ALAYA-EVOKE: FROM LINEAR-SCALING SUPERVISION TO ENDLESS WORLD

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive world models must simultaneously support persistent memory, responsive user interaction, and long-horizon generation, yet these requirements place conflicting demands on the underlying model. Maintaining history in the denoiser context or key-value cache incurs growing cost over time, forcing a trade-off between session length and retained memory, while low-latency interaction typically relies on few-step generation whose capabilities are ultimately bounded by its teacher. Alaya-EVOKE (Evoke) addresses both limitations by externalizing persistent world state and redesigning the teacher for long-horizon interactive generation. Scene geometry is maintained in an external, camera-indexed world state bank, from which only information relevant to the current view is retrieved, keeping the denoiser context bounded as the session grows. Rather than treating the teacher as a fixed high-quality generator, the teacher is explicitly designed for long-horizon supervision. Its sparse attention scheme combines chunk-wise grouping, retrieval of selected distant frames, and a linear-attention global state, yielding linear growth in activation memory and computation while enabling supervision over long temporal horizons. Such supervision exposes content drift that remains locally plausible within short windows, while per-chunk conditioning enables prompt changes and event control throughout the sequence. A 30-second long-horizon distribution-matching objective, applied under self-forced rollouts, transfers both capabilities to a three-step student that uses no classifier-free guidance (CFG), improving resistance to long-term content drift while preserving responsive conditioning. With bounded context and recurrent external memory, Evoke supports open-ended, continuously evolving generation; on a single H200 at 384×640, each $1.5\,\mathrm{s}$ chunk is generated in $2.11\,\mathrm{s}$. With three sampling steps, Evoke ranks first on the WBench navigation split and achieves competitive results on VBench-2.0, VBench-Long, and VBench-I2V.

1 INTRODUCTION

Interactive world models must jointly support persistent memory, responsive interaction, and long-horizon generation (Bruce et al., 2024; Ball et al., 2025; He et al., 2025b; Sun et al., 2025; Mao et al., 2025). A revisited scene should remain consistent with its earlier appearance; camera motions and text instructions should take effect promptly; and a session should extend from seconds to minutes or even hours. While substantial progress has been made on each capability in isolation, combining all three remains challenging. Their conjunction exposes a systems-level question that model scaling alone does not resolve (Yin et al., 2025b): how to maintain persistent world state without an ever-growing context, and how to equip a low-latency generator with the capabilities required for long-horizon interaction.

Current interactive world models typically place two demanding requirements on the denoiser: preserving persistent world state and generating responses at interactive latency. When past observations are retained as additional context frames or an accumulated key-value cache (Chen et al., 2024; Zhang & Agrawala, 2025; Li et al., 2025a; Yang et al., 2025), the cost of each denoising step grows with the session history. Windowing and cache eviction can bound this cost only by discarding information (Cui et al., 2025), while retrieval and streaming conditioning still operate under a finite denoiser-side context budget (Hong et al., 2025; Wang et al., 2026; Feng et al., 2025). Interactive generation introduces a different constraint: low latency requires inference in only a few denoising

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.

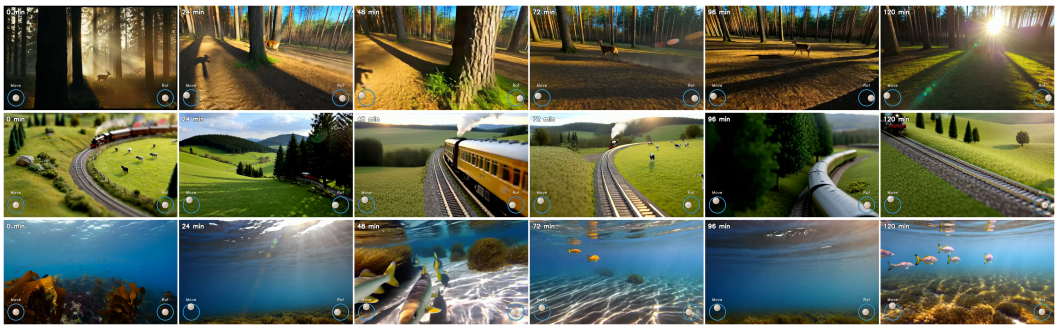

Figure 1: **Two-hour generation under continuous camera control.** Each row shows six evenly spaced frames from a three-step, CFG-free rollout.

steps, which is typically achieved by distillation from a slower teacher (Yin et al., 2024b;a; 2025a; Huang et al., 2025a). The resulting student is therefore constrained by the temporal horizon and conditioning capabilities represented in its teacher supervision, limiting both long-term consistency and responsiveness to changes introduced during a session. These limitations suggest that persistent state and interactive capabilities need not both be carried by the deployed denoiser. **Evoke decouples them: persistent spatial state is maintained in an external world state bank, while the teacher is explicitly designed to provide the long-horizon and dynamically conditioned supervision required by the few-step student.** This separation forms the central design principle of Evoke and enables bounded-cost, long-running interaction.

For persistent world state, Evoke uses a bounded world state bank indexed by camera pose. Following geometry-based conditioning (Wang & He, 2026; Ren et al., 2025; Shen et al., 2026; Wu et al., 2026), generated observations are lifted into scene geometry and rendered from the current viewpoint when revisited. Storing geometry externally keeps the denoiser's history and memory context fixed as the session grows. The bank supports bounded read, write, and eviction, providing access to previously observed surfaces within its retained temporal coverage rather than indefinite memory of every visited location.

For transferable generation capabilities, Evoke treats the teacher itself as a design variable. Distillation is commonly used to accelerate inference (Yin et al., 2024b;a; Lin et al., 2025b; Yin et al., 2025a); beyond the teacher–student causality mismatch studied in prior work (Zhu et al., 2026), we focus on two complementary axes: supervision horizon and conditioning schedule. Short-window supervision can detect local *degradation drift*, such as exposure or texture changes, yet miss *content drift*: scene identity or spatial layout may change gradually while each short segment remains plausible (Cui et al., 2025; Liu et al., 2025). Longer supervision windows expose inconsistencies across distant moments. Independently, a single global prompt cannot demonstrate how instructions introduced mid-session should affect a rollout. Per-chunk conditioning makes these changes visible to the teacher, enabling supervision of both long-range consistency and dynamic interaction.

These requirements lead to the Evoke Teacher, whose chunk-wise sparse attention makes long-sequence supervision practical with attention cost that grows approximately linearly with sequence length. Its per-chunk text conditioning represents changing instructions within a single rollout. A 30-second distribution-matching objective (Yin et al., 2024b;a), applied to self-forced rollouts (Huang et al., 2025a), transfers both capabilities to a three-step student without classifier-free guidance (Ho & Salimans, 2022). The student can then respond to scheduled prompt changes, introducing or withdrawing elements during an ongoing session—a capability we call *evocation* (Fig. 2). Bounded recurrent inference keeps each generation step independent of total session length; Section 3 details the formulation and supervision design.

With the world state bank enabled, Evoke denoises a 36-frame chunk in approximately 2 s on a single H200 at 384×640. Figure 1 shows three uninterrupted two-hour rollouts under continuous camera control. With three sampling steps, Evoke ranks first among the compared systems on WBench navigation (Ying et al., 2026) and achieves competitive VBench-2.0 (Zheng et al., 2025), VBench-Long, and VBench-I2V (Huang et al., 2024; 2025b) scores. Our contributions can be summarized as follows.



Figure 2: **Timed prompts change sky content while preserving anchored scene structure.** All rows share the input image, camera trajectory, and three-step CFG-free student. Rows 1–3 add a fireball, balloon, and fireworks, respectively, at chunk 2 and remove them at chunk 4; amber borders mark active prompts. Row 4 uses a static prompt.

(i) We formulate interactive world generation as a bounded recurrent process that decouples persistent world state from the denoiser: scene geometry is maintained in an external, camera-indexed world state bank, keeping the denoiser context and positional range independent of session duration.

(ii) We redesign the teacher for long-horizon interactive supervision through chunk-wise sparse attention and per-chunk conditioning, enabling efficient supervision over long temporal horizons while exposing both long-range content drift and mid-sequence instruction changes.

(iii) We develop Evoke, a three-step, CFG-free interactive world model that achieves state-of-the-art performance on the WBench navigation split, competitive results on VBench-2.0, VBench-Long, and VBench-I2V, and hour-scale generation under continuous camera control.

2 Related work

Persistent memory and bounded generation. Interactive world models must retain history while controlling the cost of continued generation (Bruce et al., 2024; Ball et al., 2025; Zhang et al., 2025). Context frames and key-value caches preserve past observations (Chen et al., 2024; Zhang & Agrawala, 2025; Li et al., 2025a), while windowing, eviction, and retrieval manage the trade-off between retained information and per-step cost (Cui et al., 2025; Hong et al., 2025; Wang et al., 2026; Feng et al., 2025). Geometry-based methods instead represent observations outside the denoiser and render them into the target view (Wang & He, 2026; Ren et al., 2025; Shen et al., 2026; Wu et al., 2026). This representation supports camera-conditioned generation and spatial memory together: camera pose determines how stored scene content is projected into the current view, providing guidance for camera motion and scene revisits. Evoke follows this approach with an external world state bank under fixed retention and retrieval budgets, supporting recall within retained geometric coverage while keeping per-step generation cost independent of session duration.

Few-step generation and teacher supervision. Few-step distillation reduces inference latency by transferring a slower teacher's capabilities to a student (Salimans & Ho, 2022; Luo et al., 2023; Song et al., 2023). Distribution matching (Yin et al., 2024b;a; 2025a), training on self-generated histories (Huang et al., 2025a), and teacher–student causality alignment (Zhu et al., 2026) address aspects of the training objective, conditioning distribution, and attention structure. Building on these directions, we focus on supervision horizon and conditioning schedule. Extended generation (Yang et al., 2025; Chen et al., 2025a; Li et al., 2025b) does not by itself imply that the training objective jointly constrains distant content; the relevant distinction is the temporal span covered by supervision. Likewise, instructions introduced during a session need to be represented in that supervision. Evoke therefore jointly designs long-horizon teacher supervision and per-chunk text conditioning, transferring these capabilities to the few-step student through distribution-matching distillation.

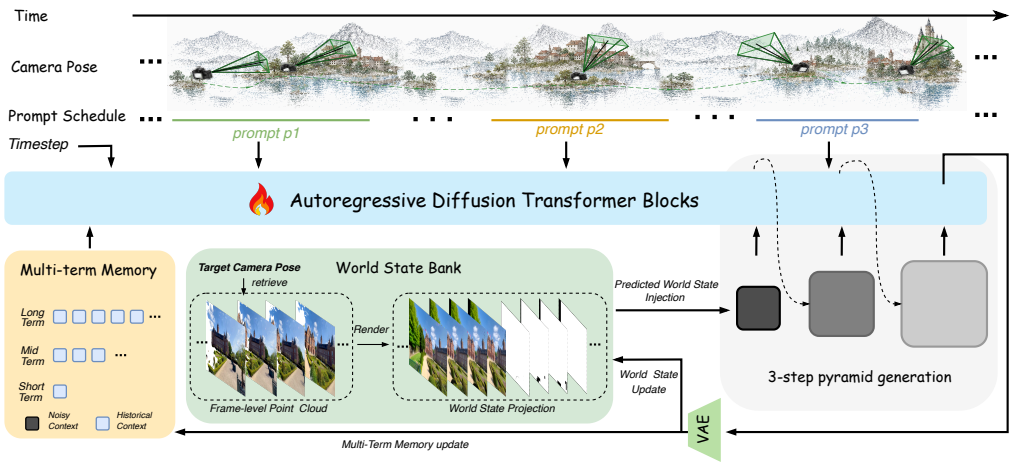

Figure 3: **Bounded recurrent inference in Evoke.** Camera-aligned geometry retrieved from the world state bank conditions the coarsest stage of three-step generation, together with local history and text. Generated chunks update the bank and history; fixed retention and retrieval budgets keep per-step cost independent of session length.

3 EVOKE

3.1 PRELIMINARIES

An interactive world model generates future observations under changing camera and text controls while preserving scene state across a session. Open-ended interaction additionally requires bounded per-step computation. We formulate this process as recurrent, state-conditioned generation and instantiate it in Evoke through an external geometric state bank and a three-step generator (Fig. 3). At step k, Evoke generates a chunk x_k of $F = 9$ latent frames, corresponding to 36 pixel frames or 1.5 s at 24 fps. Its inputs are a camera trajectory $\mathcal{P}_k$ with known intrinsics and extrinsics (He et al., 2024; 2025a), a text condition c_k, local history h_k, and geometry retrieved from world state M_k:

$$r_k = \mathrm{Read}(M_k, \mathcal{P}_k), \qquad x_k \sim p_\theta(\cdot \mid r_k, h_k, c_k), \qquad M_{k+1} = \mathrm{Write}(M_k, x_k, \mathcal{P}_k). \tag{1}$$

Camera and text controls may change between chunks. The state bank provides spatial recall under fixed budgets (Sec. 3.2), while long-horizon teacher supervision and few-step distillation train the generator for temporal consistency and responsive control (Secs. 3.3–3.4).

3.2 EXTERNALIZING WORLD STATE FOR BOUNDED GENERATION

The denoiser retains only 19 latent history frames, approximately 3.2 s, organized into long-range, mid-range, and short-range history containing 16, 2, and 1 frames, respectively. Evoke stores observations beyond this local context as external geometry (Wu et al., 2026), rather than retaining recalled history as cache entries or memory tokens (Hong et al., 2025; Wang et al., 2026). Camera pose serves both as a generation control and as an address for retrieving scene content, coupling camera-conditioned generation with spatial recall.

Writing and reading geometry. Let $\mathcal{S}_k = (h_k, M_k)$ be the world-model state and $(\mathcal{P}_k, c_k)$ the camera and text controls. Rendering co-visible source views under a fixed retrieval budget R with z-buffering (Wang & He, 2026; Ren et al., 2025) gives an observation r_k and visibility mask v_k:

$$(r_k, v_k) = \mathrm{Render}(\mathrm{Select}_R(M_k, \mathcal{P}_k), \mathcal{P}_k)\,, \qquad x_k \sim p_\theta(\cdot \mid h_k, r_k, v_k, c_k). \tag{2}$$

A depth model (Lin et al., 2025a) estimates 12 frames per chunk; Lift unprojects them using known camera parameters. The sampled chunk updates both state components:

$$\mathcal{S}_{k+1} = (\mathrm{Update}_H(h_k, x_k),\ \mathrm{Evict}_B[M_k \oplus \mathrm{Lift}(x_k, \mathcal{P}_k)])\,. \tag{3}$$

Here $H = 19$ bounds local history, B bounds retained source views, and $\oplus$ appends geometry independently, avoiding scale drift from cross-chunk fusion.

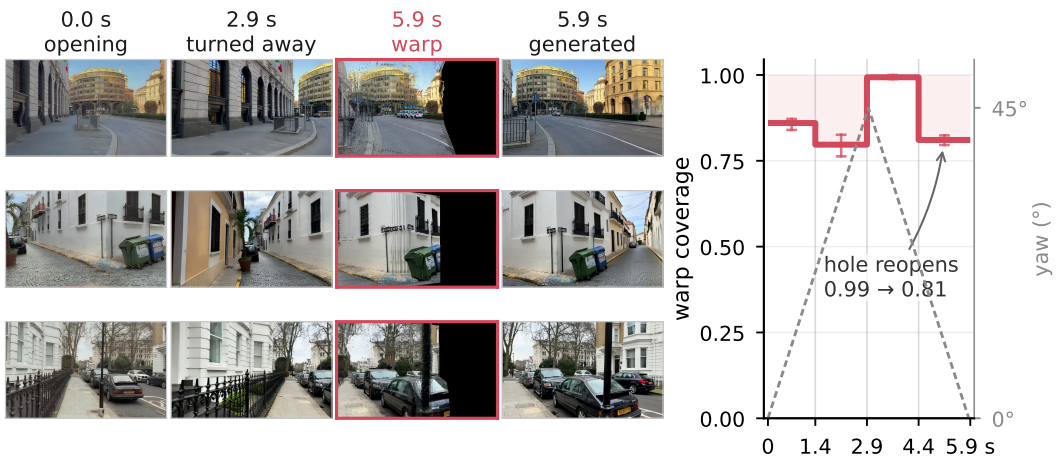

Figure 4: **Geometric memory supports scene revisits.** *Left:* initial view, turn endpoint, retrieved geometry, and generated return view. *Right:* geometric coverage during the camera turn and return.

Geometric conditioning and three-step generation. The visibility mask suppresses unreliable warp conditioning: pixels with visibility below 0.5 receive noise level $\sigma = 1$, while supported pixels use $\sigma \in [0, 0.135]$. Patch-level visibility also removes unsupported history tokens (Wang & He, 2026), leaving uncovered regions available for synthesis. Evoke uses three CFG-free (Ho & Salimans, 2022) denoising evaluations over a coarse-to-fine latent pyramid (Jin et al., 2024), at resolutions 12×20, 24×40, and 48×80. Geometry is injected only at the coarsest stage; subsequent stages refine appearance and detail. Figure 4 illustrates how retained geometry supports a camera return to a previously observed region.

Bounded operation. Fixed storage, retrieval, and history budgets, together with reuse of local positional indices (Su et al., 2024), keep the denoiser context, positional span, and per-step geometric workload independent of session duration. Recall is limited to retained geometric coverage, rather than all previously visited locations. Section 4.2 evaluates runtime and geometric-conditioning overhead. This mechanism preserves accessible spatial state; consistency of ongoing content evolution additionally depends on the student's supervision.

3.3 Long-Horizon Interactive Teacher

Supervision requirements. Local *degradation drift*, such as exposure or texture changes, can be visible within short windows, yet arise from errors accumulated over many autoregressive steps. Short-rollout supervision underrepresents the corrupted histories encountered during prolonged inference, creating a training–inference distribution mismatch. Long-horizon supervision over self-generated rollouts exposes these accumulated errors and the conditioning histories they induce. *Content drift* additionally changes scene identity or layout while each short segment remains plausible (Cui et al., 2025; Liu et al., 2025), requiring supervision across distant moments. Separately, a teacher conditioned on one global prompt cannot demonstrate responses to instructions introduced mid-session. These requirements motivate a teacher that jointly supports long sequences and time-varying text conditions.

Efficient long-sequence architecture. The Evoke Teacher adapts the 14B Wan2.2 A14B transformer (Peebles & Xie, 2023; Wan et al., 2025a;b), retaining timestep-selected high- and low-noise experts. For each nine-frame latent chunk k, sparse attention (Jia et al., 2025; Cai et al., 2025) accesses indices $\mathcal{I}_k$ comprising a first-frame sink, local context with one-frame overlap, compressed nearby frames, and selected distant frames:

$$A_k = \mathrm{softmax}\left(\frac{Q_k K_{\mathcal{I}_k}^{\top}}{\sqrt{d}}\right) V_{\mathcal{I}_k}, \qquad O_k = (A_k + gG_k)W_o. \tag{4}$$

The complementary global branch uses linear attention (Yang et al., 2023; Team et al., 2025; Chen et al., 2025b). Within each layer, all sequence tokens contribute to a compact state, queried by each

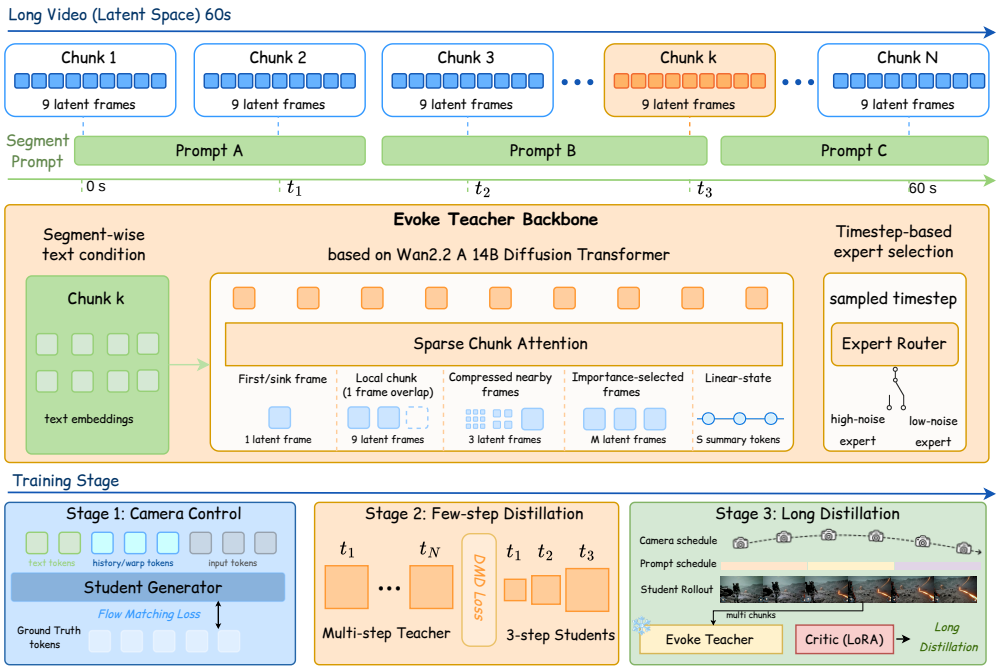

Figure 5: **Long-horizon teacher architecture and training pipeline.** For a T-chunk training sequence, bounded sparse attention and global linear attention yield approximately $O(T)$ attention cost. Chunks use segment-aligned text conditions; diffusion timesteps select high- or low-noise experts. Training progresses from camera control and few-step distillation to long-rollout distribution matching with a shared teacher/critic backbone.

chunk:

$$S = \sum_{t=1}^{N} \phi(\bar{k}_t)\bar{v}_t^\top, \qquad z = \sum_{t=1}^{N} \phi(\bar{k}_t),$$
$$G_k = \left[\frac{\phi(\bar{q}_{ki})^\top S}{\phi(\bar{q}_{ki})^\top z + \epsilon}\right]_{i\in k} W_g, \quad \phi(u) = \mathrm{ELU}(u) + 1. \tag{5}$$

N counts sequence tokens, d is the head dimension, bars denote separate global projections, and g is a learned gate. Head and layer indices are omitted, and overlapping outputs are blended before W_o. Fixed sparse budgets and compressed global state yield approximately linear attention cost (Li et al., 2025c). Unlike the external geometric bank, (S, z) is recomputed per layer over the teacher sequence: it provides a non-causal global path alongside past local and retrieved context (Zhu et al., 2026).

Dynamic text conditioning. Each chunk separately cross-attends to its text embedding $E(c_k)$. For features $\widetilde{H}_k$ entering the text branch,

$$H_k^{\text{text}} = \widetilde{H}_k + \mathrm{CrossAttn}\Big(\mathrm{LN}(\widetilde{H}_k), E(c_k)\Big). \tag{6}$$

Each segment of a long training video has a corresponding caption, which conditions the latent chunks within that segment, exposing prompt transitions within one sequence. Timed prompts use the same interface at inference (Fig. 2); Appendix Fig. A5 shows four-minute rollouts with twelve 20 s instructions. Distillation transfers this capability to Evoke.

3.4 Long-Horizon Distillation

Starting from a camera-controllable, few-step student, Evoke applies long-horizon distribution matching (Yin et al., 2024b;a) under self-forced rollouts (Huang et al., 2025a), followed by a short post-distillation continuation (training stages in Fig. 5). The teacher and critic share a backbone:

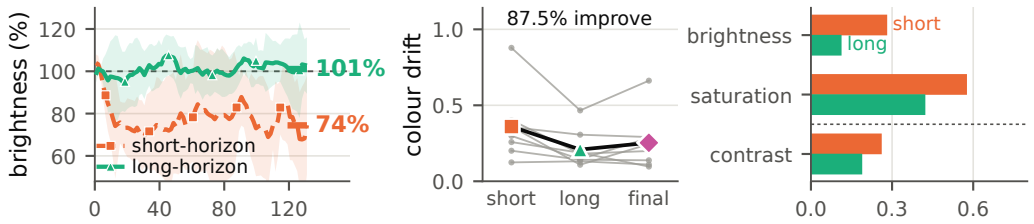

Figure 6: **Long-horizon supervision improves student photometric stability.** Three-step students are compared under matched distillation and evaluation settings. **(a)** Brightness stabilizes at 74% versus 101% of its initial value for short- versus long-horizon teachers. **(b)** 87.5% of clips improve (Wilcoxon $p = 0.016$). **(c)** Photometric drift across brightness, saturation, and contrast.

enabling a LoRA adapter (Hu et al., 2021) yields the critic, while disabling it yields the teacher, with both scores computed using the same timestep-selected expert.

Rollout and supervision window. Let q_θ and p denote student and data trajectory distributions. Supervision over windows of W consecutive chunks takes the form

$$\mathcal{L}_W(\theta) = \mathbb{E}_k \left[D\left(q_\theta^{(k:k+W-1)} \,\middle\|\, p^{(k:k+W-1)} \right) \right], \tag{7}$$

where the teacher and critic scores estimate the distribution-matching gradient. The window determines which temporal dependencies the objective can constrain. We use a ground-truth prefix x_0^{gt} followed by 20 self-generated chunks $x_1, \ldots, x_{20}$: 189 latent frames or 753 pixel frames, approximately 31.4 s. The student uses 60 denoising evaluations to construct this rollout. Teacher and critic jointly score the full sequence with text conditions aligned to each chunk, rather than scoring isolated short windows.

Distribution-matching update. Given student prediction $\hat{x}_0$ and teacher and critic predictions s_{real} and s_{fake}, we compute

$$\Delta s = s_{\mathrm{fake}} - s_{\mathrm{real}}, \qquad \nu = \mathrm{mean}_\Omega \left[|\hat{x}_0 - s_{\mathrm{real}}| \right], \qquad \mathcal{L}_{\mathrm{gen}} = \frac{1}{2} \left\| \hat{x}_0 - \left(\hat{x}_0 - \frac{\Delta s}{\nu} \right)^{\mathrm{detach}} \right\|_2^2. \tag{8}$$

The region Ω excludes the ground-truth prefix and first generated chunk x_1 from both the loss and normalization. At this boundary, teacher scoring follows the image-model first-frame distribution while the student performs video continuation; excluding it avoids the boundary flicker observed with direct matching.

Long supervision with bounded gradients. History is detached between chunks, allowing independent chunk-level backward graphs while teacher and critic score the complete rollout. Thus, the *gradient horizon* remains one chunk even though the *supervision horizon* spans 31.4 s. Sequence parallelism (Jacobs et al., 2023) and activation recomputation further control training memory. The training window need not grow with the eventual session length, which is extended through repeated inference calls. Figure 6 compares photometric stability after distillation from short- and long-horizon teachers. Distribution matching alone does not explicitly enforce the requested camera trajectory. We therefore retain the supervised warp-conditioning objective from the preceding stage as an auxiliary camera-control regularizer, while long-horizon distribution matching transfers temporal consistency and dynamic conditioning. The resulting student operates within the geometric recurrent interface of Sec. 3.2, combining learned generation capabilities with explicit recall of retained scene state.

4 Experiments

Experimental setup. The Evoke student is built on Helios (Yuan et al., 2026), and the Evoke Teacher adapts the 14B Wan2.2 A14B transformer (Wan et al., 2025a;b). Training uses Sekai (Li et al., 2026) and additional internal video data, progressing from camera-controllable generation and few-step inference to long-horizon distillation and a short post-distillation continuation. Unless stated otherwise, we evaluate the resulting student with three denoising evaluations per chunk and

Table 1: WBench navigation split, $n = 158$ cases, identical for all systems. All scores $\in [0, 100]$, higher is better; **bold** is best, underline second best. Group rows are the unweighted mean over their group's metrics; peer numbers are as reported by their authors.

Metric	Yume 1.5	Matrix-Game 2.0	HY-World 1.5 *(ar-distill)*	HY-GameCraft	LingBot-World *(fast)*	LingBot-World v2 *(fast)*	Genie 3	Happy Oyster	JoyAI-Echo-1.5 *(flash)*	Evoke (ours)
Video Quality										
Aesthetic	58.7	54.0	60.1	52.6	62.6	64.4	51.6	56.6	64.2	**65.2**
Imaging	63.3	60.3	65.4	58.7	63.8	**67.5**	59.3	63.9	66.3	**67.5**
Flickering	93.0	94.6	93.5	93.7	92.4	91.4	**95.0**	94.0	92.8	94.3
Dynamic	**96.8**	94.9	91.1	**96.8**	95.6	96.2	92.4	94.2	94.3	90.5
Smoothness	97.0	**98.2**	98.1	97.6	96.0	96.5	97.8	97.0	97.2	97.6
HPSv3-Norm	57.0	41.0	60.5	38.3	65.7	74.6	55.2	58.3	71.8	**76.3**
Quality avg. (6)	77.63	73.83	78.12	72.95	79.35	81.77	75.22	77.33	81.10	**81.90**
Setting										
Scene	53.1	49.4	53.5	50.6	63.4	66.7	61.1	57.4	64.8	**71.3**
Subject	91.7	84.9	90.8	82.5	92.4	86.9	83.8	91.1	90.1	**92.8**
Setting avg. (2)	72.40	67.15	72.15	66.55	77.90	76.80	72.45	74.25	77.45	**82.05**
Interaction										
Navigation	72.0	80.6	87.5	67.8	79.4	82.8	73.3	85.1	**87.9**	83.9
Consistency										
Background	90.3	86.9	92.7	86.5	90.9	92.5	90.7	91.4	**93.6**	92.9
Spatial	71.5	64.5	**90.6**	60.5	77.2	82.3	79.9	77.7	86.8	88.8
Gated Spatial	71.4	64.5	**84.9**	60.5	76.9	78.7	78.4	75.8	79.8	82.7
Segment	99.4	21.0	**100.0**	99.4	98.1	98.1	93.6	96.2	98.7	**100.0**
Perspective	48.0	29.2	62.5	17.9	82.8	**84.5**	54.5	75.0	83.1	71.9
Subject	88.8	87.2	89.1	82.6	88.6	88.9	90.4	91.5	90.0	**91.9**
Geometric	88.0	86.1	92.0	88.3	85.4	87.1	88.6	87.2	92.5	**94.3**
Photometric	83.3	81.3	83.1	**85.0**	79.1	79.8	84.5	79.8	82.0	82.5
Consistency avg. (8)	80.09	65.09	86.86	72.59	84.88	86.49	82.58	84.33	**88.31**	88.13
Physical										
Causal Fidelity	72.7	59.3	74.0	68.3	72.5	76.7	71.7	69.3	79.6	**86.1**
Visual Plausibility	57.7	55.0	58.6	56.5	58.8	61.4	59.7	57.6	60.7	**61.9**
Physical avg. (2)	65.20	57.15	66.30	62.40	65.65	69.05	65.70	63.45	70.15	**74.00**

no classifier-free guidance. Runtime measurements use a single H200 at 384×640 with the full VAE decoder, without key-value caching, compilation, quantization, or a distilled decoder. Experiment-specific protocols are described below; additional qualitative rollouts appear in Appendix Fig. A6.

4.1 Overall Performance

Interactive world-model performance. On the 158-case navigation split of WBench (Ying et al., 2026), Evoke ranks first overall among the systems in Table 1, based on the unweighted mean of the five group scores. It leads Video Quality (81.90), Setting (82.05), and Physical (74.00), and ranks second in Consistency (88.13). Compared with existing few-step systems (Mao et al., 2025; He et al., 2025b; HunyuanWorld, 2025; Li et al., 2025a; Robbyant Team et al., 2026; Gao et al., 2026; Ball et al., 2025; Happy Oyster, 2026), these results demonstrate strong visual quality, scene consistency, and physical plausibility. Navigation and Perspective remain below the best reported scores. Section 4.3 examines the contributions and limitations of individual mechanisms.

General video quality. Evoke-Turbo scores 66.32 on VBench-2.0 (Zheng et al., 2025), 84.69 on VBench-Long, and 88.17 on VBench-I2V (Huang et al., 2024; 2025b). Its VBench-2.0 score ranks second among 13 entries in the 2026-09-20 comparison, below Veo 3 (66.72), and reflects prompt tuning on known evaluation cases with fixed seeds. These results indicate competitive video quality with three CFG-free steps; comparisons use each system's own sampler. Appendix E provides the rankings, complete dimension scores, and evaluation settings.

4.2 Long-Horizon Generation and Efficiency

Long-session stability. Figure 1 demonstrates uninterrupted two-hour generation under continuous camera control. Quantitative evaluation uses eight 65.5-minute rollouts of 2,619 chunks each, with the world state bank restricted to 90 s of retained observations. Photometric statistics stabilize after an initial transient and exhibit little subsequent drift. Content descriptors change more rapidly near the beginning and then evolve more slowly; real-video controls exhibit comparable decorrelation as

the camera encounters new content. These measurements support resistance to progressive long-session degradation, without establishing permanent scene-identity preservation. Appendix Fig. A3 illustrates a representative rollout, and Appendix C details the evaluation protocol and interpretation.

Bounded runtime and geometric cost. Once the retention budget is filled, the active geometric source pool no longer grows with elapsed time, and per-step cost remains stable. Longer sessions increase the number of recurrent calls without increasing the state or computation required by an individual call. The student denoises each $1.5\,\mathrm{s}$ chunk in $2.11\,\mathrm{s}$ on a single H200; this measures diffusion wall clock rather than end-to-end latency. Geometry rendering uses a bounded set of source views, and unsupported warp tokens are removed before denoising. The additional conditioning cost therefore varies with retrieved geometric coverage rather than session duration. Appendix B reports the wall-clock decomposition and the dependence of conditioning overhead on coverage.

4.3 Spatial Memory, Temporal Supervision, and Interactive Control

Geometric memory. We test spatial recall using leave-and-return trajectories, measuring revisit PSNR between two $12\,\mathrm{s}$ windows rendered at identical camera poses. When the retention window is shorter than the time spent away, recall remains near the far-pose floor measured within the same run. It improves once retained geometry covers the revisit: 20 of 21 controlled comparisons follow this transition, with adequate retention yielding a 2.3–$3.2\,\mathrm{dB}$ gain over shorter windows. The resulting 15.4–$17.8\,\mathrm{dB}$ plateau indicates recognizable, but not pixel-faithful, reconstruction. Figure 4 illustrates recall in three held-out scenes along a $5.9\,\mathrm{s}$ trajectory with a 45° turn and return, using three-step, CFG-free Evoke. Retained geometry restores previously observed regions; black areas lack coverage and require synthesis. Restricting retention to $1.5\,\mathrm{s}$ reduces return-view coverage by 17.3 percentage points. This measures available geometric support, not reconstruction fidelity.

Long-horizon supervision. We compare students distilled from short- and long-horizon teachers using otherwise matched training recipes, before the final post-distillation continuation. Across paired long-rollout evaluations, the Evoke Teacher yields substantially stronger photometric stability (Fig. 6), supporting the transfer of resistance to accumulated appearance degradation. In Fig. 6(b), "final" denotes the student after post-training (the post-distillation continuation). The measured content descriptor does not significantly separate the students, so the demonstrated improvement is specific to photometric stability. Scoring-window sweeps further show that the benefit cannot be explained simply by making perturbations detectable over a longer window: detectability changes little beyond a short horizon. We therefore attribute the observed benefit to the trained teacher and rollout supervision process as a whole. Detailed sweeps and additional teacher ablations appear in Appendix C.

Dynamic interaction. In the controlled evocation evaluation, a text clause introduced mid-session is realized in 67% of cases when it targets previously unanchored content, compared with 4% when it requires replacing geometry supported by the world state bank (Appendix Fig. A4). Qualitative examples likewise show elements being introduced and withdrawn while anchored scene structure remains stable (Fig. 2). This reveals both the capability and its boundary: timed text can change content that remains free to evolve, while persistent geometry resists overwriting retained observations. Matched short- and long-teacher variants do not significantly differ in event-realization rates. Per-chunk conditioning thus provides the interface for timed control, but this comparison does not establish that the teacher horizon alone causes the capability.

5 Conclusion

Evoke combines external geometric memory with long-horizon teacher supervision and three-step, CFG-free generation. It supports spatial recall, changing camera and text controls, and two-hour rollouts with bounded per-step computation. Evoke ranks first among the compared systems on WBench navigation and achieves competitive VBench-2.0, VBench-Long, and VBench-I2V scores. Controlled comparisons demonstrate improved photometric stability from long-horizon supervision. Fine-grained identity preservation, dynamic world-state modeling, and lower end-to-end latency remain directions for improvement.

Reproducibility Statement

We describe the model interface, geometric conditioning, teacher architecture, and training stages in Section 3, and report the training data sources, sampling configuration, hardware, and experimental settings in Section 4. Appendices A–C provide training curves, runtime breakdowns, and additional evaluation and ablation details. Appendices E and F document the benchmark comparisons, including generation settings, prompt-tuning conditions, score aggregation, and per-dimension results, to clarify how the reported results were obtained.

AI use statement

During the research process, generative AI tools were used for literature discovery and review. Method implementation, data processing, and experimental analyses were performed by the researchers. During manuscript preparation, OpenAI Codex assisted with drafting and language refinement, organizing the paper, clarifying methodological descriptions and mathematical notation, refining the presentation and interpretation of experimental results, editing figure captions, formatting tables, and checking LaTeX formatting and cross-references. All AI-assisted content has been manually reviewed by the authors. All authors take responsibility for the final content of this work, including text, claims, and artifacts produced with AI assistance.

References

Philip J. Ball, Jakob Bauer, Frank Belletti, et al. Genie 3: A new frontier for world models, 2025. URL https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *ICML*, 2024.

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. *arXiv preprint arXiv:2508.21058*, 2025.

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.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025a.

Junsong Chen, Yuyang Zhao, Jincheng Yu, Ruihang Chu, Junyu Chen, Shuai Yang, Xianbang Wang, Yicheng Pan, Daquan Zhou, Huan Ling, et al. Sana-video: Efficient video generation with block linear diffusion transformer. *arXiv preprint arXiv:2509.24695*, 2025b.

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

Tianrui Feng, Zhi Li, Shuo Yang, Haocheng Xi, Muyang Li, Xiuyu Li, Lvmin Zhang, Keting Yang, Kelly Peng, Song Han, et al. Streamdiffusionv2: A streaming system for dynamic and interactive video generation. *arXiv preprint arXiv:2511.07399*, 2025.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, et al. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026.

Happy Oyster. Happy oyster: Real-time world model for interactive creation, 2026. URL https://www.happyoyster.com.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. In *CVPR*, 2025a.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025b.

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

Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, Kalyan Sunkavalli, Feng Liu, Zhengqi Li, and Hao Tan. Relic: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040*, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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

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

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b.

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

Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. Deepspeed ulysses: System optimizations for enabling training of extreme long sequence transformer models. *arXiv preprint arXiv:2309.14509*, 2023.

Weinan Jia, Yuning Lu, Mengqi Huang, Hualiang Wang, Binyuan Huang, Nan Chen, Mu Liu, Jidong Jiang, and Zhendong Mao. Moga: Mixture-of-groups attention for end-to-end long video generation. *arXiv preprint arXiv:2510.18692*, 2025.

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

Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-gamecraft: High-dynamic interactive game video generation with hybrid history condition. *arXiv preprint arXiv:2506.17201*, 2025a.

Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable video infinity: Infinite-length video generation with error recycling. *arXiv preprint arXiv:2510.09212*, 2025b.

Xingyang Li, Muyang Li, Tianle Cai, Haocheng Xi, Shuo Yang, Yujun Lin, Lvmin Zhang, Songlin Yang, Jinbo Hu, Kelly Peng, et al. Radial attention: Sparse attention with energy decay for long video generation. *arXiv preprint arXiv:2506.19852*, 2025c.

Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, et al. Sekai: A video dataset towards world exploration. *Advances in Neural Information Processing Systems*, 38, 2026.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025a.

Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. Diffusion adversarial post-training for one-step video generation. *arXiv preprint arXiv:2501.08316*, 2025b.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, Tong He, Jiangmiao Pang, Yu Qiao, and Kaipeng Zhang. Yume-1.5: A text-controlled interactive world generation model. *arXiv preprint arXiv:2512.22096*, 2025.

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

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132. IEEE, 2025.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, Jie Liu, Yansong Cheng, Yao Yao, Jiayi Zhu, Yihao Meng, Kecheng Zheng, Qingyan Bai, Jingye Chen, Zehong Shen, Yue Yu, Xing Zhu, Yujun Shen, and Hao Ouyang. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, et al. Lyra 2.0: Explorable generative 3d worlds. *arXiv preprint arXiv:2604.13036*, 2026.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

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 modeling. *arXiv preprint arXiv:2512.14614*, 2025.

Kimi Team, Yu Zhang, Zongyu Lin, Xingcheng Yao, Jiaxi Hu, Fanqing Meng, Chengyin Liu, Xin Men, Songlin Yang, Zhiyuan Li, et al. Kimi linear: An expressive, efficient attention architecture. *arXiv preprint arXiv:2510.26692*, 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*, 2025a.

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*, 2025b.

Yifan Wang and Tong He. Warp-as-history: Generalizable camera-controlled video generation from one training video. *arXiv preprint arXiv:2605.15182*, 2026.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *Advances in Neural Information Processing Systems*, 38:49371–49393, 2026.

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

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*, 2023.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, 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 *CVPR*, 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22963–22974, 2025a.

Yuanyang Yin, Yaqi Zhao, Mingwu Zheng, Ke Lin, Jiarong Ou, Rui Chen, Victor Shea-Jay Huang, Jiahao Wang, Xin Tao, Pengfei Wan, et al. Towards precise scaling laws for video diffusion transformers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18155–18165. IEEE, 2025b.

Kaining Ying, Hengrui Hu, Siyu Ren, Jiamu Li, Fengjiao Chen, Ziwen Wang, Xuezhi Cao, Xunliang Cai, and Henghui Ding. Wbench: A comprehensive multi-turn benchmark for interactive video world model evaluation. *arXiv preprint arXiv:2605.25874*, 2026.

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.

Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. *arXiv preprint arXiv:2504.12626*, 2025.

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*, 2025.

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.