arXiv:2602.24289v2 [cs.CV] 23 Aug 2026

Mode Seeking meets Mean Seeking for Fast Long Video Generation

Shengqu Cai [1 2] Weili Nie [* 2] Chao Liu [* 2] Julius Berner [2] Lvmin Zhang [1] Nanye Ma [3] Hansheng Chen [1]
Maneesh Agrawala [1] Leonidas Guibas [1] Gordon Wetzstein [1] Arash Vahdat [2]

https://primecai.github.io/mmm/

Abstract

Scaling video generation from seconds to minutes faces a critical bottleneck: while short-video data is abundant and high-fidelity, coherent long-form data is scarce and limited to narrow domains. To address this, we propose a training paradigm where Mode Seeking meets Mean Seeking, decoupling local fidelity from long-term coherence based on a unified representation via a Decoupled Diffusion Transformer (Wang et al., 2026b). Our approach utilizes a global Flow Matching head trained via supervised learning on long videos to capture narrative structure, while simultaneously employing a local Distribution Matching head that aligns sliding windows to a frozen short-video teacher via a mode-seeking reverse-KL divergence. This strategy enables the synthesis of minute-scale videos that learns long-range coherence and motions from limited long videos via supervised flow matching, while inheriting local realism by aligning every sliding-window segment of the student to a frozen short-video teacher, resulting in a few-step fast long video generator. Automatic evaluations, human ratings, and one-minute stress tests show that our method effectively closes the fidelity–horizon gap by jointly improving local sharpness/motion and long-range consistency.

1. Introduction

The generative frontier is rapidly advancing from static images to the dynamic, complex realm of videos. An important goal is to synthesize arbitrarily long, coherent, and high-fidelity video streams that mirror the rich temporal structure of our world. Such long-horizon generation could enable applications like interactive world modeling for embodied agents and games, long-form story/film generation with persistent characters and scenes, and controllable video editing or animation that maintains identity and style over extended time. Recent progress in diffusion models and transformers has yielded remarkable success in generating short video clips, typically lasting only a few seconds (Team Wan et al., 2025; Hong et al., 2023; Yang et al., 2025b; Wu et al., 2025a).

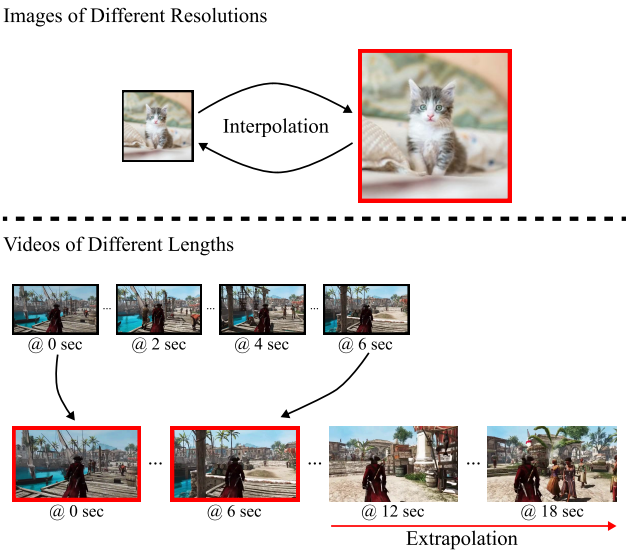

Figure 1. **Video length is not analogous to image resolution.** *Top:* For images, moving from low to high resolution is largely an interpolation of the same underlying local patch distribution. *Bottom:* For videos, moving from short clips to long sequences is temporal extrapolation. The model must introduce new events and causal structure beyond the short-clip horizon, which is fundamentally harder than multi-resolution image training.

This success is enabled by a simple data reality: seconds-long clips are abundant online and thus available at web scale, e.g., in widely used video–text datasets (Bain et al., 2021; Wang et al., 2024; Fan et al., 2025). Scaling from seconds to minutes breaks this recipe. High-quality long-form videos that are minute-scale sequences with

[*]Equal contribution [1]Stanford University, California, USA [2]NVIDIA Research, California, USA [3]NYU Courant, New York, USA. Correspondence to: Shengqu Cai <shengqu@cs.stanford.edu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

sustained events and context are far scarcer at a comparable scale, more heterogeneous, and substantially more expensive to curate and filter for generative training (and post training) (Li et al., 2024). As a result, the strongest available prior over realistic short-timescale dynamics often lives in a short-video generator—one that has already absorbed web-scale diversity and has been pushed toward high-fidelity modes through extensive refinement, while long-video training typically operates in a much more data-constrained regime and is extremely expensive.

A common practice borrowed from image generation is to train a single model on a "soup" of videos of varying lengths (Yang et al., 2025b; Kong et al., 2024; Wu et al., 2025a; Chen et al., 2025; Team Seawead et al., 2025; Gao et al., 2025b; Team Seedance et al., 2025; Gao et al., 2025a; Lin et al., 2025b). The implicit hope is that the model will smoothly interpolate across temporal horizons, much as image models learn across spatial resolutions. We argue that this assumption is fundamentally flawed, as shown in Fig. 1: the temporal dimension is not analogous to image resolution—a 1024×1024 image is a higher-resolution interpolation of a 256×256 image; its patches share essentially the same underlying distribution. A one-minute video, however, is not an "interpolation" of a 5-second video: it is an "extrapolation" that adds new events, causal chains, and narrative structure, encoding substantially more information, in some sense, similar to panorama generation but requires significantly stronger context understanding ability.

This mismatch induces a critical and underappreciated failure mode. Even when trained on mixed-duration data, a model that can generate longer sequences often comes at the expense of losing the sharp local dynamics characteristic in expert short-video teachers. The outputs can look softer, less detailed, and less "alive". In other words, the model is forced to relearn a high-fidelity short-video prior from a regime where the data and compute are most constrained, as evidenced in our experiment section.

We propose a training paradigm that decouples local fidelity from long-term coherence using a Decoupled Diffusion Transformer (Wang et al., 2026b) (DDT). A long-video "student" model learns global, minute-scale narrative structure through a mean-seeking supervised finetuning (SFT) objective on limited long-video data. Simultaneously, the student maintains local realism by aligning every generated sliding-window segment with a frozen, expert short-video teacher. This alignment utilizes a mode-seeking reverse-KL divergence, allowing the student to inherit high-fidelity short-timescale priors by querying the teacher only on the student's own rollouts. To resolve the mathematical conflict between mean-seeking flow matching and mode-seeking distribution matching, we utilize two separate lightweight velocity heads that share a unified long-context encoder. A Flow Matching head handles global coherence from real long videos, while a Distribution Matching head focuses on local realism distilled from the teacher. At inference time, the Distribution Matching head serves as a fast, few-step sampler that produces videos with both sharp local dynamics and consistent long-range context.

Putting these pieces together, we close the fidelity–horizon gap with a simple decoupling: scarce long videos supervise global minute-scale structure, while a frozen expert short-video teacher provides a prior for local realism on every sliding window. Moreover, because our teacher alignment is implemented through DMD (Yin et al., 2024b;a)-style distillation with a mode-seeking reverse-KL, the resulting DM head serves as a few-step sampler, enabling fast long-video synthesis at inference time. We summarize our key contributions as follows:

- We align every sliding-window segment of a long-video student to a frozen short-video teacher via a mode-seeking reverse KL, without requiring any additional short-video data.
- We use a DDT (Wang et al., 2026b) with separate Flow Matching (long-video supervised finetuning) and Distribution Matching (teacher matching) heads, and show that such decoupled targets, decoded from a unified intermediate representation, are helpful for aligning long-context and local-quality targets.
- By using only the DM head for inference, we unlock fast, few-step long-video inference. Experiments show improved local quality/motion while preserving long-range consistency.

2. Related Work

Long video generation. Diffusion dominates modern video synthesis, but scaling to long horizons remains challenging (Zhang et al., 2026b; Guo et al., 2024; 2025b). Training-free length extrapolation stretches pretrained models beyond their training horizon via noise rescheduling or temporal-frequency rebalancing (Qiu et al., 2024; Lu et al., 2024; Lu & Yang, 2025; Zhao et al., 2025; Ruhe et al., 2024). A complementary line blends diffusion with causal prediction—via noise-injected autoregressive rollouts, long-context AR designs (e.g., flexible RoPE (Su et al., 2024)), or teacher-to-student distillation—and appears in both research systems and production-scale deployments (Chen et al., 2024; Song et al., 2025; Gu et al., 2025; Su et al., 2024; Yin et al., 2025; Kodaira et al., 2026; Henschel et al., 2025; Chen et al., 2025; Sand.ai et al., 2025; Guo et al., 2025a; Po et al., 2026). To mitigate AR drift, rollout-aware training (Huang et al., 2025b) and extensions further improve length generalization and real-time rollout quality (Huang

et al., 2025b; Cui et al., 2026; Liu et al., 2026; Yang et al., 2026; Yi et al., 2026; Yesiltepe et al., 2026; Lu et al., 2025; Chen et al., 2026b; Lv et al., 2026; Wu et al., 2026b).

Context learning and compression. Long-context persistence is crucial as generation extends beyond a few seconds. Retrieval-based memories ground prediction in relevant history (Field-of-View-, geometry-, or view-indexed) (Xiao et al., 2025; Yu et al., 2025a; Li et al., 2025a; Huang et al., 2025a; Wu et al., 2025b; Fu et al., 2026; Xiao et al., 2026b; Huang et al., 2026a), while learned routers/policies sparsify attention by selecting salient context chunks or token groups (Cai et al., 2026; Jia et al., 2026; Meng et al., 2026; Wang et al., 2026a; Zhang et al., 2025i; Zhan et al., 2025; Zhang et al., 2025d; Wen et al., 2026; Xie et al., 2025b; Zhang et al., 2026c;a). Orthogonally, history can be compressed into a fixed-size state via latent packing or recurrent/state-space dynamics, with lightweight test-time adaptation providing another learned context representation (Zhang et al., 2025e; Xiao et al., 2026a; Savov et al., 2025; Po et al., 2025; Dalal et al., 2025; Zhang et al., 2026e;d; Ji et al., 2025; Li et al., 2026; Yu et al., 2025b).

Efficient video diffusion designs. Large spatiotemporal contexts are compute-bound, motivating kernel-level optimizations such as FlashAttention (Dao et al., 2022; Dao, 2024) and structured sparsity (e.g., sliding/tiling windows or radial masks) with training- or inference-time pruning (Zhang et al., 2025g; Li et al., 2025b; Xi et al., 2025; Yang et al., 2025a; Xia et al., 2025; Zhang et al., 2025h). Learned sparse routing/selection further keeps only salient token pairs or blocks (Wu et al., 2026a; Zhang et al., 2025f;b;a;c), while complementary directions reduce token/latent size or adopt multiscale and linear/block-linear attention to control memory growth (Bolya et al., 2023; Lee et al., 2024; Bachmann et al., 2025; HaCohen et al., 2024; Jin et al., 2025; Xie et al., 2025a; Chen et al., 2026a).

Diffusion distillation. Few-step diffusion distillation splits along the divergence used to align student and teacher: mean-seeking trajectory regression matches deterministic teacher targets via halved-step composition, PF-ODE self-consistency, average velocity, step-size conditioning, or rectified-flow straightening (Salimans & Ho, 2022; Song et al., 2023; Luo et al., 2023a; Lu & Song, 2025; Geng et al., 2025; Frans et al., 2025; Liu et al., 2023; 2024; Tong et al., 2026), while mode-seeking distribution matching minimizes a reverse-KL objective with score-difference gradients (Poole et al., 2023; Wang et al., 2023; Luo et al., 2023b; Zhou et al., 2024; Sauer et al., 2024) and underpins DMD-style image (Yin et al., 2024b;a) and video (Yin et al., 2025; Huang et al., 2025b) distillation. On highly multimodal distributions such as long videos, isotropic regression in pixel or weakly regularized latent space collapses modes and blurs high-frequency detail, motivating hybrids that interpolate the f-divergence family (Xu et al., 2025), replace MSE with a learned discriminator (Lin et al., 2026), or fuse score distillation with consistency models (Zheng et al., 2026). We adopt a different decomposition: instead of blending the two divergences on a single output, we route them to disjoint content axes, using mean-seeking flow matching for global long-video structure and mode-seeking alignment for local sliding windows.

3. Method

3.1. Preliminaries: Rectified Flow for Video Generation

We work in the latent space of a video VAE and follow the rectified flow parameterization (Lipman et al., 2023; Liu et al., 2023) used in recent video models. Let $x_0 \in \mathbb{R}^{T\times H\times W\times C}$ denote a clean video latent (short or long), and let $z \sim \pi \triangleq \mathcal{N}(0, I)$ be a prior sample. We construct a deterministic noising path

$$x_t \;=\; I_t(x_0, z) \;\triangleq\; (1-t)\,x_0 + t\,z, \quad t \in [0,1], \tag{1}$$

which induces a corresponding generative ordinary differential equation (ODE)

$$\frac{dx_t}{dt} = -u(x_t, t), \qquad x_1 \sim \pi, \tag{2}$$

where $u : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ is the marginal velocity field and d is the latent dimensionality. A standard flow-matching objective trains a network u_θ to approximate u via

$$\mathcal{L}_{\text{FM}}(\theta) \;=\; \mathbb{E}_{x_0,z,t}\left\|u_\theta(x_t, t, c) - u(x_t, t \mid x_0, z)\right\|_2^2, \tag{3}$$

where $u(x_t, t \mid x_0, z) = x_0 - z$, and c denotes conditioning (e.g., text prompt). Sampling then corresponds to solving (2) backward from $t = 1$ to $t = 0$ using u_θ. In our setting, we assume the existence of an expert short-video teacher, e.g., a pre-trained 5-second model u_{teacher} (e.g., a DiT (Peebles & Xie, 2023)-style flow model whose velocity field $u_{\text{teacher}}(x_t, t, c)$ is available at arbitrary (x_t, t, c), but whose original short-clip training data and post-training strategies need not be accessible.

3.2. Sliding-window View of Long Videos

Given a long video latent x_0^{long} of T_{long} frames and a window length L (corresponding to ≈ 5 s), we denote by

$$\text{crop}_k(x_0^{\text{long}}) \in \mathbb{R}^{L\times H\times W\times C}, \tag{4}$$

the contiguous L-frame segment starting at frame index k, where $k \in \{0, 1, \ldots, \lceil T_{\text{long}}/L\rceil - 1\}$. Let $q_\Phi^{(k)}$ denote the marginal distribution over such windows induced by the long-video student model with parameters Φ, and let p_{teacher} denote the short-video teacher distribution. Our goal is to regularize

$$\mathcal{L}_{\text{seg}}(\Phi) \;=\; \mathbb{E}_k\big[D_{\text{KL}}\big(q_\Phi^{(k)} \,\|\, p_{\text{teacher}}\big)\big], \tag{5}$$

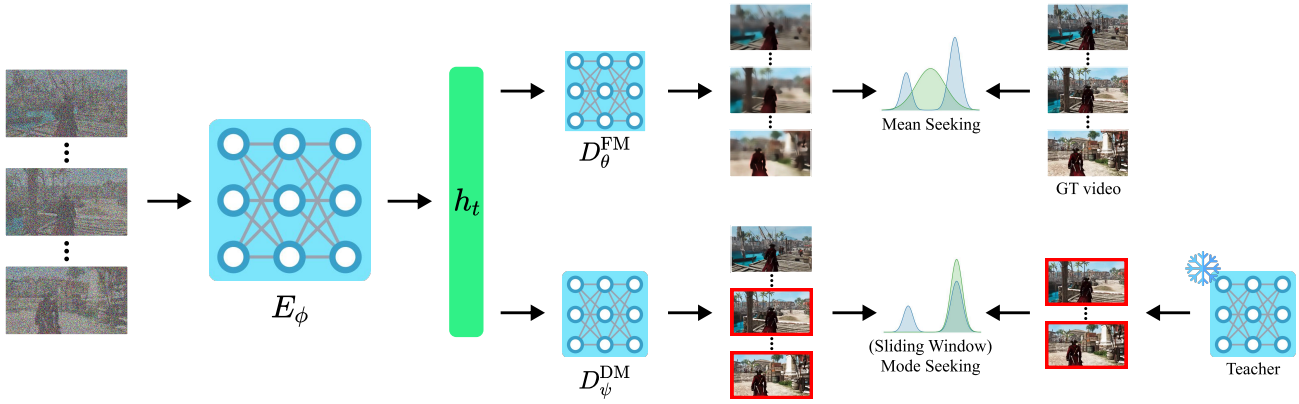

Figure 2. **Overview: mode seeking meets mean seeking.** A shared long-context condition encoder E_ϕ maps a noisy long-video latent x_t^{long} (with timestep t and conditioning c) to a unified representation h_t. Two lightweight decoder heads read out velocities from h_t: the long context Flow Matching head D_θ^{FM} is trained with supervised flow matching on real long videos (mean-seeking), while the segment-wise Distribution Matching head D_ψ^{DM} is trained via on-policy sliding-window reverse-KL alignment to an expert short-video teacher using DMD (Yin et al., 2024b;a)/VSD (Wang et al., 2023)-style gradients (mode-seeking). Both objectives update the shared encoder, but each head receives only its corresponding signal.

i.e., every sliding window produced by the long-video generator should match the short-video teacher distribution. Because we use the reverse KL, $D_{\text{KL}}(q \,\|\, p)$, this objective is mode-seeking: it encourages the student to concentrate its mass on the teacher's high-fidelity modes rather than averaging over them (Yin et al., 2025; Huang et al., 2025b). It in addition views a long generator as inducing sliding-window marginals and enforces Eq. (5) on windows sampled on-policy from long rollouts, using the short teacher as a local realism critic beyond its native horizon.

3.3. Decoupled Long-Video Diffusion Transformer

When viewing the reverse KL term in Eq. (5) as an alignment target, it introduces a fundamental tension with the standard flow-matching objective in Eq. (3). Denote by f_θ the clean-video prediction induced by the velocity head. Under the squared flow-matching/SFT objective, the point-wise optimum is mean-seeking: $f^*_{\text{SFT}}(x_t) = \mathbb{E}[x_0 \mid x_t]$. For a multimodal conditional distribution $q(x_0 \mid x_t)$, which is typical under noise ($t > 0$), this expectation averages across diverse modes and can lie off the data manifold, producing blurry or unrealistic states. Conversely, the reverse-KL/DMD objective is mode-seeking and drives the prediction toward a specific high-density teacher mode, $f^*_{\text{DMD}}(x_t) \approx x_0^{(\text{mode})}$. Because the mean of multiple modes is generally not itself a mode, $f^*_{\text{SFT}}(x_t) \neq f^*_{\text{DMD}}(x_t)$ except in degenerate cases such as an effectively unimodal conditional distribution or vanishing noise. Naïvely applying both signals to a single velocity predictor therefore yields gradient interference: the long-video SFT objective pulls the head toward a conditional mean, whereas the teacher term pulls it toward a sharp high-density mode. This motivates us to instantiate the long-video student as a decoupled architecture, where a feature encoder is followed by two decoder heads, inspired by DDT (Wang et al., 2026b). An overview of our design is demonstrated in Fig. 2.

Condition encoder. Given a noisy long video x_t^{long}, conditioning c, and timestep t, the encoder

$$h_t \;=\; E_\phi(x_t^{\text{long}}, t, c) \tag{6}$$

produces a spatiotemporal feature tensor h_t. Architecturally, E_ϕ is a video diffusion transformer with full-range temporal dependencies with full attention, and forms the shared backbone between all heads.

Two velocity heads. On top of h_t we attach two lightweight transformer decoders:

$$u_\theta(x_t^{\text{long}}, t, c) = D_\theta^{\text{FM}}(h_t, t, c), \tag{7}$$

$$v_\psi(x_t^{\text{long}}, t, c) = D_\psi^{\text{DM}}(h_t, t, c). \tag{8}$$

The Flow Matching (FM) head D_θ^{FM} parameterizes the student's global velocity field u_θ used for long-horizon training and sampling on ground truth long videos (Sec. 3.5) using Eq. (3). The Distribution Matching (DM) head D_ψ^{DM} is a few-step generator that uses teacher alignment on local windows. Sharing the encoder E_ϕ but decoupling the heads has two advantages: (i) the long-context representation h_t is learned and reused across objectives, and (ii) the short-video generation capability is distilled from the teacher distribution without forgetting when training on scarce long video data. Next, we discuss how the two heads are trained.

3.4. Local Reverse-KL via DMD/VSD

We now introduce our local distribution matching loss, which regularizes the student's local dynamics and quality using the short-video teacher. Conceptually, we would

like each window marginal $q_\Phi^{(k)}$ (where $\Phi \triangleq (\phi, \psi)$) of the student to match the teacher distribution p_{teacher} via a mode-seeking reverse KL, as shown in Eq. (5). However, directly evaluating Eq. (5) is intractable. We instead follow the DMD/VSD literature (Yin et al., 2024b;a; Wang et al., 2023), which shows that for diffusion/flow models, the gradient of a reverse-KL between student and teacher can be expressed in terms of their score/velocity differences on noisy states. We adapt this to videos with sliding local windows.

Reverse-KL gradient on noised windows. Let $q_{\Phi,t}^{(k)}$ denote the distribution of the noised window $\hat{x}_t^{(k)} = (1-t)\,\hat{x}_0^{(k)} + t\,\epsilon$ where $\hat{x}_0^{(k)}$ is a sample from the student $q_\Phi^{(k)}$ and $\epsilon \sim \mathcal{N}(0, I)$ is fresh noise. DMD shows that the reverse-KL gradient can be written as an expectation over the student distribution, involving the difference between the teacher's score and the student's own ("fake") score on the same noisy state. Concretely, using the linear interpolation in Eq. (1) (where $\alpha_t = 1 - t$), we use the window-level gradient surrogate

$$\widehat{\nabla}\,\mathcal{L}_{\text{seg}} = \mathbb{E}_{t,\,k}\Big[\lambda(t)\,\big(v_{\text{fake}}(\hat{x}_t^{\text{long}}, t, c) - u_{\text{teacher}}(\hat{x}_t^{(k)}, t, c)\big)^\top \nabla\,\hat{x}_0^{(k)}\Big]. \tag{9}$$

where $\hat{x}_0^{(k)}$ is the generated student window, and v_{fake} is a fake score estimator trained on the student's window predictions $\hat{x}_0^{(k)}$ using score matching for 5 steps in between the student updates. We treat $(v_{\text{fake}} - u_{\text{teacher}})$ as a stop-gradient term and backpropagate only through $\hat{x}_0^{(k)}$. Here $\lambda(t)$ absorbs the standard DMD/VSD weighting (including score-to-velocity scaling), $u_{\text{teacher}}(\cdot)$ is the short-video teacher velocity query, and $v_\psi^{(k)}(\cdot)$ is the student's window velocity predicted by the mode-seeking head, obtained as follows.

Cropping the mode-seeking head. We obtain this generated window by cropping the mode-seeking head's output on the noised long video latent sequences:

$$x_t^{\text{long}} = (1-t)\,x_0^{\text{long}} + t\,\epsilon, \tag{10}$$

$$h_t = E_\phi(x_t^{\text{long}}, t, c), \tag{11}$$

$$v_\psi^{\text{long}}(x_t^{\text{long}}, t, c) = D_\psi^{\text{DM}}(h_t, t, c), \tag{12}$$

$$v_\psi^{(k)}(x_t^{\text{long}}, t, c) = \text{crop}_k\Big(v_\psi^{\text{long}}(x_t^{\text{long}}, t, c)\Big), \tag{13}$$

where $\hat{x}_t^{(k)} = \text{crop}_k(\hat{x}_t^{\text{long}})$ and $\epsilon \sim \mathcal{N}(0, I)$. The teacher term $u_{\text{teacher}}(\hat{x}_t^{(k)}, t, c)$ is evaluated on the cropped K-frame window using the short-video teacher. Sliding Window DMD (Yin et al., 2024b) is non-trivial to implement on modern video generation models. In particular, their latent space often includes both image latents and video frame latents to better support image-video joint training (Team Wan et al., 2025; Kong et al., 2024). When applying DMD with sliding windows on long latent sequences, a window cropped from the middle begins with a video latent, while the teacher expects the first latent of a clip to be an image latent, leading to a semantic mismatch at window boundaries. To address this, we follow LongLive (Yang et al., 2026): for any window starting at offset $p > 0$, we decode the latent prefix $[0, ..., p-1]$ with the frozen VAE, take the last decoded RGB frame, and re-encode it into an image latent; we then prepend this reconstructed image latent to the student's windowed video latents before computing the DMD loss, masking out the reconstructed latent to avoid backpropagating through the VAE. We find this strategy to be effective not only for causal AR models (Yang et al., 2026), but also for non-causal bidirectional models like ours.

3.5. SFT Flow-Matching Anchor on Long Videos

Distribution matching teacher alignment alone cannot teach global, minute-scale coherence: the short-video teacher does not model long-range structure by design. To learn long-range dynamics from the limited long-video corpus, we train the FM head u_θ with a supervised flow-matching (SFT) objective on full-length videos. Let $x_0^{\text{long}} \sim p_{\text{long}}$ be a real long video latent, and let x_t^{long} be constructed as in (1) using a global prior $z^{\text{long}} \sim \pi$. We train u_θ with

$$\mathcal{L}_{\text{SFT}}(\phi, \theta) = \mathbb{E}_{x_0^{\text{long}}, z^{\text{long}}, t}\Big[\big\|u_\theta(x_t^{\text{long}}, t, c) - \big(x_0^{\text{long}} - z^{\text{long}}\big)\big\|_2^2\Big], \tag{14}$$

where gradients flow through both the condition encoder E_ϕ and the FM head D_θ^{FM}. This objective anchors the student's global velocity field to real long video trajectories, encouraging correct long-horizon temporal dependencies and narrative structure.

3.6. Joint Objective and Training Procedure

Our training combines two complementary signals: (i) supervised long-video flow matching on the global FM head, and (ii) local reverse-KL alignment on the DM head via the DMD/VSD gradient surrogate. Concretely, we optimize the long-video model with

$$\mathcal{L}_{\text{total}}(\phi, \theta, \psi) \;=\; \mathcal{L}_{\text{SFT}}(\phi, \theta) \;+\; \lambda_{\text{seg}}\,\mathcal{L}_{\text{seg}}(\phi, \psi), \tag{15}$$

where λ_{seg} is a scalar weight, $\mathcal{L}_{\text{SFT}}$ is the supervised FM objective on real long videos in Eq. (14), and $\mathcal{L}_{\text{seg}}$ denotes the conceptual window-level reverse KL in Eq. (5). Since $\mathcal{L}_{\text{seg}}$ is intractable to evaluate, we implement its effect by injecting the DMD/VSD gradient surrogate in Eq. (9). The shared encoder E_ϕ receives gradients from both terms, while the two heads are updated by their respective signals:

$$\nabla_\phi \mathcal{L}_{\text{total}} = \nabla_\phi \mathcal{L}_{\text{SFT}} + \lambda_{\text{seg}}\,\widehat{\nabla}_\phi \mathcal{L}_{\text{seg}}, \tag{16}$$

$$\nabla_\theta \mathcal{L}_{\text{total}} = \nabla_\theta \mathcal{L}_{\text{SFT}}, \tag{17}$$

$$\nabla_\psi \mathcal{L}_{\text{total}} = \lambda_{\text{seg}}\,\widehat{\nabla}_\psi \mathcal{L}_{\text{seg}}. \tag{18}$$

Method	NFE	Subject Consistency ↑	Background Consistency ↑	Motion Smoothness ↑	Dynamic Degree ↑	Aesthetic Quality ↑	Image Quality ↑	VLM Consistency ↑
Long-context SFT	50	0.9685	0.9533	0.9866	0.9375	0.4973	0.6303	77.28
Mixed-length SFT	50	0.9667	0.9541	0.9874	0.8906	0.5467	0.6683	74.63
CausVid (Yin et al., 2025)	4	0.9736	0.9614	0.9789	0.8594	0.6044	0.6305	39.30
Self Forcing (Huang et al., 2025b)	4	0.9489	0.9451	0.9805	0.9063	0.5556	0.6278	37.60
InfinityRoPE (Yesiltepe et al., 2026)	4	0.9689	0.9573	0.9812	0.7188	0.5342	0.6871	68.61
LongLive (Yang et al., 2026)	–	0.9681	0.9565	0.9838	0.6422	0.5401	0.6684	62.92
Ours	4	0.9682	0.9548	0.9863	0.9453	0.5735	0.6982	75.42

Table 1. **Quantitative comparisons.** The first, second, and third values are highlighted. NFE is reported for diffusion/flow samplers in our setup; LongLive (Yang et al., 2026) is a frame-level AR real-time system, so its NFE is not directly comparable. We note that rollout-based methods can get over-saturated (CausVid (Yin et al., 2025)) or static/conservative (InfinityRoPE (Yesiltepe et al., 2026), LongLive (Yang et al., 2026)), which in turn significantly boosts some consistency metrics. Even so, our method still stands out as the best overall performing model.

This realizes the intended decoupling: long-video supervision updates the global FM head, while local short-video teacher distribution matching is handled by the DM head. At the same time, they both act through and update the shared representation. Each step uses two minibatches: (1) a batch of real long videos to compute $\mathcal{L}_{\text{SFT}}$ and update (ϕ, θ); (2) a batch of on-policy student rollouts to sample windows and apply the DMD/VSD surrogate update in Eq. (9) to (ϕ, ψ).

3.7. Inference

At inference time, we discard the FM head and generate long videos using the DM head v_ψ. We note that the DM head, while supervised only on sliding windows similar to APT2 (Lin et al., 2025a), can directly generate long videos end-to-end bidirectionally. The shared encoder ensures that the representation used by v_ψ has been shaped as follows: any sliding window of the generated long video must reside in the same local distributional modes as the expert short-video teacher, while long-range coherence and narrative structure are learned directly from scarce long videos via $\mathcal{L}_{\text{SFT}}$.

Aforementioned decoupled design allows the model to simultaneously optimize for local realism and long-horizon coherence: the short-video teacher provides a mode-seeking regularizer on every local segment, and the long-video SFT objective concentrates the limited long-form data and capacity on what the teacher cannot provide: minute-scale temporal coherence. In addition, we no longer need multi-stage training and distillation; the output model with DM head is directly capable of few-step inference, unlocking fast minute-scale video generation.

4. Results

4.1. Experimental Details

Base models. We trained our methods on both the Wan (Team Wan et al., 2025) 1.3B model and the Wan (Team Wan et al., 2025) 14B model. For fair comparison, we use Wan (Team Wan et al., 2025) 2.1 1.3B throughout our quantitative and qualitative comparisons, both as the student and as the teacher whenever involved. We refer to our supplementary website for qualitative results on the 14B model.

Data. We collect and use data from several sources. From publicly available datasets, we use all videos available from the Sekai dataset (Li et al., 2025c), and a subset from MiraData (Ju et al., 2024). We also filter out single-shot videos from randomly collected internet videos. These datasets include more than 100k videos ranging from 10 seconds to minutes, with an average of 31 seconds. We set the temporal upper bound to 61 seconds and subsample when a video exceeds it.

Implementation details. We train our model on A100 and GB200 GPUs. We use dynamic batching, treating training videos as variable-length sequences and pre-processing them into length-based buckets for sampling. Variable-length training is used across training, reducing padding waste and IDLE time. To train for long contexts, we use the DeepSpeed Ulysses (Jacobs et al., 2023) sequence-parallelism strategy. We use a sequence-parallelism group size of 4 for A100 GPUs, and 2 for GB200 GPUs. We implemented everything on top of the FastGen (Nie et al., 2026) repository.

Baselines. We set up the following two baselines that are widely used in modern long video models/video world models: *(1) Long-context SFT*: the very basic long tuning strategy, where long video clips are collected and used to finetune the short pretrained model; *(2) Mixed-lengths SFT*: the advanced setup used by industrial-level model training, when videos of different lengths are jointly trained, hoping to achieve interpolation among modes of different lengths. We also compare our method against popular auto-regressive video generation methods, CausVid (Yin et al., 2025), Self-Forcing (Huang et al., 2025b), where we use the extrapolation method provided

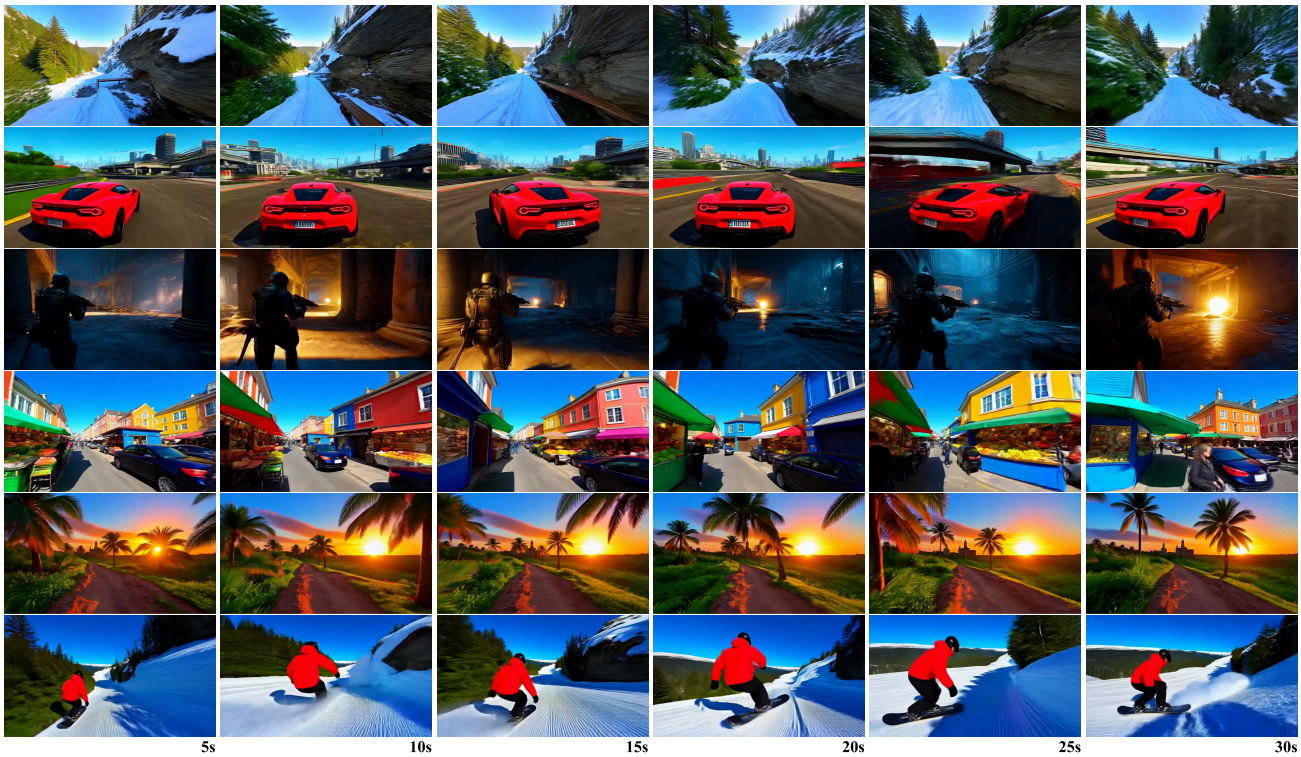

Figure 3. **Qualitative results.** Our method generalizes well to various scenarios, producing long videos that maintain local fidelity and global coherence. All results are obtained using the Wan (Team Wan et al., 2025) 1.3B model as both the student and the teacher, demonstrating how our decoupled training effectively extends short-video capabilities to long-horizon generation. We refer to our supplemental website for the videos and results from the Wan (Team Wan et al., 2025) 14B model.

in CausVid to generate long videos. Additionally, we compare with a popular Self-Forcing extrapolation method, InfinityRoPE (Yesiltepe et al., 2026). We further include LongLive (Yang et al., 2026), a recent real-time frame-level autoregressive long-video generator. We note that our method is orthogonal to causal AR models and can be used alongside them, which is an interesting follow-up.

Evaluation metric. We use 200 prompts describing long videos and events as our test set, and generate 30-second videos. We adopt the metrics and evaluation protocol used in VBench-Long (Huang et al., 2024; 2026b; Zheng et al., 2025), which includes subject consistency, background consistency, motion smoothness, temporal flickering, dynamic degree, aesthetic quality, and imaging quality. We also report consistency evaluated by Gemini-3-Pro, an advanced MLLM capable of taking video inputs. The Gemini instruction prompt is provided in our supplementary Sec. A.

4.2. Quantitative Results

We present the quantitative results in Tab. 1, and provide analysis for each class of methods in the following sections. We additionally include 60-second automatic results and human ratings in the supplementary material (Sec. B and Sec. C).

Analysis on SFT-only methods. Supervised Fine-Tuning (SFT), particularly the Mixed-Length SFT strategy used in industrial models, remains a robust baseline for establishing temporal coherence. By training jointly on videos of varying lengths, the model learns to interpolate and extend motion patterns beyond the short-clip horizon. However, the effectiveness of SFT is fundamentally capped by the scarcity of high-quality training data. While short video clips are abundant, high-resolution, single-shot long videos (e.g., $>$ 30 seconds) with sustained narrative continuity are extremely rare. Thus, while SFT is important for learning long-context global structure, it is insufficient for maintaining high fidelity across diverse domains due to data gaps, as evidenced by their relatively low quality scores.

Analysis on teacher-only methods. Methods that rely exclusively on distilling a short-video teacher or rolling out a local generator (e.g., CausVid (Yin et al., 2025), Self-Forcing (Huang et al., 2025b), InfinityRoPE (Yesiltepe et al., 2026), and LongLive (Yang et al., 2026)) face a different set of challenges when scaled to minute-level generation. While these approaches succeed in ensuring local realism for short durations ($\approx$ 5 seconds), they lack a mechanism

Figure 4. **Qualitative comparison.** "LongSFT" and "MixSFT" refer to long-context supervised finetuning (SFT) and mixed-length SFT, respectively. SFT-only methods (LongSFT, MixSFT) achieve decent long context and narrative, but appear to be blurry. Teacher-only methods (CausVid (Yin et al., 2025), Self-Forcing (Huang et al., 2025b)) suffer from quality degradation over long range. While InfinityRoPE (Yesiltepe et al., 2026) extends the temporal horizon, it does not have the ability to process long context and tends to generate still contents. Our method stands out as the best-performing model overall in terms of quality, motion, and long-horizon consistency. We refer to our supplementary website for video results, including extended comparisons with LongLive (Yang et al., 2026).

to ground long-term narrative structure. First, because these methods typically employ autoregressive (AR) rollouts, they suffer from error accumulation; slight deviations in early frames compound over dozens of autoregressive steps, leading to significant quality degradation in minute-scale outputs. Even with extension techniques like InfinityRoPE (Yesiltepe et al., 2026), the fundamental issue remains: the model is trained without any ground-truth long-video supervision. Critically, a short-video teacher is strictly "blind" to long-context concepts. We also find "sink"- or cache-based rollout methods can become conservative over long horizons, shown by the dynamic degree scores and by the static-artifact ratings in Sec. C.

Analysis on our method. These observations validate our decoupled design. By utilizing SFT on the global Flow Matching head, we provide the necessary "long-context anchor" that allows the model to understand long-scale prompts and temporal dependencies—capabilities that the short teacher cannot provide. Simultaneously, by applying Teacher Distribution Matching exclusively via the mode-seeking head on local sliding windows, we inject the high-fidelity texture and motion priors of the teacher without requiring the teacher to understand long-range context.

4.3. Qualitative Results

We showcase qualitative results in Fig. 3, and present comparisons with the baselines using representative frames in Fig. 4. We refer to our supplemental material for

Method	Consistency ↑	Motion ↑	Quality ↑
No DDT dual heads	0.9427	0.9449	0.5298
No Sliding-window DMD	0.9604	0.9621	0.6075
No SFT	0.9579	0.9690	0.5862
Full Model	0.9615	0.9685	0.6359

Table 2. **Ablation study.** The first, second, and third values are highlighted. Our full model achieves the best overall performance, highlighting the necessity of all components.

dynamic video comparisons. Long-context SFT (LongSFT) and Mixed-length SFT (MixSFT) typically preserve the rough scene layout over time, but the outputs exhibit a characteristic loss of local realism: fine textures are washed out, edges appear softened, and the foreground subject can collapse into an under-defined silhouette (top example). In the street-walk example (bottom), SFT baselines also show signs of weakened long-range camera/scene continuity, with noticeable viewpoint/scene inconsistencies across frames. These failure modes are consistent with the intuition that, under scarce long-video supervision, a single mean-seeking objective tends to average away high-frequency detail even within short temporal windows. Teacher-only or rollout-based methods, such as CausVid (Yin et al., 2025), Self-Forcing (Huang et al., 2025b), InfinityRoPE (Yesiltepe et al., 2026), and LongLive (Yang et al., 2026), can inherit stronger local contrast from the short-video teacher, but often struggle to maintain realistic evolution over extended horizons. This appears to drift toward over-saturation and reduced dynamics, or conservative motion to avoid autoregressive error accumulation, which can manifest as visually stable yet less eventful sequences. The motion degradation is especially notable in InfinityRoPE (Yesiltepe et al., 2026), which uses "sink" mechanisms to attach generation to certain frames. Our method, by decoupling global mean-seeking supervision (SFT flow matching on long videos) from local mode-seeking alignment (reverse-KL distribution matching on sliding windows), simultaneously maintains long-range scene/narrative consistency and preserves crisp local appearance. Concretely, the foreground subject remains sharper and more consistently rendered across time, while backgrounds evolve smoothly without the abrupt viewpoint shifts seen in SFT-only training or the motion collapse/pattern repetition commonly induced by teacher-only long rollouts. Overall, Fig. 4 qualitatively supports our design goal: mode seeking meets mean seeking yields long-video generation that retains short-horizon realism while extending temporal coherence.

4.4. Ablation

As shown in Tab. 2, we ablate three key ingredients of our approach: (i) the decoupled DDT dual-head design, (ii) sliding-window teacher Distribution Matching via DMD (Yin et al., 2024b;a), and (iii) SFT on real long clips. Removing the DDT dual heads and naively training a single velocity predictor with both mean-seeking SFT and mode-seeking teacher alignment leads to the largest drop across all metrics, validating our motivation that the two objectives produce gradient interference when forced through one head. Disabling sliding-window DMD degrades the model into SFT-only. Removing SFT and relying only on sliding-window teacher alignment yields competitive motion but worse global consistency and overall quality, highlighting that a short-video teacher is inherently blind to minute-scale narrative structure and cannot substitute for long-video supervision. Taken together, these results show that SFT provides the long-horizon coherence anchor, teacher distribution matching restores short-horizon fidelity, and the decoupled dual-head architecture is essential for combining them effectively.

5. Conclusion

We presented Mode Seeking meets Mean Seeking, a simple training paradigm for scaling video diffusion from seconds to minutes under scarce long-video data. We use Decoupled Diffusion Transformer, which separates objectives with a shared long-context encoder and two heads: a mean-seeking SFT flow-matching head that learns minute-scale coherence from real long clips, and a mode-seeking sliding-window distribution-matching head that aligns student windows to a frozen short-video teacher via reverse-KL distribution matching, preserving short-horizon realism without requiring additional short-video supervision. Across quantitative, qualitative, and ablation results, this decoupling closes the fidelity–horizon gap by jointly improving local sharpness/motion and long-range consistency.

Limitations. Our method improves the fidelity–horizon trade-off, but it does not solve all forms of long-horizon planning. In failure cases, the model can satisfy local visual realism while misinterpreting compositional temporal instructions, such as changing a subject's action state (e.g., walking to flying) or revisiting a previously seen landmark rather than generating a genuinely cyclic route. This suggests that stronger event-level planning, explicit memory of entities/places, and evaluation protocols that test temporal causality and revisitation remain important future work. We refer to Sec. D for further discussion on limitations and future work.

Acknowledgements

We thank Shuai Yang for the fruitful discussion and explanation regarding LongLive (Yang et al., 2026). We thank Zhen Li to provide support regarding the Sekai (Li et al., 2025c) dataset. We thank the NVIDIA Academic Grant Program for providing part of the GPU resources for this project.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Bachmann, R., Allardice, J., Mizrahi, D., Fini, E., Kar, O. F., Amirloo, E., El-Nouby, A., Zamir, A., and Dehghan, A. Flextok: Resampling images into 1d token sequences of flexible length. In *ICML*, 2025.

Bain, M., Nagrani, A., Varol, G., and Zisserman, A. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021.

Bolya, D., Fu, C.-Y., Dai, X., Zhang, P., Feichtenhofer, C., and Hoffman, J. Token merging: Your vit but faster. In *ICLR*, 2023.

Cai, S., Yang, C., Zhang, L., Guo, Y., Xiao, J., Yang, Z., Xu, Y., Yang, Z., Yuille, A., Guibas, L., Agrawala, M., Jiang, L., and Wetzstein, G. Mixture of contexts for long video generation. In *ICLR*, 2026.

Chen, B., Martí Monsó, D., Du, Y., Simchowitz, M., Tedrake, R., and Sitzmann, V. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *NeurIPS*, 2024.

Chen, G., Lin, D., Yang, J., Lin, C., Zhu, J., Fan, M., Zhang, H., Chen, S., Chen, Z., Ma, C., Xiong, W., Wang, W., Pang, N., Kang, K., Xu, Z., Jin, Y., Liang, Y., Song, Y., Zhao, P., Xu, B., Qiu, D., Li, D., Fei, Z., Li, Y., and Zhou, Y. Skyreels-v2: Infinite-length film generative model. In *arXiv*, 2025.

Chen, J., Zhao, Y., Yu, J., Chu, R., Chen, J., Yang, S., Wang, X., Pan, Y., Zhou, D., Ling, H., Liu, H., Yi, H., Zhang, H., Li, M., Chen, Y., Cai, H., Fidler, S., Luo, P., Han, S., and Xie, E. Sana-video: Efficient video generation with block linear diffusion transformer. In *ICLR*, 2026a.

Chen, S., Wei, C., Sun, S., Shen, T., Nie, P., Zou, K., Zhang, G., Yang, M.-H., and Chen, W. Context forcing: Consistent autoregressive video generation with long context. In *ICML*, 2026b.

Cui, J., Wu, J., Li, M., Yang, T., Li, X., Wang, R., Bai, A., Ban, Y., and Hsieh, C.-J. Self-forcing++: Towards minute-scale high-quality video generation. In *ICLR*, 2026.

Dalal, K., Koceja, D., Xu, J., Zhao, Y., Han, S., Cheung, K. C., Kautz, J., Choi, Y., Sun, Y., and Wang, X. One-minute video generation with test-time training. In *CVPR*, 2025.

Dao, T. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *ICLR*, 2024.

Dao, T., Fu, D. Y., Ermon, S., Rudra, A., and Ré, C. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *NeurIPS*, 2022.

Fan, W., Si, C., Song, J., Yang, Z., He, Y., Zhuo, L., Huang, Z., Dong, Z., He, J., Pan, D., et al. Vchitect-2.0: Parallel transformer for scaling up video diffusion models. In *arXiv*, 2025.

Frans, K., Hafner, D., Levine, S., and Abbeel, P. One step diffusion via shortcut models. In *ICLR*, 2025.

Fu, X., Tang, S., Shi, M., Liu, X., Gu, J., Liu, M.-Y., Lin, D., and Lin, C.-H. Plenoptic video generation. In *CVPR*, 2026.

Gao, X., Hu, L., Hu, S., Huang, M., Ji, C., Meng, D., Qi, J., Qiao, P., Shen, Z., Song, Y., Sun, K., Tian, L., Wang, G., Wang, Q., Wang, Z., Xiao, J., Xu, S., Zhang, B., Zhang, P., Zhang, X., Zhang, Z., Zhou, J., and Zhuo, L. Wan-s2v: Audio-driven cinematic video generation. In *arXiv*, 2025a.

Gao, Y., Guo, H., Hoang, T., Huang, W., Jiang, L., Kong, F., Li, H., Li, J., Li, L., Li, X., Li, X., Li, Y., Lin, S., Lin, Z., Liu, J., Liu, S., Nie, X., Qing, Z., Ren, Y., Sun, L., Tian, Z., Wang, R., Wang, S., Wei, G., Wu, G., Wu, J., Xia, R., Xiao, F., Xiao, X., Yan, J., Yang, C., Yang, J., Yang, R., Yang, T., Yang, Y., Ye, Z., Zeng, X., Zeng, Y., Zhang, H., Zhao, Y., Zheng, X., Zhu, P., Zou, J., and Zuo, F. Seedance 1.0: Exploring the boundaries of video generation models. In *arXiv*, 2025b.

Geng, Z., Deng, M., Bai, X., Kolter, J. Z., and He, K. Mean flows for one-step generative modeling. In *NeurIPS*, 2025.

Gu, Y., Mao, W., and Shou, M. Z. Long-context autoregressive video modeling with next-frame prediction. In *arXiv*, 2025.

Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., and Dai, B. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024.

Guo, Y., Yang, C., He, H., Zhao, Y., Wei, M., Yang, Z., Huang, W., and Lin, D. End-to-end training for autoregressive video diffusion via self-resampling. In *arXiv*, 2025a.

Guo, Y., Yang, C., Yang, Z., Ma, Z., Lin, Z., Yang, Z., Lin, D., and Jiang, L. Long context tuning for video generation. In *ICCV*, 2025b.

HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al. Ltx-video: Realtime video latent diffusion. In *arXiv*, 2024.

Henschel, R., Khachatryan, L., Poghosyan, H., Hayrapetyan, D., Tadevosyan, V., Wang, Z., Navasardyan, S., and Shi, H. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *CVPR*, 2025.

Hong, W., Ding, M., Zheng, W., Liu, X., and Tang, J. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *ICLR*, 2023.

Huang, J., Hu, X., Han, B., Shi, S., Tian, Z., He, T., and Jiang, L. Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft. In *arXiv*, 2025a.

Huang, K., Huang, Y., Li, Y., Bai, J., Wang, X., Lin, Z., Ning, X., Yu, J., Wang, Y., and Liu, X. Cinescene: Implicit 3d as effective scene representation for cinematic video generation. In *CVPR*, 2026a.

Huang, X., Li, Z., He, G., Zhou, M., and Shechtman, E. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *NeurIPS*, 2025b.

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., and Liu, Z. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Huang, Z., Zhang, F., Xu, X., He, Y., Yu, J., Dong, Z., Ma, Q., Chanpaisit, N., Si, C., Jiang, Y., Wang, Y., Chen, X., Chen, Y.-C., Wang, L., Lin, D., Qiao, Y., and Liu, Z. Vbench++: Comprehensive and versatile benchmark suite for video generative models. In *TPAMI*, 2026b.

Jacobs, S. A., Tanaka, M., Zhang, C., Zhang, M., Song, S. L., Rajbhandari, S., and He, Y. Deepspeed ulysses: System optimizations for enabling training of extreme long sequence transformer models. In *arXiv*, 2023.

Ji, S., Chen, X., Yang, S., Tao, X., Wan, P., and Zhao, H. Memflow: Flowing adaptive memory for consistent and efficient long video narratives. In *arXiv*, 2025.

Jia, W., Lu, Y., Huang, M., Wang, H., Huang, B., Chen, N., Liu, M., Jiang, J., and Mao, Z. Moga: Mixture-of-groups attention for end-to-end long video generation. In *ICLR*, 2026.

Jin, Y., Sun, Z., Li, N., Xu, K., Xu, K., Jiang, H., Zhuang, N., Huang, Q., Song, Y., Mu, Y., and Lin, Z. Pyramidal flow matching for efficient video generative modeling. In *ICLR*, 2025.

Ju, X., Gao, Y., Zhang, Z., Yuan, Z., Wang, X., Zeng, A., Xiong, Y., Xu, Q., and Shan, Y. Miradata: A large-scale video dataset with long durations and structured captions. In *NeurIPS Datasets and Benchmarks*, 2024.

Kodaira, A., Hou, T., Hou, J., Georgopoulos, M., Juefei-Xu, F., Tomizuka, M., and Zhao, Y. Streamdit: Real-time streaming text-to-video generation. In *CVPR*, 2026.

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. In *arXiv*, 2024.

Lee, S.-H., Wang, J., Zhang, Z., Fan, D., and Li, X. Video token merging for long video understanding. In *NeurIPS*, 2024.

Li, C., Huang, D., Lu, Z., Xiao, Y., Pei, Q., and Bai, L. A survey on long video generation: Challenges, methods, and prospects. In *arXiv*, 2024.

Li, R., Torr, P., Vedaldi, A., and Jakab, T. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In *ICCV*, 2025a.

Li, W., Pan, W., Luan, P.-C., Gao, Y., and Alahi, A. Stable video infinity: Infinite-length video generation with error recycling. In *ICLR*, 2026.

Li, X., Li, M., Cai, T., Xi, H., Yang, S., Lin, Y., Zhang, L., Yang, S., Hu, J., Peng, K., Agrawala, M., Stoica, I., Keutzer, K., and Han, S. Radial attention: $\mathcal{O}(n \log n)$ sparse attention with energy decay for long video generation. In *NeurIPS*, 2025b.

Li, Z., Li, C., Mao, X., Lin, S., Li, M., Zhao, S., Xu, Z., Li, X., Feng, Y., Sun, J., Li, Z., Zhang, F., Ai, J., Wang, Z., Wu, Y., He, T., Pang, J., Qiao, Y., Jia, Y., and Zhang, K. Sekai: A video dataset towards world exploration. In *NeurIPS*, 2025c.

Lin, S., Yang, C., He, H., Jiang, J., Ren, Y., Xia, X., Zhao, Y., Xiao, X., and Jiang, L. Autoregressive adversarial post-training for real-time interactive video generation. In *NeurIPS*, 2025a.

Lin, S., Yang, C., Lin, Z., Chen, H., and Fan, H. Continuous adversarial flow models. In *arXiv*, 2026.

Lin, W., Chen, R., Liu, B., Yan, S., Feng, R., Wei, J., Zhang, Y., Zhou, Y., Feng, C., Ran, J., Wu, Q., Liu, Z., and Guo, M. Contentv: Efficient training of video generation models with limited compute. In *arXiv*, 2025b.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. In *ICLR*, 2023.

Liu, K., Hu, W., Xu, J., Shan, Y., and Lu, S. Rolling forcing: Autoregressive long video diffusion in real time. In *ICLR*, 2026.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

Liu, X., Zhang, X., Ma, J., Peng, J., and Liu, Q. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *ICLR*, 2024.

Lu, C. and Song, Y. Simplifying, stabilizing and scaling continuous-time consistency models. In *ICLR*, 2025.

Lu, Y. and Yang, Y. Freelong++: Training-free long video generation via multi-band spectralfusion. In *arXiv*, 2025.

Lu, Y., Liang, Y., Zhu, L., and Yang, Y. FreeLong: Training-free long video generation with SpectralBlend temporal attention. In *NeurIPS*, 2024.

Lu, Y., Zeng, Y., Li, H., Ouyang, H., Wang, Q., Cheng, K. L., Zhu, J., Cao, H., Zhang, Z., Zhu, X., Shen, Y., and Zhang, M. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. In *arXiv*, 2025.

Luo, S., Tan, Y., Huang, L., Li, J., and Zhao, H. Latent consistency models: Synthesizing high-resolution images with few-step inference. In *arXiv*, 2023a.

Luo, W., Hu, T., Zhang, S., Sun, J., Li, Z., and Zhang, Z. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. In *NeurIPS*, 2023b.

Lv, C., Shi, Y., Huang, Y., Gong, R., Ren, S., and Wang, W. Light forcing: Accelerating autoregressive video diffusion via sparse attention. In *ICML*, 2026.

Meng, Y., Ouyang, H., Yu, Y., Wang, Q., Wang, W., Cheng, K. L., Wang, H., Ma, S., Li, Y., Chen, C., Zeng, Y., Zhu, X., Shen, Y., and Qu, H. Holocine: Holistic generation of cinematic multi-shot long video narratives. In *CVPR*, 2026.

Nie, W., Berner, J., Liu, C., and Vahdat, A. Nvidia fastgen: Fast generation from diffusion models. In *GitHub*, 2026.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *ICCV*, 2023.

Po, R., Nitzan, Y., Zhang, R., Chen, B., Dao, T., Shechtman, E., Wetzstein, G., and Huang, X. Long-context state-space video world models. In *ICCV*, 2025.

Po, R., Chan, E. R., Chen, C., and Wetzstein, G. Bagger: Backwards aggregation for mitigating drift in autoregressive video diffusion models. In *CVPR*, 2026.

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023.

Qiu, H., Xia, M., Zhang, Y., He, Y., Wang, X., Shan, Y., and Liu, Z. FreeNoise: Tuning-free longer video diffusion via noise rescheduling. In *ICLR*, 2024.

Ruhe, D., Heek, J., Salimans, T., and Hoogeboom, E. Rolling diffusion models. In *ICML*, 2024.

Salimans, T. and Ho, J. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022.

Sand.ai, Teng, H., Jia, H., Sun, L., Li, L., Li, M., Tang, M., Han, S., Zhang, T., Zhang, W. Q., Luo, W., Kang, X., Sun, Y., Cao, Y., Huang, Y., Lin, Y., Fang, Y., Tao, Z., Zhang, Z., Wang, Z., Liu, Z., Shi, D., Su, G., Sun, H., Pan, H., Wang, J., Sheng, J., Cui, M., Hu, M., Yan, M., Yin, S., Zhang, S., Liu, T., Yin, X., Yang, X., Song, X., Hu, X., Zhang, Y., and Li, Y. Magi-1: Autoregressive video generation at scale. In *arXiv*, 2025.

Sauer, A., Lorenz, D., Blattmann, A., and Rombach, R. Adversarial diffusion distillation. In *ECCV*, 2024.

Savov, N., Kazemi, N., Zhang, D., Paudel, D. P., Wang, X., and Gool, L. V. Statespacediffuser: Bringing long context to diffusion world models. In *NeurIPS*, 2025.

Song, K., Chen, B., Simchowitz, M., Du, Y., Tedrake, R., and Sitzmann, V. History-guided video diffusion. In *ICML*, 2025.

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. Consistency models. In *ICML*, 2023.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. In *Neurocomputing*, 2024.

Team Seawead, Yang, C., Lin, Z., Zhao, Y., Lin, S., Ma, Z., Guo, H., Chen, H., Qi, L., Wang, S., Cheng, F., Zuo, F., Zeng, X., Yang, Z., Kong, F., Wei, M., Qing, Z., Xiao, F., Hoang, T., Zhang, S., Zhu, P., Zhao, Q., Yan, J., Gui, L., Bi, S., Li, J., Ren, Y., Wang, R., Li, H., Xiao, X., Liu, S., Ling, F., Zhang, H., Wei, H., Kuang, H., Duncan, J., Zhang, J., Zheng, J., Sun, L., Zhang, M., Sun, R., Zhuang, X., Li, X., Xia, X., Chi, X., Peng, Y., Wang, Y., Wang, Y., Zhao, Z., Chen, Z., Song, Z., Yang, Z., Feng, J., Yang, J., and Jiang, L. Seaweed-7b: Cost-effective training of video generation foundation model. In *arXiv*, 2025.

Team Seedance, Chen, H., Chen, S., Chen, X., Chen, Y., Chen, Y., Chen, Z., Cheng, F., Cheng, T., Cheng, X., Chi, X., Cong, J., Cui, J., Cui, Q., Dong, Q., Fan, J., Fang, J., Fang, Z., Feng, C., Feng, H., Gao, M., Gao, Y., Guo, D., Guo, Q., Hao, B., Hao, Q., He, B., He, Q., Hoang, T., Hu, R., Hu, X., Huang, W., Huang, Z., Huang, Z.,

Ji, D., Jiang, S., Jiang, W., Jiang, Y., Jiang, Z., Kim, A., Kong, J., Lai, Z., Lao, S., Leng, Y., Li, A., Li, F., Li, G., Li, H., Li, J., Li, L., Li, M., Li, S., Li, T., Li, X., Li, X., Li, X., Li, X., Li, Y., Li, Y., Li, Y., Liang, C., Liang, H., Liang, J., Liang, Y., Liang, Z., Liao, W., Liao, Y., Lin, H., Lin, K., Lin, S., Lin, X., Lin, Z., Ling, F., Liu, F., Liu, G., Liu, J., Liu, J., Liu, J., Liu, S., Liu, S., Liu, S., Liu, S., Liu, X., Liu, X., Liu, Y., Liu, Z., Liu, Z., Lyu, J., Lyu, L., Lyu, Q., Mu, H., Nie, X., Ning, J., Pan, X., Peng, Y., Qin, L., Qu, X., Ren, Y., Shen, K., Shi, G., Shi, L., Song, Y., Song, Y., Sun, F., Sun, L., Sun, R., Sun, Y., Sun, Z., Tang, W., Tang, Y., Tao, Z., Wang, F., Wang, F., Wang, J., Wang, J., Wang, K., Wang, K., Wang, Q., Wang, R., Wang, S., Wang, S., Wang, T., Wang, W., Wang, X., Wang, Y., Wang, Y., Wang, Y., Wang, Y., Wang, Z., Wei, G., Wei, W., Wu, D., Wu, G., Wu, H., Wu, J., Wu, J., Wu, R., Wu, X., Wu, Y., Xia, R., Xiang, L., Xiao, F., Xiao, X., Xie, P., Xie, S., Xu, S., Xue, J., Yan, S., Yang, B., Yang, C., Yang, J., Yang, R., Yang, T., Yang, Y., Yang, Y., Yang, Z., Yang, Z., Yao, S., Yao, Y., Ye, Z., Yu, B., Yu, J., Yuan, C., Yuan, L., Zeng, S., Zeng, W., Zeng, X., Zeng, Y., Zhang, C., Zhang, H., Zhang, J., Zhang, K., Zhang, L., Zhang, L., Zhang, M., Zhang, T., Zhang, W., Zhang, X., Zhang, X., Zhang, Y., Zhang, Y., Zhang, Z., Zhao, F., Zhao, H., Zhao, Y., Zheng, H., Zheng, J., Zheng, X., Zheng, Y., Zheng, Y., Zhou, J., Zhu, J., Zhu, K., Zhu, S., Zhu, W., Zou, B., and Zuo, F. Seedance 1.5 pro: A native audio-visual joint generation foundation model. In *arXiv*, 2025.

Team Wan, 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., and Liu, Z. Wan: Open and advanced large-scale video generative models. In *arXiv*, 2025.

Tong, S., Ma, N., Xie, S., and Jaakkola, T. Flow map distillation without data. In *CVPR*, 2026.

Wang, Q., Shi, X., Li, B., Bian, W., Liu, Q., Lu, H., Wang, X., Wan, P., Gai, K., and Jia, X. Multishotmaster: A controllable multi-shot video generation framework. In *CVPR*, 2026a.

Wang, S., Tian, Z., Huang, W., and Wang, L. Ddt: Decoupled diffusion transformer. In *CVPR*, 2026b.

Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Li, X., Chen, G., Chen, X., Wang, Y., Luo, P., Liu, Z., Wang, Y., Wang, L., and Qiao, Y. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *ICLR*, 2024.

Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., and Zhu, J. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *NeurIPS*, 2023.

Wen, Q., Huang, Z., Meng, X., He, W., and Li, C.-G. Mixture-of-Top-k attention: Efficient attention via scalable fast weights. In *arXiv*, 2026.

Wu, B., Zou, C., Li, C., Huang, D., Yang, F., Tan, H., Peng, J., Wu, J., Xiong, J., Jiang, J., Linus, Patrol, Zhang, P., Chen, P., Zhao, P., Tian, Q., Liu, S., Kong, W., Wang, W., He, X., Li, X., Deng, X., Zhe, X., Li, Y., Long, Y., Peng, Y., Wu, Y., Liu, Y., Wang, Z., Dai, Z., Peng, B., Li, C., Gong, G., Xiao, G., Tian, J., Lin, J., Liu, J., Zhang, J., Lian, J., Pan, K., Wang, L., Niu, L., Chen, M., Chen, M., Zheng, M., Yang, M., Hu, Q., Yang, Q., Xiao, Q., Wu, R., Xu, R., Yuan, R., Sang, S., Huang, S., Gong, S., Huang, S., Guo, W., Yuan, X., Chen, X., Hu, X., Sun, W., Wu, X., Ren, X., Yuan, X., Mi, X., Zhang, Y., Sun, Y., Lu, Y., Li, Y., Huang, Y., Tang, Y., Li, Y., Deng, Y., Zhou, Y., Hu, Z., Liu, Z., Yang, Z., Yang, Z., Lu, Z., Zhou, Z., and Zhong, Z. Hunyuanvideo 1.5 technical report. In *arXiv*, 2025a.

Wu, J., Hou, L., Yang, H., Tao, X., Tian, Y., Wan, P., Zhang, D., and Tong, Y. Vmoba: Mixture-of-block attention for video diffusion models. In *ICLR*, 2026a.

Wu, R., He, X., Cheng, M., Yang, T., Zhang, Y., Kang, Z., Cai, X., Wei, X., Guo, C., Li, C., and Cheng, M.-M. Infinite-world: Scaling interactive world models to 1000-frame horizons via pose-free hierarchical memory. In *arXiv*, 2026b.

Wu, X., Zhang, G., Xu, Z., Zhou, Y., Lu, Q., and He, X. Pack and force your memory: Long-form and consistent video generation. In *arXiv*, 2025b.

Xi, H., Yang, S., Zhao, Y., Xu, C., Li, M., Li, X., Lin, Y., Cai, H., Zhang, J., Li, D., Chen, J., Stoica, I., Keutzer, K., and Han, S. Sparse video-gen: Accelerating video diffusion transformers with spatial-temporal sparsity. In *ICML*, 2025.

Xia, Y., Ling, S., Fu, F., Wang, Y., Li, H., Xiao, X., and Cui, B. Training-free and adaptive sparse attention for efficient long video generation. In *ICCV*, 2025.

Xiao, J., Yang, C., Zhang, L., Cai, S., Zhao, Y., Guo, Y., Wetzstein, G., Agrawala, M., Yuille, A., and Jiang, L. Captain cinema: Towards short movie generation. In *ICLR*, 2026a.

Xiao, Z., Lan, Y., Zhou, Y., Ouyang, W., Yang, S., Zeng, Y., and Pan, X. Worldmem: Long-term consistent world simulation with memory. In *NeurIPS*, 2025.

Xiao, Z., Zhao, Y., Li, L., Lan, Y., Yu, N., Garg, R., Cooper, R., Taghavi, M. H., and Pan, X. Video4spatial: Towards visuospatial intelligence with context-guided video generation. In *CVPR Findings*, 2026b.

Xie, E., Chen, J., Chen, J., Cai, H., Tang, H., Lin, Y., Zhang, Z., Li, M., Zhu, L., Lu, Y., and Han, S. SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *ICLR*, 2025a.

Xie, Y., Gu, T., Li, Z., Zhang, C., Song, G., Zhao, X., Liang, C., Jiang, J., Xu, H., and Luo, L. X-streamer: Unified human world modeling with audiovisual interaction. In *arXiv*, 2025b.

Xu, Y., Nie, W., and Vahdat, A. One-step diffusion models with f-divergence distribution matching. In *arXiv*, 2025.

Yang, S., Xi, H., Zhao, Y., Li, M., Zhang, J., Cai, H., Lin, Y., Li, X., Xu, C., Peng, K., et al. Sparse videogen2: Accelerate video generation with sparse attention via semantic-aware permutation. In *NeurIPS*, 2025a.

Yang, S., Huang, W., Chu, R., Xiao, Y., Zhao, Y., Wang, X., Li, M., Xie, E., Chen, Y., Lu, Y., Han, S., and Chen, Y. Longlive: Real-time interactive long video generation. In *ICLR*, 2026.

Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., Yin, D., Zhang, Y., Wang, W., Cheng, Y., Xu, B., Gu, X., Dong, Y., and Tang, J. CogVideoX: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025b.

Yesiltepe, H., Meral, T. H. S., Akan, A. K., Oktay, K., and Yanardag, P. Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. In *CVPR*, 2026.

Yi, J., Jang, W., Cho, P. H., Nam, J., Yoon, H., and Kim, S. Deep forcing: Training-free long video generation with deep sink and participative compression. In *ICML*, 2026.

Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., and Freeman, W. T. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, 2024a.

Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W. T., and Park, T. One-step diffusion with distribution matching distillation. In *CVPR*, 2024b.

Yin, T., Zhang, Q., Zhang, R., Freeman, W. T., Durand, F., Shechtman, E., and Huang, X. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, 2025.

Yu, J., Bai, J., Qin, Y., Liu, Q., Wang, X., Wan, P., Zhang, D., and Liu, X. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In *SIGGRAPH Asia*, 2025a.

Yu, Y., Wu, X., Hu, X., Hu, T., Sun, Y., Lyu, X., Wang, B., Ma, L., Ma, Y., Wang, Z., and Qi, X. Videossm: Autoregressive long video generation with hybrid state-space memory. In *arXiv*, 2025b.

Zhan, C., Li, W., Shen, C., Zhang, J., Wu, S., and Zhang, H. Bidirectional sparse attention for faster video diffusion training. In *arXiv*, 2025.

Zhang, J., Huang, H., Zhang, P., Wei, J., Zhu, J., and Chen, J. SageAttention2: Efficient attention with thorough outlier smoothing and per-thread INT4 quantization. In *ICML*, 2025a.

Zhang, J., Wei, J., Zhang, P., Zhu, J., and Chen, J. Sageattention: Accurate 8-bit attention for plug-and-play inference acceleration. In *International Conference on Learning Representations (ICLR)*, 2025b.

Zhang, J., Xiang, C., Huang, H., Wei, J., Xi, H., Zhu, J., and Chen, J. Spargeattention: Accurate and training-free sparse attention accelerating any model inference. In *ICML*, 2025c.

Zhang, J., Jiang, K., Xiang, C., Feng, W., Hu, Y., Xi, H., Chen, J., and Zhu, J. Spargeattention2: Trainable sparse attention via hybrid top-k+top-p masking and distillation fine-tuning. In *arXiv*, 2026a.

Zhang, J., Jiang, M., Dai, N., Lu, T., Uzunoglu, A., Zhang, S., Wei, Y., Wang, J., Patel, V. M., Liang, P. P., Khashabi, D., Peng, C., Chellappa, R., Shu, T., Yuille, A., Du, Y., and Chen, J. World-in-world: World models in a closed-loop world. In *ICLR*, 2026b.

Zhang, J., Wang, H., Jiang, K., Zheng, K., Jiang, Y., Stoica, I., Chen, J., Zhu, J., and Gonzalez, J. E. Sla2: Sparse-linear attention with learnable routing and qat. In *arXiv*, 2026c.

Zhang, K., Jiang, L., Wang, A., Fang, J. Z., Zhi, T., Yan, Q., Kang, H., Lu, X., and Pan, X. StoryMem: Multi-shot long video storytelling with memory. In *arXiv*, 2025d.

Zhang, L., Cai, S., Li, M., Wetzstein, G., and Agrawala, M. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *NeurIPS*, 2025e.

Zhang, L., Cai, S., Li, M., Zeng, C., Lu, B., Rao, A., Han, S., Wetzstein, G., and Agrawala, M. Tinyhistory: Lightweight video history embeddings via two-stage context learning. In *ECCV*, 2026d.

Zhang, P., Chen, Y., Huang, H., Lin, W., Liu, Z., Stoica, I., Xing, E. P., and Zhang, H. H. Faster video diffusion with trainable sparse attention. In *NeurIPS*, 2025f.

Zhang, P., Chen, Y., Su, R., Ding, H., Stoica, I., Liu, Z., and Zhang, H. Fast video generation with sliding tile attention. In *ICML*, 2025g.

Zhang, T., Bi, S., Hong, Y., Zhang, K., Luan, F., Yang, S., Sunkavalli, K., Freeman, W. T., and Tan, H. Test-time training done right. In *ICLR*, 2026e.

Zhang, Y., Xing, J., Xia, B., Liu, S., Peng, B., Tao, X., Wan, P., Lo, E., and Jia, J. Training-free efficient video generation via dynamic token carving. In *NeurIPS*, 2025h.

Zhang, Z., Mao, J., Chang, S., He, Y., Han, Y., Tang, J., Wang, F., and Zhuang, B. Bife: Better interaction, fewer errors for minute-long video generation. In *arXiv*, 2025i.

Zhao, M., He, G., Chen, Y., Zhu, H., Li, C., and Zhu, J. Riflex: A free lunch for length extrapolation in video diffusion transformers. In *ICML*, 2025.

Zheng, D., Huang, Z., Liu, H., Zou, K., He, Y., Zhang, F., Zhang, Y., He, J., Zheng, W.-S., Qiao, Y., and Liu, Z. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. In *arXiv*, 2025.

Zheng, K., Wang, Y., Ma, Q., Chen, H., Zhang, J., Balaji, Y., Chen, J., Liu, M.-Y., Zhu, J., and Zhang, Q. Large scale diffusion distillation via score-regularized continuous-time consistency. In *ICLR*, 2026.

Zhou, M., Zheng, H., Wang, Z., Yin, M., and Huang, H. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *ICML*, 2024.

Method	Subject Consistency ↑	Background Consistency ↑	Motion Smoothness ↑	Dynamic Degree ↑	Aesthetic Quality ↑	Image Quality ↑	VLM Consistency ↑
CausVid (Yin et al., 2025)	0.9702	0.9589	0.9771	0.5078	0.5596	0.6264	41.82
Self-Forcing (Huang et al., 2025b)	0.9516	0.9467	0.9784	0.4609	0.5524	0.6291	40.37
InfinityRoPE (Yesiltepe et al., 2026)	0.9683	0.9568	0.9810	0.7891	0.5415	0.6826	66.08
LongLive (Yang et al., 2026)	0.9679	0.9572	0.9827	0.8203	0.5704	0.6908	62.74
Ours	0.9676	0.9569	0.9839	0.9297	0.5861	0.7015	73.08

Table 3. **Additional 60-second quantitative comparison.** The first, second, and third values are highlighted within this 60-second stress test. Our method achieves the strongest motion smoothness, dynamic degree, aesthetic quality, image quality, and VLM consistency, while remaining competitive on subject/background consistency.

A. Gemini Evaluation Prompt.

```
You are evaluating a video generation result.

Task: score the *semantic consistency* of the video on a 0-100 scale.

Semantic consistency means: objects, identities, attributes, and the overall
scene remain coherent over time;

penalize sudden unrealistic changes, object identity swaps, content drift, or
contradictions between frames.

IMPORTANT: account for whether the video is trivially consistent because it is
static.

If the video is essentially a still image / frozen frame(s) with little-to-no
motion or temporal change, do NOT give a high score.

In that case, assign a low score because it does not demonstrate temporal
consistency under motion.

Return ONLY a single integer between 0 and 100 (inclusive). No other words.
```

B. Additional Experiment on One-Minute Generation

To test whether the fidelity–horizon trade-off persists beyond the 30-second setting in the main paper, we additionally evaluate 60-second generation using the same automatic protocol as Sec. 4. Because fully retraining the SFT-only baselines at this horizon requires a separate long-horizon tuning run, this additional evaluation focuses on rollout and context-extension baselines for which 60-second generation was available. Thus, the table should be read as a one-minute stress test rather than a replacement for the complete 30-second benchmark in Tab. 1.

The one-minute results reinforce the interpretation of the quantitative results in Sec. 4. CausVid and LongLive remain strong on frame-level subject/background consistency, but their lower dynamic degree and VLM consistency indicate more conservative temporal evolution. In contrast, our model obtains the best dynamic degree (0.9297), image quality (0.7015), and VLM consistency (73.08), suggesting that the decoupled mean-seeking and mode-seeking objectives preserve both local quality and long-horizon evolution at the longer generation horizon.

C. User Study

We conducted a user study with 31 participants. Each participant viewed randomized sets of baseline and ours videos: 25 sets at 30 seconds and 15 sets at 60 seconds. Participants could scrub through each video and rated overall quality, global consistency, and static artifacts. Scores are averaged over participants; higher is better for overall quality and global consistency, while lower is better for static artifacts. As with Sec. B, the 60-second setting excludes SFT-only baselines because they require separate retraining for this longer native horizon.

D. Further Discussion on Limitations and Future Work

We also note that our method is orthogonal to causal autoregressive methods (Yin et al., 2025; Huang et al., 2025b). A natural follow-up is to use it as a base model for causal AR training, or, conversely, to distill our long-context bidirectional model into a causal sampler, which could be as simple as adding a causal attention mask during training. Because our student

Horizon	Method	Overall Quality ↑	Global Consistency ↑	Static Artifact ↓
30 sec	LongSFT	2.85	3.09	2.11
	MixedSFT	2.66	3.42	1.64
	CausVid (Yin et al., 2025)	1.98	1.92	2.98
	Self-Forcing (Huang et al., 2025b)	2.71	1.75	2.82
	InfinityRoPE (Yesiltepe et al., 2026)	3.23	2.99	4.12
	LongLive (Yang et al., 2026)	3.12	3.27	4.03
	Ours	3.68	3.59	1.86
60 sec	CausVid (Yin et al., 2025)	2.06	1.90	4.08
	Self-Forcing (Huang et al., 2025b)	2.69	1.98	3.91
	InfinityRoPE (Yesiltepe et al., 2026)	3.11	2.86	3.03
	LongLive (Yang et al., 2026)	3.24	3.06	3.22
	Ours	3.59	3.41	2.08

Table 4. **User study.** The first, second, and third values are highlighted, with the ordering reversed for static artifacts because lower is better. Our method receives the highest overall quality and global consistency ratings for both 30-second and 60-second videos, and substantially reduces static artifacts relative to rollout-based baselines.

is trained with substantially longer native temporal context than typical short-clip teachers, we expect it to extrapolate well to minute-scale (and longer) generation, especially when combined with rollout-robustness and context-extension techniques such as Rolling Forcing (Liu et al., 2026), LongLive (Yang et al., 2026), or longer-context positional embedding schemes (e.g., InfinityRoPE (Yesiltepe et al., 2026)). Finally, our 60-second supplemental benchmark is a stress test against available rollout/context-extension baselines, not a complete replacement for retrained SFT-only baselines at that horizon. More broadly, the native long-context encoder provides a persistent history representation in the spirit of Genie-style models, and adding interaction/action conditioning on top of this representation is an especially promising direction.