WORLDSTATE: SCALABLE IMPLICIT MEMORY FOR INTERACTIVE VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive video world models generate videos conditioned on user inputs and camera movements. Long-horizon interaction requires these models to preserve previously generated environments and recover historically consistent content during spatial revisits. Explicit memories either require growing storage and retrieval or rely on potentially inaccurate geometric estimates. Recurrent linear memory avoids these costs but suffers from imbalanced frame-wise updates and fixed capacity: spatial normalization weakens erasure relative to writing, while a single state cannot scale with video length, forcing a growing amount of historical information to compete for limited memory capacity. To address these limitations, we propose **WorldState**, a scalable implicit memory model built upon a hybrid attention backbone. WorldState decouples memory erasing and writing for fine-grained memory editing and uniformly organizes history into a logarithmically growing set of temporally isolated states. Completed historical states remain subject to channel-wise decay but are excluded from subsequent erasure and writing, reducing repeated overwriting while preserving adaptive forgetting. Context-aware memory routing further retrieves relevant historical states according to the current generation context. Extensive experiments demonstrate that WorldState substantially improves long-term memory retention and spatial revisit consistency while maintaining scalable long-horizon generation efficiency, enabling high-fidelity and spatially coherent video generation over extended interaction horizons.

1 INTRODUCTION

Recent advances in video generation (Hong et al., 2022; HaCohen et al., 2026; Wan et al., 2025) have enabled interactive video world models that generate visual observations conditioned on user inputs and camera movements, supporting applications in autonomous driving simulation (Zhang et al., 2025a; Wu et al., 2025; Russell et al., 2025), embodied intelligence (Yuan et al., 2026), and interactive gaming (Valevski et al., 2025; Yu et al., 2025b). Over long interaction horizons, these models must preserve information about previously explored environments so that revisited regions remain consistent in appearance and spatial structure. Scalable storage and retrieval of long-term spatial information is therefore a central challenge for interactive video world models.

Existing video world models primarily leverage historical information through either explicit or implicit memory. Explicit memory preserves historical frames or 3D spatial representations for subsequent generation (Team et al., 2026c; Gao et al., 2026; Yu et al., 2025a; Xiao et al., 2026; Yu et al., 2026; Zhao et al., 2025; Li et al., 2025). However, frame-based methods are often restricted to limited contexts due to computational costs, making long-range consistency difficult to maintain, while 3D-based methods rely on depth and camera pose estimation, whose errors may accumulate during continual updates. In contrast, implicit memory (Cai et al., 2026; Zhang et al., 2025b; Wang et al., 2026b) compresses history into internal model states without maintaining an ever-growing history cache (Wei et al., 2026; Liu et al., 2025; Dalal et al., 2025; Yu et al., 2025c; Zhu et al., 2026). In particular, recurrent linear attention enables linear-time sequence modeling, providing an efficient foundation for long-horizon generation. Recent large language models, such as Kimi K3 (Team et al., 2026b; Hatamizadeh et al., 2026) and Qwen3-Next (Team, 2026), further combine linear and full attention to balance long-context efficiency with precise information access. This raises a natural question: *can efficient implicit recurrent states address the more challenging problem of long-term spatial revisitation in interactive video world models?*

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.

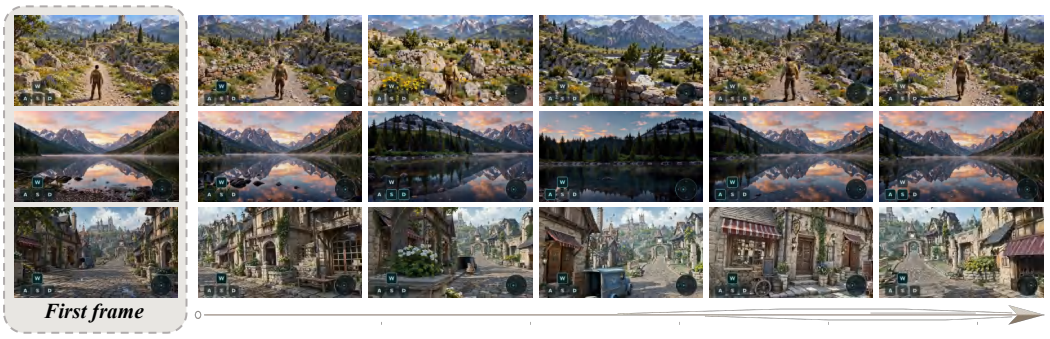


Figure 1: **Long-horizon interactive world simulation by WorldState via implicit memory.** WorldState preserves historical scenes through temporally isolated recurrent states and dynamically retrieves relevant memories based on the current context. This allows the model to maintain natural dynamics and responsive interaction while preserving stable scene structure and visual consistency over long-horizon exploration and spatial revisits.

Recurrent linear memory faces two limitations in long-horizon video generation. In the standard Gated DeltaNet (GDN) (Yang et al., 2025) formulation used for frame-wise video recurrence, a shared update gate controls both memory erasure and writing. Spatial-token normalization further introduces an asymmetric scaling between these operations: erasure depends quadratically on the key magnitude, whereas writing depends on it linearly. Effective erasure is therefore weakened relative to writing, making outdated associations difficult to remove selectively (Zhu, 2026). The model consequently relies more heavily on global decay, which indiscriminately attenuates historical information. An additional limitation stems from representing the entire history with a single fixed-capacity state. As the sequence grows, old and new information increasingly compete for limited capacity, causing distant memories to be disrupted by subsequent writes.

Based on this analysis, we propose **WorldState**, which redesigns recurrent linear memory along three key dimensions: memory editing, preservation, and retrieval. First, we decouple erasing from writing and introduce channel-wise decay to frame-wise video recurrence, enabling finer-grained control over memory updates. We further expand the single state into a logarithmically growing set of temporally isolated states and uniformly distribute historical observations among them. Once a state is completed, it retains channel-wise decay but is excluded from subsequent erasure and writing. This decay-only historical update prevents later observations from directly overwriting completed memories while preserving adaptive forgetting. Our uniform allocation does not determine the importance of historical observations solely by their temporal distance. To avoid cross-period interference caused by indiscriminately aggregating all states, we further introduce context-aware memory routing, which dynamically adjusts the read weights of different historical states according to the current observation and selectively retrieves information relevant to the current scene. As illustrated in Fig. 1, WorldState transforms a single continuously overwritten recurrent state into a scalable implicit memory that supports flexible editing, long-term preservation, and selective retrieval. Extensive experiments demonstrate that WorldState maintains high generation efficiency and visual quality while achieving superior or competitive performance against strong baselines on SANA-Bench (Zhu et al., 2026) and RealEstate10K (Zhou et al., 2018).

Our main contributions are as follows:

- We introduce a decoupled memory editing mechanism that removes the tight coupling between erasing and writing, allowing recurrent memory to independently control the removal of historical information and the incorporation of new content.
- We develop a scalable implicit memory architecture that organizes history into logarithmically growing, temporally isolated states and dynamically routes relevant historical states according to the current context, enabling long-term preservation and selective retrieval during spatial revisits.
- Extensive experiments demonstrate that WorldState substantially improves long-term memory retention and spatial revisit consistency while maintaining competitive generation efficiency and visual quality.

2 Related Work

2.1 Video World Models

Video world models (Jiang et al., 2026; Team et al., 2026a) aim to predict and simulate future visual states from historical observations and control signals such as actions, camera trajectories, or user inputs. With the rapid development of diffusion (Nie et al., 2025; 2026) and video generation models (Lipman et al., 2023; HaCohen et al., 2026; Hong et al., 2022), this line of research has progressed from short-horizon video prediction toward autoregressive and interactive generation, with applications in games, autonomous driving, and embodied simulation (Mao et al., 2025; Team et al., 2025; HY-World et al., 2026; Gao et al., 2026). Diffusion Transformers (DiTs) have been widely adopted as backbones in recent video generation models (Peebles & Xie, 2023), yet their standard full attention scales quadratically with sequence length, making long-horizon interaction prohibitively expensive. To improve scalability, recent work has explored sliding-window attention, autoregressive block-wise generation, cache compression and linear attention (Chen et al., 2025; 2026; Yu et al., 2025c). Despite these advances in generation length and efficiency, maintaining distant history and consistent world states under long interactions, large camera motions (Zhang et al., 2026), and spatial revisits remains a major challenge.

2.2 Memory in Long Video Generation

Memory is essential for maintaining long-term consistency and supporting spatial revisits in interactive world models. Existing approaches can be broadly categorized into explicit and implicit memory. Explicit methods directly preserve historical frames (Wang et al., 2026d; Yu et al., 2025a; Xiao et al., 2026; Yu et al., 2026; Sun et al., 2026), KV features (Ji et al., 2025), or structured spatial representations such as 3D point clouds (Zhao et al., 2025; Shen et al., 2026; Huang et al., 2025) and surfels (Li et al., 2025) constructed from depth and camera poses. These memories can be reused through retrieval, field-of-view (FOV) overlap, or geometric projection, providing direct access to past observations and strong multi-view consistency. However, they often rely on accurate geometric estimation, while their storage and retrieval costs may increase with generation length.

To reduce the cost of long-term memory, another line of work adopts compressed or recurrent implicit memory (Wei et al., 2026; Huang et al., 2026; Wu et al., 2026), including temporal packing (Zhang et al., 2025b), learned memory compressors (Liu et al., 2025) and linear attention (Yu et al., 2025c; Dalal et al., 2025; Po et al., 2025; Savov et al., 2025). These methods summarize growing histories into compact representations without requiring explicit 3D structures, offering better computational and memory efficiency. Nevertheless, finite context windows inevitably discard distant observations, while fixed-capacity recurrent states must continuously absorb new information, making remote memories vulnerable to overwriting and interference (Guo et al., 2026). Our work follows the implicit-memory paradigm and focuses on improving how long-term history is edited, preserved, and accessed for long-horizon generation and spatial revisits in interactive world models.

3 Method

3.1 Preliminary

Interactive video world models need to continuously maintain historical states during autoregressive generation. In SANA-WM (Zhu et al., 2026), each linear-attention head maintains a recurrent state $\mathbf{S}_t \in \mathbb{R}^{D\times D}$, which is updated by a gated delta rule:

$$\mathbf{S}_t = \mathbf{S}_{t-1}\mathbf{M}_t + \mathbf{U}_t, \qquad \mathbf{M}_t = \gamma_t \left(\mathbf{I} - \hat{\mathbf{K}}_t \boldsymbol{\beta}_t \hat{\mathbf{K}}_t^\top\right), \qquad \mathbf{U}_t = \mathbf{V}_t \boldsymbol{\beta}_t \hat{\mathbf{K}}_t^\top. \tag{1}$$

Here, γ_t is a frame-level decay factor, and $\boldsymbol{\beta}_t$ denotes the per-token update gate. A key limitation of this formulation is that $\boldsymbol{\beta}_t$ simultaneously controls the erasure of old information and the writing of new content, preventing the two operations from being adjusted independently.

In SANA-WM, the keys are further rescaled as

$$\hat{\mathbf{K}}_t = c\,\bar{\mathbf{K}}_t, \qquad c = (DS)^{-1/2}, \tag{2}$$

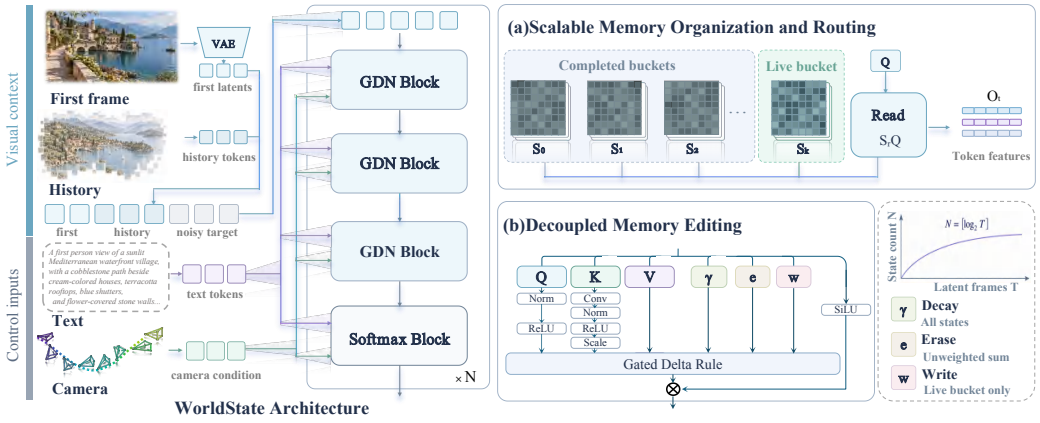


Figure 2: **Overview of WorldState architecture. (a) Scalable memory organization and routing:** History is uniformly partitioned into logarithmically growing temporal states, with context-aware routing for memory retrieval. **(b) Decoupled memory editing:** Independent erase and write gates with channel-wise decay enable fine-grained recurrent memory updates.

where D is the head dimension and S is the number of spatial tokens per frame. The erasure term depends quadratically on the key magnitude, whereas the write term depends only linearly on it. Consequently, when $\hat{\mathbf{K}}_t$ is scaled by a factor of c, the write strength is attenuated by c, while the erasure strength is attenuated by c^2. Since video frames contain a large number of spatial tokens, $c \ll 1$, causing targeted erasure to be substantially suppressed relative to writing (Zhu, 2026).

As a result, the recurrent state is not incapable of forgetting, but instead relies more heavily on the global decay factor γ_t for indiscriminate forgetting. As the sequence grows, distant information is either increasingly interfered with by subsequent writes or gradually weakened by global decay, which exacerbates memory interference during long-horizon spatial revisits.

3.2 Decoupled Memory Editing

As shown in equation 1, $\mathbf{M}_t$ controls the retention and erasure of the previous state, while $\mathbf{U}_t$ writes new information from the current frame into memory. In the original formulation, both operations share the same gate $\boldsymbol{\beta}_t$, coupling erasure and writing to the same update strength. However, these two operations correspond to fundamentally different memory-editing decisions: erasure acts on the key side to determine which existing memory associations should be removed, whereas writing acts on the value side to determine which new content should be injected into the state. Sharing a single gate therefore unnecessarily restricts the flexibility of recurrent memory editing.

We decouple the shared gate into an erase gate $\mathbf{e}_t$ and a write gate $\mathbf{w}_t$, which independently control the removal of old information and the injection of new content (Fig. 2(b)). Meanwhile, the original frame-level scalar decay applies a uniform forgetting rate to all memory channels and therefore cannot accommodate their different retention requirements. We further extend it to a channel-wise decay vector $\boldsymbol{\gamma}_t$, whose corresponding diagonal decay matrix $\mathrm{Diag}(\boldsymbol{\gamma}_t)$ allows individual memory channels to learn distinct temporal scales for finer-grained retention and forgetting. Keeping the original recurrent formulation unchanged, we redefine the state-transition and write terms as

$$\mathbf{M}_t = \mathrm{Diag}(\boldsymbol{\gamma}_t)\big[\mathbf{I} - (\mathbf{e}_t \odot \hat{\mathbf{K}}_t)\hat{\mathbf{K}}_t^\top\big], \qquad \mathbf{U}_t = (\mathbf{w}_t \odot \mathbf{V}_t)\hat{\mathbf{K}}_t^\top. \tag{3}$$

Here, $\odot$ denotes element-wise multiplication. The erase gate selectively modulates historical associations along the key dimension, while the write gate controls which new information is injected along the value dimension. Together with channel-wise decay, this formulation provides finer-grained control over memory retention, erasure, and writing, enabling more flexible and precise recurrent memory editing.

3.3 Scalable Temporal State Isolation

Decoupled memory editing improves the update of an active recurrent state, but it does not resolve the fundamental capacity limitation of fixed-size recurrent memory. As the sequence grows,

an increasing amount of historical information must compete within the same fixed-capacity state, leading to accumulated memory interference and degradation of distant information. On the other hand, allowing memory capacity to grow linearly with sequence length approaches the opposite extreme of explicitly retaining the full history, sacrificing the efficiency advantage of recurrent linear attention for long-sequence modeling.

To address this limitation, we allow memory capacity to grow logarithmically with the generation horizon. We follow the standard bounded-horizon setting of controllable video world models, where the requested generation duration is specified before rollout. Given T latent frames after VAE encoding, WorldState maintains $N = \lceil \log_2 T \rceil$ recurrent states. This logarithmic growth provides a natural trade-off between fixed-capacity recurrent memory and explicit memory that grows linearly with history. Compared with a single recurrent state, the memory capacity increases from $O(1)$ to $O(\log T)$, allowing long-term memory to gradually expand with the generation horizon. Meanwhile, since only a logarithmic number of states need to be accessed at each step, the overall computational complexity remains $O(T \log T)$, avoiding the quadratic cost of modeling the full history.

We further partition the T latent frames uniformly into N contiguous temporal buckets:

$$|\mathcal{B}_i| \in \left\{ \left\lfloor \frac{T}{N} \right\rfloor, \left\lceil \frac{T}{N} \right\rceil \right\}, \qquad \sum_{i=1}^{N} |\mathcal{B}_i| = T. \tag{4}$$

Here, $\mathcal{B}_i$ denotes the contiguous temporal interval assigned to the i-th recurrent state. In spatial revisit scenarios, regions observed early in the interaction may become important again after long periods of exploration. We therefore do not assume that the importance of historical information monotonically decreases with temporal distance. Instead, uniform partitioning allocates approximately equal memory capacity to different historical periods, avoiding coarser representations solely because an observation occurred earlier. Comparisons with alternative temporal allocation strategies are provided in the App. C. For a frame t belonging to temporal bucket $\mathcal{B}_j$, the live state receives the full memory update, whereas completed historical states retain only channel-wise decay. To account for their contributions to erasure without directly modifying them, we update the states as

$$\mathbf{S}_t^{\text{live}} = \mathbf{S}_{t-1}^{\text{live}} \mathbf{M}_t + \mathbf{U}_t + \sum_{i<j} \mathbf{S}_{t-1}^{(i)} \left[\mathbf{M}_t - \text{Diag}(\boldsymbol{\gamma}_t) \right], \qquad \mathbf{S}_t^{(i)} = \mathbf{S}_{t-1}^{(i)} \, \text{Diag}(\boldsymbol{\gamma}_t), \quad i < j. \tag{5}$$

The summation term incorporates erase corrections from completed states. Thus, erasure is computed from both the live state and all completed historical states, while the correction and new write are applied only to the live state. Completed states remain subject to channel-wise decay, allowing their contributions to attenuate over time, but are not directly modified by subsequent erasure or writing operations. At each step, completed historical states are weighted by the context-aware routing mechanism introduced in the following section and jointly read with the active state.

3.4 Context-Aware Memory Routing

The memory routing module in Fig. 2(a) uses the current-frame query features to assign context-dependent read weights to historical temporal buckets. For each attention head, the routing module takes spatially averaged queries and produces positive weights $r_{t,i}$ for the N_s completed buckets, normalized to have unit mean. The weighted historical states are combined with the live state and read using the transformed current-frame queries $\hat{\mathbf{Q}}_t$, yielding the readout numerator $\mathbf{O}_t^{\text{num}}$ before normalization:

$$\frac{1}{N_s} \sum_{i=1}^{N_s} r_{t,i} = 1, \qquad \mathbf{O}_t^{\text{num}} = \left(\mathbf{S}_t^{\text{live}} + \sum_{i=1}^{N_s} r_{t,i} \mathbf{S}_t^{(i)} \right) \hat{\mathbf{Q}}_t. \tag{6}$$

The live state is always assigned a weight of 1, while the historical bucket weights are initialized to 1 so that learning starts from uniform readout. All spatial tokens within the same frame and attention head share the same bucket-level routing weights, but each token still uses its own query to read from the aggregated memory.

The routing mechanism affects only memory readout and does not participate in memory updates. The erase correction is computed from the unweighted sum of the decayed live and completed historical states. This correction and the new write are applied only to the live state, while completed

historical states undergo channel-wise decay only. The normalization stream is also excluded from bucket-level routing. This separation allows the model to adaptively retrieve the most relevant historical periods without letting retrieval preferences directly alter how memory is stored or updated.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Implementation details. WorldState is built upon SANA-WM, a world model with a hybrid attention architecture. We train the model on RealEstate10K (Zhou et al., 2018) and ABot-World (Jiang et al., 2026). The training resolution is 1280×704, and each sample contains 257 RGB frames, corresponding to 33 latent frames after VAE encoding with a temporal downsampling factor of 8. We train on 8 H100 GPUs using the AdamW optimizer with a learning rate of 1×10^{-4} and a global batch size of 8. For results with refinement, we apply the same independent LTX-2 refiner, consisting of 48 Transformer layers. Further training and inference details are provided in the App. A.

Method	Res	#G	Simple-Trajectory			Hard-Trajectory			Efficiency	
			PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	Mem↓	Tput↑
Gen3C	**720p**	1	15.460	0.3918	0.4349	14.620	0.3397	0.4836	60.59	5.83
Spatia	**720p**	1	15.750	0.3971	0.4204	14.880	0.3599	0.4652	75.10	4.26
Infinite-World	480p	1	12.380	0.2540	0.5930	12.100	0.2463	0.6007	61.04	6.37
Matrix-Game 3.0	**720p**	8	11.520	0.3086	0.5509	12.500	0.3352	0.5066	130.08	3.67
SANA-WM	**720p**	1	14.270	0.3411	0.4617	14.330	0.3355	0.4759	42.2	25.5
SANA-WM + refiner	**720p**	1	14.710	0.3000	0.4720	14.930	0.3109	0.4662	66.5	23.3
WorldState	**720p**	1	16.760	0.4145	0.3439	16.250	0.3942	0.3637	49.1	23.7
WorldState + refiner	**720p**	1	16.390	0.3653	0.3798	16.420	0.3705	0.3690	73.2	21.3

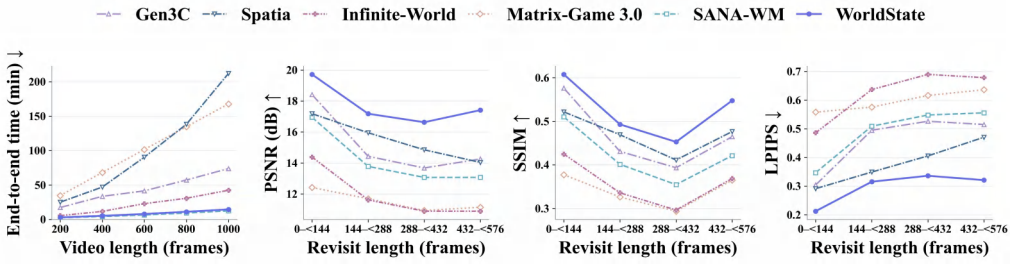
Table 1: Quantitative comparison of long-term revisit consistency and inference efficiency on the SANA 1-min benchmark. Mem. denotes peak GPU memory usage in GB, and Tput. denotes throughput in videos per hour.

Evaluation. We compare our method against Gen3C (Ren et al., 2025), Spatia (Zhao et al., 2025), Infinite-World (Wu et al., 2026), Matrix-Game 3.0 (Wang et al., 2026d) and SANA-WM (Zhu et al., 2026). Our main evaluation is conducted on SANA-Bench (Zhu et al., 2026), a long-video TI2V benchmark containing 80 scenes across four categories, with simple and hard camera trajectories. Given a text prompt, an initial frame, and a complete camera trajectory, each method generates a 961-frame video at 16 FPS under the same protocol. We evaluate spatial memory using revisit pairs and compute PSNR, SSIM, and LPIPS between frames generated at revisited viewpoints. Camera controllability follows SANA-WM, using Pi3 (Wang et al., 2026c) to estimate poses and reporting RotErr, TransErr, and CamMC. Generation quality is evaluated using eight VBench (Huang et al., 2023) dimensions, covering subject consistency (SC), background consistency (BC), temporal flickering (TF), motion smoothness (MS), aesthetic quality (AQ), imaging quality (IQ), dynamic degree (DD), and overall consistency (OC), as detailed in App. B.

We further conduct ground-truth evaluation on RealEstate10K. Following Spatia (Zhao et al., 2025), we randomly sample 200 videos from the test split and evaluate from two perspectives. **Novel View Synthesis** measures PSNR, SSIM, and LPIPS between frames generated along the reference camera trajectory and the corresponding ground-truth frames. **Closed-Loop** constructs camera trajectories that eventually return to the initial viewpoint and compares the generated return frame with the initial observation using PSNR_C, SSIM_C, and LPIPS_C, evaluating the recovery of previously observed scene content after long-horizon exploration.

4.2 QUANTITATIVE RESULTS

Long-horizon revisit consistency. Table 1 shows that WorldState substantially improves long-term revisit consistency under both Simple and Hard trajectory settings. Compared with SANA-WM, WorldState increases PSNR from 14.27 to 16.76 on the Simple split and from 14.33 to 16.25 on the Hard split, while also achieving substantially lower LPIPS. These results indicate that temporally isolated recurrent states effectively alleviate long-term memory overwriting and interference, en-



(a) Latency comparison on H100 **(b) Revisit quality across video lengths**

Figure 3: Scaling performance across increasing generation horizons. (a) End-to-end video generation time under the same resolution. (b) Long-horizon scene consistency measured between frames generated at matched revisited viewpoints.

abling more reliable recovery of historical scene content during spatial revisits. Meanwhile, WorldState introduces only moderate inference overhead and remains competitive with existing baselines in terms of inference efficiency and memory consumption.

Method	Camera Pose ↓			VBench ↑								
	R	T	CMC	SC	BC	TF	MS	AQ	IQ	DD	OC	Overall
Simple-Trajectory Split												
Gen3C	14.51	1.92	1.99	0.8150	0.8811	0.9678	0.9912	0.5076	0.5982	0.7375	0.2298	0.7674
Spatia	5.85	1.63	1.65	0.8803	0.9150	0.9628	0.9787	0.5629	0.7005	0.4500	0.2465	0.7797
Infinite-World	16.47	2.23	2.41	0.8369	0.8568	0.9578	0.9856	0.5256	0.6927	0.8875	0.2487	0.7882
Matrix-Game 3.0	11.72	1.76	1.84	0.7918	0.8788	0.9347	0.9760	0.4257	0.7173	0.9750	0.2038	0.7650
SANA-WM	5.17	1.55	1.57	0.8746	0.9177	0.9524	0.9780	0.5557	0.6938	0.6375	0.2479	0.7869
SANA-WM + refiner	5.15	1.40	1.42	0.8875	0.9308	0.9633	0.9862	0.5866	0.7226	0.6125	0.2485	0.8080
WorldState	3.16	1.25	1.16	0.8856	0.9244	0.9579	0.9816	0.5572	0.6966	0.6375	0.2422	0.7951
WorldState + refiner	4.21	1.38	1.39	0.8926	0.9350	0.9666	0.9881	0.5871	0.7232	0.6125	0.2481	0.8124
Hard-Trajectory Split												
Gen3C	19.97	1.83	1.97	0.7985	0.8740	0.9591	0.9902	0.4971	0.6002	0.9750	0.2337	0.7759
Spatia	16.34	1.71	1.83	0.8654	0.9080	0.9572	0.9773	0.5489	0.6903	0.7625	0.2458	0.7929
Infinite-World	20.29	1.77	1.93	0.8329	0.8823	0.9570	0.9750	0.5401	0.6713	0.9500	0.2477	0.7904
Matrix-Game 3.0	21.83	1.77	1.94	0.7998	0.8799	0.9368	0.9762	0.4276	0.7136	0.9875	0.2102	0.7684
SANA-WM	7.61	1.51	1.56	0.8591	0.9076	0.9457	0.9757	0.5356	0.6891	0.9250	0.2484	0.7963
SANA-WM + refiner	8.31	1.39	1.44	0.8712	0.9214	0.9573	0.9847	0.5670	0.7145	0.8875	0.2473	0.8167
WorldState	3.99	1.06	1.08	0.8709	0.9117	0.9515	0.9799	0.5393	0.6913	0.8500	0.2453	0.7990
WorldState + refiner	5.12	1.12	1.15	0.8805	0.9242	0.9601	0.9868	0.5714	0.7157	0.8250	0.2480	0.8173

Table 2: Camera pose accuracy and VBench results on the 1-min benchmark.

Camera controllability and generation quality. WorldState also demonstrates strong camera controllability and generation quality in Table 2. On the Simple-Trajectory split, compared with SANA-WM, WorldState reduces RotErr, TransErr, and CamMC from 5.17, 1.55, and 1.57 to 3.16, 1.25, and 1.16, respectively, demonstrating more accurate trajectory following. Meanwhile, WorldState achieves competitive VBench Overall performance and outperforms Matrix-Game 3.0 and the other unrefined baselines, particularly in subject and background consistency, while remaining competitive in visual quality. Using the same long-video refiner as SANA-WM further improves the visual quality of WorldState. The refiner serves primarily as a visual enhancement module and does not participate in long-term memory modeling. These results demonstrate that WorldState strengthens long-term memory without compromising camera controllability or generation quality.

Ground-truth evaluation. As shown in Table 3, WorldState achieves substantial improvements on both novel-view synthesis and closed-loop revisit evaluation on RealEstate10K. For novel-view synthesis, higher PSNR and SSIM together with lower LPIPS indicate that WorldState generates scene content more consistent with the corresponding ground-truth observations along the reference trajectory. For closed-loop evaluation, WorldState more accurately recovers the scene at the initial viewpoint after long-horizon camera motion, further demonstrating its ability to preserve and reuse distant spatial information during generation.

Scaling Analysis. Figure 3(a) compares the end-to-end generation time at different video lengths on a single H100 GPU under the same settings. WorldState maintains favorable inference efficiency as the generation horizon increases. Fig. 3(b) further reports long-term memory performance under different revisit distances. Specifically, we select historical videos of varying lengths from RealEstate10K and provide the corresponding observations and camera trajectories to revisit previously observed regions. Although revisit performance gradually degrades for all methods as the revisit distance increases, WorldState consistently maintains stronger and more stable performance, indicating its ability to better preserve and recover early scene information over extended horizons.

Method	Novel View Synthesis			Closed-Loop		
	PSNR↑	SSIM↑	LPIPS↓	PSNR_C↑	SSIM_C↑	LPIPS_C↓
Gen3C	16.217	0.495	0.409	16.340	0.489	0.416
Spatia	16.381	0.486	0.329	16.330	0.481	0.348
Infinite-World	12.833	0.373	0.573	11.800	0.350	0.637
Matrix-Game 3.0	11.935	0.346	0.573	10.799	0.315	0.646
SANA-WM	15.186	0.446	0.437	13.240	0.386	0.545
WorldState	**18.305**	**0.542**	**0.269**	**16.404**	**0.509**	**0.347**

Table 3: Quantitative comparison on RealEstate10K. All metrics are computed against ground truth for novel-view synthesis and closed-loop evaluation, while PSNR_C, SSIM_C, and LPIPS_C measure closed-loop spatial consistency between the initial and final revisited frames.

4.3 Qualitative Results

We present a visual comparison with representative world models in Fig. 4 under one-minute long-horizon interaction and spatial revisit scenarios. Gen3C and Spatia rely on explicit geometric or spatial memories, but remain susceptible to errors in geometry estimation and historical reprojection during long rollouts. Infinite-World supports long-horizon generation through compressed historical representations, yet details from distant historical scenes may gradually deteriorate as generation proceeds. Matrix-Game 3.0 and SANA-WM generate coherent dynamics and respond to camera motion, but exhibit noticeable drift when revisiting previously observed regions. In contrast, WorldState preserves historical scene information through scalable, temporally partitioned recurrent states and selectively accesses relevant states through query-conditioned routing. It maintains global landmarks after hundreds of frames and preserves scene consistency during long-range loop closures where baseline methods exhibit noticeable drift, illustrating the effectiveness of scalable implicit memory in mitigating long-term memory overwriting and interference. Additional qualitative results and full-length videos are provided in App. D and the supplementary material, respectively.

4.4 Ablation Studies

We conduct controlled ablations on a representative subset of both SANA-Bench trajectory splits to evaluate each WorldState component and different temporal memory allocation strategies. All variants and our base model are trained using the same training setup for a fair comparison.

E	I	R	PSNR↑	SSIM↑	LPIPS↓	VBench↑
			14.66	0.3835	0.3928	0.6737
✓			15.53	0.3809	0.4067	0.6785
✓	✓		16.02	0.3950	0.3877	0.7806
✓	✓	✓	**16.28**	**0.4170**	**0.3756**	**0.7852**

(a) Component ablation.

Setting	#Sn	Decay	Erase	Write	PSNR↑	SSIM↑	LPIPS↓	VBench↑
Single-State	1	✓	✓	✓	14.66	0.3835	0.3928	0.6737
Capacity-Matched	1	✓	✓	✓	16.08	0.4112	0.3894	0.7819
True Sealing	N	×	×	×	14.88	0.3516	0.4821	0.6919
Historical Erasure	N	✓	✓	×	16.15	**0.4189**	0.3828	0.7801
Decay-Only	N	✓	×	×	**16.28**	0.4170	**0.3756**	**0.7852**

(b) Historical-state update ablation.

Table 4: Ablation studies of WorldState. E, I, and R denote decoupled memory editing, temporal state isolation, and context-aware routing, respectively; #Sn denotes the number of recurrent states. For E+I, the trained router is disabled only at inference by setting all routing weights to one.

Effectiveness of long-term memory design. Table 4(a) summarizes the component ablation of WorldState. The complete model achieves the best overall performance, increasing PSNR from 14.66 to 16.28 and VBench from 0.6737 to 0.7852 over the single-state baseline. These results demonstrate the complementary roles of decoupled memory editing, temporal state isolation, and context-aware routing in improving memory updates, reducing cross-period interference, and selectively retrieving relevant historical information.



Figure 4: One-minute visual comparison with five representative baselines. WorldState maintains scene structure and appearance across long temporal gaps and out-of-view intervals.

Memory capacity and historical-state updates. Table 4(b) shows that enlarging a single global state to match total memory capacity outperforms the fixed-capacity single-state baseline but still underperforms temporally partitioned WorldState. This demonstrates that the improvement does not arise solely from increased memory capacity; temporal isolation is also crucial for reducing cross-period interference. More importantly, a monolithic state has a fixed size and cannot allocate capacity on demand for different generation horizons, preventing adaptive growth. Completely sealing historical states prevents early memories from being overwritten, but also permanently preserves redundant and outdated associations at their original scale. As more historical states accumulate, these stale components persist in the aggregated readout. Channel-wise decay instead gradually suppresses unreliable components without introducing new content. In contrast, directly applying key-conditioned erasure to completed states allows future frames to repeatedly modify stored memories, thereby undermining temporal isolation. WorldState therefore retains only channel-wise decay for completed states, achieving the best overall balance between limiting stale-information accumulation and preventing continual modification of historical content.

5 Conclusion

We propose WorldState, a scalable implicit memory architecture for interactive video world models, designed to mitigate memory interference and forgetting arising from continually writing new observations into a shared, fixed-capacity recurrent state. WorldState enables stable preservation and accurate recovery of long-term memories by decoupling memory erasure from writing, isolating historical states from different temporal periods, and selectively retrieving relevant information based on the current context. Experimental results demonstrate that, while maintaining efficient generation, WorldState substantially improves long-term spatial revisit consistency and visual quality, providing an effective approach toward interactive world models with persistent memory.

LLM USE STATEMENT

Generative AI tools were used to assist with code development, support the analysis of experimental results and provide suggestions for improvement, and finally refine the completed manuscript to improve its clarity and readability.

ETHICS STATEMENT

This work studies long-term memory mechanisms for interactive video world models. It does not involve human-subject experiments or the collection of new personal data. Our experiments use existing research datasets and generated visual content. We encourage responsible use of the technology, compliance with applicable dataset licenses, and the adoption of appropriate safeguards, such as content provenance, watermarking, and access control, in real-world applications.

REPRODUCIBILITY STATEMENT

We describe the WorldState architecture and provide the complete formulations of decoupled memory editing, temporal state isolation, and context-aware memory routing in Sec. 3. Sec. 4.1 specifies the training datasets, compared methods, evaluation benchmarks, and metrics. Detailed data preprocessing, backbone configurations, tensor dimensions, gate parameterization, numerical precision, optimization settings, and the two-stage training procedure are provided in App. A.1. The sampling configuration and evaluation protocols for spatial revisit consistency, camera controllability, visual quality, and long-horizon degradation are documented in App. B. We also provide component, memory-update, capacity, and temporal-allocation ablations to clarify the contribution of each design choice.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan Yuille, Leonidas Guibas, Maneesh Agrawala, Lu Jiang, and Gordon Wetzstein. Mixture of contexts for long video generation. In *ICLR*, 2026.

Junsong Chen, Yuyang Zhao, Jincheng Yu, Ruihang Chu, Junyu Chen, Shuai Yang, Xianbang Wang, Yicheng Pan, Daquan Zhou, Huan Ling, Haozhe Liu, Hongwei Yi, Hao Zhang, Muyang Li, Yukang Chen, Han Cai, Sanja Fidler, Ping Luo, Song Han, and Enze Xie. Sana-video: Efficient video generation with block linear diffusion transformer, 2025. URL https://arxiv.org/abs/2509.24695.

Junsong Chen, Jincheng Yu, Yitong Li, Shuchen Xue, Haozhe Liu, Jingyu Xin, Yuyang Zhao, Tian Ye, Zhangjie Wu, Zian Wang, Daquan Zhou, Ping Luo, Song Han, and Enze Xie. Sana-video 2.0: Hybrid linear attention with attention residuals for efficient video generation, 2026. URL https://arxiv.org/abs/2607.21553.

Karan Dalal, Daniel Koceja, Gashon Hussein, Jiarui Xu, Yue Zhao, Youjin Song, Shihao Han, Ka Chun Cheung, Jan Kautz, Carlos Guestrin, Tatsunori Hashimoto, Sanmi Koyejo, Yejin Choi, Yu Sun, and Xiaolong Wang. One-minute video generation with test-time training, 2025. URL https://arxiv.org/abs/2504.05298.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, Ka Leong Cheng, Haojie Zhang, Jian Gao, Tianrui Feng, Yuzheng Liu, Yao Yao, Yinghao Xu, Xing Zhu, Yujun Shen, and Hao Ouyang. Infinite worlds with versatile interactions, 2026. URL https://arxiv.org/abs/2607.07534.

Han Guo, Songlin Yang, Tarushii Goel, Eric P. Xing, Tri Dao, and Yoon Kim. Log-linear attention, 2026. URL https://arxiv.org/abs/2506.04761.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. Ltx-2: Efficient joint audio-visual foundation model, 2026. URL https://arxiv.org/abs/2601.03233.

Ali Hatamizadeh, Yejin Choi, and Jan Kautz. Gated deltanet-2: Decoupling erase and write in linear attention, 2026. URL https://arxiv.org/abs/2605.22791.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers, 2022. URL https://arxiv.org/abs/2205.15868.

Junchao Huang, Xinting Hu, Boyao Han, Shaoshuai Shi, Zhuotao Tian, Tianyu He, and Li Jiang. Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft, 2025. URL https://arxiv.org/abs/2510.03198.

Kaiyi Huang, Yukun Huang, Yu Li, Jianhong Bai, Xintao Wang, Zinan Lin, Xuefei Ning, Jiwen Yu, Pengfei Wan, Yu Wang, and Xihui Liu. Cinescene: Implicit 3d as effective scene representation for cinematic video generation, 2026. URL https://arxiv.org/abs/2602.06959.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models, 2023. URL https://arxiv.org/abs/2311.17982.

Team HY-World, Chenjie Cao, Xuhui Zuo, Zhenwei Wang, Yisu Zhang, Junta Wu, Zhenyang Liu, Yuning Gong, Yang Liu, Bo Yuan, Chao Zhang, Coopers Li, Dongyuan Guo, Fan Yang, Haiyu Zhang, Hang Cao, Jianchen Zhu, Jiaxin Lin, Jie Xiao, Jihong Zhang, Junlin Yu, Lei Wang, Lifu Wang, Lilin Wang, Linus, Minghui Chen, Peng He, Penghao Zhao, Qi Chen, Rui Chen, Rui Shao, Sicong Liu, Wangchen Qin, Xiaochuan Niu, Xiang Yuan, Yi Sun, Yifei Tang, Yifu Sun, Yihang Lian, Yonghao Tan, Yuhong Liu, Yuyang Yin, Zhiyuan Min, Tengfei Wang, and Chunchao Guo. Hy-world 2.0: A multi-modal world model for reconstructing, generating, and simulating 3d worlds, 2026. URL https://arxiv.org/abs/2604.14268.

Sihui Ji, Xi Chen, Shuai Yang, Xin Tao, Pengfei Wan, and Hengshuang Zhao. Memflow: Flowing adaptive memory for consistent and efficient long video narratives, 2025. URL https://arxiv.org/abs/2512.14699.

Fan Jiang, Zhaoxu Sun, Mengchao Wang, Ziyu Zhu, Chiyu Wang, Yunpeng Zhang, Wenlin Liu, Yun Wang, Xue Zheng, Rui Sun, Junfeng Ni, Hongyu Pan, Zhongxu Sun, Fei Yu, Zengye Ge, Mengmeng Du, Nianfei Fan, Mingchao Sun, Yu Liu, Yongchang, Yanqing Zhu, Jiahang Wang, Ning Ying, Yuze Xuan, Di Yang, Zhicheng Liu, Zhe Gao, Tingbing Xu, Jiacheng Sui, Wenjin Yang, Junnan Lai, Shufeng Liu, Yuan Liu, Zheng Zhou, Yingliang Peng, Dawei Cao, Kaifeng Sheng, Yuxiang Cai, Fei Lu, Mu Xu, and Ning Guo. Abot-world-0: Infinite interactive world rollout on a single desktop gpu, 2026. URL https://arxiv.org/abs/2607.19191.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory, 2025. URL https://arxiv.org/abs/2506.18903.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Zhiheng Liu, Xueqing Deng, Shoufa Chen, Angtian Wang, Qiushan Guo, Mingfei Han, Zeyue Xue, Mengzhao Chen, Ping Luo, and Linjie Yang. Worldweaver: Generating long-horizon video worlds via rich perception, 2025. URL https://arxiv.org/abs/2508.15720.

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, 2025. URL https://arxiv.org/abs/2512.22096.

Wenbo Nie, Lang Nie, Chunyu Lin, Jingwen Chen, Ke Xing, Jiyuan Wang, and Kang Liao. Beyond wide-angle images: Structure-to-detail video portrait correction via unsupervised spatiotemporal adaptation, 2025. URL https://arxiv.org/abs/2504.00401.

Wenbo Nie, Zixiang Li, Renshuai Tao, Bin Wu, Yunchao Wei, and Yao Zhao. Cocodiff: Correspondence-consistent diffusion model for fine-grained style transfer, 2026. URL https://arxiv.org/abs/2602.14464.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Ryan Po, Yotam Nitzan, Richard Zhang, Berlin Chen, Tri Dao, Eli Shechtman, Gordon Wetzstein, and Xun Huang. Long-context state-space video world models, 2025. URL https://arxiv.org/abs/2505.20171.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models, 2025. URL https://arxiv.org/abs/2410.13720.

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, 2025. URL https://arxiv.org/abs/2503.03751.

Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving, 2025. URL https://arxiv.org/abs/2503.20523.

Nedko Savov, Naser Kazemi, Deheng Zhang, Danda Pani Paudel, Xi Wang, and Luc Van Gool. Statespacediffuser: Bringing long context to diffusion world models, 2025. URL https://arxiv.org/abs/2505.22246.

Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, Sanja Fidler, Jiahui Huang, Huan Ling, Jun Gao, and Xuanchi Ren. Lyra 2.0: Explorable generative 3d worlds, 2026. URL https://arxiv.org/abs/2604.13036.

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, 2026. URL https://arxiv.org/abs/2512.14614.

HunyuanWorld Team, Zhenwei Wang, Yuhao Liu, Junta Wu, Zixiao Gu, Haoyuan Wang, Xuhui Zuo, Tianyu Huang, Wenhuan Li, Sheng Zhang, Yihang Lian, Yulin Tsai, Lifu Wang, Sicong Liu, Puhua Jiang, Xianghui Yang, Dongyuan Guo, Yixuan Tang, Xinyue Mao, Jiaao Yu, Junlin Yu, Jihong Zhang, Meng Chen, Liang Dong, Yiwen Jia, Chao Zhang, Yonghao Tan, Hao Zhang, Zheng Ye, Peng He, Runzhou Wu, Minghui Chen, Zhan Li, Wangchen Qin, Lei Wang, Yifu Sun, Lin Niu, Xiang Yuan, Xiaofeng Yang, Yingping He, Jie Xiao, Yangyu Tao, Jianchen Zhu, Jinbao Xue, Kai Liu, Chongqing Zhao, Xinming Wu, Tian Liu, Peng Chen, Di Wang, Yuhong Liu, Linus, Jie Jiang, Tengfei Wang, and Chunchao Guo. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels, 2025. URL https://arxiv.org/abs/2507.21809.

InSpatio Team, Donghui Shen, Guofeng Zhang, Haomin Liu, Haoyu Ji, Hujun Bao, Hongjia Zhai, Jialin Liu, Jing Guo, Nan Wang, Siji Pan, Weihong Pan, Weijian Xie, Xianbin Liu, Xiaojun Xiang, Xiaoyu Zhang, Xinyu Chen, Yifu Wang, Yipeng Chen, Zhenzhou Fan, Zhewen Le, Zhichao Ye, and Ziqiang Zhao. Inspatio-world: A real-time 4d world simulator via spatiotemporal autoregressive modeling, 2026a. URL https://arxiv.org/abs/2604.07209.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, M. C., Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y. Charles, H. S. Che, Guanduo Chen, Guangyu Chen, Guanzheng Chen, Huarong Chen, Jia Chen, Jianlong Chen, Jun Chen, Kexin Chen, Peng Chen, Ruijue Chen, Wentao Chen, Xin Chen, Yang Chen, Yanru Chen, Yifei Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Dazhi Cheng, Yean Cheng, Jialei Cui, Jingbing Cui, Anqi Dai, Jiaqi Deng, Hao Ding, Rui Ding, Shaofeng Ding, Mengfan Dong, Mengnan Dong, Yuhao Dong, Yuxin Dong, Angang Du, Chenzhuang Du, Dikang Du, Jusen Du, Yulun Du, Yu Fan, Jing Feng, Qiulin Feng, Yichen Feng, Kelin Fu, Qiang Fu, Fuxuan Gao, Hongcheng Gao, Jingyue Gao, Tong Gao, Weijia Gao, Shangyi Geng, Jie Gong, Linhu Gong, Shengao Gong, Xiaochen Gong, Qizheng Gu, Yicheng Gu, Shuhao Guan, Haiqing Guo, Shiqi Guo, Xiang Guo, Zhengyan Guo, Beixi Hao, Wenxin Hao, Xiaoru Hao, Dailan He, Haotian He, Lehan He, Qi He, Weiran He, Xinran He, Xinyi He, Yibo He, Yunjia He, Chao Hong, Tiange Hong, Hao Hu, Jiaxi Hu, Ruikun Hu, Weiming Hu, Yangyang Hu, Zhenxing Hu, Liang Hua, Jinbin Huang, Ke Huang, Ruiyuan Huang, Siying Huang, Weixiao Huang, Yan Huang, Zhengjie Huang, Zhiqi Huang, Yulong Hui, Chaobo Jia, Yutong Jiang, Zhejun Jiang, Zuoyou Jiang, Wenyi Jin, Xinyi Jin, Yu Jing, Huanjun Kong, Guokun Lai, Aidi Li, Cheng Li, Chengyuan Li, Cong Li, Fang Li, Guanyu Li, Haoyang Li, Jia Li, Junxiong Li, Lei Li, Letian Li, Lincan Li, Weihong Li, Wentao Li, Xintong Li, Yang Li, Yishen Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zhaoxi Li, Zheming Li, Zhengxiao Li, Zhiyuan Li, Jiawei Lin, Xiaohan Lin, Yibo Lin, Zichao Lin, Ziyan Lin, Bill Liu, Boxiao Liu, Chuan Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu, Yanming Liu, Yibo Liu, Yipeng Liu, Zhengying Liu, Zhiheng Liu, Enzhe Lu, Haoyu Lu, Linqiang Lu, Tingzhan Lu, Zhiyuan Lu, Aotian Luo, G. Luo, Junyu Luo, Yifan Luo, B. Lyu, Wenzhou Lyu, Shaoguang Mao, Yuan Mei, Xin Men, Minqing Ni, Yixuan Niu, Siyuan Pan, Shujun Peng, Zhangyang Qi, Ruoyu Qin, ZeChao Qin, Zeyu Qin, Haiquan Qiu, Jianxin Qiu, Jiezhong Qiu, Bowen Qu, Yuhao Qu, Zeyu Shang, Youbo Shao, Han Shen, Jincheng Shi, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Wingchun Siu, Pengwei Song, Xiaoxi Song, Jianlin Su, Yunfeng Su, Zhaochen Su, Lin Sui, Jingsong Sun, Junyao Sun, Shaoning Sun, Shuzhe Sun, Tongyu Sun, Yujun Sun, Yunpeng Tai, Chuning Tang, Heyi Tang, Sirui Tang, Zecheng Tang, Chaoran Tian, Rongpeng Tian, Yu Tian, Wei Tu, Chensi Wang, Chuang Wang, Chunjie Wang, Dinglu Wang, Feng Wang, Hailong Wang, Haiming Wang, Hao Wang, Hao Wang, Huaqing Wang, Hui Wang, Jiayi Wang, Jinglong Wang, Jinhong Wang, Jiuzheng Wang, Linian Wang, Shaobo Wang, Shenzhi Wang, Shuyi Wang, Si Wang, Siyuan Wang, Tianfu Wang, Wenjue Wang, Xingran Wang, Xinmei Wang, Xinyuan Wang, Xusheng Wang, Yalin Wang, Yangkun Wang, Yao Wang, Yaoyu Wang, Yejie Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang, Zhengtao Wang, Zhenhao Wang, Zhongsheng Wang, Zifan Wang, Chu Wei, Ming Wei, Shouxin Wei, Zichen Wen, Fan Wu, Haoning Wu, Rucong Wu, Wenhao Wu, Xiaoxue Wu, Yingcong Wu, Yongqi Wu, Yuxin Wu, Zijian Wu, Xinglang Xian, Chenxuan Xiang, Yuye Xiang, Bocheng Xiao, Chenjun Xiao, Xin Xiao, Jin Xie, Xiaotong Xie, Yifeng Xie, Zhe Xie, Bowei Xing, Yiming Xiong, Baosheng Xu, Boyu Xu, Jiale Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Qingtao Xu, Shuyao Xu, Suting Xu, Tiantian Xu, Tianxiang Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ye Xu, Yueni Xu, Ziyao Xu, Haonan Xue, Junjie Yan, Yaoyao Yan, Fan Yang, Guangyao Yang, Hao Yang, Junwei Yang, Ruoyu Yang, Wenjie Yang, Xiaofei Yang, Xinyu Yang, Yi Yang,

Yiling Yang, Ying Yang, Yuchen Yang, Zhen Yang, Zhilin Yang, Zian Yang, Zuhao Yang, Haotian Yao, Dan Ye, Haoran Ye, Wenjie Ye, Zhanbo Ye, Bohong Yin, Haoxiang Yin, Xietong Yin, Chengzhen Yu, Haozhen Yu, Longhui Yu, Shengnan Yu, Shuying Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Tongtian Yue, Wei Yue, Yang Yue, Dunyuan Zha, Haobing Zhan, B. H. Zhang, Dehao Zhang, Fei Zhang, Hao Zhang, Haoyuan Zhang, Huanyu Zhang, Jiapei Zhang, Jiaxuan Zhang, Jin Zhang, Kaiyi Zhang, Miaozhen Zhang, Puqi Zhang, Qinglei Zhang, Rong Zhang, Rui Zhang, Shaoshuai Zhang, Shiyi Zhang, Xiaobin Zhang, Xiaoyun Zhang, Y. Zhang, Yangkun Zhang, Ye Zhang, Yichi Zhang, Yikun Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Zijing Zhang, Bin Zhao, Chenguang Zhao, Feifan Zhao, Jinglun Zhao, Jinxiang Zhao, Shuai Zhao, Wenshuo Zhao, Xiangyu Zhao, Xuanle Zhao, Yikai Zhao, Zijia Zhao, Haozhi Zheng, Huabin Zheng, Ruihan Zheng, Shaojie Zheng, Tengyang Zheng, Haofeng Zhong, Lei Zhong, Longguang Zhong, M. Zhou, Qiankang Zhou, Runjie Zhou, Ruozhang Zhou, Xinyu Zhou, Yiqiao Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yangjunfeng Zhu, Yuxuan Zhu, Zhen Zhu, Chen Zhuang, Weiyu Zhuang, and Xinxing Zu. Kimi k3: Open frontier intelligence, 2026b. URL https://arxiv.org/abs/2607.24653.

Qwen Team. Qwen3.5-omni technical report, 2026. URL https://arxiv.org/abs/2604.15804.

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, 2026c. URL https://arxiv.org/abs/2601.20540.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines, 2025. URL https://arxiv.org/abs/2408.14837.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. Vggt-ω, 2026a. URL https://arxiv.org/abs/2605.15195.

Weijie Wang, Haoyu Zhao, Yifan Yang, Feng Chen, Zeyu Zhang, Yefei He, Zicheng Duan, Donny Y. Chen, Yuqing Yang, and Bohan Zhuang. Latent spatial memory for video world models, 2026b. URL https://arxiv.org/abs/2606.09828.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning, 2026c. URL https://arxiv.org/abs/2507.13347.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, Yidan Xietian, Jiangbo Pei, Liang Hu, Boyi Jiang, Hua Xue, Zidong Wang, Haofeng Sun, Wei Li, Wanli Ouyang, Xianglong He, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory, 2026d. URL https://arxiv.org/abs/2604.08995.

Zhengxuan Wei, Xu Guo, Xinghui Li, Xunzhi Xiang, Min Wei, Yiran Zhu, Qiulin Wang, Xintao Wang, Pengfei Wan, Xiangwang Hou, and Qi Fan. Geometry-aware implicit memory for video world models, 2026. URL https://arxiv.org/abs/2606.02436.

Ruiqi Wu, Xuanhua He, Meng Cheng, Tianyu Yang, Yong Zhang, Zhuoliang Kang, Xunliang Cai, Xiaoming Wei, Chunle Guo, Chongyi Li, and Ming-Ming Cheng. Infinite-world: Scaling interactive world models to 1000-frame horizons via pose-free hierarchical memory, 2026. URL https://arxiv.org/abs/2602.02393.

Yanhao Wu, Haoyang Zhang, Tianwei Lin, Lichao Huang, Shujie Luo, Rui Wu, Congpei Qiu, Wei Ke, and Tong Zhang. Generating multimodal driving scenes via next-scene prediction, 2025. URL https://arxiv.org/abs/2503.14945.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory, 2026. URL https://arxiv.org/abs/2504.12369.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule, 2025. URL https://arxiv.org/abs/2412.06464.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval, 2025a. URL https://arxiv.org/abs/2506.03141.

Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos, 2025b. URL https://arxiv.org/abs/2501.08325.

Jiwen Yu, Jianxiong Gao, Jianhong Bai, Yiran Qin, Kaiyi Huang, Quande Liu, Xintao Wang, Pengfei Wan, Kun Gai, and Xihui Liu. Memlearner: Learning to query context memory for video world models, 2026. URL https://arxiv.org/abs/2606.31734.

Yifei Yu, Xiaoshan Wu, Xinting Hu, Tao Hu, Yangtian Sun, Xiaoyang Lyu, Bo Wang, Lin Ma, Yuewen Ma, Zhongrui Wang, and Xiaojuan Qi. Videossm: Autoregressive long video generation with hybrid state-space memory, 2025c. URL https://arxiv.org/abs/2512.04519.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination?, 2026. URL https://arxiv.org/abs/2603.16666.

Cheng Zhang, Boying Li, Meng Wei, Yan-Pei Cao, Camilo Cruz Gambardella, Dinh Phung, and Jianfei Cai. Unified camera positional encoding for controlled video generation, 2026. URL https://arxiv.org/abs/2512.07237.

Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, Xun Cao, and Wei Yin. Epona: Autoregressive diffusion world model for autonomous driving, 2025a. URL https://arxiv.org/abs/2506.24113.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models, 2025b. URL https://arxiv.org/abs/2504.12626.

Jinjing Zhao, Fangyun Wei, Zhening Liu, Hongyang Zhang, Chang Xu, and Yan Lu. Spatia: Video generation with updatable spatial memory, 2025. URL https://arxiv.org/abs/2512.15716.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images, 2018. URL https://arxiv.org/abs/1805.09817.

Haoyi Zhu. Reflections on video deltanet, Aug 2026. URL https://haoyizhu.site/blog/video-delta-rule/. Blog post.

Haoyi Zhu, Haozhe Liu, Yuyang Zhao, Tian Ye, Junsong Chen, Jincheng Yu, Tong He, Song Han, and Enze Xie. Sana-wm: Efficient minute-scale world modeling with hybrid linear diffusion transformer, 2026. URL https://arxiv.org/abs/2605.15178.

A Training Data and Implementation Details

A.1 Datasets and Preprocessing

Data sources and preprocessing. We construct a camera-trajectory-conditioned training set from ABot-World and RealEstate10K, containing approximately 22,000 and 10,000 video clips, respectively. All training samples are standardized to 16 FPS while preserving temporal alignment between video frames and their corresponding camera parameters. After trajectory augmentation, each training clip contains 257 RGB frames, which are encoded using the same VAE as the generation model. With a temporal downsampling factor of 8, each clip is represented by 33 latent frames. Measured by the duration of the augmented training clips, the resulting dataset contains approximately 150 hours of video. Note that this duration is computed over the actual training sequences and therefore includes repeated observations introduced by trajectory revisits, rather than the duration of deduplicated source footage.

Visualization samples. The initial frames shown in our qualitative results are sourced from SANA-Bench and RealEstate10K, or generated using GPT Image 2. For these samples, we use Qwen3-VL (Bai et al., 2025) to generate image descriptions, which are used as text conditions for video generation.

A.2 Revisit-Trajectory Construction

To provide explicit supervision for revisiting previously observed viewpoints, we synchronously reorder video frames and their corresponding camera poses to construct two types of augmented trajectories: full revisits and asymmetric partial revisits. For a full revisit, we concatenate a forward sequence with its reversed sequence while avoiding duplication of the turning frame, such that the trajectory eventually returns to the initial viewpoint. For example, a 129-frame source clip yields a $129 + 128 = 257$-frame full-revisit sequence. For an asymmetric partial revisit, a shorter reversed segment is appended to the forward sequence, with the return segment accounting for 15%–45% of the final sequence and terminating before reaching the initial viewpoint. This produces training samples with diverse turning points and revisit intervals.

A.3 Backbone Architecture

WorldState is built upon the official SANA-WM model (Zhu et al., 2026). Its backbone is a chunk-causal Diffusion Transformer consisting of 20 Transformer blocks, including 15 frame-wise Gated DeltaNet (GDN) linear attention blocks and 5 full-attention blocks. A full-attention block is inserted after every three consecutive linear-attention blocks.

We replace all GDN layers in the backbone with the proposed linear-attention module that supports fine-grained memory editing. The model has a hidden dimension of 2,240. Each linear-attention layer contains 20 attention heads, with a head dimension of 112.

When introducing WorldState, we retain the pretrained query, key, value, and output projections of the original GDN layers. We augment each recurrent attention module with independent erase, write, and channel-wise decay projections, together with routing parameters for reading from different temporal states.

A.4 Memory Tensors and Gate Parameterization

To avoid confusion with the diffusion time step, we use t to denote the latent video-frame index and τ to denote the diffusion denoising time step.

In our implementation, the query, key, and value tensors are stored in token-first format:

$$\widetilde{\mathbf{Q}}, \widetilde{\mathbf{K}}, \widetilde{\mathbf{V}} \in \mathbb{R}^{B\times H\times F\times S\times D}, \tag{7}$$

where B denotes the batch size, $H = 20$ is the number of attention heads, F is the number of latent video frames in the current attention forward pass, S is the number of spatial tokens per frame, and $D = 112$ is the dimension of each attention head.

For consistency with the matrix notation used in the main text, we transpose the token and channel dimensions for each sample, attention head, and frame:

$$\mathbf{Q}_t = \widetilde{\mathbf{Q}}_t^\top, \qquad \mathbf{K}_t = \widetilde{\mathbf{K}}_t^\top, \qquad \mathbf{V}_t = \widetilde{\mathbf{V}}_t^\top, \tag{8}$$

such that

$$\mathbf{Q}_t, \mathbf{K}_t, \mathbf{V}_t \in \mathbb{R}^{D\times S}. \tag{9}$$

Accordingly, the recurrent memory state of each sample and attention head is represented as

$$\mathbf{S}_t \in \mathbb{R}^{D\times D}. \tag{10}$$

The memory readout in the main-text notation has shape $D \times S$ and is transposed back to the token-first format before the output projection.

For an input resolution of 1280×704, spatial compression and patch embedding produce a latent spatial grid of size 40×22, resulting in $S = 880$ spatial tokens per frame. Note that S varies with the input resolution and is not a fixed model constant.

The backbone feature computation is performed in BF16. The memory matrices, normalization states, and the main matrix operations involved in the recurrent update are accumulated in FP32 to improve numerical stability over long sequences.

The original per-token update-gate vector is defined as

$$\mathbf{b}_t = \sigma\left(\mathbf{b}_t^{\text{logit}}\right) \in \mathbb{R}^S. \tag{11}$$

In the main-text equations,

$$\boldsymbol{\beta}_t = \mathrm{Diag}(\mathbf{b}_t) \in \mathbb{R}^{S\times S} \tag{12}$$

denotes the corresponding diagonal gate matrix. This convention makes $\mathbf{K}_t\boldsymbol{\beta}_t\mathbf{K}_t^\top$ and $\mathbf{V}_t\boldsymbol{\beta}_t\mathbf{K}_t^\top$ both $D \times D$ matrices.

WorldState introduces independent erase and write residuals on top of the original gate logits. Let $\mathbf{X}_t$ denote the current-frame features in token-first format. The gates used in the main-text matrix notation are defined as

$$\mathbf{e}_t = \left[\sigma\left(\mathbf{b}_t^{\text{logit}}\mathbf{1}_D^\top + f_{\text{erase}}(\mathbf{X}_t)\right)\right]^\top, \tag{13}$$

$$\mathbf{w}_t = \left[\sigma\left(\mathbf{b}_t^{\text{logit}}\mathbf{1}_D^\top + f_{\text{write}}(\mathbf{X}_t)\right)\right]^\top, \tag{14}$$

where

$$\mathbf{e}_t, \mathbf{w}_t \in \mathbb{R}^{D\times S}. \tag{15}$$

Here, $\mathbf{1}_D \in \mathbb{R}^D$ broadcasts the original per-token gate logits across the channel dimension, while f_{erase} and f_{write} are zero-initialized linear projections with outputs in $\mathbb{R}^{S\times D}$. The erase gate controls the removal of historical associations along the key dimension, whereas the write gate controls the incorporation of new content along the value dimension.

The channel-wise decay vector is predicted from the spatially averaged features of the current frame:

$$\boldsymbol{\gamma}_t \in (0, 1]^D. \tag{16}$$

Its corresponding diagonal matrix $\mathrm{Diag}(\boldsymbol{\gamma}_t) \in \mathbb{R}^{D\times D}$ allows different memory channels to learn distinct retention rates. The implementation does not impose an explicit strictly positive lower bound on $\gamma_{t,d}$. Consequently, extremely small decay values may underflow to zero under finite-precision computation.

Queries and keys are first processed by the RMS normalization inherited from SANA-WM and are subsequently mapped to nonnegative features through ReLU. The scaling factor

$$c = (DS)^{-1/2} \tag{17}$$

is applied only to the post-ReLU keys and is not applied to the queries or values. The queries and keys in the main numerator stream are further processed with rotary positional embeddings, whereas the denominator stream uses the unrotated nonnegative queries and keys.

A.5 Training Configuration

WorldState is trained using a two-stage strategy. In the first stage, we train the full model for 6,000 iterations, enabling it to learn fine-grained recurrent memory editing through independent erase, write, and channel-wise decay operations.

In the second stage, we continue training for another 6,000 iterations after introducing temporal state partitioning and context-aware routing. This stage enables the model to organize different historical periods into temporally isolated recurrent states and selectively retrieve relevant historical information according to the current generation context.

A.6 Inference Configuration

WorldState generates videos at a resolution of 1280×704 and 16 FPS. During evaluation, each camera trajectory contains 961 RGB frames, corresponding to 121 latent frames after temporal VAE encoding. Stage-1 generation uses the chunk-flow Euler sampler with 60 denoising steps, a classifier-free guidance (CFG) scale of 5, and a flow shift of 8.

B Evaluation Protocols

Benchmark overview. SANA-Bench (Zhu et al., 2026) evaluates one-minute video generation conditioned on an initial image and a camera trajectory. Each test sample contains an initial scene image, a text prompt, and per-frame camera intrinsics and extrinsics. The benchmark consists of 80 scenes evenly distributed across four categories: game-style scenes, indoor scenes, urban outdoor scenes, and natural outdoor scenes. Each category contains 20 scenes. The same scenes are evaluated using both the Simple and Hard trajectory splits, resulting in 160 generated videos per model. Each trajectory contains 961 frames at 16 FPS, corresponding to approximately 60 seconds of video.

The Simple split covers slow curved motion with return, S-shaped motion, orbiting observation, backward motion with return, forward exploration with panning and revisiting, lateral orbiting, L-shaped motion with return, figure-eight motion, approach-and-retreat motion, and indoor exploration. The Hard split contains large-scale orbiting, stepwise ascent, long-range out-and-back motion, crossing zigzag motion, forward motion with large panning, double-loop motion, elevation changes with downward viewing, spiral descent, corner peeking, and vertical motion combined with horizontal panning. Since SANA-Bench does not provide ground-truth future videos, it primarily evaluates camera controllability, generation quality, and self-consistency during spatial revisits.

Spatial revisit consistency. Revisit consistency is evaluated using pairs of frames from the same generated video that have similar camera viewpoints but occur at different times. Candidate frame pairs are provided by the trajectory metadata and ranked according to their camera-viewpoint matching scores. We compute PSNR, SSIM, and LPIPS between the selected frame pairs to evaluate whether a model preserves consistent scene structure and visual appearance when returning to a previously observed region.

Camera controllability. Following the official protocol, we first use Pi3 (Wang et al., 2026c) to estimate the camera trajectory from each generated video. The estimated trajectory is then aligned with the target input trajectory, and RotErr, TransErr, and CamMC are reported. The video is sampled every four frames, yielding at most 241 camera viewpoints for a 961-frame video. RotErr measures the average angular error between the aligned camera orientations, TransErr measures the average aligned positional error, and CamMC jointly measures the discrepancy between the estimated and target camera rotations and translations.

Video quality and Overall score. We compute the Overall score from seven VBench quality dimensions: subject consistency (SC), background consistency (BC), temporal flickering (TF), motion smoothness (MS), aesthetic quality (AQ), imaging quality (IQ), and dynamic degree (DD). The raw score s_d of each dimension d is normalized using the official fixed lower and upper bounds:

$$\hat{s}_d = \mathrm{clip}\left(\frac{s_d - m_d^{\min}}{m_d^{\max} - m_d^{\min}}, 0, 1\right). \tag{18}$$

Dimension	$m_d^{\min}$	$m_d^{\max}$	Weight
SC	0.1462	1.0000	1.0
BC	0.2615	1.0000	1.0
TF	0.6293	1.0000	1.0
MS	0.7060	0.9975	1.0
AQ	0.0000	1.0000	1.0
IQ	0.0000	1.0000	1.0
DD	0.0000	1.0000	0.5

Table 5: Official normalization ranges and weights used to compute the VBench Overall score on SANA-Bench.

The normalization ranges and weights are summarized in Table 5.

The final Overall score is computed as

$$\text{Overall} = \frac{\hat{s}_{\text{SC}} + \hat{s}_{\text{BC}} + \hat{s}_{\text{TF}} + \hat{s}_{\text{MS}} + \hat{s}_{\text{AQ}} + \hat{s}_{\text{IQ}} + 0.5\hat{s}_{\text{DD}}}{6.5}. \tag{19}$$

We report this score on a 0–1 scale. Overall consistency (OC), spatial-revisit metrics, camera errors, and long-horizon quality-degradation metrics are reported separately and are not included in the Overall score.

C Additional Quantitative Analysis

Method	Simple-Trajectory			Hard-Trajectory		
	$\text{IQ}_{0\text{–}10} \uparrow$	$\text{IQ}_{50\text{–}60} \uparrow$	$\Delta\text{IQ} \downarrow$	$\text{IQ}_{0\text{–}10} \uparrow$	$\text{IQ}_{50\text{–}60} \uparrow$	$\Delta\text{IQ} \downarrow$
Spatia	0.7436	0.6813	0.0623	0.7387	0.6889	0.0498
SANA-WM	0.7264	0.6803	0.0461	0.7308	0.6912	0.0396
WorldState	0.7213	0.6836	0.0377	0.7251	0.6923	0.0328

Table 6: Long-horizon visual quality degradation on the 1-min benchmark. $\text{IQ}_{0\text{–}10}$ and $\text{IQ}_{50\text{–}60}$ denote the VBench imaging quality scores over the first and last 10-second segments, respectively. We define $\Delta\text{IQ} = \text{IQ}_{0\text{–}10} - \text{IQ}_{50\text{–}60}$, where a smaller value indicates less visual quality degradation over time.

Long-Horizon Visual Quality. As shown in Table 6, WorldState exhibits the smallest decrease in visual quality from the first to the last 10-second segment on both the Simple and Hard trajectory splits. Although its initial imaging quality is not the highest, it achieves the highest IQ score in the final segment, demonstrating more stable visual quality over long generation horizons.

Memory Scaling Strategy. Table 7 shows that both adaptive scaling strategies substantially outperform the fixed single-state baseline. Square-root scaling achieves slightly higher PSNR and SSIM but requires more states and longer generation time. In contrast, logarithmic scaling achieves comparable revisit performance with fewer states and faster generation, while obtaining better LPIPS, providing a more favorable trade-off between generation quality and efficiency.

Temporal Memory Allocation. We further investigate how historical observations should be distributed across multiple memory states in Tab. 8. Under the same number of memory states and the same overall memory budget, we compare three temporal allocation strategies: Fine-to-Coarse, Coarse-to-Fine, and Uniform allocation.

Fine-to-Coarse allocation preserves earlier history at finer temporal granularity while progressively increasing the temporal span covered by later memory states. Coarse-to-Fine follows the opposite design, allocating finer memory capacity to recent observations while compressing distant history more aggressively. In contrast, our Uniform allocation distributes memory capacity approximately evenly across the entire interaction history. This design is motivated by the fact that the importance of historical observations in spatial revisits does not necessarily decay with temporal distance: a

Growth Strategy	# States	Total State Compute	Gen. Time(s)↓	PSNR↑	SSIM↑	LPIPS↓
Fixed	1	$O(T)$	1302.6	14.66	0.3835	0.3928
Square Root	$\lceil\sqrt{T}\rceil$	$O(T^{3/2})$	1531.1	16.36	0.4256	0.3777
Logarithmic	$\lceil\log_2 T\rceil$	$O(T\log T)$	1462.6	16.28	0.4170	0.3756

Table 7: Comparison of different memory-state growth strategies at the longest evaluation horizon. Fixed maintains a single recurrent state regardless of the generation horizon, while Square Root and Logarithmic increase the number of states as $\lceil\sqrt{T}\rceil$ and $\lceil\log_2 T\rceil$, respectively. All variants use the same per-state dimension.

Allocation Strategy	Temporal Spans	PSNR↑	SSIM↑	LPIPS↓	VBench↑
Fine-to-Coarse	$1, 2, 4, 8, \ldots$	15.33	0.3588	0.4174	0.7720
Coarse-to-Fine	$\ldots, 8, 4, 2, 1$	15.22	0.3835	0.4128	0.7659
Uniform	Approximately Equal	16.28	0.4170	0.3756	0.7852

Table 8: Ablation on temporal memory allocation strategies.

region observed early in the interaction may become critical again after a long trajectory. This experiment therefore evaluates whether balanced temporal allocation is more suitable for long-term spatial revisits than memory organizations biased toward either early or recent history.

D Additional Qualitative Results

Long-Horizon 3D Consistency. As shown in Fig. 5, we further evaluate WorldState in terms of camera controllability and long-term geometric consistency. Specifically, we design camera trajectories that involve a 360° in-place rotation and revisiting previously observed viewpoints, and use WorldState to generate videos conditioned on these trajectories. We then apply VGGT-Omega (Wang et al., 2026a) to reconstruct the generated frames in 3D. The recovered scene geometry and camera paths show that WorldState can accurately follow complex camera motions while preserving coherent spatial layouts and 3D structures throughout long-horizon generation and viewpoint revisitation.

Figure 7 and 8 provide additional qualitative comparisons with representative explicit- and implicit-memory world models. WorldState exhibits less structural drift and better appearance consistency when revisiting previously observed regions.

More results. Figures 9 and 10 present additional 30–40-second WorldState generations featuring flowing rivers, drifting cloud layers, and swaying trees. Fig. 11 presents 40–60-second generations in game environments with character interactions. Together, these results demonstrate that WorldState preserves natural scene dynamics, coherent spatial structure, and consistent character appearance and interactions throughout extended rollouts.

E Effect of the Long-Video Refiner

To balance global content modeling and local visual quality, incorporating a dedicated refinement stage has become a common practice in video generation. One line of work assigns specialized roles to different models within the diffusion process. For example, LingBot-World (Team et al., 2026c) follows the dual-expert architecture of Wan2.2 (Wan et al., 2025), using high- and low-noise experts to model global structure and fine-grained details, respectively. Another line of work adopts a cascaded pipeline, in which an independent upsampler or refiner enhances the output of a base model. Movie Gen (Polyak et al., 2025) employs a diffusion-based upsampler for high-resolution synthesis, LTX-2 (HaCohen et al., 2026) combines latent upscaling with tiled refinement to recover high-frequency details, and SANA-WM uses a long-video refiner to improve the visual quality and long-term stability of minute-long videos. Such designs allow the base model to focus on content, motion, and temporal structure, while the refinement stage repairs local structures, reduces artifacts,

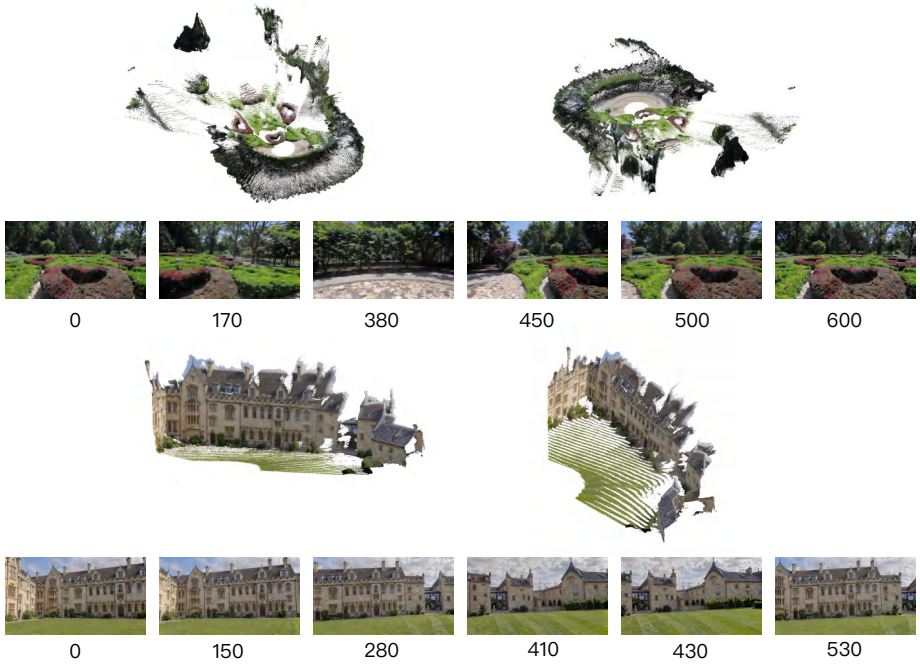


Figure 5: 3D reconstruction of WorldState rollouts. We design camera trajectories that include a 360° in-place rotation and viewpoint revisitation, and use WorldState to generate videos following these trajectories. The scene geometry and camera paths reconstructed by VGGT-Omega demonstrate that WorldState follows complex camera motion while maintaining consistent 3D structure over long-horizon generation and viewpoint revisitation.

and enhances texture details, enabling longer and higher-quality video generation under limited training resources.

Following the two-stage generation pipeline of SANA-WM, we apply the same long-video refiner to the Stage-1 outputs after WorldState completes the generation of the full video. The refiner takes the video latents generated by WorldState, which already incorporate the camera trajectory and reference-image conditioning, and enhances them under text conditioning while retaining a clean latent anchor. The resulting latents are then decoded into the final video.

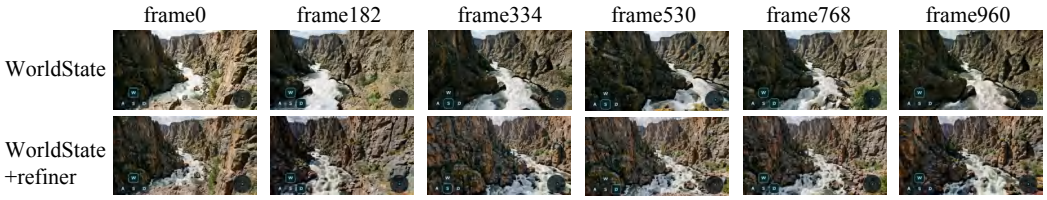


Figure 6: Visual quality comparison between WorldState and WorldState with the refiner over a 961-frame rollout. The refiner helps maintain higher visual fidelity over long horizons, reducing quality degradation as generation progresses.

The refiner is not involved in WorldState training or in memory writing, state maintenance, and memory retrieval during autoregressive generation. WorldState is responsible for generating video content and camera motion while maintaining long-term scene consistency, whereas the refiner serves primarily as a post-generation visual enhancement module. For a fair comparison, we apply the same refiner with identical inference settings to SANA-WM and WorldState. We report results both before and after refinement to distinguish the long-term memory improvements introduced by WorldState from the visual-quality gains provided by the refiner.



Figure 7: Additional qualitative comparison of Gen3C, Spatia, Matrix-Game 3.0, SANA-WM, and WorldState under long-horizon spatial revisit trajectories. WorldState better preserves scene structure and visual appearance after extended exploration and viewpoint revisitation.

Fig. 6 shows that the refiner substantially improves the overall visual quality of WorldState outputs. After refinement, scene structures such as mountains become more complete, boundaries and textures appear sharper, and blurred regions and local artifacts in the Stage-1 outputs are reduced. The improvements are particularly evident for distant objects, complex structures, and fine-grained environmental textures.



Figure 8: Additional qualitative comparison of Gen3C, Spatia, Matrix-Game 3.0, SANA-WM, and WorldState under long-horizon spatial revisit trajectories. WorldState better preserves scene structure and visual appearance after extended exploration and viewpoint revisitation.

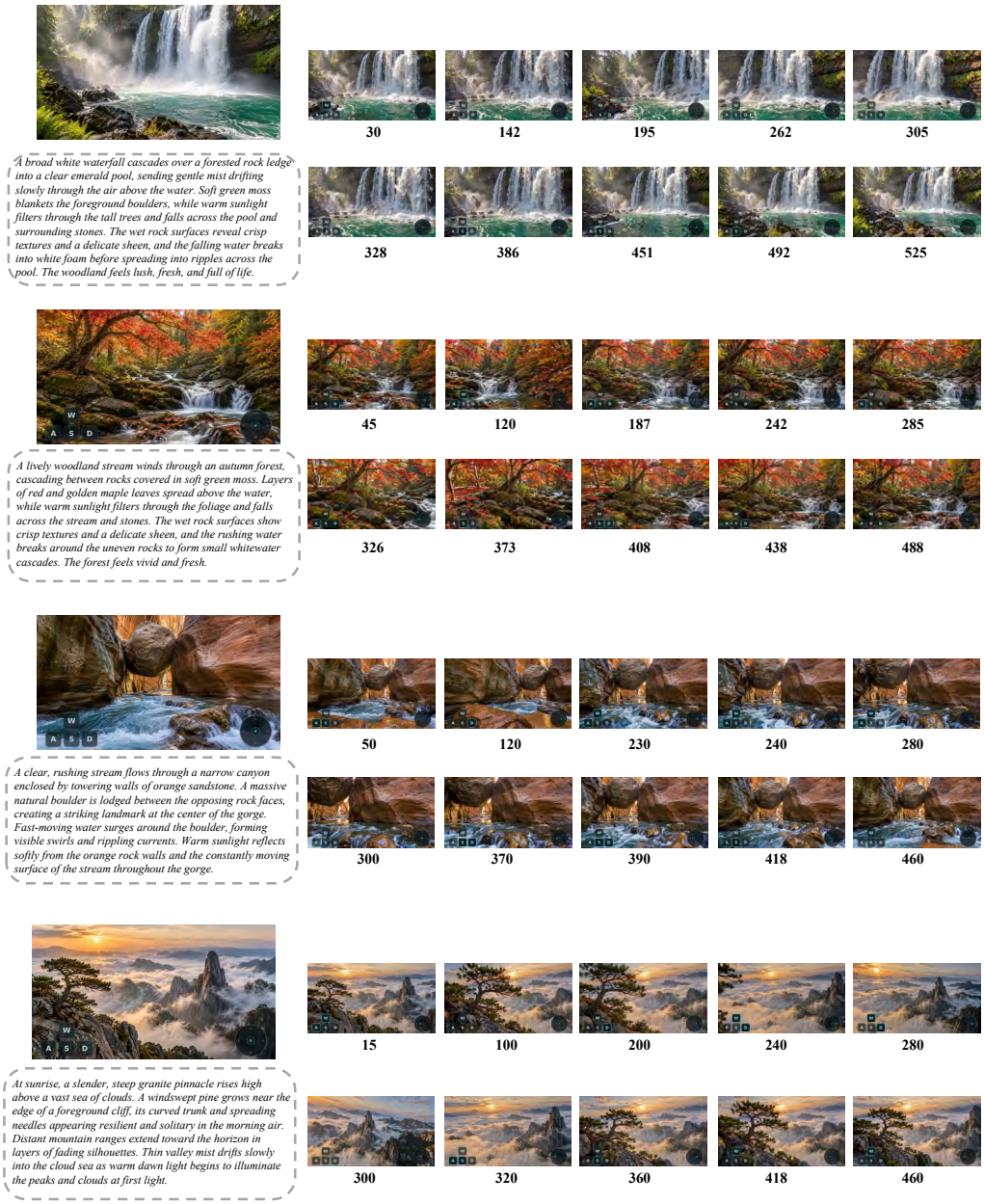


Figure 9: Additional 30–40-second generation results from WorldState.

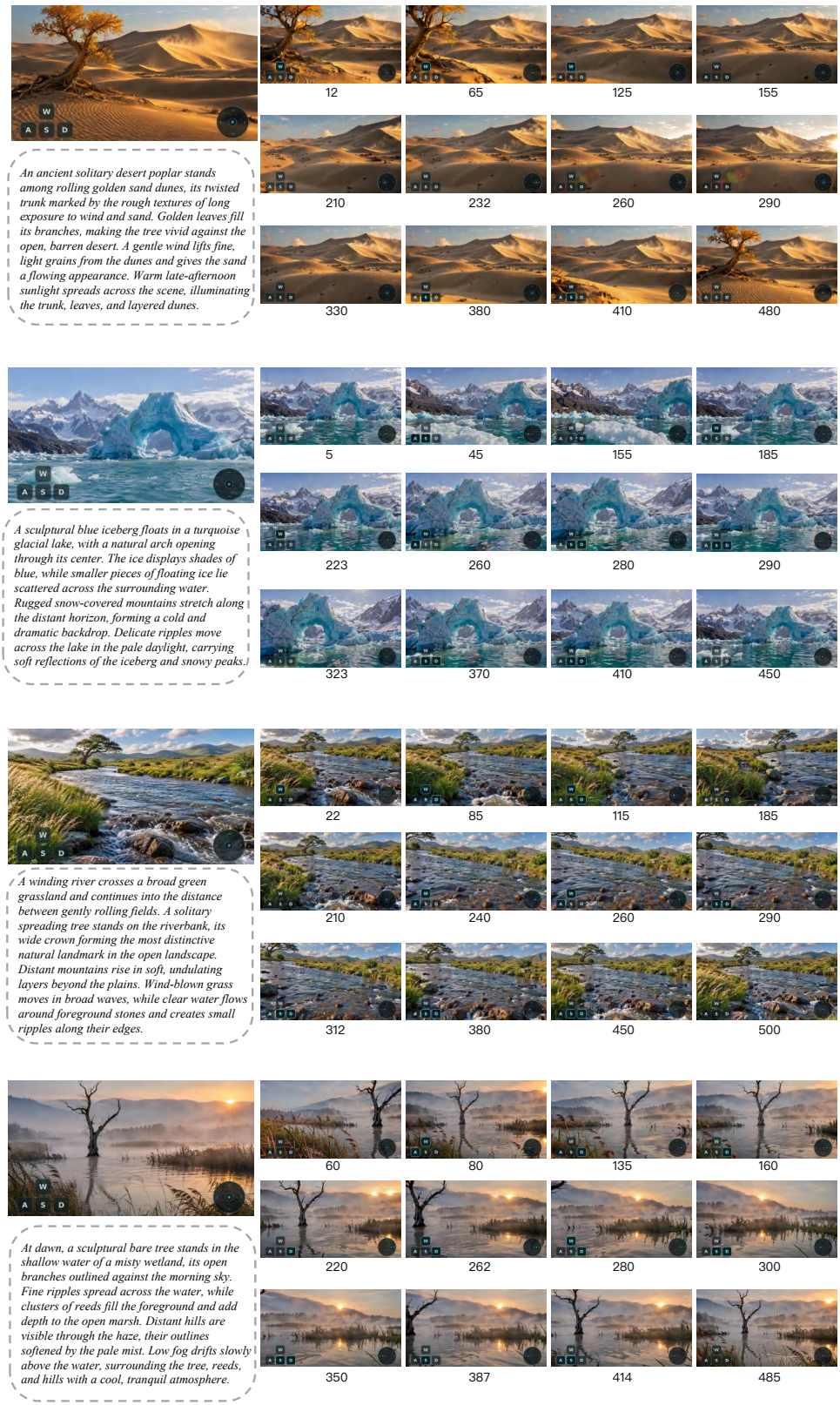


Figure 10: Additional 30–40-second generation results from WorldState.



Figure 11: Additional 40–60-second generation results from WorldState.