MOVIEROLL: ROLLING OUT LONG-FORM VIDEOS WITH HIERARCHICAL KV CACHING AND CUMULATIVE LONG TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming video generation seeks to extend high-quality short-video synthesis to continuous, interactive generation. However, the mismatch between training on short videos with ground-truth context and long-horizon inference conditioned on generated content makes quality difficult to sustain. Meanwhile, full-history KV caching incurs memory costs that grow with generation length. To address these challenges, we present **MovieRoll**, a **base-to-distilled** framework that first establishes a multi-step streaming base through supervised training and then distills it into a few-step generator through long self-rollouts. At the base stage, Window-wise Teacher Forcing jointly denoises multiple chunks with progressive noise levels within each window, preserving within-window temporal interactions and improving resistance to long-horizon degradation compared with chunk-wise training. To retain historical context efficiently, hierarchical KV-cache compression (**HieraKV**) exploits layer-wise attention sparsity in streaming DiTs, retaining the global available history in shallow layers but only the sink chunk and the most recent history chunk in deep layers, which reduces KV-cache memory by 78% and improves degradation resistance. Building on this base, cumulative long tuning with distribution matching distillation (**CuLT-DMD**) accumulates and averages gradients from different local clips throughout an entire long self-rollout. The resulting gradient spans the full sample and provides a rollout-level estimate, improving stability during long-horizon generation. Finally, at inference time, threshold-triggered history selection and KV recomputation bound the retained context, while classifier-free guidance annealing (**AnnealCFG**) progressively reduces guidance strength toward the clean endpoint during denoising to further stabilize MovieRoll-Base. In 30-second HeliosBench evaluations, MovieRoll-Base achieves competitive generation quality. MovieRoll-Distilled achieves state-of-the-art (**SOTA**) overall scores in both single- and multi-event settings, combining the highest naturalness with sufficient dynamics. Model weights and code will be publicly released after the review process.

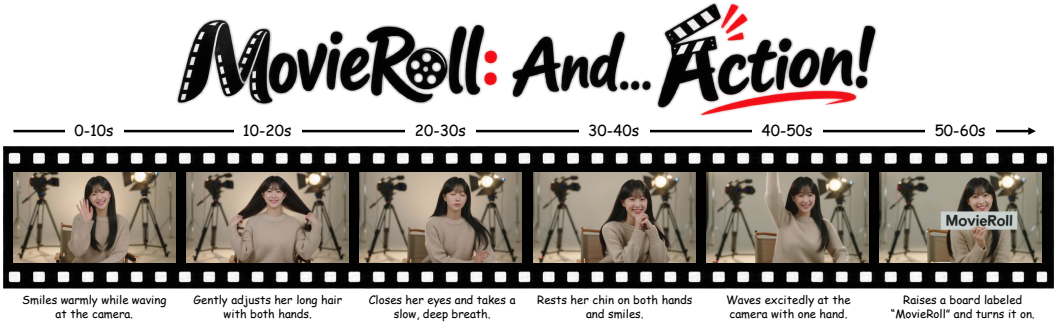


Figure 1: **MovieRoll supports minute-scale long-form video generation and multi-event inputs**. It employs multiple strategies to address error accumulation in long-video generation, enabling the synthesis of realistic videos.

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

1 INTRODUCTION

Video generation models based on bidirectional attention produce highly realistic short clips (Zheng et al., 2024; 2025b; Lin et al., 2024; Ge et al., 2026; Yang et al., 2025b; Wan et al., 2025; Kong et al., 2024; Wu et al., 2025; HaCohen et al., 2024; 2026; Polyak et al., 2024; Ge et al., 2025; Yuan et al., 2025), but joint denoising delays the first output and increases computational cost with the target duration. In contrast, streaming generation produces chunks sequentially, conditioning each chunk on previously finalized content (Chen et al., 2024; Yin et al., 2025; Yang et al., 2025a; Zhang et al., 2026a; Sun et al., 2025; HunyuanWorld, 2025; Zhang et al., 2025; Huang et al., 2026b;a), enabling continuous output and online interaction.

However, streaming video generation faces several challenges. First, models are typically trained on short videos but are expected to generate much longer sequences at inference time. This discrepancy introduces both exposure bias, as training conditions on ground-truth histories whereas inference relies on self-generated histories, and an extrapolation challenge when generation extends beyond the training horizon. Second, KV-cache memory grows continuously with the generation length, making long-horizon inference increasingly memory intensive. Existing methods like Self Forcing (Huang et al., 2026c) mitigate exposure bias by distilling a teacher distribution along self-rollout trajectories, while LongLive (Yang et al., 2025a) further introduces streaming long tuning over long self-rollout sequences to improve long-horizon inference. However, efficient self-rollout restricts these methods primarily to few-step generators, leaving a lack of sufficiently capable multi-step streaming base models for downstream fine-tuning and as initialization for distillation. Resampling Forcing (Guo et al., 2026) and Helios-Base (Yuan et al., 2026) instead explore exposure-bias mitigation for multi-step base models. However, Resampling Forcing is evaluated only for generation up to 15 seconds, leaving its long-horizon extrapolation ability unverified, while Helios-Base adopts a bidirectional-attention continuation architecture rather than causal attention, precluding KV caching and limiting the historical context available during inference. For history management, prior work has explored sliding windows (Huang et al., 2026c; Zhu et al., 2026; Liu et al., 2026; Yang et al., 2025a), retrieval (Cai et al., 2026), evolving memory (Bian et al., 2026; Yi et al., 2025), and context compression (Chen et al., 2026). These approaches either discard intermediate history directly or introduce additional mechanisms that increase implementation and scaling complexity.

To address these challenges, we present **MovieRoll**, a **base-to-distilled** framework that first establishes a multi-step streaming base through supervised training and then distills it through long self-rollouts. At the base stage, Window-wise Teacher Forcing adapts a pretrained bidirectional model using multiple rolling windows (Ruhe et al., 2024; Teng et al., 2025; Liu et al., 2026) in each forward pass. Each window conditions on clean preceding history while jointly denoising multiple noisy chunks with progressive noise levels. Compared with chunk-wise training, this design improves robustness against long-horizon degradation and provides a supervised base for subsequent adaptation and distillation. To control the continuously growing KV-cache memory during streaming generation, we further exploit the layer-wise sparsity of streaming DiTs. Specifically, we use a hierarchical KV-cache compression strategy (**HieraKV**) that retains the global history in shallow layers while keeping only the sink chunk and the most recent history chunk in deep layers. Compared with retaining the full KV history across all layers, HieraKV reduces peak KV-cache memory by approximately 78% and reduces inference time by approximately 53% (a 2.13× speedup), while also improving long-horizon generation quality.

Building on this base, cumulative long tuning with distribution matching distillation (**CuLT-DMD**) uses long self-rollouts to train a few-step generator under DMD supervision (Yin et al., 2024b;a). Unlike LongLive, which repeatedly alternates generator and critic updates within a long self-rollout and therefore makes each update dependent on the quality of the current rollout timestamp, CuLT-DMD accumulates gradients over all local clips in the rollout and performs a single update after averaging them. This yields a rollout-level gradient estimate that covers the entire long sample and is less sensitive to timestamp-specific quality fluctuations. Empirically, CuLT-DMD achieves better long-horizon generation performance than LongLive-style long tuning.

Finally, during inference, threshold-triggered history selection and KV recomputation maintain a bounded context size while providing stronger long-horizon extrapolation than sliding-window strategies. To further mitigate inference-time degradation in MovieRoll-Base, classifier-free guidance annealing (**AnnealCFG**) adjusts the CFG strength (Ho & Salimans, 2022) along the denoising

trajectory. It maintains strong guidance at high noise levels and progressively reduces it toward the clean endpoint, preserving visual quality while further improving long-horizon generation stability.

On the 30-second HeliosBench evaluations (Yuan et al., 2026), MovieRoll-Base achieves competitive performance despite training only on videos up to 15 seconds, while MovieRoll-Distilled achieves state-of-the-art overall scores among the compared methods in both single- and multi-event settings, with the highest naturalness while maintaining sufficient dynamics. Overall, our contributions are as follows. First, we establish a robust multi-step streaming base with Window-wise Teacher Forcing and HieraKV. Second, we propose CuLT-DMD for stable long-horizon self-rollout distillation. Finally, we introduce threshold-triggered KV recomputation together with AnnealCFG for efficient and stable long-horizon inference.

2 Related Work

2.1 Bidirectional Video Generation Models

Open-source video diffusion models, including Open-Sora (Zheng et al., 2024), Open-Sora Plan (Lin et al., 2024), CogVideoX (Yang et al., 2025b), HunyuanVideo (Kong et al., 2024; Wu et al., 2025), and Wan (Wan et al., 2025), jointly denoise full sequences with bidirectional temporal interactions. Full-sequence denoising increases computation and memory with video length and delays frame emission until sampling completes, limiting low-latency streaming.

2.2 Autoregressive Video Generation Models

Autoregressive diffusion models generate frames or chunks sequentially and reuse historical states through KV caching. Diffusion Forcing (Chen et al., 2024) assigns independent noise levels to temporal tokens, while CausVid (Yin et al., 2025) distills a bidirectional model into a causal few-step generator using distribution matching distillation (DMD) (Yin et al., 2024b;a). To mitigate exposure bias, Self Forcing (Huang et al., 2026c) combines training-time self-rollouts with video-level distribution matching. Causal Forcing (Zhu et al., 2026) revisits its ODE initialization using an autoregressive teacher. Rolling Forcing (Liu et al., 2026) jointly refines chunks at different noise levels in a rolling window, caching initial and recent frames. Helios (Yuan et al., 2026) instead combines history corruption and compression for continuation from a fixed historical context. Long rollouts can still degrade beyond the training horizon.

2.3 Interactive Video Generation Models

Interactive generation requires responsiveness to changing prompts while preserving temporal continuity and long-range consistency. LongLive (Yang et al., 2025a) combines autoregressive generation, streaming long tuning, short-window attention, and a persistent frame sink. At prompt switches, KV recaching rebuilds historical states under the new condition to reduce residual semantics and preserve continuity. MemFlow (Ji et al., 2025) retrieves prompt-relevant historical frames and sparsely activates useful memory tokens. Echo-Infinity (Bian et al., 2026) combines persistent anchors and recent context with evolving learnable memory that absorbs discarded history, retaining information beyond the local window under bounded computation and memory. Efficiently exploiting long-range history remains challenging for sustained interactive generation.

3 Method

3.1 Overview

Base-model training. We fine-tune the pretrained bidirectional teacher into MovieRoll-Base, a multi-step autoregressive model. Each forward pass jointly trains multiple rolling windows at progressive noise levels, conditioned on clean ground-truth histories. Prompts may vary across windows, while noisy chunks interact bidirectionally within each window. We average the flow-matching loss (Lipman et al., 2022; Liu et al., 2022) over training windows:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon},t,\mathbf{c}}\left[\left\|v_\theta(\mathbf{x}_t, t, \mathbf{x}_0^{<i}, \mathbf{c}) - \mathbf{v}_t\right\|_2^2\right], \tag{1}$$

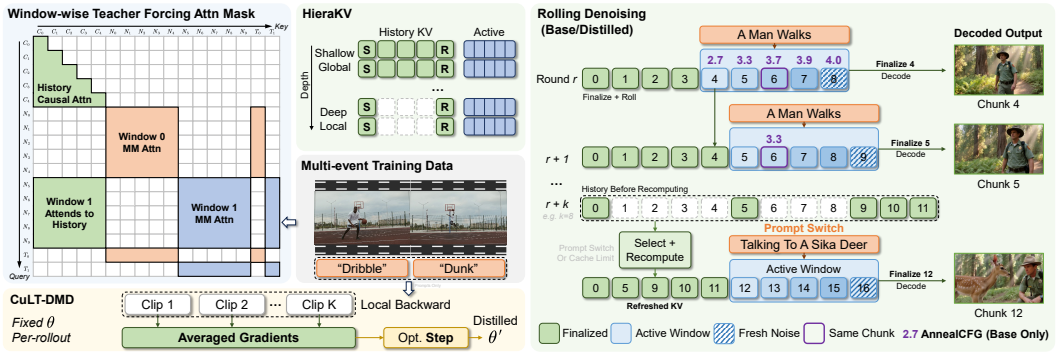


Figure 2: **Overview of MovieRoll.** The base model is trained on multi-event videos with event-specific captions using window-wise teacher forcing with cross-window attention isolation. HieraKV retains full history KV in shallow layers but only the sink chunk (S) and the most recent history chunk (R) in deep layers. CuLT-DMD distills the base model through long self-rollouts, averaging clip-level gradients for one update per rollout. Both models use rolling denoising and recompute KV from selected history chunks when the cache limit is reached or the prompt changes. AnnealCFG progressively reduces CFG during denoising in the base model only.

where $\mathbf{x}_0$ and $\mathbf{x}_t$ are the clean and noisy window latents, $\mathbf{x}_0^{<i}$ is the clean history preceding the first chunk i of the window, and $\mathbf{c}$ is the corresponding prompt. Here, $\boldsymbol{\epsilon}$ is Gaussian noise, t specifies the within-window noise levels, and $\mathbf{v}_t$ is the target velocity.

Long-horizon distillation. During distillation, the autoregressive generator G_θ instead conditions on self-generated histories. Distribution matching distillation (DMD) (Yin et al., 2024b;a) aligns the resulting window distribution with the pretrained teacher. For a generated window $\hat{\mathbf{x}}$ and its noisy version $\hat{\mathbf{x}}_t$, the DMD gradient is

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = -\mathbb{E}_{t,\hat{\mathbf{x}},\hat{\mathbf{x}}_t}\left[\left(s_{\text{real}}(\hat{\mathbf{x}}_t, t) - s_{\text{fake}}(\hat{\mathbf{x}}_t, t)\right)\frac{\partial \hat{\mathbf{x}}}{\partial \theta}\right], \tag{2}$$

where s_{real} and s_{fake} are scores estimated by the frozen bidirectional teacher and a trainable critic, respectively; conditioning arguments are omitted. To extend training beyond short rollouts, **Cumulative Long Tuning (CuLT)** trains on long samples and averages gradients across local clips, stabilizing optimization and improving long-horizon generation.

Inference. MovieRoll generates videos with a rolling denoising window while applying **HieraKV** throughout inference. Shallow blocks retain the global available history, whereas deeper blocks compress intermediate-history KV states, reducing memory consumption and improving long-horizon generalization of the base model. Once the history exceeds a predefined context threshold, representative clean chunks are selected and their KV states are recomputed to refresh the cache. During base-model inference, **AnnealCFG** further anneals classifier-free guidance over denoising timesteps to reduce the train–inference gap and accumulated errors.

3.2 HieraKV: Hierarchical KV-Cache Compression across Layers

Full-history KV caching incurs increasing memory and computation costs as video length grows, whereas uniform sliding-window caching discards intermediate history at every layer. Our inference-time attention analysis of the model trained with window-wise teacher forcing (Fig. 3) reveals a depth-dependent pattern: shallow blocks distribute attention relatively evenly across historical chunks, while deeper blocks concentrate a larger fraction of historical attention on sink and recent chunks, together with a modest decrease in historical-attention entropy.

Motivated by this pattern, **HieraKV** adopts a fixed depth split: the first L_s blocks retain all available history, while the remaining blocks cache only the sink chunk and the most recent historical chunk. The sink chunk provides a visual anchor, while the recent chunk maintains local temporal continuity with the latest finalized content. Intermediate history can still influence deeper computation through representations propagated from the shallow blocks, without storing their KV states across depths.

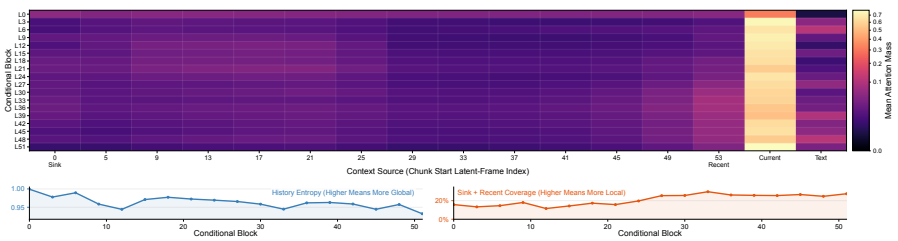


Figure 3: **Hierarchical attention patterns across layers.** We analyze attention in a model trained with window-wise teacher forcing. The heatmap shows mean attention mass across layers and context sources. History attention is more broadly distributed in shallow layers, whereas deeper layers generally exhibit lower history-attention entropy (bottom left) and higher sink-plus-recent coverage (bottom right), indicating a stronger preference for the sink and most recent history chunks.

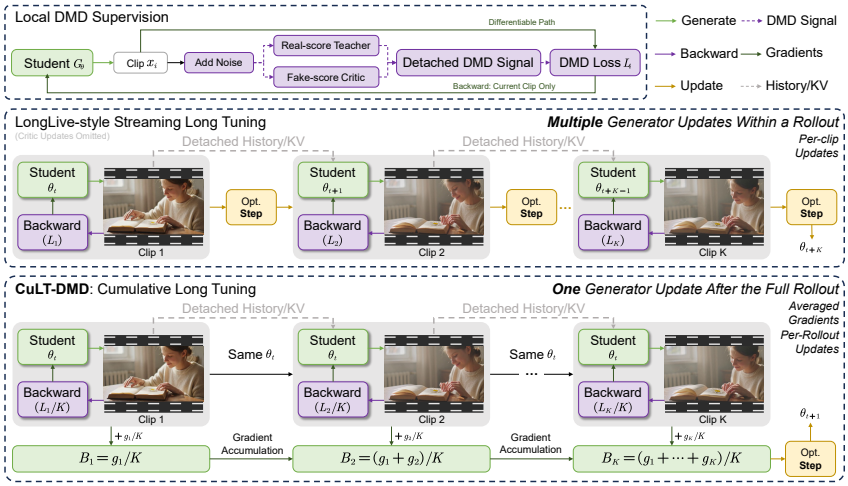


Figure 4: **Cumulative Long Tuning DMD (CuLT-DMD).** LongLive-style streaming long tuning updates the generator after applying local DMD supervision to each clip. In contrast, CuLT-DMD keeps generator parameters fixed throughout a rollout, averages clip-level gradients across all K clips, and performs a single generator update after the full rollout. Both strategies use detached history/KV, with backpropagation restricted to the current clip.

This design reduces KV-cache storage while preserving a pathway for long-range information. HieraKV requires no content-dependent retrieval and compresses only historical KV states, leaving bidirectional attention within the active denoising window unchanged.

3.3 CuLT-DMD: Cumulative Long Tuning with DMD

Short-rollout distillation provides limited exposure to long-horizon self-generated histories. LongLive (Yang et al., 2025a) extends training to long rollouts but alternates generator and critic updates multiple times within each rollout, making each update dependent on the current rollout position. In contrast, **CuLT-DMD** averages gradients across local clips within a long rollout to obtain a rollout-level gradient estimate, improving long-horizon generation.

At generator update r, we process K local windows with fixed parameters $\theta = \theta_r$. For window k, the generator predicts clean latents, concatenated with detached history and re-noised:

$$\hat{\mathbf{x}}_0^{(k)} = G_\theta\Big(\mathrm{sg}(\mathbf{x}_{\boldsymbol{\tau}}^{(k)}), \boldsymbol{\tau}; \mathrm{sg}(\hat{\mathbf{h}}^{(k)})\Big), \quad \hat{\mathbf{z}}_0^{(k)} = [\mathrm{sg}(\hat{\mathbf{h}}^{(k)}), \hat{\mathbf{x}}_0^{(k)}], \quad \hat{\mathbf{z}}_t^{(k)} = \alpha_t\, \mathrm{sg}(\hat{\mathbf{z}}_0^{(k)}) + \sigma_t \boldsymbol{\epsilon}. \quad (3)$$

Here, $\hat{\mathbf{h}}^{(k)}$ denotes the clean latents of previously generated chunks retained as history for window k, sg denotes stop-gradient, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The generator timesteps $\boldsymbol{\tau}$ differ from the DMD timestep t; prompt conditioning is omitted.

The frozen teacher and trainable critic process the full noisy clip $(\hat{\mathbf{z}}_t^{(k)}, t)$, yielding clean-latent predictions D_{real}^{γ} with CFG and D_{fake} with conditional-only evaluation, respectively. Let $a_k = \max(\text{mean}\,|\hat{\mathbf{z}}_0^{(k)} - D_{\text{real}}^{\gamma}|, 10^{-4})$, with the mean taken over all latent entries of each clip. A binary mask $\mathbf{m}^{(k)}$ selects supervised content-latent entries in the current window and excludes history and other unsupervised positions, with $N_k = \sum_j m_j^{(k)} > 0$. The detached DMD signal and masked surrogate loss are

$$\mathbf{d}_k = \text{sg}\left[\frac{D_{\text{fake}} - D_{\text{real}}^{\gamma}}{a_k}\right], \quad \mathcal{L}_{\text{DMD}}^{(k)} = \frac{1}{2N_k}\left\|\mathbf{m}^{(k)} \odot \left[\hat{\mathbf{z}}_0^{(k)} - \text{sg}\Big(\hat{\mathbf{z}}_0^{(k)} - \mathbf{d}_k\Big)\right]\right\|_2^2, \tag{4}$$

where $\odot$ denotes element-wise multiplication. History informs the DMD signal, but only supervised predictions in the current window receive gradients.

We backpropagate each $\mathcal{L}_{\text{DMD}}^{(k)}/K$ and release its graph while retaining accumulated gradients and detached history/KV states. After all K windows, we update the generator once:

$$\mathbf{g}_{\text{roll}} = \frac{1}{K}\sum_{k=1}^{K} \nabla_\theta \mathcal{L}_{\text{DMD}}^{(k)}\Big|_{\theta=\theta_r}, \quad \theta_{r+1} = \text{OptStep}(\theta_r, \mathbf{g}_{\text{roll}}). \tag{5}$$

This averages locally truncated gradients across temporal positions under a single generator snapshot. With fixed local-clip and retained-history/KV budgets, peak GPU memory remains bounded as rollout length increases.

3.4 AnnealCFG: Mitigating Long-Horizon Degradation with CFG Annealing

Classifier-free guidance (CFG) (Ho & Salimans, 2022) combines conditional and unconditional velocity predictions as

$$v_{\text{cfg}} = v_{\text{uncond}} + w(t)\,(v_{\text{cond}} - v_{\text{uncond}})\,. \tag{6}$$

For $w(t) > 1$, CFG extrapolates beyond the conditional prediction, improving visual quality and semantic adherence but potentially introducing a train–inference mismatch (Saharia et al., 2022; Patel et al., 2023; Chung et al., 2025). Unlike bidirectional models that jointly denoise all frames, streaming models immediately reuse each finalized chunk as context for subsequent generation. Guidance-induced deviations can therefore propagate through self-generated histories, amplifying this mismatch and accumulating errors over long rollouts.

To mitigate this effect, **AnnealCFG** maintains strong guidance at high noise levels and progressively attenuates it as denoising approaches the clean endpoint. For the flow-matching timestep $t \in [0, 1000]$, we define

$$w(t) = 1 + (w_{\max} - 1)\frac{t}{1000}, \tag{7}$$

where $w_{\max} > 1$ denotes the maximum guidance scale. As denoising proceeds from $t = 1000$ to $t = 0$, the guidance scale decreases linearly from $w_{\max}$ to 1, recovering the conditional prediction at the clean endpoint. This schedule retains strong guidance during early denoising while reducing guidance-induced extrapolation at later timesteps, immediately before generated chunks are incorporated into the history. AnnealCFG is applied only to MovieRoll-Base to improve robustness to degradation accumulated over long-horizon streaming generation, whereas MovieRoll-Distilled performs conditional-only inference without CFG.

4 Experiment

4.1 Implementation Details

MovieRoll-Base is initialized from the open-source HunyuanVideo-1.5-480P-T2V (Wu et al., 2025) checkpoint. The model uses 6 global-history blocks and 48 local-history blocks. Each chunk contains 4 latent frames, giving the 5-chunk rolling window a steady-state length of 20 latent frames. We perturb clean histories with probability $p = 0.9$, sampling the noise level as $\sigma \sim \mathcal{U}(0, 0.15)$. MovieRoll-Distilled is initialized via ODE regression and subsequently trained with rolling DMD on single- and two-event prompts. We use Muon (Liu et al., 2025) optimizer across all training

Table 1: **Quantitative comparison of single-event and multi-event generation on 30-second videos (721 frames).** MovieRoll-Distilled achieves state-of-the-art performance in both settings, while MovieRoll-Base also delivers strong results.

Method	Aesthetic ↑	Dynamic ↑	Smoothness ↑	Semantic ↑	Naturalness ↑	Drifting Aesthetic ↑	Drifting Smoothness ↑	Drifting Semantic ↑	Drifting Naturalness ↑	Total Score ↑
Single-event evaluation										
MemFlow	8	5	10	5	6	8	10	10	5	6.762
LongLive	8	3	10	5	5	8	10	10	5	6.447
Echo-Infinity	8	5	9	5	4	5	10	8	2	5.430
Helios-Base	8	4	10	5	8	7	10	9	2	6.747
Helios-Distilled	8	6	9	5	4	4	10	7	1	5.163
MovieRoll-Base	8	4	10	5	9	6	10	9	2	6.903
MovieRoll-Distilled	8	6	10	4	10	6	10	8	6	7.260
Multi-event evaluation										
MemFlow	8	3	10	3	7	6	10	9	1	5.754
LongLive	9	1	10	3	5	6	10	9	1	5.214
Echo-Infinity	8	3	9	3	7	5	10	9	1	5.625
Helios-Base	8	1	10	2	10	7	10	9	2	6.402
Helios-Distilled	8	3	10	2	7	5	10	9	1	5.400
MovieRoll-Base	8	2	10	2	10	6	10	9	1	6.234
MovieRoll-Distilled	8	4	10	3	10	5	10	8	8	7.044

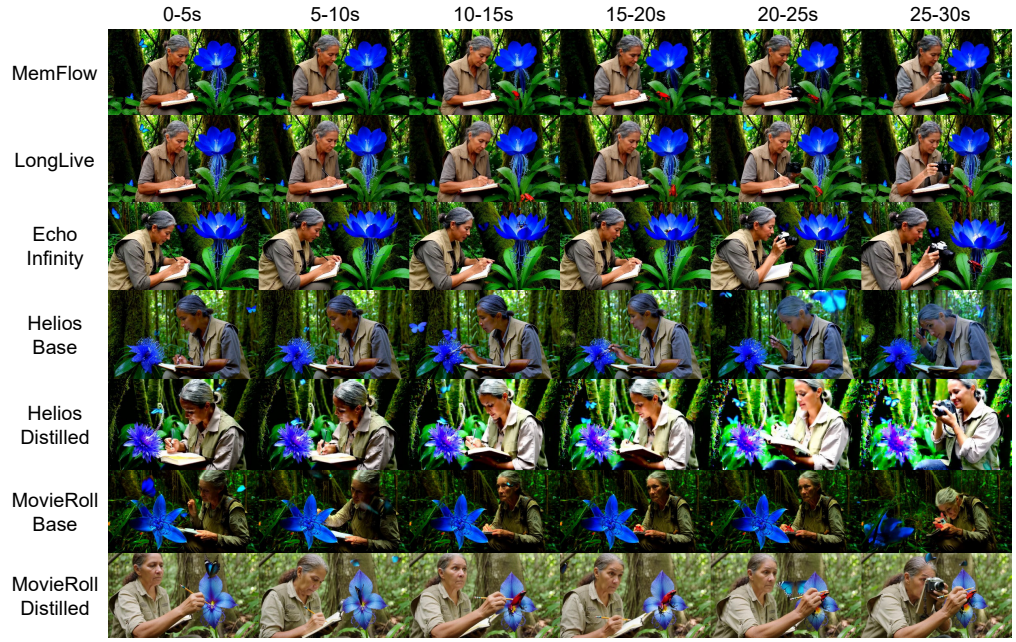


Figure 5: **Qualitative comparison of multi-event video generation.** We compare 30s videos generated from three sequential prompts, one per 10s segment. MovieRoll maintains coherent scene composition and consistent visual quality across event changes, while some competing methods exhibit noticeable color drift or scene degradation.

stages. At inference, both models trigger KV recomputation after 18 chunks. MovieRoll-Base uses 50 denoising steps with CFG annealing ($4 \rightarrow 1$), whereas MovieRoll-Distilled uses 5 steps without CFG. Both use an Euler solver with a timestep shift of 5. For single-event evaluation, we randomly sample 200 prompts from the MovieGen (Polyak et al., 2024) and VBench (Huang et al., 2024; Zheng et al., 2025a) category prompt sets. Multi-event evaluation uses 100 prompts from MemFlow (Ji et al., 2025). All prompts are rewritten using the streaming-video-prompt-rewrite skill. We report all HeliosBench metrics (Yuan et al., 2026). Additional details are provided in the appendix.

Table 2: **Ablation results.** Ablation studies show that HieraKV, AnnealCFG, and CuLT-DMD each improve the generation quality of MovieRoll. When ablating each component, all other settings are kept fixed. All metrics are evaluated on 721-frame single-event videos.

Model Variants	Aesthetic ↑	Dynamic ↑	Smoothness ↑	Semantic ↑	Naturalness ↑	Drifting Aesthetic ↑	Drifting Smoothness ↑	Drifting Semantic ↑	Drifting Naturalness ↑	Total Score ↑
HieraKV										
Global 54 (All Global)	7	4	10	5	8	5	10	8	1	6.321
Global 18 + Local 36	7	3	10	5	9	6	10	8	1	6.645
Global 12 + Local 42	7	4	10	5	9	6	10	9	1	6.774
Global 6 + Local 48	7	4	10	5	9	6	10	9	2	6.873
AnnealCFG										
Constant CFG = 6	7	5	9	5	5	3	8	8	1	5.160
Constant CFG = 4	7	4	9	5	8	6	10	9	1	6.489
Linear decay: $6 \rightarrow 1$	7	5	9	5	7	5	10	9	2	6.264
Linear decay: $4 \rightarrow 1$	7	4	10	5	9	6	10	9	2	6.873
CuLT-DMD										
LongLive-style Streaming Long Tuning	8	5	10	4	10	6	10	8	5	7.131
Cumulative Long Tuning Only On 241 Frames	8	6	10	5	10	5	10	8	4	7.218
Cumulative Long Tuning	8	6	10	4	10	6	10	8	6	7.260

4.2 Main Comparison

Quantitative Results. We compare MovieRoll with MemFlow (Ji et al., 2025), LongLive (Yang et al., 2025a), Helios-Base and Helios-Distilled (Yuan et al., 2026), and Echo-Infinity (Bian et al., 2026). As shown in Tab. 1, MovieRoll achieves the highest overall score in both settings, with particularly strong Naturalness and competitive motion dynamics. This indicates that MovieRoll generates videos that maintain sufficient motion amplitude while remaining closer to real videos, resulting in the best overall perceptual quality.

Qualitative Results. Fig. 5 presents a qualitative comparison of MovieRoll with representative methods. MemFlow, LongLive, and Echo-Infinity exhibit limited motion. Helios-Base maintains good visual quality but shows weak prompt alignment, whereas Helios-Distilled suffers from severe background degradation. In contrast, MovieRoll-Base produces more balanced compositions, while MovieRoll-Distilled generates the most realistic videos among the compared methods, maintaining sufficient motion and responding effectively to prompt switches.

4.3 Ablation Study

HieraKV. We investigate the effect of global–local block allocation in HieraKV on generation quality using the Base model. To ensure a controlled comparison, all configurations use the same training data and history-recomputation threshold. Fig. 6 reports the memory consumption and inference time under different numbers of global-history blocks. Using 6 global-history blocks reduces inference time by approximately 53% (a 2.13× speedup) and peak KV-cache memory by approximately 78% relative to the all-global configuration. As reported in Tab. 2, using 6 global-history blocks and 48 local-history blocks achieves the best 30-second performance. These results show that HieraKV provides not only efficiency gains but also improved long-horizon generation performance, suggesting that reducing direct dependence on intermediate history in deeper blocks enhances model robustness.

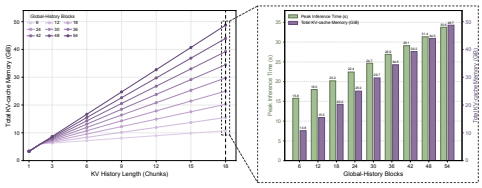


Figure 6: **KV-cache memory and peak inference speed versus global-history blocks.** Reducing the number of global-history blocks substantially lowers KV-cache memory consumption and improves inference speed. All measurements are conducted on a single NVIDIA H20 GPU.

CuLT-DMD. Starting from the same Base model, we train distilled variants with different long-tuning strategies. In the default configuration, Phase 1 is trained on 241 frames, while Phase 2 extends training to 721 frames. As shown in Tab. 2, CuLT-DMD achieves better motion dynamics and drifting metrics than LongLive-style streaming long tuning (Yang et al., 2025a). This suggests

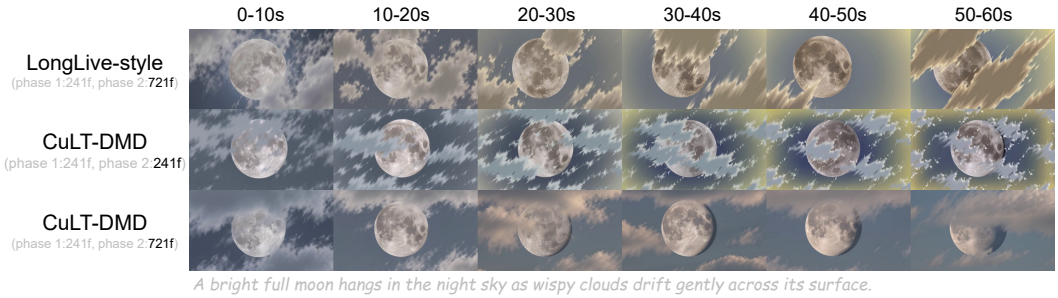


Figure 7: **Generation quality under different distillation strategies.** CuLT-DMD with long-horizon training achieves the best generation quality at extended durations. In contrast, LongLive-style streaming long tuning and CuLT-DMD trained only on short sequences both suffer from noticeable quality degradation during long-horizon inference.

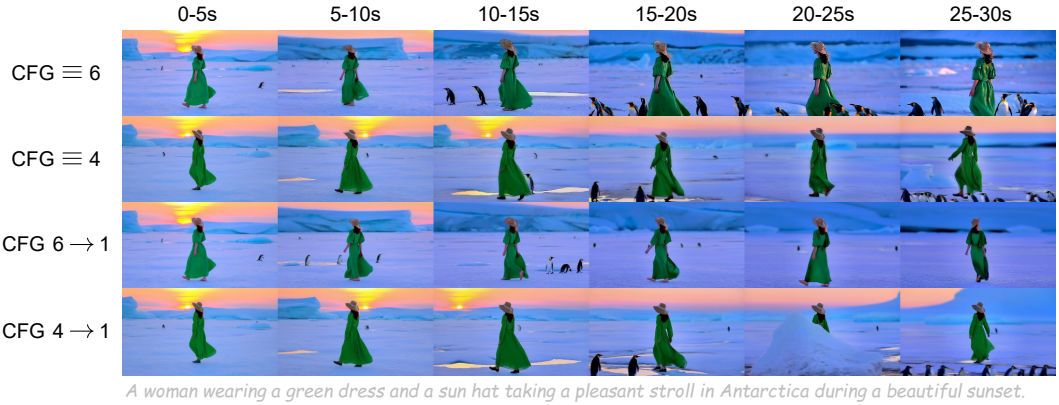


Figure 8: **Generation quality under different CFG settings.** Compared with a constant CFG schedule, AnnealCFG gradually attenuates guidance throughout denoising, significantly reducing long-horizon degradation, such as color shifts and accumulated blur.

that averaging gradients across multiple local clips within the same rollout provides a more representative rollout-level training signal, leading to better long-horizon performance. Fig. 7 further presents qualitative comparisons, where CuLT-DMD trained on long samples better preserves visual naturalness over long horizons.

AnnealCFG. We investigate the effect of CFG strategies on degradation robustness using MovieRoll-Base. As shown in Tab. 2, linearly decreasing CFG outperforms constant CFG, with particularly large improvements on the drifting metrics. Fig. 8 further shows that constant CFG leads to background degradation and visual blurring during long-horizon generation, whereas AnnealCFG substantially alleviates these artifacts, indicating that it effectively suppresses error accumulation during streaming inference.

5 Conclusion

We presented MovieRoll, a base-to-distilled framework for long-horizon streaming video generation. MovieRoll combines Window-wise Teacher Forcing for robust multi-step base training, HieraKV for efficient history management, CuLT-DMD for long-horizon self-rollout distillation, and inference-time recomputation with AnnealCFG for stable extrapolation. Experiments on 30-second HeliosBench show that MovieRoll-Base remains competitive beyond its training horizon, while MovieRoll-Distilled achieves state-of-the-art overall performance in both single- and multi-event evaluations, demonstrating the effectiveness of the proposed framework for high-quality and efficient streaming video generation.

AI USE STATEMENT

We used generative AI tools to assist with code implementation, paper writing, figure-layout brainstorming, literature search, video evaluation, and prompt rewriting. All AI-assisted code was reviewed and verified by at least two co-authors. All AI-assisted text and literature comparisons were manually reviewed and verified against the original content and sources. AI-generated figures were used solely to assist with layout design and logo creation, and all resulting content was reviewed and verified for correctness by at least two authors. The use of AI for video evaluation and prompt rewriting is further described in the Appendix. We take full responsibility for the final content, claims, code, experiments, and artifacts presented in this work.

ETHICS STATEMENT

MovieRoll is developed for research on efficient and long-horizon video generation. As with general-purpose generative video models, it may potentially be misused to create misleading or harmful synthetic content and may inherit biases from the data and pretrained models on which it builds. We encourage responsible use of the proposed techniques together with appropriate safeguards for synthetic-media generation.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed descriptions of Window-wise Teacher Forcing, HieraKV, CuLT-DMD, AnnealCFG, and the inference-time recomputation strategy in the main paper. The Appendix further reports implementation details, training and inference configurations, hyperparameters, benchmark and evaluation protocols, prompt rewriting and video evaluation settings, and additional ablations.

REFERENCES

Yuxuan Bian, Zeyue Xue, Songchun Zhang, Shiyi Zhang, Weiyang Jin, Yaowei Li, Junhao Zhuang, Haoran Li, Jie Huang, Haoyang Huang, et al. Echo-infinity: Learning evolving memory for real-time infinite video generation. *arXiv preprint arXiv:2606.04527*, 2026.

Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan Yuille, Leonidas Guibas, et al. Mixture of contexts for long video generation. In *International Conference on Learning Representations*, volume 2026, pp. 92075–92093, 2026.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. *arXiv preprint arXiv:2602.06028*, 2026.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 30824–30850, 2025.

Yunyang Ge, Xinhua Cheng, Chengshu Zhao, Xianyi He, Shenghai Yuan, Bin Lin, Bin Zhu, and Li Yuan. Flashi2v: Fourier-guided latent shifting prevents conditional image leakage in image-to-video generation. *arXiv preprint arXiv:2509.25187*, 2025.

Yunyang Ge, Xianyi He, Zezhong Zhang, Bin Lin, Bin Zhu, Xinhua Cheng, and Li Yuan. OSP-Next: Efficient high-quality video generation with sparse sequence parallelism, hif8 quantization, and reinforcement learning. *arXiv preprint arXiv:2605.28691*, 2026.

Yuwei Guo, Ceyuan Yang, Hao He, Yang Zhao, Meng Wei, Zhenheng Yang, Weilin Huang, and Dahua Lin. End-to-end training for autoregressive video diffusion via self-resampling. In *European Conference on Computer Vision*, pp. 324–344. Springer, 2026.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. LTX-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Lianghua Huang, Zhi-Fan Wu, Yupeng Shi, Wei Wang, Mengyang Feng, Junjie He, Chen-Wei Xie, Yu Liu, Jingren Zhou, Ang Wang, et al. Wan-streamer v0. 2: Higher resolution, same latency. *arXiv preprint arXiv:2607.04443*, 2026a.

Lianghua Huang, Zhi-Fan Wu, Wei Wang, Yupeng Shi, Mengyang Feng, Junjie He, Chen-Wei Xie, Yu Liu, Jingren Zhou, Ang Wang, et al. Wan-streamer v0. 1: End-to-end real-time interactive foundation models. *arXiv preprint arXiv:2606.25041*, 2026b.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026c.

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

Team HunyuanWorld. Hy-world 1.5: A systematic framework for interactive world modeling with real-time latency and geometric consistency. *arXiv preprint*, 2025.

Sihui Ji, Xi Chen, Shuai Yang, Xin Tao, Pengfei Wan, and Hengshuang Zhao. Memflow: Flowing adaptive memory for consistent and efficient long video narratives. *arXiv preprint arXiv:2512.14699*, 2025.

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

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-Sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations*, volume 2026, pp. 91177–91196, 2026.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

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

Niket Patel, Luis Salamanca, and Luis Barba. Bridging the gap: Addressing discrepancies in diffusion model training for classifier-free guidance. *arXiv preprint arXiv:2311.00938*, 2023.

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

David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling diffusion models. *arXiv preprint arXiv:2402.09470*, 2024.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world model. *arXiv preprint*, 2025.

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

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

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. HunyuanVideo-1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.

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

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025b.

Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

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

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

Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12978–12988. IEEE, 2025.

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

Jintao Zhang, Kai Jiang, Jintao Chen, Xu Wang, Yang Luo, Yuji Wang, Dechuang Chen, Jungang Li, Chengyang Ye, Marco Chen, et al. Vidu s1: A real-time interactive video generation model. *arXiv preprint arXiv:2607.03118*, 2026a.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. *Advances in Neural Information Processing Systems*, 38:30546–30566, 2026b.

Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, et al. Matrix-game: Interactive world foundation model. *arXiv preprint arXiv:2506.18701*, 2025.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025a.

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

Zangwei Zheng, Xiangyu Peng, Yuxuan Lou, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, et al. Open-Sora 2.0: Training a commercial-level video generation model in $200 k. *arXiv preprint arXiv:2503.09642*, 2025b.

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

MOVIEROLL: ROLLING OUT LONG-FORM VIDEOS WITH HIERARCHICAL KV CACHING AND CUMULATIVE LONG TUNING
APPENDIX

A IMPLEMENTATION DETAILS OF MAIN EXPERIMENTS

In Sec. 4.1, we briefly describe the training and inference settings of MovieRoll-Base and MovieRoll-Distilled. In this section, we provide their detailed configurations.

Tab. 3, Tab. 4 and Tab. 5 list the model configuration and all training and inference hyperparameters. Below we briefly describe the parts that the tables do not explain.

Training data. MovieRoll is trained on an internal video corpus that cannot be publicly released due to confidentiality restrictions. We instead provide detailed stage-wise statistics of the training data, including data scale, sample duration, resolution, and task composition, in Tab. 5. Model weights and code will be released after the review process.

Streaming architecture. The clean history is stored in a KV cache that is read by all 54 blocks. The first 6 blocks attend to the full (compressed) history, while the other 48 blocks attend only to the first chunk, which serves as an attention sink, and to the most recent chunk. Temporal RoPE positions do not grow with the video length: in both groups of blocks, the cached chunks and the current window are always indexed contiguously from zero.

Base model training. Within a rolling window, the chunks receive increasing timesteps: $[0, 1]$ is split at sorted random points, and every chunk takes the same logit-normal relative position within its sub-interval (with a small probability, all chunks share one timestep). The clean history is perturbed with noise to resemble self-generated history at test time. In Stage 2, each chunk of a multi-event clip is conditioned on the caption of the event that it overlaps.

Distillation. During ODE initialization, we run the Stage 2 base model with AnnealCFG: the 50-step teacher uses a CFG scale linearly annealed from 4 to 1. The 5-step student regresses the clean endpoints of the teacher's rolling trajectories, thereby distilling both the accelerated sampling trajectory and the annealed CFG behavior. In rolling DMD, the generator rolls out the whole video with its 5-step schedule. Each window's prediction is concatenated with the preceding frames to form a score window, and the DMD loss is applied only to the current window. To regularize predictions at different noise levels (Liu et al., 2026), with probability 0.5, we re-noise the window to a shared timestep and pass it through the generator again before computing the loss.

Inference. When the KV cache reaches the recompute threshold, it is rebuilt from a few selected chunks (Tab. 4), which bounds memory usage and keeps positions within the RoPE range. For multi-event prompts, the cache is additionally rebuilt at each event switch. Intermediate chunks are retained for only the five most recent events, with those from earlier events discarded.

Ablation settings. For the HieraKV ablation, we train all block-allocation variants on the same data for 10K steps, using the same learning rate, weight decay, and other training settings as Stage 1 of MovieRoll-Base. At inference, all settings follow the main experiments except that the KV recomputation threshold is set to 15 chunks. For the AnnealCFG ablation, we use the MovieRoll-Base checkpoint obtained after Stage 1 and evaluate different CFG schedules, again with a recomputation threshold of 15 chunks while keeping all other inference settings unchanged.

For the CuLT-DMD ablation, we compare three distillation strategies starting from the same Base model. The default CuLT-DMD configuration and the LongLive-style baseline both use 241-frame rollouts in Phase 1 and 721-frame rollouts in Phase 2. The short-rollout variant instead uses 241-frame rollouts in both phases. Our LongLive-style baseline adapts the shared-rollout training schedule of LongLive (Yang et al., 2025a) to our architecture. The generator and critic process different supervision windows in temporal order along a shared rollout, retaining detached history and KV states across optimizer updates. Once the rollout is exhausted, a new prompt is sampled to begin the next rollout. Each phase consists of 600 global steps, with one critic update at every step and one

Table 3: **Model and streaming configuration.**

Backbone	MMDiT, 8.3B parameters, 54 dual-stream blocks
Hidden size / heads	2048 / 16
Patch size	$1\times1\times1$
VAE	causal 3D VAE, 16×16 spatial and $4\times$ temporal compression, 32 channels
Text encoders	Qwen2.5-VL-7B (up to 1,000 tokens), Glyph-ByT5 (up to 256 tokens)
Frame rate	24 fps
Chunk size	4 latent frames (16 frames); the first chunk has 5 latent frames
Rolling window	5 chunks at staggered noise levels
Full-history blocks	1–6
Sink + recent blocks	7–54 (first chunk + most recent chunk)

Table 4: **Inference settings.** Denoising passes are counted per output chunk. Each pass processes the full rolling window, and the conditional and unconditional branches of CFG count as two passes. The KV-encoding pass shared by both models for each new chunk is excluded. During KV recomputation, the retained chunks are re-encoded at t=0 without text conditioning.

	Base model	Distilled model
Resolution / frame rate	480×864 / 24 fps	
Chunk size / window size	4 latent frames / 5 chunks	
Sampling steps	50 (10 per round)	5 (1 per round)
Timesteps	50-step schedule, shift 5	$\{1000, 952, 882, 769, 556\}$
CFG	per-frame, linear $4 \rightarrow 1$	none
Denoising passes per chunk	20	1
KV context noise σ	0.05	0.15
KV recompute threshold	18 chunks ($\approx$12 s)	
Chunks kept after recompute	sink + middle chunk of each event + 3 most recent	

generator update every third step, resulting in 600 critic updates and 200 generator updates. This update schedule is independent of rollout boundaries. On steps involving both networks, the generator and critic process successive supervision windows, and both optimizer updates are deferred until their backward passes are complete. Each generator update in the LongLive-style baseline therefore uses supervision from a single window. In contrast, CuLT-DMD keeps the generator parameters fixed throughout each rollout and accumulates gradients across all K supervision windows before averaging them for a single update, where $K = 3$ for 241-frame rollouts and $K = 9$ for 721-frame rollouts. All distillation variants use the same inference settings as the main experiments.

B HeliosBench Evaluation Protocol

We use the long-video version of HeliosBench, which contains five full-video quality metrics and four long-range drift metrics, as shown in Tab. 7. Full-video metrics evaluate the average quality over the entire video. Drift metrics measure the absolute score difference between the first and last $\rho = 0.15$ fraction of the video (Zhang et al., 2026b):

$$d_m(V) = |q_m(V_{\text{start}}) - q_m(V_{\text{end}})|\,,$$

where q_m is the corresponding base metric. Higher values are preferred for full-video metrics, while lower values are preferred for drift metrics.

For aesthetic evaluation, frames are resized and center-cropped to 224×224 before CLIP encoding. For semantic evaluation, frames are resized to width 640 while preserving the aspect ratio. For naturalness evaluation, the long side is resized to 512 pixels and the VLM assigns an integer score $s \in \{1, \ldots, 5\}$, where larger values indicate more natural videos. The naturalness implementation used here employs three Gemini-2.5-Pro calls per video; the original HeliosBench implementation used one GPT-5.2 call.

Raw-score conversion. Raw scores are converted to ratings $r_m \in \{1, \ldots, 10\}$ using the thresholds in Tab. 8. For higher-is-better metrics, the first satisfied threshold gives ratings 10 through 2

Table 5: **Training hyperparameters.** Frame counts are at 24 fps. Base-stage clips are bucketed in steps of 8 frames and their batch size depends on the clip length (6 clips per data-parallel group at 33 frames, decreasing to 1 clip at $\geq$121 frames). Context noise: noise level added to the clean history (applied with probability p in the base stages). G/F: generator/fake score. P1/P2: DMD phase 1/2.

	Base model		Distillation	
	Stage 1	Stage 2	ODE init.	Rolling DMD
Initialization	Bidirectional T2V	Stage 1 (EMA)	Stage 2 (EMA)	ODE student
Training data	620K single-event clips	738K single-event clips + 560K multi-event clips	1.1K teacher trajectories	5K single-event prompts (P1) 5K two-event prompts (P2)
Frames per sample	33–241	33–361	241	241 (P1) / 721 (P2)
Duration	1.4–10 s	1.4–15 s	10 s	10 s / 30 s
Resolution	multi-aspect, $\approx 640^2$ pixels		480×864, 864×480	480×864, 864×480
Teacher / real score	–	–	Stage 2, 50 steps CFG 4 $\rightarrow$ 1	Bidirectional T2V CFG 4
Text condition	single event	multi events (0.9) single event (0.1)	single event	single event (P1) / two events (P2)
Task (T2V/I2V)	0.8 / 0.2	0.8 / 0.2	T2V only	0.8 / 0.2
Context noise	$\mathcal{U}(0, 0.15)$ per chunk, p=0.9		$\mathcal{U}(0, 0.15)$ per sample	0.15 (fixed)
Null-text dropout	0.05 per sample + 0.05 per window		–	–
Objective	flow matching		ODE regression	DMD
Timestep shift	3	3	5	5
Optimizer	Muon	Muon	Muon	Muon (G & F)
Learning rate	2×10^{-5}	5×10^{-6}	1×10^{-5}	1×10^{-5} (G & F)
Weight decay	0.01	0.01	0.01	0.01
LR schedule	constant, no warmup			
Gradient clipping	1.0	1.0	1.0	10.0
EMA decay	0.9993	0.9993	–	0.99 (from step 200)
Training steps	10K	2K	1K	600 (P1) + 600 (P2)
Batch size (videos)	32–192	16–96	32	32
Precision	bf16 mixed precision			
GPUs (SP degree)	64 (2)	64 (4)	32 (1)	128 (4)

Table 6: **Rolling DMD settings.** Frame counts are pixel frames at 24 fps; latent frame counts are given in parentheses.

	Phase 1	Phase 2
Frames	241 (61)	721 (181)
Chunks / supervised windows K	15 / 3	45 / 9
Events per video	1	2 (switch at frame 129–593)
Training steps	600	600
Generator schedule	$\{1000, 800, 600, 400, 200\}$, shift 5, 1 step per round	
Fake score initialization	bidirectional T2V model	
Fake-score / generator updates	3 : 1	
Score window	31 latent frames (121 frames)	
Uniform-timestep regularization	probability 0.5; shared $t \in \{952, 882, 769, 556\}$	
KV recompute (training)	at 18 chunks; sink + 1–5 random + 3 most recent	

in order; for lower-is-better metrics, the inequality is reversed. If none of the nine thresholds is satisfied, the rating is set to 1.

The final HeliosBench score is the Total Weighted Rating (TWR):

$$\mathrm{TWR} = \sum_{m=1}^{9} w_m r_m,$$

where w_m is the metric weight in Tab. 7. The weights sum to approximately 0.996, resulting in a theoretical maximum TWR of 9.96.

Multi-event Evaluation. For multi-event videos, semantic and drifting-semantic scores are computed per event using the event-specific prompt. The video-level semantic score is the mean of the segment scores, while drifting semantic is the mean of the absolute beginning–end difference within each segment.

Table 7: **Metrics used in the long-video HeliosBench evaluation.**

Metric	Model	Measurement	Dir./Weight
Aesthetic	CLIP ViT-L/14 + LAION head	All-frame aesthetic score divided by 10 and averaged.	↑ / 0.030
Dynamic	Farneback optical flow	Mean pixel-wise flow magnitude over adjacent frames.	↑ / 0.030
Motion smoothness	AMT-S	One frame is interpolated between each adjacent pair; score is $(255 - \Delta_{\mathrm{VFI}})/255$.	↑ / 0.030
Semantic alignment	ViCLIP (InternVid-10M-FLT)	Cosine similarity between the prompt and eight middle-sampled frames.	↑ / 0.255
Naturalness	Gemini-2.5-Pro	VLM score from 16 uniformly sampled frames; $(s-1)/4$, averaged over three calls.	↑ / 0.255
Drifting aesthetic	Same as aesthetic	Absolute difference between the first and last 15% window scores.	↓ / 0.099
Drifting motion smoothness	Same as motion smoothness	Absolute difference between the first and last 15% window scores.	↓ / 0.099
Drifting semantic	Same as semantic alignment	Absolute difference between the first and last 15% window scores.	↓ / 0.099
Drifting naturalness	Same as naturalness	Absolute difference between the first and last 15% window scores.	↓ / 0.099

Table 8: **Mapping from raw metric scores to HeliosBench ratings.**

Metric	Rate 10	Rate 9	Rate 8	Rate 7	Rate 6	Rate 5	Rate 4	Rate 3	Rate 2	Rate 1
Aesthetic (↑)	$[0.70, \infty)$	$[0.65, 0.70)$	$[0.60, 0.65)$	$[0.55, 0.60)$	$[0.50, 0.55)$	$[0.45, 0.50)$	$[0.40, 0.45)$	$[0.35, 0.40)$	$[0.30, 0.35)$	$(-\infty, 0.30)$
Dynamic (↑)	$[0.45, \infty)$	$[0.40, 0.45)$	$[0.35, 0.40)$	$[0.30, 0.35)$	$[0.25, 0.30)$	$[0.20, 0.25)$	$[0.15, 0.20)$	$[0.10, 0.15)$	$[0.05, 0.10)$	$(-\infty, 0.05)$
Motion smoothness (↑)	$[0.99, \infty)$	$[0.98, 0.99)$	$[0.97, 0.98)$	$[0.96, 0.97)$	$[0.95, 0.96)$	$[0.94, 0.95)$	$[0.93, 0.94)$	$[0.92, 0.93)$	$[0.91, 0.92)$	$(-\infty, 0.91)$
Semantic alignment (↑)	$[0.30, \infty)$	$[0.29, 0.30)$	$[0.28, 0.29)$	$[0.27, 0.28)$	$[0.26, 0.27)$	$[0.25, 0.26)$	$[0.24, 0.25)$	$[0.23, 0.24)$	$[0.22, 0.23)$	$(-\infty, 0.22)$
Naturalness (↑)	$[0.65, \infty)$	$[0.60, 0.65)$	$[0.55, 0.60)$	$[0.50, 0.55)$	$[0.45, 0.50)$	$[0.40, 0.45)$	$[0.30, 0.40)$	$[0.25, 0.30)$	$[0.20, 0.25)$	$(-\infty, 0.20)$
Drifting aesthetic (↓)	$(-\infty, 0.01]$	$(0.01, 0.02]$	$(0.02, 0.03]$	$(0.03, 0.04]$	$(0.04, 0.05]$	$(0.05, 0.06]$	$(0.06, 0.07]$	$(0.07, 0.08]$	$(0.08, 0.09]$	$(0.09, \infty)$
Drifting motion smoothness (↓)	$(-\infty, 0.01]$	$(0.01, 0.02]$	$(0.02, 0.03]$	$(0.03, 0.04]$	$(0.04, 0.05]$	$(0.05, 0.06]$	$(0.06, 0.07]$	$(0.07, 0.08]$	$(0.08, 0.09]$	$(0.09, \infty)$
Drifting semantic (↓)	$(-\infty, 0.01]$	$(0.01, 0.02]$	$(0.02, 0.03]$	$(0.03, 0.04]$	$(0.04, 0.05]$	$(0.05, 0.06]$	$(0.06, 0.07]$	$(0.07, 0.08]$	$(0.08, 0.09]$	$(0.09, \infty)$
Drifting naturalness (↓)	$(-\infty, 0.06]$	$(0.06, 0.08]$	$(0.08, 0.10]$	$(0.10, 0.12]$	$(0.12, 0.14]$	$(0.14, 0.16]$	$(0.16, 0.18]$	$(0.18, 0.20]$	$(0.20, 0.22]$	$(0.22, \infty)$

Choice of the Naturalness Evaluator. We compare GPT-5.2 and Gemini-2.5-Pro on the same 2,400 videos from the 30-second (721-frame) single-event benchmark, covering 12 methods with 200 videos per method. For this comparison, both evaluators score each video three times, and the resulting scores are averaged.

As shown in Tab. 9, the two evaluators exhibit strong agreement in method-level naturalness rankings, with a Spearman correlation of $\rho = 0.930$. MovieRoll-Distilled and the evaluated MovieRoll-Base checkpoint rank first and second, respectively, under both evaluators. Thus, their leading naturalness rankings in this setting are preserved when the evaluator is changed.

Across all statistical measures of method-level separation in Tab. 9, Gemini-2.5-Pro yields greater between-method score separation in this comparison. We therefore adopt Gemini-2.5-Pro as the naturalness evaluator.

C Choice of Extrapolation Strategy

Because training videos are restricted to a fixed maximum duration, e.g., 10s, generating longer videos at inference time depends on the effectiveness of the extrapolation strategy. Previous methods commonly use a sliding-window strategy, whereas MovieRoll performs recomputation when the cached history reaches a predefined threshold. Using the same model weights trained on videos with a maximum duration of 10s, we compare three extrapolation strategies: retaining the full KV history, using a sliding window that preserves one sink chunk and the 14 most recent chunks, and applying threshold-triggered recomputation with a threshold of 15 chunks. Once recomputation is triggered, we retain the sink chunk, the three most recent chunks, and the chunk at the midpoint of the remaining history. Their quantitative results are reported in Tab. 10. The results show that

Table 9: **Comparison of naturalness evaluators on the same 30-second (721-frame) single-event videos.** Each method contains 200 videos, and both evaluators average three calls per video. The method-level rankings show strong agreement (Spearman's $\rho = 0.930$).

	GPT-5.2		Gemini-2.5-Pro	
Method	**Score**	**Rank**	**Score**	**Rank**
CausVid	0.1446	11	0.0208	12
Self Forcing	0.1596	10	0.0392	11
Causal Forcing	0.2508	9	0.1421	9
Rolling Forcing	0.3217	8	0.3625	8
MemFlow	0.3325	7	0.4612	4
LongLive	0.3575	5	0.4417	5
Echo Infinity	0.3412	6	0.3667	7
Krea	0.1117	12	0.0617	10
Helios-Base	0.4271	3	0.5592	3
Helios-Distilled	0.4179	4	0.3862	6
MovieRoll-Base	0.4825	2	0.6096	2
MovieRoll-Distilled	**0.6096**	**1**	**0.7979**	**1**
Method-level score separation				
Mean range	0.498		**0.777**	
Mean std	0.141		**0.235**	
η^2	0.202		**0.286**	
F statistic	55.0		**86.8**	

Table 10: **Effects of different extrapolation strategies on video generation quality.**

Extrapolation Strategy	**Aesthetic** ↑	**Dynamic** ↑	**Smoothness** ↑	**Semantic** ↑	**Naturalness** ↑	**Drifting Aesthetic** ↑	**Drifting Smoothness** ↑	**Drifting Semantic** ↑	**Drifting Naturalness** ↑	**Total Score** ↑
Full History	7	4	10	5	8	6	10	9	1	6.519
Sliding Window	7	4	10	5	8	6	10	9	1	6.519
KV Recomputing	7	4	10	5	9	6	10	9	2	6.873

Table 11: **Effects of different autoregressive strategies on video generation quality.**

Autoregressive Strategy	**Aesthetic** ↑	**Dynamic** ↑	**Smoothness** ↑	**Semantic** ↑	**Naturalness** ↑	**Drifting Aesthetic** ↑	**Drifting Smoothness** ↑	**Drifting Semantic** ↑	**Drifting Naturalness** ↑	**Total Score** ↑
241-frame evaluation										
Chunk-wise Teacher Forcing	7	4	10	5	9	6	10	9	1	6.774
Chunk-wise Teacher Forcing with Self-resampling	7	6	10	5	10	5	10	8	1	6.891
Window-wise Teacher Forcing	8	5	10	5	10	7	10	9	2	7.287
721-frame evaluation										
Chunk-wise Teacher Forcing	7	3	9	4	6	2	10	6	1	5.001
Chunk-wise Teacher Forcing with Self-resampling	6	5	10	4	5	1	10	3	1	4.410
Window-wise Teacher Forcing	7	4	10	5	8	5	10	8	1	6.321

threshold-triggered recomputation achieves better performance beyond the training duration. We therefore adopt it as the extrapolation strategy.

D Choice of Autoregressive Training Strategy

MovieRoll adopts Window-wise Teacher Forcing to train its autoregressive base model. We compare it with Chunk-wise Teacher Forcing, training both variants on the same data for 10,000 steps. A third variant is initialized from the chunk-wise checkpoint and further trained with self-resampling (Guo et al., 2026) for 10,000 steps. To isolate the effect of the training strategy, all variants use full KV caches in every block without HieraKV and generate 30-second videos with recomputation-based extrapolation. As shown in Tab. 11, all three strategies achieve comparable quality at 241 frames. However, at 721 frames, the chunk-wise variants degrade substantially, whereas Window-wise Teacher Forcing maintains consistently high quality. We therefore adopt Window-wise Teacher Forcing for MovieRoll.

E PROMPT REWRITING FOR STREAMING VIDEO GENERATION

Streaming video generation requires producing temporally coherent and uninterrupted visual content, which imposes specific constraints on user prompts. For example, a prompt describing a zoom-in shot is unsuitable for long-duration streaming generation because the specified camera motion cannot be sustained indefinitely. We therefore use the following system prompt to rewrite the raw prompt provided by the user, making it better suited for streaming video generation.

Skill: `streaming-video-prompt-rewrite`

Description. Rewrites short user inputs into structured, single-paragraph prompts for streaming (indefinitely extended) video generation, in single-event and interactive multi-event settings.
Goal. Expand a short user input into an objective, detailed prompt that describes a scene which can be extended chunk by chunk indefinitely, without reaching a dead end.
Modes.

- *Single:* one input sentence → one prompt.
- *Sequential (interactive):* given the finalized rewrites of all previous events and one new raw event, rewrite only the new event. Future events are never visible.
- *Batch:* N known events → N prompts, separated by `@@@EVENT_BREAK@@@`.

Hard constraints.

1. *Camera.* Exactly one base behavior, stated once in a canonical sentence: static ("The camera remains static, keeping *[subject]* centered.") or tracking ("The camera tracks *[subject]*, continuously keeping *[subject]* centered."). Use tracking when the subject moves through space and static otherwise; convert any other movement (pan, zoom, dolly, tilt, orbit, crane, roll) to the closer of the two. A static camera never reframes.
2. *Centering.* The main subject stays at the frame center throughout; close-ups only for stationary subjects.
3. *Open-ended action.* Describe ongoing or repeating motion. Never use ending words ("finally," "eventually," "in the end," "at last," "wraps up"); after the opening beats, switch to continuity framing ("Throughout," "Meanwhile").
4. *Natural pace.* Real-world speed unless the input specifies otherwise; no stacked intensifiers; summarize crowds instead of animating many small elements. Under a static camera the subject must still visibly move (breathing, swaying, gesturing).
5. *Fidelity.* Keep every entity, exact count, event order, requested style, and on-screen text; never invent or drop subjects.
6. *Style.* Default to cinematic realism, also for fantastical subjects, unless the input names another style. End with a single short style tag, e.g., "Cinematic photography realistic style."
7. *Format.* Output only the prompt as a single paragraph in the present tense, without self-referential phrases (e.g., "this video shows"). Use the same language as the input; if the language is unclear or unsupported, default to English. Keep the output to 150–300 words, or an equivalent length for languages where word counts are not directly comparable.

Prompt structure. Summary sentence naming all subjects → subject appearance → chronological action (*Initially / Then / Next → Throughout*) → environment, where every sound has a visible source → restatement of the subject's fixed traits → camera sentence, shot size, lighting, mood → style tag.
Multi-event consistency.

- One camera setup (behavior, viewpoint, angle, shot size) for the whole sequence; new subjects enter the established frame; switch between static and tracking only when an event requires it.
- Recurring entities keep identical descriptions; the setting stays the same unless the input moves it.
- No cross-event pronouns: each prompt reintroduces entities by their full noun phrase ("It glides toward the diver." → "The giant manta ray glides toward the diver in the black wetsuit.").
- Each event continues from the previous state and receives the full length budget.

Validation. Every output is checked automatically for length, single paragraph, ending words, self-reference, forbidden camera motion, exactly one camera sentence, and explicit centering under tracking. Failed outputs are regenerated with the violations appended as feedback, for up to 3 attempts.

Table 12: **Inference efficiency comparison on 60-second videos with 1441 frames.** We report both end-to-end (E2E) runtime, which includes the complete generation pipeline, and denoising-only runtime, which excludes VAE decoding, text encoding, and other auxiliary computation. FPS is computed from the corresponding runtime. All measurements are conducted on a single NVIDIA H20 GPU.

Method	Params	Resolution ($H \times W$)	Inference Time (s)		FPS ↑	
			E2E	Denoising	E2E	Denoising
CausVid	1.3B	480×832	232	131	6.21	11.00
Self Forcing	1.3B	480×832	455	361	3.17	3.99
Causal Forcing	1.3B	480×832	429	336	3.36	4.29
Rolling Forcing	1.3B	480×832	500	407	2.88	3.54
MemFlow	1.3B	480×832	338	253	4.26	5.70
LongLive	1.3B	480×832	264	175	5.46	8.23
Echo-Infinity	1.3B	480×832	304	210	4.74	6.86
Krea	14B	480×832	1270	1161	1.13	1.24
Helios-Distilled	14B	384×640	470	407	3.07	3.54
MovieRoll-Distilled	**8.3B**	480×864	974	787	1.48	1.83

Table 13: **Single-event evaluations across different video durations.** MovieRoll-Distilled maintains the best overall performance under all evaluation settings.

Method	Aesthetic ↑	Dynamic ↑	Smoothness ↑	Semantic ↑	Naturalness ↑	Drifting Aesthetic ↑	Drifting Smoothness ↑	Drifting Semantic ↑	Drifting Naturalness ↑	Total Score ↑
241-frame evaluation										
CausVid	8	4	9	4	1	7	10	8	2	4.578
Self Forcing	8	4	10	5	4	7	10	9	1	5.628
Causal Forcing	8	10	9	5	4	6	10	8	1	5.580
Rolling Forcing	8	2	10	5	7	8	10	10	2	6.630
MemFlow	8	4	10	5	8	8	10	10	4	7.143
LongLive	8	3	10	5	7	8	10	9	4	6.759
Echo-Infinity	8	5	9	5	6	7	10	9	2	6.237
Krea	8	10	9	6	1	4	10	7	1	4.773
Helios-Base	8	4	10	5	8	8	10	9	5	7.143
Helios-Distilled	8	5	9	5	8	8	10	9	4	7.044
MovieRoll-Base	8	5	10	5	10	7	10	9	4	7.485
MovieRoll-Distilled	8	6	10	5	10	6	10	8	6	7.515
721-frame evaluation										
CausVid	8	4	9	4	1	6	10	8	6	4.875
Self Forcing	7	3	9	5	1	1	10	5	1	3.783
Causal Forcing	7	6	9	5	1	5	10	8	1	4.566
Rolling Forcing	8	2	10	5	4	7	10	9	1	5.568
MemFlow	8	5	10	5	6	8	10	10	5	6.762
LongLive	8	3	10	5	5	8	10	10	5	6.447
Echo-Infinity	8	5	9	5	4	5	10	8	2	5.430
Krea	7	10	9	5	1	1	10	5	1	3.993
Helios-Base	8	4	10	5	8	7	10	9	2	6.747
Helios-Distilled	8	6	9	5	4	4	10	7	1	5.163
MovieRoll-Base	8	4	10	5	9	6	10	9	2	6.903
MovieRoll-Distilled	8	6	10	4	10	6	10	8	6	7.260
1441-frame evaluation										
CausVid	8	3	9	3	1	6	10	7	8	4.689
Self Forcing	7	3	9	3	1	1	10	1	1	2.877
Causal Forcing	7	5	9	5	1	3	10	7	1	4.239
Rolling Forcing	8	2	10	5	1	5	10	8	1	4.506
MemFlow	8	4	10	5	4	8	10	10	2	5.925
LongLive	8	3	10	5	4	9	10	10	3	6.093
Echo-Infinity	7	6	9	5	4	5	10	8	3	5.529
Krea	7	10	9	4	1	1	10	4	2	3.738
Helios-Base	8	4	10	6	6	6	10	9	2	6.393
Helios-Distilled	7	6	9	4	3	1	10	5	1	4.128
MovieRoll-Distilled	7	6	10	4	10	6	10	8	4	7.032

F Limitations in Inference Speed

Despite the substantial performance gains of MovieRoll-Distilled over existing streaming methods, the strategies responsible for its long-horizon robustness introduce nontrivial inference overhead. In particular, our ablation studies demonstrate that rolling-window inference, recomputation-based

Table 14: **Multi-event evaluations on 60-second (1441-frame) videos.** MovieRoll-Distilled continues to achieve the best overall performance among the compared methods.

Method	Aesthetic ↑	Dynamic ↑	Smoothness ↑	Semantic ↑	Naturalness ↑	Drifting Aesthetic ↑	Drifting Smoothness ↑	Drifting Semantic ↑	Drifting Naturalness ↑	Total Score ↑
MemFlow	8	2	10	2	4	5	10	9	1	4.605
LongLive	8	1	10	3	4	5	10	9	1	4.830
Echo-Infinity	8	3	9	3	5	3	10	9	1	4.917
Helios-Base	8	2	10	2	9	5	10	9	1	5.880
Helios-Distilled	8	3	10	1	5	1	10	8	1	4.140
MovieRoll-Distilled	8	4	10	2	10	4	10	8	6	6.492

extrapolation, and HieraKV history management substantially mitigate error accumulation and improve generation stability over long horizons. Rolling-window inference requires activating a much larger set of noisy frames at each step than chunk-wise inference, whereas threshold-triggered recomputation and HieraKV history management further increase latency relative to a sliding-window implementation. This overhead is further amplified by MovieRoll's 8.3B MMDiT backbone, which is larger than the 1.3B backbones used by several competing streaming methods. As a result, the MovieRoll-Distilled model, without sparse attention, low-bit quantization, or other inference acceleration techniques, does not yet support real-time generation beyond 16 fps. We report the inference times of MovieRoll-Distilled and competing methods in Tab. 12. Although Helios-Distilled approaches the inference speed of the 1.3B methods, it is restricted to a resolution of 384×640, and our evaluation indicates that its long-horizon stability remains limited.

G Additional Comparisons with Other Methods

In Sec. 4.2, we compare MovieRoll with existing methods on the 30-second HeliosBench, reporting quantitative results for both single-event and multi-event settings, as well as qualitative examples for the multi-event setting. In this section, we provide additional comparisons. For the single-event experiments, beyond MemFlow (Ji et al., 2025), LongLive (Yang et al., 2025a), Echo-Infinity (Bian et al., 2026), Helios-Base, and Helios-Distilled (Yuan et al., 2026), we further include CausVid (Yin et al., 2025), Self Forcing (Huang et al., 2026c), Causal Forcing (Zhu et al., 2026), Rolling Forcing (Liu et al., 2026), and Krea (Millon, 2025). We report the corresponding comparisons with MovieRoll in this section. Although MovieRoll-Base substantially improves degradation resistance, it remains insufficiently stable for 1441-frame long-form generation. We therefore report only the performance of MovieRoll-Distilled on the 1441-frame benchmark.

Single-event Comparison. Tab. 13 reports the complete benchmark scores for MovieRoll and competing methods at 241, 721, and 1441 frames. MovieRoll achieves the best performance at all three video lengths. Fig. 9, Fig. 10, and Fig. 11 further present qualitative comparisons at 10s (241 frames), 30s (721 frames), and 60s (1441 frames), respectively. MovieRoll consistently exhibits stronger visual quality across all durations.

Multi-event Comparison. In the main paper, we evaluate methods supporting multi-event generation under the 721-frame, 3-event setting. In this section, we extend the comparison to 1441 frames with 6 events, as reported in Tab. 14. MovieRoll remains the best-performing method in this longer setting. Qualitative results in Fig. 12 further show that MovieRoll-Distilled preserves strong semantic consistency across events while achieving the highest visual quality.

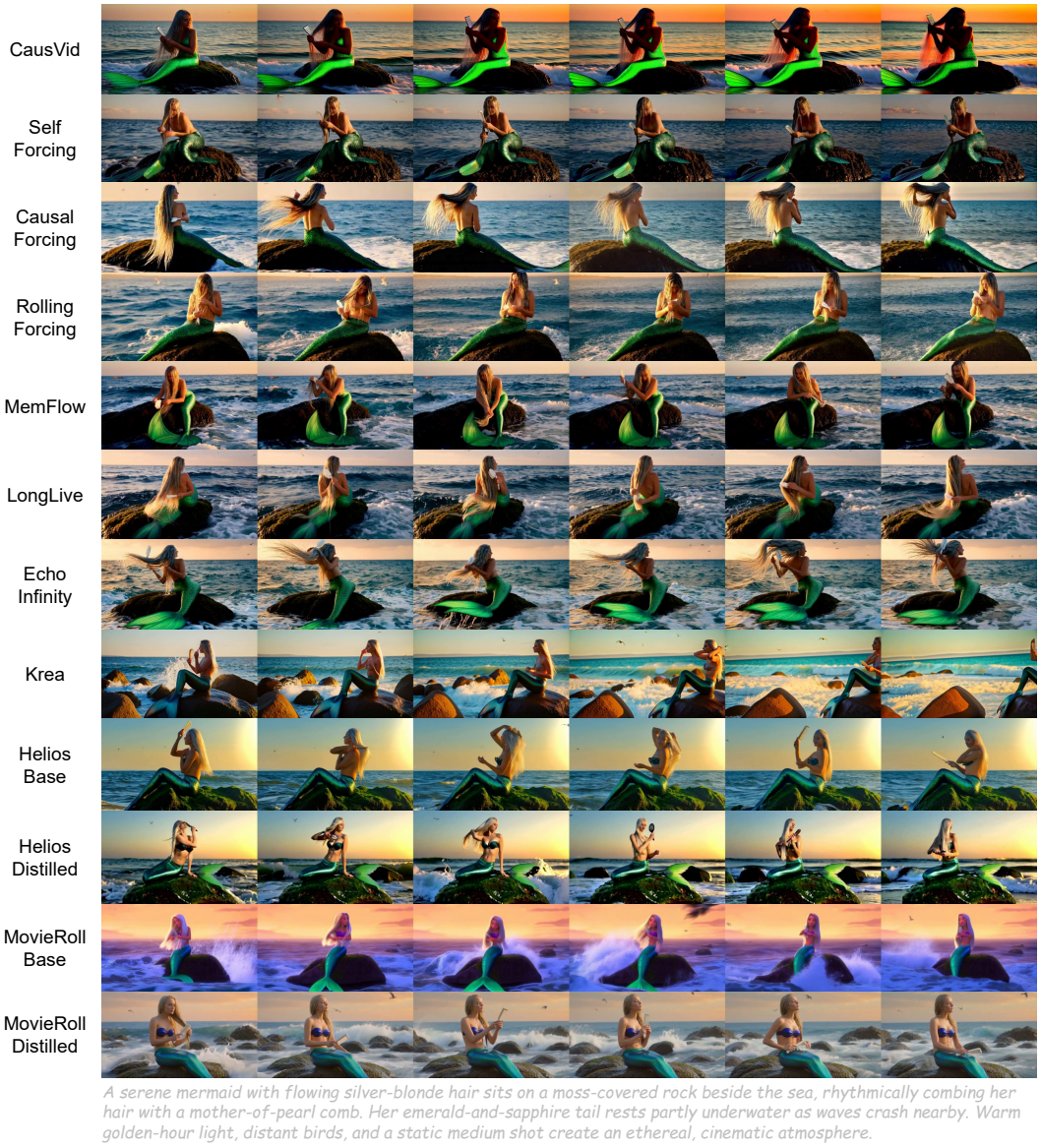


Figure 9: Generation quality of different methods under 241-frame single-event evaluation.

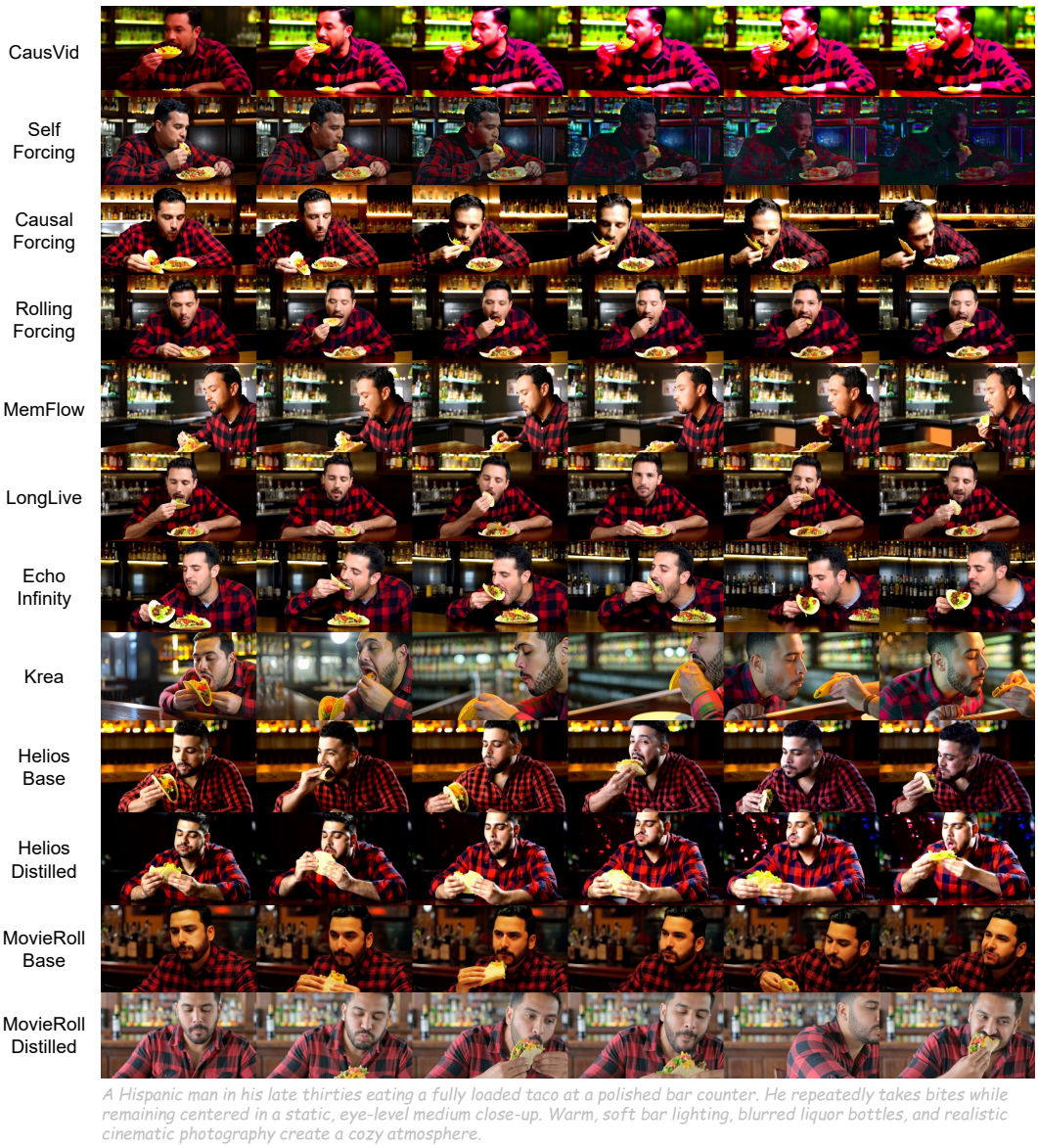


Figure 10: Generation quality of different methods under 721-frame single-event evaluation.

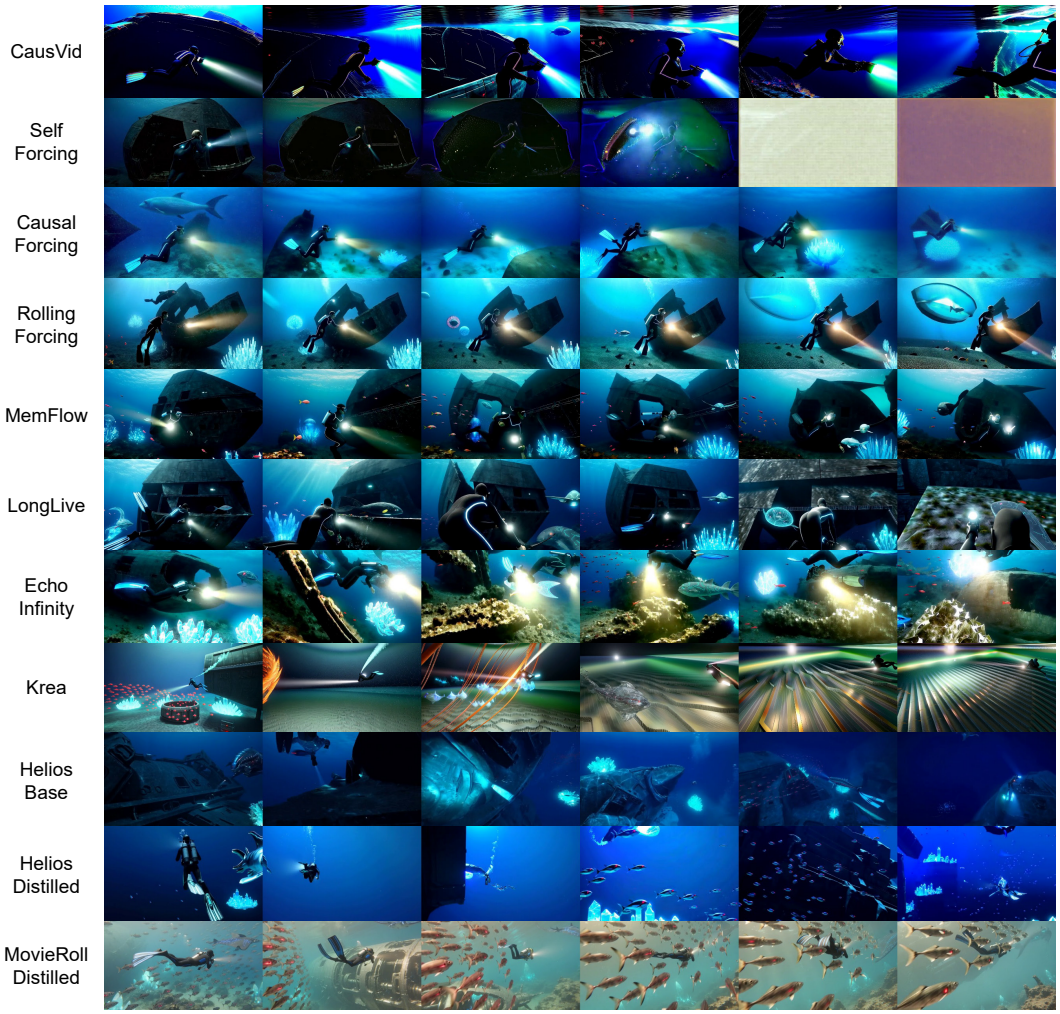
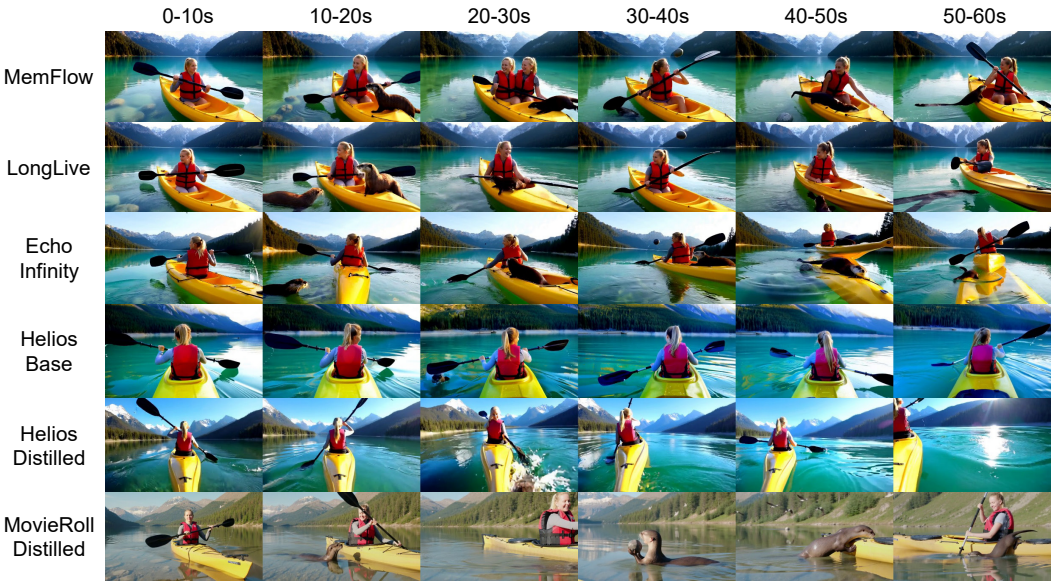


Figure 11: Generation quality of different methods under 1441-frame single-event evaluation.



Prompt 1 (0-10s): A young woman steadily kayaks across a crystal-clear mountain lake in a yellow kayak.
Prompt 2 (10-20s): A playful otter surfaces beside the kayak, interrupting the woman's rhythmic paddling.
Prompt 3 (20-30s): The otter circles the kayak playfully while the woman watches and laughs.
Prompt 4 (30-40s): The otter floats on its back and rhythmically juggles a small rock.
Prompt 5 (40-50s): The otter approaches the kayak as the woman observes its playful interaction at close range.
Prompt 6 (50-60s): The otter dives beneath the lake, and the woman watches a quiet farewell.

Figure 12: Generation quality of different methods under 1441-frame multi-event evaluation.