency	Dynamic Quality	Overall Preference
Rolling Forcing	8.6	8.8	8.1	6.1	8.2
LongLive	9.1	8.9	9.0	6.6	8.9
Self-Forcing	8.3	7.1	6.2	8.0	5.1
Ours	8.7	9.0	8.7	8.6	8.5

that the proposed correction strategy not only improves quantitative metrics, but also leads to perceptually more stable and temporally coherent videos.

6 Conclusion

In this paper, we propose Test-Time Correction, a training-free test-time method for stabilizing distilled autoregressive diffusion models in long-horizon video generation. The proposed approach addresses error accumulation by introducing training-free, reference-based correction along the stochastic sampling process, allowing corrected predictions to be smoothly inherited by subsequent denoising steps. Without modifying model parameters or requiring additional training, our method effectively suppresses temporal drift while preserving the original generation behavior. Extensive experiments demonstrate that Test-Time Correction consistently improves long-horizon stability across multiple distilled video generation models, extending the achievable generation length to 30 seconds with a slight computational overhead and competitive visual quality.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China, under Grant Nos. 62192783, 62276128, and 62406140; the Young Elite Scientists Sponsorship Program by China Association for Science and Technology, under Grant No. 2023QNRC001; the Key Research and Development Program of Jiangsu Province, under Grant No. BE2023019; and the Jiangsu Natural Science Foundation, under Grant Nos. BK20221441 and BK20241200. The authors would like to thank the support of Huawei Ascend Cloud Ecological Development Project.

References

1. Balestrieri, R., Ballas, N., Rabbat, M., LeCun, Y.: Gaussian embeddings: How jepas secretly learn your data density. CoRR (2025)

2. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: CVPR (2023)
3. Cai, S., Yang, C., Zhang, L., Guo, Y., Xiao, J., Yang, Z., Xu, Y., Yang, Z., Yuille, A., Guibas, L., et al.: Mixture of contexts for long video generation. CoRR (2025)
4. Chen, B., Martí Monsó, D., Du, Y., Simchowitz, M., Tedrake, R., Sitzmann, V.: Diffusion forcing: Next-token prediction meets full-sequence diffusion. NeurIPS (2024)
5. Chen, G., Lin, D., Yang, J., Lin, C., Zhu, J., Fan, M., Zhang, H., Chen, S., Chen, Z., Ma, C., et al.: Skyreels-v2: Infinite-length film generative model. CoRR (2025)
6. Chen, Y., Liang, Y., Wang, J., Chen, T., Cheng, J., Gu, Z., Huang, Y., Jiang, Z., Li, W., Li, T., et al.: Teleworld: Towards dynamic multimodal synthesis with a 4d world model. CoRR (2025)
7. Cui, J., Wu, J., Li, M., Yang, T., Li, X., Wang, R., Bai, A., Ban, Y., Hsieh, C.J.: Self-forcing++: Towards minute-scale high-quality video generation. CoRR (2025)
8. Cui, J., Wu, J., Li, M., Yang, T., Li, X., Wang, R., Bai, A., Ban, Y., Hsieh, C.J.: Lol: Longer than longer, scaling video generation to hour. CoRR (2026)
9. Deng, H., Pan, T., Diao, H., Luo, Z., Cui, Y., Lu, H., Shan, S., Qi, Y., Wang, X.: Autoregressive video generation without vector quantization. In: ICLR (2025)
10. Eyring, L., Karthik, S., Dosovitskiy, A., Ruiz, N., Akata, Z.: Noise hypernetworks: Amortizing test-time compute in diffusion models. CoRR (2025)
11. Guo, Y., Yang, C., He, H., Zhao, Y., Wei, M., Yang, Z., Huang, W., Lin, D.: End-to-end training for autoregressive video diffusion via self-resampling. CoRR (2025)
12. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In: ICLR (2024)
13. He, H., Liang, J., Wang, X., Wan, P., Zhang, D., Gai, K., Pan, L.: Scaling image and video generation via test-time evolutionary search. CoRR (2025)
14. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In: ICLR (2023)
15. Hong, Y., Liu, B., Wu, M., Zhai, Y., Chang, K.W., Li, L., Lin, K., Lin, C.C., Wang, J., Yang, Z., Wu, Y.N., Wang, L.: Slowfast-VGen: Slow-fast learning for action-driven long video generation. In: ICLR (2025)
16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: ICLR (2022)
17. Hu, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In: CVPR (2024)
18. Huang, J., Hu, X., Han, B., Shi, S., Tian, Z., He, T., Jiang, L.: Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft. CoRR (2025)
19. Huang, T., Zheng, W., Wang, T., Liu, Y., Wang, Z., Wu, J., Jiang, J., Li, H., Lau, R.W.H., Zuo, W., Guo, C.: Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. SIGGRAPH (2025)
20. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. In: NeurIPS (2025)
21. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: CVPR (2024)
22. HunyuanWorld, T.: Hy-world 1.5: A systematic framework for interactive world modeling with real-time latency and geometric consistency. arXiv preprint (2025)

23. Ji, S., Chen, X., Yang, S., Tao, X., Wan, P., Zhao, H.: Memflow: Flowing adaptive memory for consistent and efficient long video narratives. *CoRR* (2025)
24. Jia, W., Lu, Y., Huang, M., Wang, H., Huang, B., Chen, N., Liu, M., Jiang, J., Mao, Z.: Moga: Mixture-of-groups attention for end-to-end long video generation. *CoRR* (2025)
25. Jia, W., Lu, Y., Huang, M., Wang, H., Huang, B., Chen, N., Liu, M., Jiang, J., Mao, Z.: Moga: Mixture-of-groups attention for end-to-end long video generation. *CoRR* (2025)
26. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. *CoRR* (2024)
27. Liang, J., He, R., Tan, T.: A comprehensive survey on test-time adaptation under distribution shifts. *IJCV* (2025)
28. Liu, F., Wang, H., Cai, Y., Zhang, K., Zhan, X., Duan, Y.: Video-t1: Test-time scaling for video generation. *CoRR* (2025)
29. Liu, K., Hu, W., Xu, J., Shan, Y., Lu, S.: Rolling forcing: Autoregressive long video diffusion in real time. *CoRR* (2025)
30. Liu, M., Liu, J., Zhang, Y., Li, J., Yang, M.Y., Nex, F., Cheng, H.: 4dstr: Advancing generative 4d gaussians with spatial-temporal rectification for high-quality and consistent 4d generation. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 40, pp. 7224–7232 (2026)
31. Liu, M., Zhang, D., Liu, J., Cui, J., Xie, H., Chen, G., Ye, H., Yang, M.Y., Nex, F., Cheng, H.: Driveva: Video action models are zero-shot drivers. *arXiv preprint arXiv:2604.04198* (2026)
32. Lu, Y., Zeng, Y., Li, H., Ouyang, H., Wang, Q., Cheng, K.L., Zhu, J., Cao, H., Zhang, Z., Zhu, X., et al.: Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. *CoRR* (2025)
33. Ma, X., Wang, Y., Chen, X., Jia, G., Liu, Z., Li, Y., Chen, C., Qiao, Y.: Latte: Latent diffusion transformer for video generation. *TMLR* (2025)
34. Ma, Y., Liu, C., Wang, J., Liu, J., Huang, H., Wu, Z., Zhang, C., Li, X.: Tempomaster: Efficient long video generation via next-frame-rate prediction. *CoRR* (2025)
35. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. *TMLR* (2024)
36. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: *ICCV* (2023)
37. Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie gen: A cast of media foundation models. *CoRR* (2024)
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: *ICML* (2021)
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *CVPR* (2022)
40. Sun, W., Zhang, H., Wang, H., Wu, J., Wang, Z., Wang, Z., Wang, Y., Zhang, J., Wang, T., Guo, C.: Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *CoRR* (2025)
41. Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M.: Test-time training with self-supervision for generalization under distribution shifts. In: *ICML* (2020)

42. Teng, H., Jia, H., Sun, L., Li, L., Li, M., Tang, M., Han, S., Zhang, T., Zhang, W., Luo, W., et al.: Magi-1: Autoregressive video generation at scale. CoRR (2025)
43. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. CoRR (2025)
44. Wang, R., Sun, Y., Tandon, A., Gandelsman, Y., Chen, X., Efros, A.A., Wang, X.: Test-time training on video streams. JMLR (2025)
45. Xiang, X., Chen, Y., Zhang, G., Wang, Z., Gao, Z., Xiang, Q., Shang, G., Liu, J., Huang, H., Gao, Y., et al.: Macro-from-micro planning for high-quality and parallelized autoregressive long video generation. CoRR (2025)
46. Xiang, X., Xue, H., Dai, Z., Wang, D., Li, M., Yue, Y., Ma, F., Yu, W., Chang, H., Yu, F.R.: Remask-animate: Refined character image animation using mask-guided adapters. In: AAAI (2025)
47. Yang, S., Huang, W., Chu, R., Xiao, Y., Zhao, Y., Wang, X., Li, M., Xie, E., Chen, Y., Lu, Y., et al.: Longlive: Real-time interactive long video generation. CoRR (2025)
48. Yesiltepe, H., Meral, T.H.S., Akan, A.K., Oktay, K., Yanardag, P.: Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollback. CoRR (2025)
49. Yi, J., Jang, W., Cho, P.H., Nam, J., Yoon, H., Kim, S.: Deep forcing: Training-free long video generation with deep sink and participative compression. CoRR (2025)
50. Yin, S., Wu, C., Yang, H., Wang, J., Wang, X., Ni, M., Yang, Z., Li, L., Liu, S., Yang, F., et al.: Nuwa-xl: Diffusion over diffusion for extremely long video generation. In: ACL (2023)
51. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast autoregressive video diffusion models. In: CVPR (2025)
52. Yu, J., Bai, J., Qin, Y., Liu, Q., Wang, X., Wan, P., Zhang, D., Liu, X.: Context as memory: Scene-consistent interactive long video generation with memory retrieval. In: SIGGRAPH Asia (2025)
53. Yu, Z., Hayakawa, A., Ishii, M., Yu, Q., Shibuya, T., Zhang, J., Mitsufuji, Y.: Autorefiner: Improving autoregressive video diffusion models via reflective refinement over the stochastic sampling path. CoRR (2025)
54. Zhang, G., Shi, C., Jiang, Z., Xiang, X., Qian, J., Shi, S., Jiang, L.: Proteus-id: Id-consistent and motion-coherent video customization. CoRR (2025)
55. Zhang, L., Cai, S., Li, M., Wetzstein, G., Agrawala, M.: Frame context packing and drift prevention in next-frame-prediction video diffusion models. In: NeurIPS (2025)
56. Zhang, P., Chen, Y., Su, R., Ding, H., Stoica, I., Liu, Z., Zhang, H.: Fast video generation with sliding tile attention. In: ICML (2025)
57. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
58. Zhu, S., Chen, J.L., Dai, Z., Dong, Z., Xu, Y., Cao, X., Yao, Y., Zhu, H., Zhu, S.: Champ: Controllable and consistent human image animation with 3d parametric guidance. In: ECCV (2024)