Cycle-World: Mitigating Error Accumulation in Long-term Video World Models via Reverse-Prediction Cycle Consistency

Zihan Su[1⋆], Teng Hu[1⋆], Jiangning Zhang[2], Ruiyan Wang[1], Ran Yi[1†], Lizhuang Ma[1†], and Dacheng Tao[3]

[1] School of Computer Science, Shanghai Jiao Tong University, Shanghai 200240, China
[2] Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou, China
[3] Nanyang Technological University, Singapore

Abstract. Autoregressive diffusion models have enabled high-quality video generation, yet their sequential nature inherently suffers from error accumulation. In long-horizon video synthesis, minor prediction deviations compound over time, inevitably leading to unconstrained generative drift, structural collapse, and severe visual degradation. To address this, we propose Cycle-World, a novel framework designed for stable and temporally consistent long-video generation. Our approach tackles error drift by enforcing strict temporal reversibility across both the training and inference phases. Theoretically, we demonstrate that forward generative drift can be strictly bottlenecked by a cycle-consistency objective. During training, we integrate an efficient reverse-prediction model to implicitly embed causal constraints into the forward generator, compelling it to produce reversible sequences that tightly adhere to the natural video manifold. At inference time, we repurpose this frozen reverse model as a runtime corrector. Through gradient-based cycle guidance, it iteratively refines the generated latent representations, actively suppressing accumulated errors before they are committed to the historical context. Extensive experiments on the VBench benchmark demonstrate that Cycle-World's dual-phase synergy significantly mitigates error drift, achieving state-of-the-art overall generation quality and long-horizon temporal consistency in 60-second synthesis.

Keywords: Video generation · Cycle consistency · Error accumulation

1 Introduction

The field of video generation has witnessed unprecedented advancements recently, driven by powerful models such as Sora [36, 37], Seedance [12, 41], and

⋆ Equal contribution, † Corresponding authors.

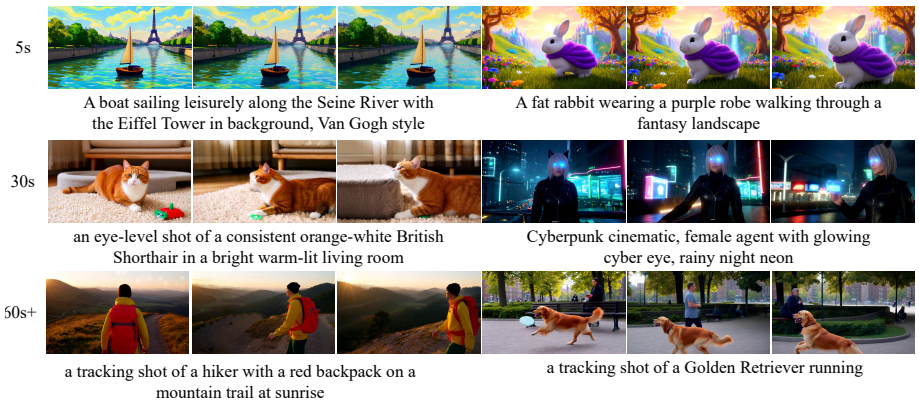

Fig. 1: High-fidelity, long-horizon video generation with Cycle-World. By effectively bottlenecking generative drift via temporal reversibility, our framework strictly suppresses structural hallucinations inherent in forward-only models. Cycle-World maintains state-of-the-art visual quality, strict physical conservation, and temporally consistent object states over extended generation horizons.

Kling [44], alongside pioneering open-source efforts like Wan [45], Hunyuan-Video [29]. While these models primarily rely on bidirectional or full-sequence architectures that achieve remarkable visual quality, their non-causal nature fundamentally limits their flexibility for open-ended, sequential, and interactive generation. As the community pivots towards the paradigm of World Models [17, 25, 34, 42], there is a critical consensus that real-time interactivity, continuous generation, and causal reasoning are indispensable. Consequently, the field is experiencing a paradigm shift towards causal, autoregressive generation models [16, 23, 56, 60].

However, this shift towards causal autoregressive models introduces severe bottlenecks in long video generation. A primary challenge is the train-inference mismatch, often leading to rapid accumulation of errors. Traditionally, models are trained using *Teacher-Forcing*, which strictly relies on ground-truth past frames, causing significant exposure bias: during inference, the autoregressive model relies on its own past predictions, where minor deviations — unseen during training — propagate and amplify. To bridge this gap and mitigate drift, the community has continuously evolved towards more robust training paradigms. This progression spans from the introduction of *Diffusion Forcing* [5] to recent advanced train-inference alignment strategies, such as *Self-Forcing* [23] and *LongLive* [50].

Nevertheless, we observe that even with these sophisticated mitigations, forward-only generation still suffers from a more fundamental and destructive issue: *structural hallucinations*. While existing progressive methods effectively suppress generic noise accumulation, they operate strictly in a unidirectional, forward-time manner. Crucially, they lack **temporal cycle consistency**—the

explicit temporal constraints required to ensure that a generated state logically allows its past to have happened, such that past states could be reversely predicted from future states. Without this bidirectional verification, the models lack the physical constraints necessary to maintain continuous object states. Consequently, they are prone to severe video distortion and non-physical artifacts—such as characters clipping through solid objects, entities spontaneously appearing, or objects vanishing without a trace. Unlike generic noise, these structural hallucinations represent irreversible violations of real-world physics that abruptly destroy the integrity of the generated sequence.

To fundamentally address these irreversible structural hallucinations, we propose **Cycle-World**, a unified framework that enforces physical consistency via temporal reversibility. Our core motivation is grounded in a fundamental premise: *if a generated causal sequence obeys real-world dynamics, it must be temporally reversible*. Based on this insight, we first establish the **Cycle-Bounded Drift (CBD)** theorem, formally proving that the unconstrained error accumulation in forward autoregressive synthesis can be strictly bottlenecked by minimizing the reverse reconstruction error. Guided by this theoretical guarantee, we translate the mathematical bound into a practical **Cycle-Consistent Learning (CCL)** paradigm. By introducing a reverse-prediction branch, we explicitly constrain the forward generator to produce inherently reversible latents, enabling it to foresee and suppress non-physical artifacts during training. Furthermore, to extend the applicability of our theory to pre-trained models and resource-constrained scenarios, we propose **Cycle-Guided Inference (CGI)**. This inference-time strategy repurposes the reverse model as a runtime critic to iteratively refine latents, offering a plug-and-play solution that significantly boosts stability without the need for expensive architectural modifications.

Extensive experiments conducted on VBench validate the effectiveness of our framework. Cycle-World significantly mitigates error drift, achieving state-of-the-art visual fidelity, semantic consistency, and overall temporal stability in long-video generation, as shown in Fig. 1.

In summary, our main contributions are threefold:

- We establish the **Cycle-Bounded Drift (CBD)** theorem, a theoretical framework demonstrating that unconstrained generative drift and structural hallucinations in forward autoregressive synthesis can be strictly bottlenecked by enforcing temporal reversibility.
- We propose a **Cycle-Consistent Learning (CCL)** paradigm. By introducing a novel reverse-prediction cycle-consistency loss ($\mathcal{L}_{cycle}$), we explicitly constrain the forward causal generator to internalize physical conservation and maintain long-term structural integrity.
- We introduce **Cycle-Guided Inference (CGI)**, a zero-shot runtime optimization strategy. By repurposing the frozen reverse model as a runtime critic, CGI utilizes iterative gradient-based latent refinement to actively rectify non-physical artifacts, significantly enhancing long-term generation quality without architectural modifications to the forward model.

2 Related Work

2.1 Video Generation

Diffusion models [14,35,43], particularly Diffusion Transformers (DiT) [38], have established the prevailing paradigm for video generation [21,29,45,51,59], achieving remarkable visual fidelity. This rapid progress spans various specialized capabilities, including multimodal customized generation [4, 19, 20], and native high-resolution synthesis [22,48], and spatiotemporally consistent video processing [27, 32]. Furthermore, recent advancements have expanded into joint audio-visual generation, leveraging cross-modal interactions and cross-task synergy to achieve synchronized multi-sensory synthesis [18, 30, 33, 46, 57]. However, despite these visual and multimodal achievements, their reliance on non-causal, bidirectional attention for concurrent frame denoising restricts their flexibility for open-ended generation. Conversely, pure autoregressive (AR) models [15, 28] offer sequential flexibility via discrete next-token prediction but suffer from irreversible information loss during latent compression and suboptimal generation diversity.

To bridge this gap, recent works explore hybrid Autoregressive Diffusion architectures [10, 11, 23, 26, 56], which temporally decompose generation by conditioning future frames on past context. Despite their promise, these models suffer from severe exposure bias. During iterative inference, conditioning on imperfect prior predictions causes minor deviations to amplify continuously. This catastrophic error accumulation constitutes the primary bottleneck our work addresses.

2.2 Long Video Generation

Early long-video approaches relied on generating overlapping clips [8, 39, 47] or performing temporal interpolation between sparse keyframes [3, 53]. While extending the generation window, these heuristic designs fail to achieve true, infinitely long streaming synthesis. Consequently, the focus has shifted toward native causal modeling. However, traditional training paradigms for these models often employ Teacher Forcing [40, 49], which inevitably introduces a severe distribution discrepancy between the training and inference stages. To alleviate this issue, Diffusion Forcing [5] proposes joint denoising optimization for tokens with independent noise levels, which SkyReels-V2 [6] further integrates with Multi-modal Large Language Models to facilitate infinite-length, cinematic video synthesis. CausVid [56] extends distribution matching distillation [54, 55] to the video domain to mitigate error accumulation in AR generation. To resolve exposure bias and bridge the train-test gap, Self-Forcing [23] innovatively simulates inference conditions directly during training by executing autoregressive rollouts with a Key-Value cache, optimizing the model conditioned on its own historically generated outputs. Most recent cutting-edge works build upon this foundation. LongLive [50], for instance, adopts streaming long-sequence fine-tuning to strictly maintain end-to-end "train-long-test-long" consistency.

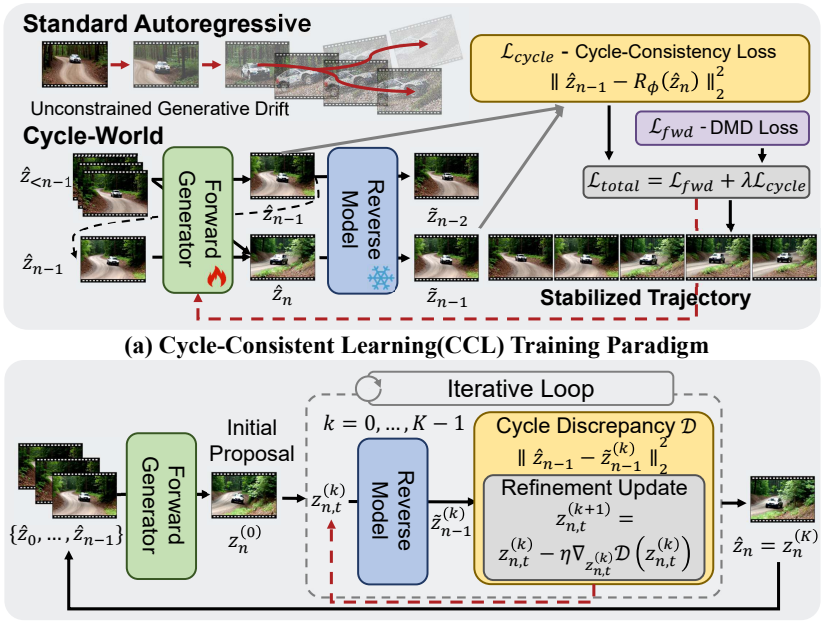

Fig. 2: Overview of Cycle-World. We mitigate generative drift in autoregressive video synthesis by enforcing temporal reversibility. **(a) Training (CCL):** The forward generator G_θ and reverse model R_ϕ are jointly optimized. A latent cycle-consistency loss ($\mathcal{L}_{cycle}$) explicitly penalizes physically irreversible trajectories. **(b) Inference (CGI):** The frozen R_ϕ acts as a runtime corrector. By evaluating cycle discrepancy ($\mathcal{D}$), it performs iterative gradient-based latent refinement to actively prune accumulated errors before they enter the historical context.

3 Methodology

The overarching goal of **Cycle-World** is to mitigate the unconstrained generative drift and structural hallucinations inherent in long-term autoregressive video synthesis. To achieve this, we introduce a novel framework grounded in the principle of temporal reversibility—the intuition that physically valid and structurally sound video dynamics must be accurately reversible. As illustrated in Figure 2, our approach addresses this challenge through a cohesive pipeline encompassing theoretical grounding, cycle-consistent learning, and inference-time optimization. Specifically, we first establish the **Cycle-Bounded Drift (CBD)** theorem (Sec. 3.1), a theoretical foundation demonstrating that the forward generative error can be strictly constrained by minimizing a reverse-prediction cycle-consistency error. Guided by this insight, we propose a **Cycle-Consistent Learning (CCL)** paradigm (Sec. 3.2). In this phase, we introduce a reverse-prediction model R_ϕ to enforce a cycle-consistency loss ($\mathcal{L}_{cycle}$) on the forward causal generator G_θ, explicitly penalizing irreversible structural hallucinations.

Finally, to combat out-of-distribution perturbations during open-ended generation, we introduce **Cycle-Guided Inference** (Sec. 3.3). Under this strategy, we repurpose the frozen reverse model as a runtime corrector. By actively evaluating cycle discrepancy, it performs iterative gradient-based latent refinement to prune non-physical artifacts before they are committed to the historical context.

3.1 Theoretical Foundation: Bounding Generative Drift via Temporal Reversibility

Let $Z = \{z_1, z_2, \ldots, z_N\}$ denote the ground-truth latent sequence of a natural video, where each z_n is a latent chunk. In standard autoregressive synthesis, a forward causal generator G_θ sequentially predicts $\hat{z}_n$ conditioned on the previously generated context $\hat{z}_{<n}$. Because this sequential generation is inherently unconstrained, minor prediction deviations compound at each step n, leading to unbounded generative drift and structural collapse in long video synthesis.

We theoretically argue that this generative drift can be strictly constrained by enforcing temporal reversibility. Because natural video dynamics obey physical laws and spatiotemporal causality, they are inherently reversible. If a generated frame $\hat{z}_n$ severely deviates from the natural manifold, it loses the causal information necessary to reconstruct its history.

Let $e_n = \|\hat{z}_n - z_n\|$ be the accumulated generative drift at step n. To constrain this, we introduce a reverse-prediction model R_ϕ mapping a state at n back to $n-1$. We establish our theoretical guarantees based on two mild assumptions:

Assumption 1 (Reverse Predictability of Natural Dynamics) *The reverse model R_ϕ is comprehensively trained under a self-forcing paradigm [23]. Given this aligned training scheme, its approximation error for short-horizon reverse prediction is bounded by a small constant $\epsilon_R > 0$. That is, $\|R_\phi(z_n) - z_{n-1}\| \leq \epsilon_R$.*

Assumption 2 (Reverse-Lipschitz Continuity) *The learned reverse mapping R_ϕ preserves the distance properties within the relevant latent manifold, satisfying a reverse-Lipschitz condition. There exists a constant $C > 0$ such that for any two states z_a and z_b, $\|z_a - z_b\| \leq C \|R_\phi(z_a) - R_\phi(z_b)\|$.*

Under these conditions, we demonstrate that the forward generative error e_n is constrained by the cycle-consistency objective and the error from the previous steps. We formally define the single-step cycle-consistency distance as $d_{cycle}^{(n)} = \|\hat{z}_{n-1} - R_\phi(\hat{z}_n)\|$.

Due to space constraints, all detailed proofs for the theoretical results presented in this section are deferred to Supplementary Material.

Theorem 1 (Cycle-Bounded Drift). *Under Assumptions 1 and 2, the forward generative drift e_n satisfies:*

$$e_n \leq C \left(d_{cycle}^{(n)} + e_{n-1} + \epsilon_R \right). \tag{1}$$

To understand the compounding drift over a long horizon, we recursively unroll this step-wise recurrence.

Corollary 1 (Long-Horizon Error Bound). *Assuming a worst-case upper limit on the single-step cycle-consistency distance,* $\max_i d_{cycle}^{(i)} \leq \delta_{cycle}$*, the total generative drift at step* n *(for* $C \neq 1$*) is explicitly governed by:*

$$e_n \leq C^n e_0 + (\delta_{cycle} + \epsilon_R) C \frac{C^n - 1}{C - 1}. \tag{2}$$

Building on Corollary 1, next we show Cycle-World framework is theoretically superior to unconstrained baselines. Let $\delta_{unc} = \max_i \|\hat{z}_{i-1}^{unc} - R_\phi(\hat{z}_i^{unc})\|$ represent the maximum local cycle error of a standard, unconstrained generator, and δ_{cycle} represent our constrained error, where $\delta_{cycle} \ll \delta_{unc}$.

Proposition 1 (Theoretical Advantage over Unconstrained Baselines). *Given the identical initial drift condition* e_0 *and the same pre-trained reverse model* R_ϕ *(with properties defined in Assumptions 1 and 2), let* E_n^{unc} *and* E_n^{ours} *denote the theoretical upper limits of the generative drift at step* n *for the unconstrained baseline and our constrained method, respectively. The explicit drift reduction (the gap between these theoretical limits) achieved by our method is:*

$$\Delta E_n = E_n^{unc} - E_n^{ours} = (\delta_{unc} - \delta_{cycle}) C \frac{C^n - 1}{C - 1} > 0. \tag{3}$$

Remark. Proposition 1 is the theoretical cornerstone of our method. The term $(\delta_{unc} - \delta_{cycle})$ represents the single-step advantage gained by explicitly enforcing temporal reversibility. Crucially, the multiplier $C\frac{C^n-1}{C-1}$ implies that this advantage is not merely additive, but scales with the sequence length n. This proves that while standard generation and our method might perform similarly for very short clips, the unconstrained baseline will inevitably suffer from a much faster error explosion over time. By minimizing the single-step cycle error upper bound δ_{cycle}, Cycle-World effectively suppresses the base magnitude of the accumulated generative drift, theoretically guaranteeing significantly better structural preservation in long-video generation.

3.2 Cycle-Consistent Learning: Enforcing Cycle Consistency via Reverse Prediction

Constraining the Forward Model via Temporal Cycle Consistency. Building upon the theoretical guarantees established in Section 3.1—specifically Proposition 1, which proves that bounding single-step cycle error mathematically curtails compounding generative drift—we translate this insight into a practical learning framework. we extend a forward autoregressive video generator G_θ predicting latent chunks $\hat{z}_n$ given history $\hat{z}_{<n}$. Since standard maximum likelihood training optimizes solely forward prediction, unconstrained reverse consistency causes unbounded error accumulation scaling with $C\frac{C^n-1}{C-1}$ over sequence length n (Corollary 1). To minimize this bound and mitigate structural hallucinations, we introduce a reverse-prediction model R_ϕ enforcing temporal cycle consistency.

The Pixel-Latent Mismatch Bottleneck. Implementing R_ϕ naively by training a separate autoregressive model on reversed videos incurs severe dual distillation overhead and faces a prohibitive structural barrier: pixel-latent mismatch. Because Video VAEs temporally compress continuous frame sequences into a compact latent representation, the latent features of a forward sequence do not exhibit simple temporal symmetry with those of a reversed sequence. This mismatch precludes the direct computation of the cycle-consistency distance, $d_{cycle}^{(n)} = \|\hat{z}_{n-1} - R_\phi(\hat{z}_n)\|$ defined in Theorem 1. Aligning the forward-ordered chunk $\hat{z}_{n-1}$ and a reverse-predicted chunk residing in disjoint manifolds requires an exorbitant Decode-Flip-Encode loop of decoding, temporally flipping, and re-encoding, rendering joint training computationally infeasible.

Our Solution: Intrinsic Latent Reversibility. To fundamentally circumvent this bottleneck and the inherent feature mismatch, we propose *Intrinsic Latent Reversibility*, redefining the cycle consistency objective directly on the intrinsic latent manifold rather than mapping back to the extrinsic pixel domain. R_ϕ need not predict decoded, temporally flipped frames. Instead, we formulate the reverse task as learning the inverse transition dynamics within the latent space. We set the optimization target of $R_\phi(\hat{z}_n)$ to the forward-ordered history $\hat{z}_{n-1}$, learning the inverse probability $P(z_{n-1}|\hat{z}_n)$ directly. This approach eliminates the need for "Decode-Flip-Encode" loop. Consequently, the cycle-consistency distance $d_{cycle}^{(n)}$ reduces to a direct vector subtraction in the shared latent space. This allows us to enforce strict structural constraints with negligible computational overhead, making joint training feasible.

Forward Generation via Self-Forcing. To bridge the train-inference discrepancy, we train G_θ via self-forcing to predict the current chunk $\hat{z}_n$ given its generated history $\hat{z}_{<n}$ instead of ground truth. Since Distribution Matching Distillation (DMD) lacks paired regression constraints, optimizing this process updates the forward objective $\mathcal{L}_{fwd}$ using its score-difference gradient:

$$\nabla_\theta \mathcal{L}_{fwd} = -\mathbb{E}_{t,\epsilon}\left[(s_{real}(\hat{z}_{n,t}, t) - s_{fake}(\hat{z}_{n,t}, t))\,\frac{\partial G_\theta(\hat{z}_{<n})}{\partial \theta}\right], \tag{4}$$

where $\hat{z}_{n,t}$ is the noisy latent at diffusion timestep t. While this gradient update ensures high-fidelity chunk generation, it lacks explicit penalties for the structural drift that accumulates over long sequences.

Latent Cycle-Consistency Objective. To bound this generative drift, we implement the proposed Intrinsic Latent Reversibility via a reverse-prediction model R_ϕ. As established, R_ϕ operates directly on the latent manifold, tasked with reconstructing the *preceding* latent z_{n-1} given the current generation $\hat{z}_n$. Formally, let $\hat{z}_n$ be the latent chunk synthesized by the forward generator. The reverse model attempts to predict the immediate history $\tilde{z}_{n-1} = R_\phi(\hat{z}_n)$. The cycle-consistency objective is defined as the expected squared Euclidean distance between the *actual autoregressive conditioning context* used by the forward model and the *reconstructed history* inferred by the reverse model:

$$\mathcal{L}_{cycle} = \mathbb{E}_{\hat{z}_n \sim G_\theta}\left[\|\hat{z}_{n-1} - R_\phi(\hat{z}_n)\|_2^2\right]. \tag{5}$$

Minimizing this objective explicitly enforces the invertibility assumption (Assumption 2), ensuring that the generated $\hat{z}_n$ retains sufficient causal information to recover its origin, thereby preventing error accumulation.

Joint Optimization. The final training objective seamlessly integrates the distribution-level supervision with our structural cycle constraint:

$$\mathcal{L}_{total} = \mathcal{L}_{fwd} + \lambda \mathcal{L}_{cycle}, \tag{6}$$

where λ is a hyperparameter scaling the penalty strength. Crucially, during backpropagation, the gradients from $\mathcal{L}_{cycle}$ flow through the reverse model R_ϕ and back into the forward generator G_θ. This mechanism implicitly endows G_θ with *foresight*—it penalizes the generation of physically implausible artifacts (like disappearing objects) that, while locally reasonably under $\mathcal{L}_{fwd}$, fail to accurately reconstruct their history, explicitly pruning divergent trajectories during the learning phase.

3.3 Cycle-Guided Inference: Optimizing Generative Latents via Runtime Guidance

While the proposed Cycle-Consistent Learning paradigm ensures the generator intrinsically preserves temporal reversibility, it requires full-scale parameter retraining. In the era of large-scale foundation models, such retraining is often computationally prohibitive or practically infeasible due to closed-source weights. To address this limitation and extend the benefits of temporal reversibility to broader scenarios, we introduce **Cycle-Guided Inference (CGI)**, a zero-shot, training-free latent optimization strategy. Crucially, this strategy is model-agnostic: as long as the target model and the reverse corrector operate within the same latent manifold (sharing the same Video VAE), CGI can be seamlessly plugged into *any* off-the-shelf autoregressive video generator, enabling structural hallucinations mitigation without updating a single model parameter.

Cycle Guidance in Latent Space. During the autoregressive inference phase, the weights of both the forward generator G_θ and the reverse model R_ϕ are strictly frozen. While earlier sections abstract the generation of the n-th block as a single output z_n, actual synthesis in diffusion models involves an iterative denoising process over timesteps $\{t_T, \ldots, t_1\}$. Unlike conventional decoding that passively accepts forward predictions, we actively rectify accumulated errors at intermediate diffusion timesteps committing them to the historical context buffer (KV cache). Let $z_{n,t}$ denote the latent state at diffusion timestep t. First, the forward generator G_θ predicts the corresponding clean latent $\hat{z}_{n|t} = G_\theta(z_{n,t}, t, \mathcal{H})$ based on the historical context $\mathcal{H}$. To evaluate the physical plausibility of this prediction, we compute the cycle discrepancy $\mathcal{D}$. Specifically, the reverse evaluation is formulated as a two-stage autoregressive process. The predicted clean latent is temporally flipped, denoted by the operator $\mathcal{F}(\cdot)$, and processed by the reverse model at a fixed context noise level t_{ctx} to construct a reverse contextual cache $\mathcal{H}_{\text{rev}}$. Subsequently, the reverse model utilizes this newly constructed cache to predict the clean predecessor state from standard Gaussian noise ϵ,

conditioned on the initial diffusion timestep t_T. The discrepancy is defined as the Euclidean distance between the causal condition $\hat{z}_{n-1}$ (the confirmed output of the previous block) and the time-flipped reverse reconstruction:

$$\begin{aligned} R_\phi(\mathcal{F}(\hat{z}_{n|t}^{(k)}), t_{\text{ctx}}, \mathcal{H}_{\text{rev}}^{(k)}), \quad \tilde{z}_{n-1}^{(k)} = \mathcal{F}\left(R_\phi(\epsilon, t_T, \mathcal{H}_{\text{rev}}^{(k)})\right), \\ \mathcal{D}(z_{n,t}^{(k)}) = \left\| \hat{z}_{n-1} - \tilde{z}_{n-1}^{(k)} \right\|_2^2, \end{aligned} \tag{7}$$

where $\hat{z}_{n|t}^{(k)} = G_\theta(z_{n,t}^{(k)}, t, \mathcal{H})$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, and k is optimization iteration index.

Gradient-Based Latent Refinement. Since forward prediction and reverse reconstruction are fully differentiable processes, we can iteratively refine the latent state via cycle guidance. To maximize efficiency and structural impact, this optimization is strategically applied only during a specific window of early denoising timesteps (from T_{start} to T_{end}). We compute the gradient of the cycle discrepancy with respect to the current state $z_{n,t}^{(k)}$ and perform gradient descent:

$$z_{n,t}^{(k+1)} = z_{n,t}^{(k)} - \eta \nabla_{z_{n,t}^{(k)}} \mathcal{D}(z_{n,t}^{(k)}), \tag{8}$$

where η is the optimization step size. After K iterations, the refined state $z_{n,t}^{(K)}$ is detached from the computation graph. This optimized state is then passed to the standard diffusion transition function Ψ to obtain the latent state for the subsequent timestep. Upon reaching the final denoising step t_1, the resulting clean latent $\hat{z}_n$ is appended to the historical context buffer to guide the generation of the next block. This active rectification mechanism acts as a structural bottleneck, preventing inherent drift from manifesting as visual degradation. The complete inference procedure is summarized in Supplementary Material.

4 Experiments

4.1 Experiment Settings

Baselines. We evaluate the proposed method against several video generation baselines, categorized by their architectural paradigms. For bidirectional diffusion and transformer models, we compare with LTX-Video [13] and Wan2.1 [45]. Within the autoregressive family, we evaluate general-purpose models of varying scales, including NOVA [10] (0.6B), SkyReels-V2 [6] (1.3B), Pyramid Flow [26] (2B), and MAGI-1 [11] (4.5B). To ensure a direct comparison with models sharing our foundational architecture and efficient training setup, we include the 1.3B parameter distilled few-step generators CausVid [56] and chunk-wise Self Forcing [23]. Finally, to assess performance over extended sequences, we compare against models designed for long video generation: LongLive [50], Self Forcing++ [9], Rolling Forcing [31], Infinity-RoPE [52], and Context Forcing [7].

Evaluation Metrics. We report the performance on VBench [24, 58] following [7, 23, 50]. To assess physical consistency, we use two metrics. The first is

Table 1: Quantitative comparison on the VBench for 5s and 60s video generation.

Model	#Params	Evaluation scores on 5s ↑			Evaluation scores on 60s ↑		
		Total	Quality	Semantic	Total	Quality	Semantic
Bidirectional models							
LTX-Video [13]	1.9B	80.00	82.30	70.79	-	-	-
Wan2.1 [45]	1.3B	84.26	85.30	80.09	-	-	-
Autoregressive models							
SkyReels-V2 [6]	1.3B	82.67	84.70	74.53	70.47	75.30	51.15
MAGI-1 [11]	4.5B	79.18	82.04	67.74	69.87	76.12	44.87
CausVid [56]	1.3B	81.20	84.05	69.80	71.04	76.80	48.01
NOVA [10]	0.6B	80.12	80.39	79.05	65.25	70.25	45.24
Pyramid Flow [26]	2B	81.72	84.74	69.62	-	-	-
Self Forcing, chunk-wise [23]	1.3B	84.31	85.07	81.28	71.86	77.20	50.51
Long autoregressive models							
LongLive [50]	1.3B	83.72	85.42	76.95	82.62	84.53	74.97
Self Forcing++ [9]	1.3B	83.11	83.79	80.37	-	-	-
Rolling Forcing [31]	1.3B	81.22	84.08	69.78	79.31	81.87	67.69
Infinity-RoPE [52]	1.3B	81.79	83.27	75.87	79.99	80.81	74.30
Context Forcing [7]	1.3B	83.44	84.98	77.29	82.45	83.55	76.10
Ours	1.3B	**84.36**	**86.14**	77.25	**82.88**	84.39	**76.86**

Physical Commonsense (PC) from the VideoPhy benchmark [1, 2], which measures adherence to real-world physical laws. The second is the Physical Alignment and Consistency Evaluation (PACE), an LLM-as-a-Judge metric powered by Gemini. PACE scores videos from 0 to 100, penalizing physical hallucinations based on prompt compliance and four criteria: gravity and mass representation, collision dynamics, motion continuity (avoiding sudden teleportation or unnatural warping), and long-term temporal coherence.

4.2 Comparison Results

Quantitative Evaluation on VBench. We comprehensively evaluate our method against state-of-the-art bidirectional, standard autoregressive, and long-video specific autoregressive models on the VBench benchmark. As reported in Table 1, we evaluate both short-horizon (5s) and long-horizon (60s) video generation to demonstrate our model's robustness against generative drift.

For 5-second generation, our model achieves the highest Total score and Quality score, outperforming strong baselines such as LongLive and Self-Forcing. The superiority of our approach becomes overwhelmingly evident in the extremely long-horizon (60s) setting. Standard autoregressive models suffer from catastrophic error accumulation over extended contexts; for instance, the Total score of Self-Forcing plummets from 84.31 (5s) to 71.86 (60s), and SkyReels-V2 drops from 82.67 to 70.47. In stark contrast, our method effectively bounds this generative drift, maintaining a remarkable Total score and achieving the highest Semantic score at 60 seconds. This minimal performance degradation over time confirms that our cycle-consistency framework successfully preserves the structural integrity and visual fidelity of the generated sequence over infinite horizons.

Physical Consistency Evaluation. To further validate whether our model accurately captures the underlying physical rules of the visual world, we evalu-

Fig. 3: Qualitative comparison of long video generation. We compare our proposed method against several baseline models. The figure displays sampled frames at 0, 30, and 60 seconds for two distinct scenes. While the baseline methods struggle with subject loss, severe structural distortion, or identity shifts over the extended timeframe, our method successfully preserves temporal consistency, visual quality, and subject identity throughout the entire 60-second duration.

ate it on Physical Commonsense (PC) and Physical Alignment and Consistency Evaluation (PACE) metrics. As shown in Table 2, our method significantly outperforms all baseline approaches, achieving the highest average score of 75.66. By enforcing temporal reversibility, our model inherently internalizes causal constraints, endowing it with a superior understanding of complex physical dynamics compared to standard autoregressive predictors.

Qualitative Comparison. Figure 3 visualizes the 60-second generation quality of our model compared to the baselines. As the autoregressive steps accumulate, the baseline methods exhibit severe structural distortion, loss of the main subject, and prominent identity shifts. Our Cycle-World framework, however, acts as a strict temporal regularizer. It consistently preserves the subject's identity, maintains sharp background details, and ensures temporal continuity from the first frame to the very last, translating the quantitative resilience observed in Table 1 into striking visual stability.

Table 2: Quantitative evaluation of physical consistency. Our proposed method outperforms all baseline approaches across both metrics, demonstrating a comprehensive understanding of complex physical dynamics.

Method	PC	PACE	Avg.
Self-Forcing	55.97	78.57	67.27
LongLive	61.94	78.03	69.99
RollingForcing	67.91	75.05	71.48
Infinity-Rope	60.45	78.58	69.52
Ours	69.40	81.91	75.66

Table 3: Ablation study of the proposed components on the VBench benchmark. We evaluate the individual and synergistic effects of the training-time cycle loss ($\mathcal{L}_{cycle}$) and the inference-time cycle guidance on 5s video generation. The best results are highlighted in **bold**.

Method	Tot.	Qual.	Sem.	PC	PACE
Baseline	83.72	85.42	76.95	61.94	78.03
+ CCL	83.89	85.72	76.55	68.70	79.5
+ CGI	**84.51**	**86.37**	**77.05**	65.67	78.80
Cycle-World	84.36	86.14	77.25	**69.40**	**81.91**

4.3 Ablation Studies

Effectiveness of Cycle-World Components. To evaluate the contributions of our proposed modules, we conduct an ablation study on visual quality and physical consistency. Table 3 reports the quantitative metrics on 5-second video generation, while Figure 4 illustrates the qualitative long-horizon stability over 30 seconds.

Quantitative Synergy in Short-Horizon Generation. As shown in Table 3, the Baseline achieves a Total VBench score of 83.72 while scoring 61.94 on PC. Adding the Cycle-Consistent Learning (+ CCL) provides a learned parametric prior. While it moderately improves general visual metrics, its primary benefit is in physical consistency, raising PC by nearly 7 points. This suggests that penalizing reverse-prediction errors embeds causal constraints into the generator. Conversely, applying the runtime corrector solely during inference (+ CGI) acts as an instance-level regularizer against visual degradation, achieving the highest VBench Quality and Total scores. However, relying purely on gradient-based inference optimization without a learned parametric physical prior yields sub-optimal physical consistency.

The full Cycle-World model combines both mechanisms. While its general VBench scores are marginally lower than the CGI-only variant, it achieves the highest performance in physical consistency. This combined strategy ensures the generated videos maintain both visual quality and physical accuracy.

Qualitative Results in Long-Horizon Generation. The benefit of this combination is more evident when extending the generation to 30 seconds (Fig. 4). In long-horizon synthesis, the Baseline exhibits rapid trajectory drift and background degradation. The +CCL variant preserves the core structural identity but struggles to suppress high-frequency temporal artifacts over extended autoregressive steps. The +CGI variant maintains aesthetic coherence locally but eventually drifts from the physical manifold due to the lack of an intrinsic causal prior. The full Cycle-World model combines these strengths. By using CCL to

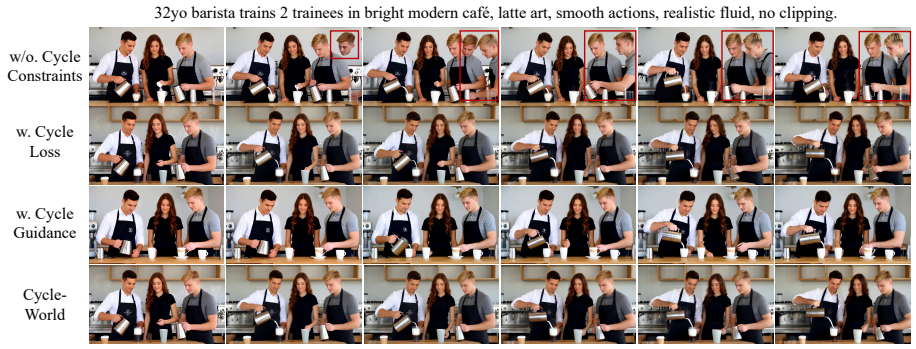

Fig. 4: Qualitative ablation study of the proposed Cycle-World components. We compare the visual quality and temporal consistency of different model variants. While the training cycle loss ($\mathcal{L}_{cycle}$) establishes a robust parametric foundation for structural stability, and the runtime corrector acts as an active safeguard against temporal artifacts, their combination achieves a powerful dual-phase synergy. The full Cycle-World model effectively prevents trajectory drift and background degradation, yielding superior results in long video generation.

keep initial forward proposals close to the reversible manifold, the runtime corrector (CGI) performs more accurate latent refinement. This approach eliminates spatial distortion and temporal identity shifts, producing visually stable and physically grounded long videos.

5 Conclusion

In this paper, we presented Cycle-World, a novel autoregressive video generation framework designed to tackle the pervasive issue of unconstrained generative drift in long-horizon synthesis. Based on the theoretical insight that forward prediction errors can be strictly bottlenecked by enforcing temporal reversibility, we proposed a unified strategy that maintains causal consistency across both the training and inference phases. During training, we introduce a reverse-prediction cycle loss alongside the Distribution Matching Distillation (DMD) objective. This explicitly embeds causal constraints into the generator's parametric weights, establishing a robust foundation for structurally stable frame generation. During inference, we repurpose the frozen reverse model as a fully differentiable runtime corrector. By leveraging gradient-based cycle guidance, this mechanism acts as an active, instance-level safeguard, dynamically pruning out-of-distribution trajectory drift and temporal artifacts before they compound. Extensive experiments on the VBench benchmark demonstrate that this dual-phase synergy effectively prevents the structural collapse and aesthetic degradation typically observed in prolonged autoregressive generation. Consequently, Cycle-World achieves state-of-the-art performance in producing highly consistent, smooth, and high-fidelity 60-second videos.

Acknowledgement

This work was supported by National Natural Science Foundation of China (No. 62302297, 625B2115, 62272447, 62472285, 72192821, 62472285), the Fundamental Research Funds for the Central Universities (YG2023QNB17, YG2024QNA44). [Re Prof Tao] This project is supported by the National Research Foundation, Singapore, under its NRF Professorship Award No. NRF-P2024-001.

References

1. Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.W., Grover, A.: Videophy: Evaluating physical commonsense for video generation. arXiv preprint arXiv:2406.03520 (2024) 11
2. Bansal, H., Peng, C., Bitton, Y., Goldenberg, R., Grover, A., Chang, K.W.: Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. arXiv preprint arXiv:2503.06800 (2025) 11
3. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22563–22575 (2023) 4
4. Cai, Y., Zhang, H., Chen, X., Xing, J., Hu, Y., Zhou, Y., Zhang, K., Zhang, Z., Kim, S.Y., Wang, T., Zhang, Y., Yang, X., Lin, Z., Yuille, A.: Omnivcus: Feedforward subject-driven video customization with multimodal control conditions. In: Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N. (eds.) Advances in Neural Information Processing Systems. vol. 38, pp. 115404–115423. Curran Associates, Inc. (2025), https://proceedings.neurips.cc/paper_files/paper/2025/file/a79054a9da91d73ed3cb1a9e87d7cd2d-Paper-Conference.pdf 4
5. Chen, B., Martí Monsó, D., Du, Y., Simchowitz, M., Tedrake, R., Sitzmann, V.: Diffusion forcing: Next-token prediction meets full-sequence diffusion. Advances in Neural Information Processing Systems **37**, 24081–24125 (2024) 2, 4
6. 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. arXiv preprint arXiv:2504.13074 (2025) 4, 10, 11
7. Chen, S., Wei, C., Sun, S., Nie, P., Zhou, K., Zhang, G., Yang, M.H., Chen, W.: Context forcing: Consistent autoregressive video generation with long context. arXiv preprint arXiv:2602.06028 (2026) 10, 11
8. Chen, X., Wang, Y., Zhang, L., Zhuang, S., Ma, X., Yu, J., Wang, Y., Lin, D., Qiao, Y., Liu, Z.: Seine: Short-to-long video diffusion model for generative transition and prediction. In: The Twelfth International Conference on Learning Representations (2023) 4
9. 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. arXiv preprint arXiv:2510.02283 (2025) 10, 11
10. Deng, H., Pan, T., Diao, H., Luo, Z., Cui, Y., Lu, H., Shan, S., Qi, Y., Wang, X.: Autoregressive video generation without vector quantization. arXiv preprint arXiv:2412.14169 (2024) 4, 10, 11

11. Dobrosotskaya, I.Y., James, G.L.: Magi-1 interacts with β-catenin and is associated with cell–cell adhesion structures. Biochemical and biophysical research communications **270**(3), 903–909 (2000) 4, 10, 11
12. Gao, Y., Guo, H., Hoang, T., Huang, W., Jiang, L., Kong, F., Li, H., Li, J., Li, L., Li, X., et al.: Seedance 1.0: Exploring the boundaries of video generation models. arXiv preprint arXiv:2506.09113 (2025) 1
13. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., Panet, P., Weissbuch, S., Kulikov, V., Bitterman, Y., Melumian, Z., Bibi, O.: Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2024) 10, 11
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 4
15. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868 (2022) 4
16. Hou, S., Wang, C., Zhuang, W., Chen, Y., Wang, Y., Bao, H., Chai, J., Xu, W.: A causal convolutional neural network for multi-subject motion modeling and generation. Computational Visual Media **10**(1), 45–59 (2024). https://doi.org/10.1007/s41095-022-0307-3 2
17. Hu, T., Lu, M., Wang, Y., Zhang, J., Hao, J., Pan, Y., Yi, R., Ma, L., Tao, D.: Metaworld: Scaling multi-agent video world model from single-view video data (2026), https://arxiv.org/abs/2606.02753 2
18. Hu, T., Yu, Z., Zhang, G., Su, Z., Zhou, Z., Zhang, Y., Zhou, Y., Lu, Q., Yi, R.: Harmony: Harmonizing audio and video generation through cross-task synergy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16085–16095 (June 2026) 4
19. Hu, T., Yu, Z., Zhou, Z., Liang, S., Zhou, Y., Lin, Q., Lu, Q.: Hunyuancustom: A multimodal-driven architecture for customized video generation (2025), https://arxiv.org/abs/2505.04512 4
20. Hu, T., Yu, Z., Zhou, Z., Zhang, J., Zhou, Y., Lu, Q., Yi, R.: Polyvivid: Vivid multi-subject video generation with cross-modal interaction and enhancement. In: Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N. (eds.) Advances in Neural Information Processing Systems. vol. 38, pp. 49394–49420. Curran Associates, Inc. (2025), https://proceedings.neurips.cc/paper_files/paper/2025/file/4683beb6bab325650db13afd05d1a14a-Paper-Conference.pdf 4
21. Hu, T., Zhang, J., Huang, H., Yi, R., Su, Z., Weng, J., Xue, Z., Ma, L., Yang, M.H., Tao, D.: Evolution of video generative foundations (2026), https://arxiv.org/abs/2604.06339 4
22. Hu, T., Zhang, J., Su, Z., Yi, R.: Ultragen: High-resolution video generation with hierarchical attention (2025), https://arxiv.org/abs/2510.18775 4
23. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009 (2025) 2, 4, 6, 10, 11
24. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: VBench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024) 10

25. HunyuanWorld, T.: Hy-world 1.5: A systematic framework for interactive world modeling with real-time latency and geometric consistency. arXiv preprint (2025) 2
26. Jin, Y., Sun, Z., Li, N., Xu, K., Jiang, H., Zhuang, N., Huang, Q., Song, Y., Mu, Y., Lin, Z.: Pyramidal flow matching for efficient video generative modeling. arXiv preprint arXiv:2410.05954 (2024) 4, 10, 11
27. Karacan, L., Sarıgül, M.: Full-frame video stabilization via spatiotemporal transformers. Computational Visual Media **11**(3), 655–667 (2025). https://doi.org/10.26599/CVM.2025.9450416 4
28. Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Schindler, G., Hornung, R., Birodkar, V., Yan, J., Chiu, M.C., et al.: Videopoet: A large language model for zero-shot video generation. arXiv preprint arXiv:2312.14125 (2023) 4
29. 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. arXiv preprint arXiv:2412.03603 (2024) 2, 4
30. Liu, K., Zheng, Y., Wang, K., Wu, S., Zhang, R., Luo, J., Hatzinakos, D., Liu, Z., Fei, H., Chua, T.S.: Javisdit++: Unified modeling and optimization for joint audio-video generation. In: The Fourteenth International Conference on Learning Representations (2026) 4
31. Liu, K., Hu, W., Xu, J., Shan, Y., Lu, S.: Rolling forcing: Autoregressive long video diffusion in real time. arXiv preprint arXiv:2509.25161 (2025) 10, 11
32. Liu, Y., Zhao, H., Chan, K.C.K., Wang, X., Loy, C.C., Qiao, Y., Dong, C.: Temporally consistent video colorization with deep feature propagation and self-regularization learning. Computational Visual Media **10**(2), 375–395 (2024). https://doi.org/10.1007/s41095-023-0342-8 4
33. Low, C., Wang, W., Katyal, C.: Ovi: Twin backbone cross-modal fusion for audio-video generation (2025), https://arxiv.org/abs/2510.01284 4
34. Mao, X., Li, Z., Li, C., Xu, X., Ying, K., He, T., Pang, J., Qiao, Y., Zhang, K.: Yume-1.5: A text-controlled interactive world generation model. arXiv preprint arXiv:2512.22096 (2025) 2
35. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International conference on machine learning. pp. 8162–8171. PMLR (2021) 4
36. OpenAI: Sora. https://openai.com/sora (2024) 1
37. OpenAI: Sora 2. https://openai.com/index/sora-2/ (2025) 1
38. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023) 4
39. Qiu, H., Xia, M., Zhang, Y., He, Y., Wang, X., Shan, Y., Liu, Z.: Freenoise: Tuning-free longer video diffusion via noise rescheduling. arXiv preprint arXiv:2310.15169 (2023) 4
40. Rasul, K., Seward, C., Schuster, I., Vollgraf, R.: Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In: International conference on machine learning. pp. 8857–8868. PMLR (2021) 4
41. Seed, B.: Seedance 2.0. https://seed.bytedance.com/en/seedance2_0 (2026) 1
42. Skywork AI Matrix-Game Team: Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. Technical report (2026), https://github.com/SkyworkAI/Matrix-Game/blob/main/Matrix-Game-3/assets/pdf/report.pdf 2
43. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 4

44. Technology, K.: Kling. https://kling.kuaishou.com/ (2025) 2
45. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025) 2, 4, 10, 11
46. Wang, D., Zuo, W., Li, A., Chen, L.H., Liao, X., Zhou, D., Yin, Z., Dai, X., Jiang, D., Yu, G.: Universe-1: Unified audio-video generation via stitching of experts. arXiv preprint arXiv:2509.06155 (2025) 4
47. Wang, F.Y., Chen, W., Song, G., Ye, H.J., Liu, Y., Li, H.: Gen-l-video: Multi-text to long video generation via temporal co-denoising. arXiv preprint arXiv:2305.18264 (2023) 4
48. Wang, H., Ma, C.Y., Liu, Y.C., Hou, J., Xu, T., Wang, J., Juefei-Xu, F., Luo, Y., Zhang, P., Hou, T., et al.: Lingen: Towards high-resolution minute-length text-to-video generation with linear computational complexity. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2578–2588 (2025) 4
49. Williams, R.J., Zipser, D.: A learning algorithm for continually running fully recurrent neural networks. Neural computation **1**(2), 270–280 (1989) 4
50. 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. arXiv preprint arXiv:2509.22622 (2025) 2, 4, 10, 11
51. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024) 4
52. Yesiltepe, H., Meral, T.H.S., Akan, A.K., Oktay, K., Yanardag, P.: Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. arXiv preprint arXiv:2511.20649 (2025) 10, 11
53. 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: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1309–1320 (2023) 4
54. Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., Freeman, B.: Improved distribution matching distillation for fast image synthesis. Advances in neural information processing systems **37**, 47455–47487 (2024) 4
55. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6613–6623 (2024) 4
56. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22963–22974 (2025) 2, 4, 10, 11
57. Zhang, G., Zhou, Z., Hu, T., Peng, Z., Zhang, Y., Chen, Y., Zhou, Y., Lu, Q., Wang, L.: Uniavgen: Unified audio and video generation with asymmetric cross-modal interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1950–1960 (2026) 4
58. Zheng, D., Huang, Z., Liu, H., Zou, K., He, Y., Zhang, F., Zhang, Y., He, J., Zheng, W.S., Qiao, Y., Liu, Z.: VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. arXiv preprint arXiv:2503.21755 (2025) 10
59. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024) 4

60. Zhu, H., Zhao, M., He, G., Su, H., Li, C., Zhu, J.: Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. arXiv preprint arXiv:2602.02214 (2026) 2