SPATIAL MEMORY INTELLIGENCE: ENDOWING WORLD MODELS WITH UNDERSTANDING-DRIVEN LONG-TERM MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-video generation and world models have shown strong potential for interactive entertainment and embodied simulation by predicting future observations conditioned on user actions and historical memory. However, as memory sequences grow longer and their structures become increasingly complex, managing long-range spatial context becomes increasingly challenging, calling for a more intelligent and systematic memory-management strategy. Building on the advancing spatial reasoning capabilities of multimodal large language models (MLLMs) and the broader vision of unified models, we propose Spatial Memory Intelligence (SMI), the first framework to systematically employ an understanding model for spatial-memory management in long-video world models. SMI introduces four coordinated atomic operations: spatial clustering, within-cluster sparsification, action-aware retrieval, and reliability-aware filtering. Extensive experiments across multiple baselines, benchmarks, and world-model backbones demonstrate the effectiveness and generalizability of SMI, achieving comprehensive improvements in memory sparsity, generation stability, and spatial consistency.

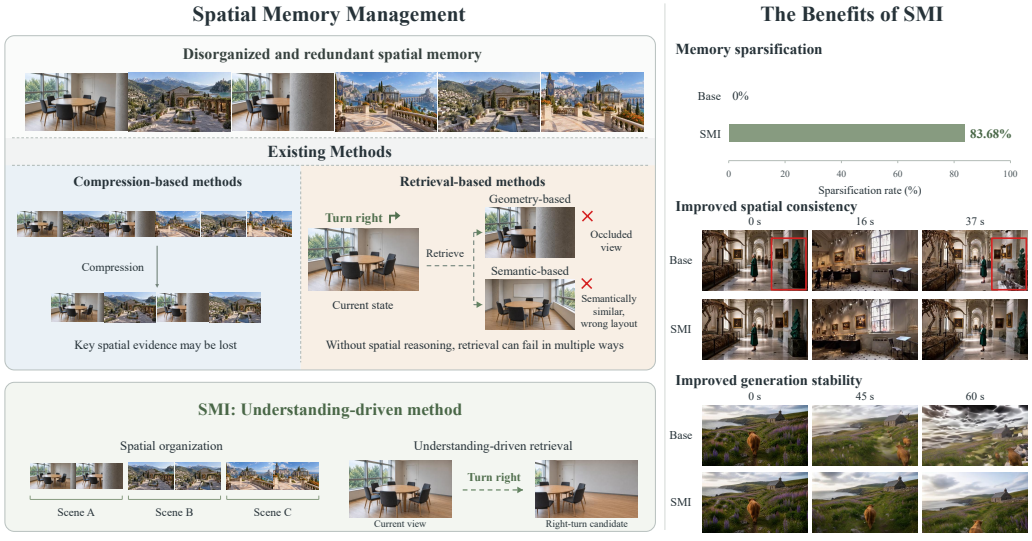

Figure 1: **Motivation and benefits of SMI.** Understanding-driven spatial-memory management organizes disorganized and redundant histories and retrieves relevant memories, enabling 83.68% memory sparsification with improved spatial consistency and generation stability.

1 INTRODUCTION

World models are driving a pivotal shift in visual and spatial intelligence, moving generative modeling toward controllable simulation of environments. Within this paradigm, video provides a natural

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.

medium for representing both the visual appearance and temporal evolution of an environment. Video world models (Parker-Holder & Fruchter, 2025; Che et al., 2025; Sun et al., 2025; Wang et al., 2026b; Mao et al., 2026; Gao et al., 2026) serve as interactive visual simulators by conditioning future visual states on historical observations, language instructions, and camera signals. This interactivity enables spatial exploration across different regions of the generated world, where newly generated observations are continuously appended to the visual history and become part of the context for future prediction. Through this iterative process, the model progressively integrates observations from different regions and viewpoints into an evolving world state that captures semantics and geometry, making video world models a promising foundation for interactive entertainment (Tang et al., 2025; Team et al., 2026a) and embodied simulation (Agarwal et al., 2026; Chen et al., 2026).

The continual evolution of the generated world poses intertwined challenges for memory efficiency, spatial consistency, and generation stability. As the observation history grows, retaining all observations introduces substantial redundancy and incurs high storage costs, motivating memory sparsification (Yu et al., 2025c; Zhang et al., 2025a). In parallel, maintaining spatial consistency requires retrieving memories of previously observed regions that provide relevant semantic and geometric evidence. Failure to retrieve such evidence can lead to geometric inconsistencies in revisited regions and unintended changes in object identities or states (King et al., 2026; Wu et al., 2026; Hu et al., 2026). Moreover, newly generated observations may themselves contain errors. If admitted into memory and reused as references, these errors can propagate and compound during autoregressive generation, undermining long-horizon stability (Liu et al., 2026b; Zhang et al., 2025a). Together, these challenges call for systematic spatial-memory management that determines what to retain, what to retrieve, and which new observations are reliable enough to store.

Existing approaches to memory management in world models primarily focus on memory compression, sparse computation, and selective retrieval. Compression-based methods (Jin et al., 2025; Savov et al., 2025; Wu et al., 2026; Hong et al., 2025; Zhang et al., 2026b; Wei et al., 2026) summarize an expanding context into a bounded representation, often discarding fine-grained visual evidence needed to maintain long-horizon consistency. Sparse-computation methods (Xi et al., 2025; Xia et al., 2025) exploit attention sparsity to reduce computational cost, but target generation efficiency rather than long-term memory management. To maintain spatial consistency, retrieval-based methods select relevant historical observations as context based on geometric relations or semantic similarity. Geometry-based retrieval (Yu et al., 2025a; Xiao et al., 2025; Wu et al., 2025c; Ren et al., 2025; Li et al., 2025c) relies on camera or geometry relevance and is sensitive to occlusion, dynamic content, semantic changes and geometry errors. Semantic retrieval (Ji et al., 2025; Wu et al., 2025d) captures high-level relevance but can struggle to distinguish similar content across different locations or viewpoints. Some methods jointly address memory compression and implicit content querying (Zhang et al., 2026b) through an auxiliary encoder and the generative model itself, but remain constrained by their limited spatial perception and semantic understanding. These limitations motivate a unified memory manager that organizes, sparsifies, and retrieves spatial context while reasoning over both geometry and semantics.

Recent MLLMs have made substantial progress in long-video understanding, spatial perception, and reasoning (Bai et al., 2025b; Chen et al., 2024b; Yang et al., 2025a;b). Meanwhile, a growing line of work explores unified multimodal architectures that support both understanding and generation (Wu et al., 2025a; Xie et al., 2025; Deng et al., 2025). **Motivated by the spatial and semantic capabilities of understanding models and by the broader vision of unified understanding and generation**, we propose Spatial Memory Intelligence (SMI), the first framework to systematically employ an understanding model for spatial-memory management in long-video world models. SMI decomposes spatial-memory management into four coordinated atomic operations (Figure 2): spatial clustering, within-cluster sparsification, action-aware retrieval, and reliability-aware filtering. Spatial clustering structures the memory by grouping spatially proximate memory chunks into the same cluster. Within-cluster sparsification removes redundant observations from each cluster to reduce storage costs. Action-aware retrieval combines the recent context with the current action to determine which historical memories provide the most useful spatial evidence for generating the next chunk. Reliability-aware filtering prevents newly generated chunks exhibiting visual drift from entering memory, limiting error propagation through subsequent reuse. Together, these atomic operations constitute a complete and unified mechanism for spatial-memory management.

For each operation, we formulate dedicated instructions and construct corresponding supervision, enabling the MLLM to learn the semantic and spatial judgments required throughout the memory-

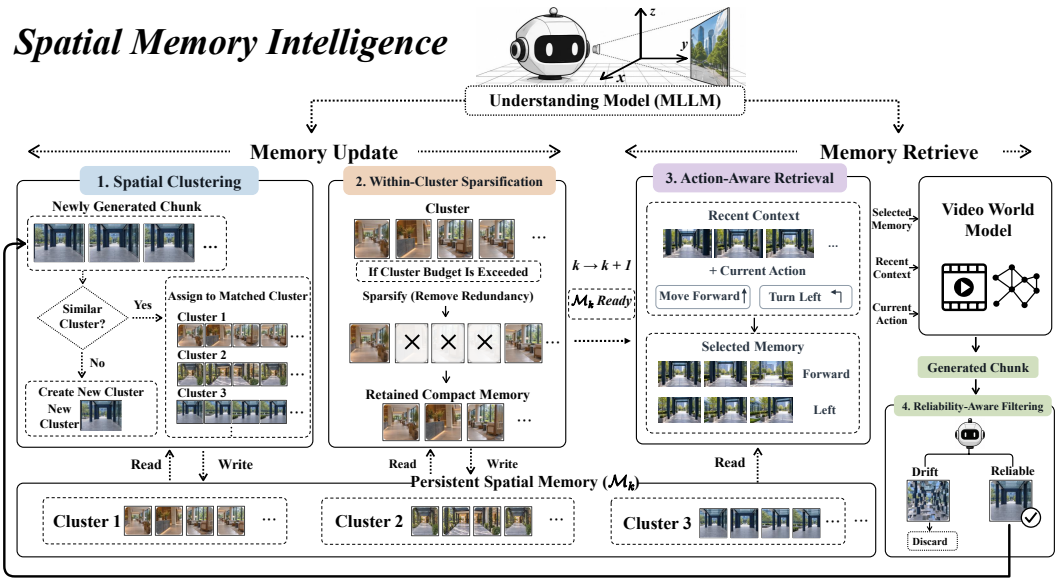

Figure 2: Overview of SMI. SMI uses an MLLM to manage spatial memory $\mathcal{M}_k$ through four atomic operations: spatial clustering, within-cluster sparsification, action-aware retrieval, and reliability-aware filtering.

management pipeline. Compared with existing approaches, SMI provides a more comprehensive memory-management framework and achieves broader performance improvements, as illustrated in Figure 1. Our contributions are summarized as follows:

- We introduce SMI, the first unified MLLM-driven framework for systematic spatial-memory management in long-horizon video world models. SMI defines four coordinated atomic operations for organizing, consolidating, filtering, and recalling spatial memory.
- We construct dedicated operation-oriented supervision datasets covering all four atomic operations and formulate these operations as learnable multimodal decision tasks, enabling the MLLM to develop unified reasoning capabilities for spatial-memory management.
- Extensive experiments across multiple baselines, benchmarks, and world-model backbones demonstrate the effectiveness and general applicability of SMI, with consistent improvements in memory sparsity, long-horizon generation stability, and visual consistency.

2 METHOD

We propose Spatial Memory Intelligence (SMI), a unified framework that employs an MLLM to manage long-term spatial memory for video world models, as illustrated in Figure 2. SMI comprises four coordinated atomic operations: spatial clustering, within-cluster sparsification, action-aware retrieval, and reliability-aware filtering. We begin by formalizing action-conditioned long-video generation in Section 2.1 and explaining the motivation for using an MLLM as the spatial-memory manager in Section 2.2. We then introduce the four spatial-memory operations in Section 2.3, followed by the operation-oriented data construction and training procedure in Section 2.4.

2.1 PRELIMINARIES

Action-conditioned video world models. Action-conditioned video world models generate frames autoregressively from selected historical observations and current actions. Given an action sequence $a_{1:K}$, the joint distribution of the generated video frames $\boldsymbol{x}_{1:K}$ is formulated as

$$p_\theta(\boldsymbol{x}_{1:K} \mid a_{1:K}) = \prod_{k=1}^{K} p_\theta\left(\boldsymbol{x}_k \mid \boldsymbol{x}_{<k}^{\text{subset}}, a_k\right), \tag{1}$$

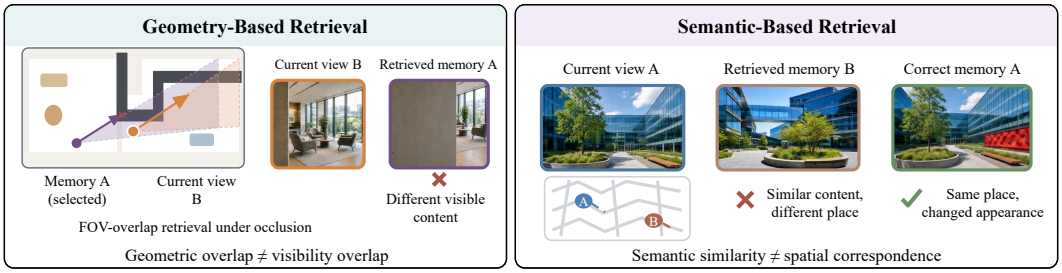

Figure 3: Failure modes of geometry- and semantic-based memory retrieval.

where θ denotes the model parameters, K is the number of generated frames, $\boldsymbol{x}_k$ is the frame generated at step k, $\boldsymbol{x}_{<k}^{\text{subset}}$ denotes the selected historical frames before step k, and a_k is the current action. SMI manages the historical observations used in this process.

2.2 Why Use an MLLM?

Despite leveraging geometric or semantic information, existing memory-management strategies still exhibit retrieval limitations across diverse scenes, as illustrated in Figure 3. Geometry-based retrieval can fail to identify useful memories under spatial occlusion, whereas semantic retrieval may overlook spatial and viewpoint changes. The fundamental limitation is that these approaches lack the ability to understand the relationships between spatial observations and to reason based on these relationships. Over long-horizon interaction, **managing an increasingly complex observation history requires reasoning over historical states, tracking semantic and spatial changes during interaction, and relating observations across viewpoints and time.** Meeting these demands remains challenging for both fixed heuristics and the generator's implicit memory mechanisms.

These limitations call for a dedicated memory manager that can reason over the evolving history and determine how memory should be organized, retained, and retrieved for the current context. Scalable, data-driven MLLMs provide a natural basis for such a manager, as they can learn memory-management strategies from diverse multimodal data. Moreover, their growing capabilities in contextual understanding and semantic reasoning (Alayrac et al., 2022; McKinzie et al., 2024; Bai et al., 2025a), together with rapid advances in spatial perception and reasoning (Chen et al., 2024b; Yang et al., 2025a; Ma et al., 2025b), make MLLMs well suited to spatial-memory management in video world models. More broadly, progress toward unified understanding-and-generation architectures (Wu et al., 2025a; Xie et al., 2025; Deng et al., 2025) reflects a growing trend toward closer integration of multimodal understanding and generation. In this work, we establish an initial framework for managing spatial memory with an MLLM and define four atomic operations.

2.3 Atomic Spatial Operations

SMI decomposes spatial-memory management into four coordinated atomic operations: spatial clustering, within-cluster sparsification, action-aware retrieval, and reliability-aware filtering. At generation step k, we define C_k as the newly generated video chunk and $\mathcal{M}_k = \{m_i\}_{i=1}^{N_k}$ as the spatial memory bank, where m_i is the i-th retained video memory chunk and $N_k = |\mathcal{M}_k|$ is the number of retained chunks. When camera parameters are available, π_i denotes the camera parameters associated with m_i and serves as an auxiliary geometric cue.

Spatial clustering. Temporal proximity does not imply spatial proximity. During inference, unpredictable user actions can cause the camera to revisit different regions in arbitrary order, interleaving spatially distinct observations in temporal memory. As shown in Figure 4, the resulting memory is temporally ordered but spatially disorganized and redundant, making sparsification difficult and causing generation errors to accumulate. To address these issues, SMI organizes the streaming-generated chunks into spatial clusters. At step k, let the j-th spatial cluster be

$$\mathcal{S}_j = \{m_{j,q}\}_{q=1}^{n_j} \subseteq \mathcal{M}_k, \tag{2}$$

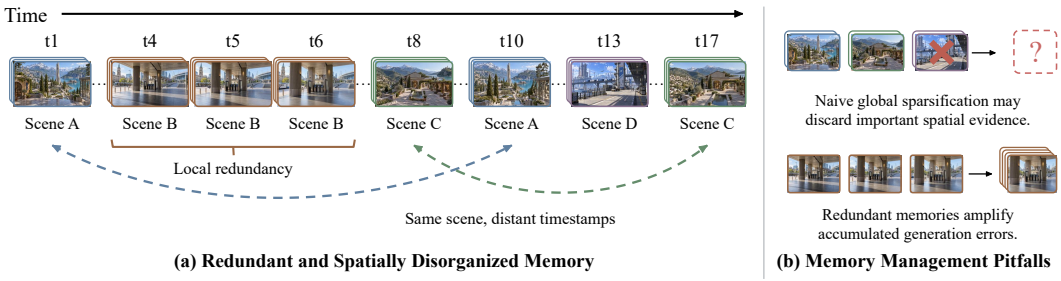

Figure 4: Motivation for spatial clustering: redundancy and spatial disorder in long-term memory.

where j indexes the clusters, q indexes the memory items in cluster j, and n_j is the number of memory items in that cluster. We select the memory item with the largest projection coverage from the remaining chunks in the same cluster as its prototype:

$$p_j = \underset{m_{j,q} \in \mathcal{S}_j}{\arg\max} \operatorname{Cov}\left(m_{j,q}, \mathcal{S}_j - \{m_{j,q}\}\right). \tag{3}$$

Here, $\mathcal{S}_j - \{m_{j,q}\}$ denotes all memory items in cluster j except $m_{j,q}$, Cov denotes their average camera-projection coverage of $m_{j,q}$, and p_j is the prototype of cluster j.

For a newly generated chunk C_k, we first use camera position and orientation to shortlist a set of spatially closest cluster prototypes. The MLLM then determines whether C_k matches one of these candidate prototypes. If a match exists, C_k is assigned to the corresponding spatial cluster; otherwise, a new cluster is created:

$$\operatorname{Assign}(C_k) = \begin{cases} \mathcal{S}_j, & \text{if the MLLM matches } C_k \text{ to prototype } p_j, \\ \mathcal{S}_{\text{new}}, & \text{if no matching prototype exists.} \end{cases} \tag{4}$$

Here, $\mathcal{S}_{\text{new}}$ denotes the new cluster created for C_k.

Within-cluster sparsification. After spatial clustering, the memory items within each cluster are ordered by generation time:

$$\mathcal{S}_j = \left(m_{j,1}, \ldots, m_{j,n_j}\right). \tag{5}$$

When n_j reaches a predefined sparsification trigger threshold T_s, SMI initiates within-cluster redundancy detection. The MLLM compares the intermediate memory items and identifies redundant chunks that can be safely removed. We formalize this operation as

$$\mathcal{K}_j = \operatorname{Sparse}_{\text{MLLM}}\left(\mathcal{S}_j\right), \qquad n_j \geq T_s. \tag{6}$$

Here, T_s specifies only the number of memory items required to trigger within-cluster sparsification, and $\mathcal{K}_j$ is the retained subset after redundant items are removed. The number of retained items is not fixed in advance; instead, it is determined dynamically by the MLLM according to whether each item contains new spatial structure, object states, or occlusion relations. An intermediate item is removed if it provides no additional useful information relative to its neighboring memories. Because sparsification is performed only within spatially coherent clusters, SMI removes repeated observations without discarding important evidence from different spatial regions.

Action-aware retrieval. At generation step k, SMI selects useful historical memories conditioned on the recent context and current action. For efficiency, the current camera state and action are first used to shortlist the K_g memory items with the largest spatial overlap with the current view:

$$\mathcal{G}_k = \operatorname{TopK}_{K_g}\left(\mathcal{M}_k; \boldsymbol{\pi}_k, a_k\right). \tag{7}$$

Here, k denotes the current generation step, $\boldsymbol{\pi}_k$ is the camera state at step k, a_k is the current action, K_g is the number of geometrically shortlisted candidates, and $\mathcal{G}_k$ is the resulting candidate set.

The MLLM then uses the recent context, current action, and candidate memories to determine which historical chunks provide useful information for generating the next chunk:

$$\mathcal{R}_k = \operatorname{Retrieve}_{\text{MLLM}}\left(C_{k-1}^{\text{recent}}, a_k, \mathcal{G}_k\right), \qquad 0 \leq |\mathcal{R}_k| \leq P_r. \tag{8}$$

Here, C_{k-1}^{recent} denotes the recent context, $\mathcal{R}_k$ is the retrieved historical memory set, and P_r is the maximum number of returned memory items. The MLLM may return an empty set when none of the candidates is useful, or select up to P_r memory items when multiple historical memories jointly support the next generation.

Reliability-Aware Filtering. For each newly generated video chunk C_k, the MLLM determines whether it exhibits evident visual drift and outputs a filtering decision:

$$D_k = \text{Filter}_{\text{MLLM}}(C_k) \in \{\text{admit}, \text{discard}\}. \tag{9}$$

Here, $\text{Filter}_{\text{MLLM}}$ denotes the reliability-aware filtering operation performed by the MLLM, and D_k denotes the resulting decision for C_k. If $D_k = \text{admit}$, the chunk is added to $\mathcal{M}_k$; otherwise, it is discarded, thereby reducing the accumulation of generation errors.

We provide visualizations of the different atomic operations during actual inference in Appendix A.

2.4 Operation-Oriented Data Construction

Although general-purpose MLLMs already possess a certain degree of spatial reasoning ability, these specialized spatial operations still need to be learned from data. We first collect approximately one thousand trajectories from real model inference and remove low-quality samples. Next, we manually annotate more than one hundred examples from the remaining samples. We then use these human-annotated examples as demonstrations for multiple state-of-the-art teacher MLLMs, which annotate the rest of the data. A detailed description of the data-construction pipeline and dataset composition is provided in Appendix G.

Finally, human annotators review the labels generated by the teacher MLLMs, discard predictions with evident errors, and correct inconsistent annotations. Through this multi-stage data construction pipeline with repeated cleaning and filtering, we obtain a training dataset of $N_{\mathcal{D}}$ examples:

$$\mathcal{D} = \{(\mathcal{I}_n, Y_n)\}_{n=1}^{N_{\mathcal{D}}}. \tag{10}$$

Here, $\mathcal{I}_n$ denotes the multimodal input for the corresponding atomic operation, and Y_n denotes its annotated target output. We then optimize the MLLM using standard cross-entropy. For a target output sequence Y_n with L_n tokens, the training objective is

$$\mathcal{L}_{\text{SFT}}(\phi) = -\frac{1}{N_{\mathcal{D}}} \sum_{n=1}^{N_{\mathcal{D}}} \sum_{\ell=1}^{L_n} \log p_\phi(y_{n,\ell} \mid \mathcal{I}_n, y_{n,<\ell}). \tag{11}$$

Here, ϕ denotes the trainable parameters of the MLLM, $y_{n,\ell}$ denotes the ℓ-th token in the target sequence of the n-th sample, and $y_{n,<\ell}$ denotes all target tokens preceding it.

3 Experiments

3.1 Implementation Details

Backbones and optimization. We conduct experiments using the HY1.5-8B and Wan2.2-5B world-model backbones from WorldPlay-1.5 (Sun et al., 2025). Qwen3.5-4B (Qwen Team, 2026) serves as the multimodal understanding model for the proposed spatial-memory operations and is fine-tuned with AdamW using a learning rate of 2×10^{-5} and a global batch size of 64. Additional experimental details are provided in Appendix C.

3.2 Evaluation Metrics

We assess overall performance using four complementary evaluation protocols. **(1) VBench** (Huang et al., 2024). We report background consistency, motion smoothness, and aesthetic quality to evaluate generation stability and consistency. **(2) GPT-5.6-sol evaluation.** Three independent GPT-5.6-sol evaluator agents blindly assess the generated videos. We report Camera-Constrained Scene Consistency and Overall Visual Quality, with the detailed evaluation protocol and full results provided in Appendix D. **(3) Reconstruction consistency.** Following WorldPlay, each model traverses

Table 1: Comparison of the Base, five memory-management baselines, and SMI on HY1.5. All methods use identical generation settings. Bold indicates the best reported value.

Method	VBench			GPT-5.6-sol evaluation		Reconstruction		Efficiency	
	Background Consistency↑	Motion Smoothness↑	Aesthetic Quality↑	Camera-Constrained Consistency↑	Overall Visual Quality↑	PSNR↑	LPIPS↓	Sparsity↑	Latency↓
Base (FoV) (Sun et al., 2025)	0.937983	0.991386	0.445389	51.6053	53.1538	12.478528	0.603425	0.0000%	$\times 1.00$
FramePack (Zhang et al., 2025a)	0.939390	0.989362	0.432000	43.2800	46.6474	11.718369	0.621979	**94.3820%**	$\times 1.00$
Deep Forcing (Yi et al., 2025)	**0.946473**	0.982810	0.440462	37.6760	33.6297	11.260350	0.664867	79.2135%	$\times \mathbf{0.81}$
MoC (Cai et al., 2025)	0.937999	0.990528	0.438969	48.9196	51.9717	10.802151	0.735292	0.0000%	$\times 1.15$
VMem (Li et al., 2025c)	0.939695	0.991162	0.447630	54.4727	56.5060	12.768029	0.582628	0.0000%	$\times 1.47$
MemFlow (Ji et al., 2025)	0.938170	0.990061	0.430962	43.3160	45.6500	10.296440	0.712713	49.4253%	$\times 1.27$
SMI (ours)	0.942890	**0.991999**	**0.469357**	**67.0200**	**65.3677**	**13.309393**	**0.554519**	83.6842%	$\times 1.38$

Table 2: Comparison of the Base, five memory-management baselines, and SMI on Wan2.2. All methods use identical generation settings. Bold indicates the best reported value.

Method	VBench			GPT-5.6-sol evaluation		Reconstruction		Efficiency	
	Background Consistency↑	Motion Smoothness↑	Aesthetic Quality↑	Camera-Constrained Consistency↑	Overall Visual Quality↑	PSNR↑	LPIPS↓	Sparsity↑	Latency↓
Base (FoV) (Sun et al., 2025)	0.930115	0.928419	0.424737	34.6524	37.8363	11.434590	0.596041	0.0000%	$\times 1.00$
FramePack (Zhang et al., 2025a)	**0.934067**	0.933264	0.421781	34.7984	37.9917	11.011042	0.611023	**94.3820%**	$\times 0.98$
Deep Forcing (Yi et al., 2025)	0.920607	0.929061	0.444197	46.8560	50.7300	11.198159	0.602744	79.2135%	$\times \mathbf{0.75}$
MoC (Cai et al., 2025)	0.916892	0.931185	0.432453	45.5013	49.2763	11.143642	0.610697	0.0000%	$\times 1.24$
VMem (Li et al., 2025c)	0.920369	0.929605	0.439904	43.8800	45.3613	11.157617	0.603184	0.0000%	$\times 1.34$
MemFlow (Ji et al., 2025)	0.924393	0.932660	0.435056	38.8487	40.2990	10.887503	0.610859	49.4253%	$\times 1.13$
SMI (ours)	0.917329	**0.935492**	**0.445095**	**46.9309**	**51.4909**	**11.564875**	**0.587550**	82.2073%	$\times 1.31$

a round-trip camera trajectory, and we compute PSNR and LPIPS between corresponding frames from the outbound and return paths. **(4) User study.** We conduct a user study to compare the generation stability and spatial consistency of SMI and the baselines. The evaluation protocol and results are provided in Appendix E. For VBench and GPT-5.6-sol evaluation, we use 100 randomly sampled one-minute rollouts. For reconstruction consistency, we use a separate set of 100 one-minute rollouts containing revisit trajectories. **For fair evaluation, all test scenes are held out from the training set. We will release all datasets and model weights for reproducibility.**

Efficiency. We define memory sparsity as the proportion of memory items removed from the total memory. The Base performs no memory sparsification and therefore has 0% sparsity. Latency is measured as end-to-end generation time relative to the Base, which is normalized to $\times 1.00$. All measurements use the same hardware and generation settings.

3.3 Baselines

Besides the HY1.5 and Wan2.2 baselines that use a field-of-view (FoV)-based retrieval strategy, we compare SMI with five memory-management methods: FramePack and Deep Forcing for compression, Mixture of Contexts (MoC) for sparse-routing-based retrieval, VMem for geometry-based retrieval, and MemFlow for semantic retrieval. These methods cover four complementary strategies for long-term memory management: compression, sparsification, geometric retrieval, and semantic retrieval. For a fair comparison, all methods use the same initial observations, action sequences, random seeds, context budget, and generation configuration. More details are provided in Appendix F.

3.4 Results

3.4.1 Quantitative Results

Tables 1 and 2 present quantitative comparisons between SMI and the baselines on HY1.5 and Wan2.2, respectively. On HY1.5, SMI improves Aesthetic Quality from 0.445389 to 0.469357, GPT-evaluated Scene Consistency from 51.6053 to 67.0200, and PSNR from 12.4785 to 13.3093 dB over the Base, while maintaining a high memory sparsity of 83.6842%. A similar trend holds on Wan2.2. Overall, SMI substantially sparsifies memory while improving spatial consistency and generation stability, achieving more comprehensive improvements than the other methods. Although the additional understanding model introduces time overhead, we expect this overhead to decrease substantially as generation and understanding are unified through shared representations.

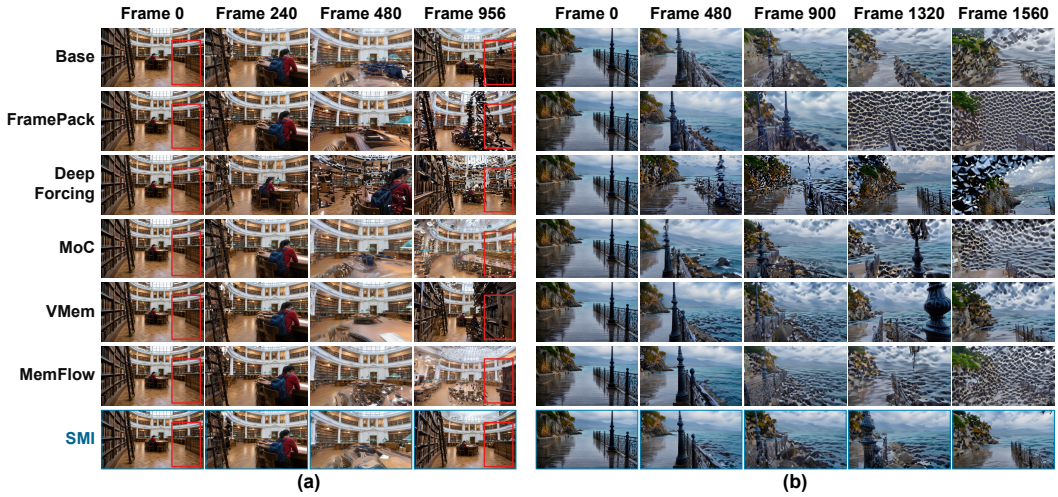

Figure 5: Qualitative comparisons on HY1.5: (a) spatial consistency and (b) generation stability. Red boxes highlight corresponding regions in the initial and final frames.

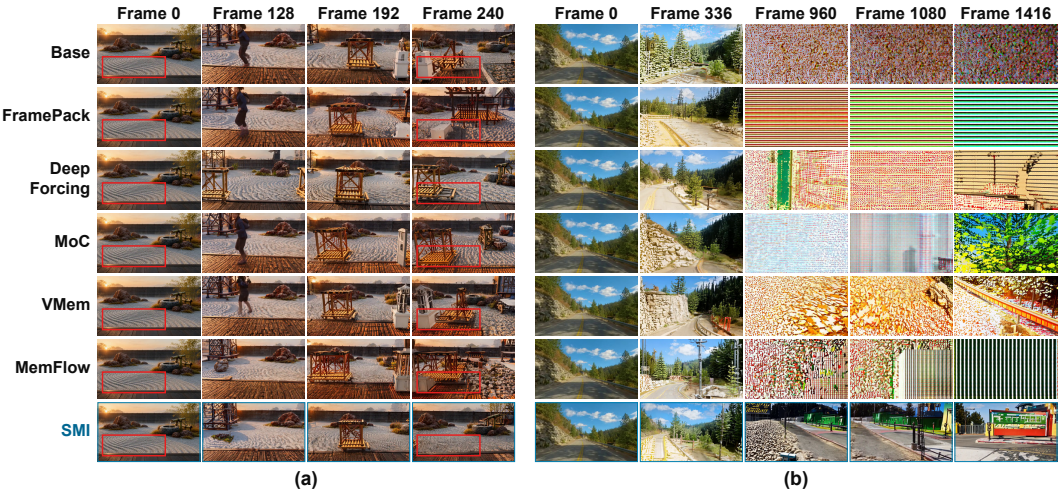

Figure 6: Qualitative comparisons on Wan2.2: (a) spatial consistency and (b) generation stability. Red boxes highlight corresponding regions in the initial and final frames.

3.4.2 Qualitative Results

Figures 5 and 6 illustrate the generation stability and spatial consistency of different methods on HY1.5 and Wan2.2. SMI demonstrates the best spatial consistency and generation stability among the compared methods on both backbones, achieving comprehensive improvements in both aspects. More qualitative results are provided in Appendix J.

3.5 Ablation Studies

3.5.1 Component Ablation

We conduct ablations on HY1.5 by disabling different combinations of spatial clustering, within-cluster sparsification, and reliability-aware filtering, while retaining action-aware retrieval in all variants. Table 3 and Figure 7(1) present the quantitative and qualitative results. Removing components degrades generation stability and consistency to varying degrees. In particular, without spatial clustering, sparsification can discard important spatial evidence, leading to a decline in consistency. The reconstruction scores and the visual inconsistencies in Figure 7(1) support this observation.

Table 3: Component ablation on HY1.5. A checkmark denotes an enabled module. Latency is relative to the Base. Bold indicates the best value.

Setting	Modules			VBench			Reconstruction		Efficiency	
	Spatial Clustering	Within-cluster Sparsification	Reliability-Aware Filtering	Background Consistency↑	Motion Smoothness↑	Aesthetic Quality↑	PSNR↑	LPIPS↓	Sparsity↑	Latency↓
Full	✓	✓	✓	**0.942890**	**0.991999**	**0.469357**	**13.309393**	**0.554519**	**83.6842%**	×1.38
(a)	✓	✓		0.941184	0.991616	0.455922	13.232422	0.561628	81.5789%	×1.36
(b)		✓	✓	0.939966	0.991568	0.450583	11.050455	0.675472	80.5263%	×1.38
(c)			✓	0.939622	0.991600	0.457186	12.985179	0.567009	35.7368%	×1.32
(d)				0.937700	0.991267	0.446470	12.812125	0.581249	0.0000%	**×1.27**

Table 4: Effect of operation-oriented fine-tuning on HY1.5. Both variants use Qwen3.5-4B. Latency is relative to the Base; bold indicates the better value.

Variant	VBench			GPT-5.6-sol evaluation		Reconstruction		Efficiency	
	Background Consistency↑	Motion Smoothness↑	Aesthetic Quality↑	Camera-Constrained Consistency↑	Overall Visual Quality↑	PSNR↑	LPIPS↓	Sparsity↑	Latency↓
SMI (full)	**0.942890**	**0.991999**	**0.469357**	**67.0200**	**65.3677**	**13.309393**	**0.554519**	83.6842%	**×1.38**
w/o fine-tuning	0.941856	0.991978	0.451266	42.8380	44.7440	9.396166	0.714904	**86.3444%**	×1.42

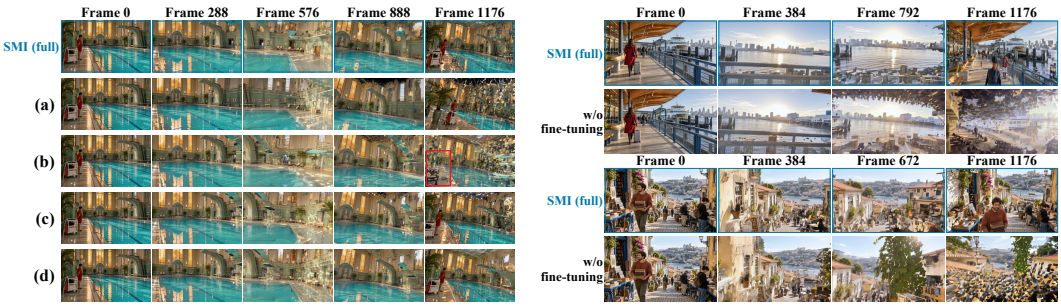

Figure 7: **Qualitative ablations on HY1.5. (1) Component ablation:** SMI (full) is compared with variants without (a) reliability-aware filtering, (b) spatial clustering, (c) spatial clustering and within-cluster sparsification, and (d) all three components. The red box highlights an inconsistent poolside region upon revisiting. **(2) Operation-oriented training:** SMI with the fine-tuned Qwen3.5-4B is compared with its counterpart without fine-tuning on two scenes.

3.5.2 Effect of Operation-Oriented Training

We compare the original Qwen3.5-4B with its operation-oriented fine-tuned counterpart in SMI. As shown in Table 4 and Figure 7(2), the understanding model without fine-tuning struggles to reliably perform the atomic operations, resulting in substantially lower generation quality and consistency. Fine-tuning on high-quality data enables the understanding model to adapt to these operations and manage complex spatial memory.

4 Conclusion

In this work, we presented SMI, the first unified framework that uses an understanding model to manage spatial memory in long-horizon video world models. SMI organizes memory through four coordinated atomic operations: spatial clustering, within-cluster sparsification, action-aware retrieval, and reliability-aware filtering. By sparsifying redundant memory observations while preserving useful spatial evidence, SMI substantially reduces memory overhead and improves long-horizon generation stability and spatial consistency. Extensive experiments across multiple baselines, benchmarks, and world-model backbones validate the effectiveness of SMI. Looking ahead, we hope this work takes a step toward a more general framework for memory management in future unified models.

AI USE STATEMENT

We used AI tools to assist with data annotation, manuscript translation and polishing, code development, and the evaluation of generated videos. The annotation and evaluation procedures are described in the corresponding sections of the paper. We reviewed the AI-assisted text and annotations and checked the generated code. We take full responsibility for the final content of this paper, including its claims, results, and associated artifacts.

ETHICS STATEMENT

SMI investigates spatial-memory management to support more stable and spatially consistent long-horizon video generation. Our operation-oriented training data are collected from model-generated trajectories and curated through manual annotation, model-assisted labeling, and human review. We remove low-quality samples and review and correct model-generated annotations to improve supervision quality. We encourage responsible research and use of SMI, including respect for privacy and applicable data and model licenses. Generated videos should be clearly identified as synthetic content when presented to users.

REPRODUCIBILITY STATEMENT

In Method Section 2, we describe SMI's four atomic operations and training objective. In Experiments Section 3, we present the model and optimization settings, evaluation metrics, baseline comparisons, and ablation studies. Appendices C–H provide additional implementation details, evaluation protocols, data construction, and operation-specific prompts. All test scenes are excluded from the operation-oriented training data. Within each backbone, all compared methods use the same initial observations, action sequences, random seeds, context budget, and generation settings. We will release the datasets and fine-tuned model weights to support reproduction and future research.

REFERENCES

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Jake Bruce, Michael 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 *Forty-first International Conference on Machine Learning*, 2024.

Han Cai, Junyan Li, Muyan Hu, Chuang Gan, and Song Han. Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 17302–17313, 2023.

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. *arXiv preprint arXiv:2508.21058*, 2025.

Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. In *International Conference on Learning Representations*, volume 2025, pp. 37546–37593, 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, 2024a.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024b.

Yuzhi Chen, Ronghan Chen, Dongjie Huo, Yandan Yang, Dekang Qi, Haoyun Liu, Tong Lin, Shuang Zeng, Junjin Xiao, Xinyuan Chang, et al. Abot-physworld: Interactive world foundation model for robotic manipulation with physics alignment. *arXiv preprint arXiv:2603.23376*, 2026.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, and Peter Grasch. MM-Spatial: Exploring 3D spatial understanding in multimodal LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7395–7408, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 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.

David Ha and Jurgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

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

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13504–13514, 2024.

Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, et al. Relic: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040*, 2025.

Anthony Hu, Václav Volhejn, Adrien Ramanana Rahary, Chris Mulder, Aditya Makkar, Amélie Royer, Manu Orsini, Alyx Liao, Adam Jelley, Eloi Alonso, et al. Multiplayer interactive world models with representation autoencoders. *arXiv preprint arXiv:2607.05352*, 2026.

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. *arXiv preprint arXiv:2510.03198*, 2025a.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, volume 38, pp. 167283–167308, 2025b.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

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

Yang Jin, Zhicheng Sun, Kun Xu, Kun Xu, Liwei Chen, Hao Jiang, Quzhe Huang, Chengru Song, Yuliang Liu, Di Zhang, Yang Song, Kun Gai, and Yadong Mu. Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22185–22209, 2024.

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. In *International Conference on Learning Representations*, volume 2025, pp. 23378–23402, 2025.

Wayne King, Zeyue Xue, Yuxuan Bian, Jie Huang, Haoran Li, Yaowei Li, Yaofeng Su, Yuming Li, Haoyu Wang, Shiyi Zhang, et al. Echo-memory: A controlled study of memory in action world models. *arXiv preprint arXiv:2606.09803*, 2026.

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*, 2(3):6, 2025a.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026.

Pengteng Li, Pinhao Song, Wuyang Li, Huizai Yao, Weiyu Guo, Yijie Xu, Dugang Liu, and Hui Xiong. See&Trek: Training-free spatial prompting for multimodal large language model. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25690–25699, 2025c.

Fangfu Liu, Kai He, Tianchang Shen, Tianshi Cao, Sanja Fidler, Yueqi Duan, Jun Gao, Igor Gilitschenski, Zian Wang, and Xuanchi Ren. Gamma-world: Generative multi-agent world modeling beyond two players. *arXiv preprint arXiv:2605.28816*, 2026a.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/935151cc6cb5d8b6816133b75233775a-Abstract-Conference.html.

Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. Spatial-SSRL: Enhancing spatial understanding via self-supervised reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9570–9581, 2026c.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26439–26455, 2024.

Chenyang Ma, Kai Lu, Ta-Ying Cheng, Niki Trigoni, and Andrew Markham. SpatialPIN: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3D priors. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan L. Yuille. SpatialReasoner: Towards explicit and generalizable 3D spatial reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025a.

Wufei Ma, Luoxin Ye, Celso M. de Melo, Alan Yuille, and Jieneng Chen. Spatialllm: A compound 3d-informed design towards spatially-intelligent large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17249–17260, 2025b.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, and Kaipeng Zhang. Yume1.5: A text-controlled interactive world generation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7752–7761, 2026.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, et al. Mm1: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pp. 304–323. Springer, 2024.

Jonas Pai, Liam Achenbach, Victoriano Montesinos, Benedek Forrai, Oier Mees, and Elvis Nava. mimic-video: Video-action models for generalizable robot control beyond vlas. *arXiv preprint arXiv:2512.15692*, 2025.

Jack Parker-Holder and Shlomi Fruchter. Genie 3: A new frontier for world models. Google DeepMind Blog, 2025. URL https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14313–14323, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6121–6132, 2025.

Nedko Savov, Naser Kazemi, Deheng Zhang, Danda Pani Paudel, Xi Wang, and Luc V Gool. Statespacediffuser: Bringing long context to diffusion world models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 68865–68898, 2025.

Georgy Savva, Oscar Michel, Daohan Lu, Suppakit Waiwitlikhit, Timothy Meehan, Dhairya Mishra, Srivats Poddar, Jack Lu, and Saining Xie. Solaris: Building a multiplayer video world model in minecraft. *arXiv preprint arXiv:2602.22208*, 2026.

Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James M. Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in VLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18221–18232, 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.

Junshu Tang, Jiacheng Liu, Jiaqi Li, Longhuang Wu, Haoyu Yang, Penghao Zhao, Siruis Gong, Xiang Yuan, Shuai Shao, Linfeng Zhang, et al. Hunyuan-gamecraft-2: Instruction-following interactive game world model. *arXiv preprint arXiv:2511.23429*, 2025.

DreamX Team, Yancheng Bai, Rui Chen, Xiangxiang Chu, Rujing Dang, Hao Dou, Bingjie Gao, Qiwen Gu, Siyu Hong, Jiachen Lei, et al. Dreamx-world 1.0: A general-purpose interactive world model. *arXiv preprint arXiv:2606.16993*, 2026a.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026b.

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

Zehan Wang, Sashuai Zhou, Shaoxuan He, Haifeng Huang, Lihe Yang, Ziang Zhang, Xize Cheng, Shengpeng Ji, Tao Jin, Hengshuang Zhao, and Zhou Zhao. SpatialCLIP: Learning 3D-aware image representations from spatially discriminative language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29656–29666, 2025.

Zehan Wang, Tengfei Wang, Haiyu Zhang, Xuhui Zuo, Junta Wu, Haoyuan Wang, Wenqiang Sun, Zhenwei Wang, Chenjie Cao, Hengshuang Zhao, et al. Worldcompass: Reinforcement learning for long-horizon world models. *arXiv preprint arXiv:2602.09022*, 2026a.

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*, 2026b.

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. *arXiv preprint arXiv:2606.02436*, 2026.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12966–12977, 2025a.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

Ruiqi Wu, Xuanhua He, Meng Cheng, Tianyu Yang, Yong Zhang, Zhuoliang Kang, Xunliang Cai, Xiaoming Wei, Chunle Guo, Chongyi Li, et al. Infinite-world: Scaling interactive world models to 1000-frame horizons via pose-free hierarchical memory. *arXiv preprint arXiv:2602.02393*, 2026.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. In *Advances in Neural Information Processing Systems*, volume 38, pp. 49371–49393, 2025c.

Xiaofei Wu, Guozhen Zhang, Zhiyong Xu, Yuan Zhou, Qinglin Lu, and Xuming He. Pack and force your memory: Long-form and consistent video generation. *arXiv preprint arXiv:2510.01784*, 2025d.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. Vila-u: A unified foundation model integrating visual understanding and generation. In *International Conference on Learning Representations*, 2025e.

Haocheng Xi, Shuo Yang, Yilong Zhao, Chenfeng Xu, Muyang Li, Xiuyu Li, Yujun Lin, Han Cai, Jintao Zhang, Dacheng Li, et al. Sparse videogen: Accelerating video diffusion transformers with spatial-temporal sparsity. *arXiv preprint arXiv:2502.01776*, 2025.

Yifei Xia, Suhan Ling, Fangcheng Fu, Yujie Wang, Huixia Li, Xuefeng Xiao, and Bin Cui. Training-free and adaptive sparse attention for efficient long video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15982–15993, 2025.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. In *Advances in Neural Information Processing Systems*, volume 38, pp. 49632–49652, 2025.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, 2025.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10632–10643, 2025a.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, et al. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025b.

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

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

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

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. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025a.

Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11590–11599. IEEE, 2025b.

Yifei Yu, Xiaoshan Wu, Xinting Hu, Tao Hu, Yangtian Sun, Xiaoyang Lyu, Bo Wang, Lin Ma, Yuewen Ma, Zhongrui Wang, et al. Videossm: Autoregressive long video generation with hybrid state-space memory. *arXiv preprint arXiv:2512.04519*, 2025c.

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.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Chen Sun, Ming-Hsuan Yang, Forrester Cole, Trevor Darrell, and Deqing Sun. LoGeR: Long-context geometric reconstruction with hybrid memory. *arXiv preprint arXiv:2603.03269*, 2026a.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 30546–30566, 2025a.

Lvmin Zhang, Shengqu Cai, Muyang Li, Chong Zeng, Beijia Lu, Anyi Rao, Song Han, Gordon Wetzstein, and Maneesh Agrawala. Tinyhistory: Lightweight video history embeddings via two-stage context learning. *arXiv preprint arXiv:2512.23851*, 2026b.

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

Ruosen Zhao, Zhikang Zhang, Jialei Xu, Jiahao Chang, Dong Chen, Lingyun Li, Weijian Sun, and Zizhuang Wei. SpaceMind: Camera-guided modality fusion for spatial reasoning in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16811–16822, 2026.

Yiqi Zhu, Ziyue Wang, Can Zhang, Peng Li, and Yang Liu. CoSpace: Benchmarking continuous space perception ability for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29569–29579, 2025.

APPENDIX

A QUALITATIVE EXAMPLES OF RUNTIME MEMORY MANAGEMENT

Figures 8 and 9 visualize action-aware retrieval: conditioned on the current observation and action, the MLLM performs spatial reasoning to select historical candidates that provide appropriate spatial context. Geometry-only retrieval may select irrelevant or incorrect candidates (e.g., A or B) because it does not understand the spatial relationships between observations. Figure 10 illustrates assigning a new chunk to an existing cluster or creating a new one; Figure 11 shows how chunks are accepted or rejected based on visual reliability; and Figure 12 illustrates retaining informative chunks and removing redundant ones through within-cluster sparsification.

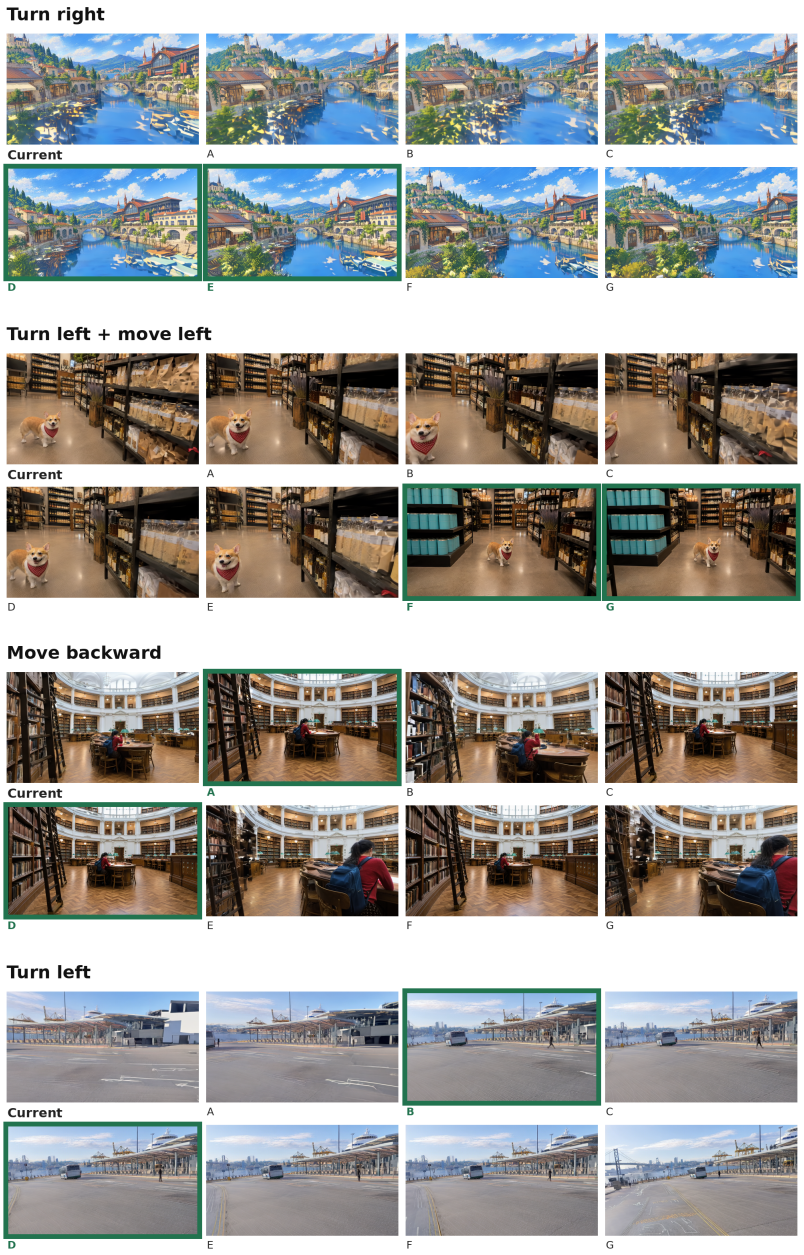

Figure 8: Action-aware retrieval examples. 'Current' denotes the current observation, and A–G denote historical candidates. Green borders mark the selected candidates.

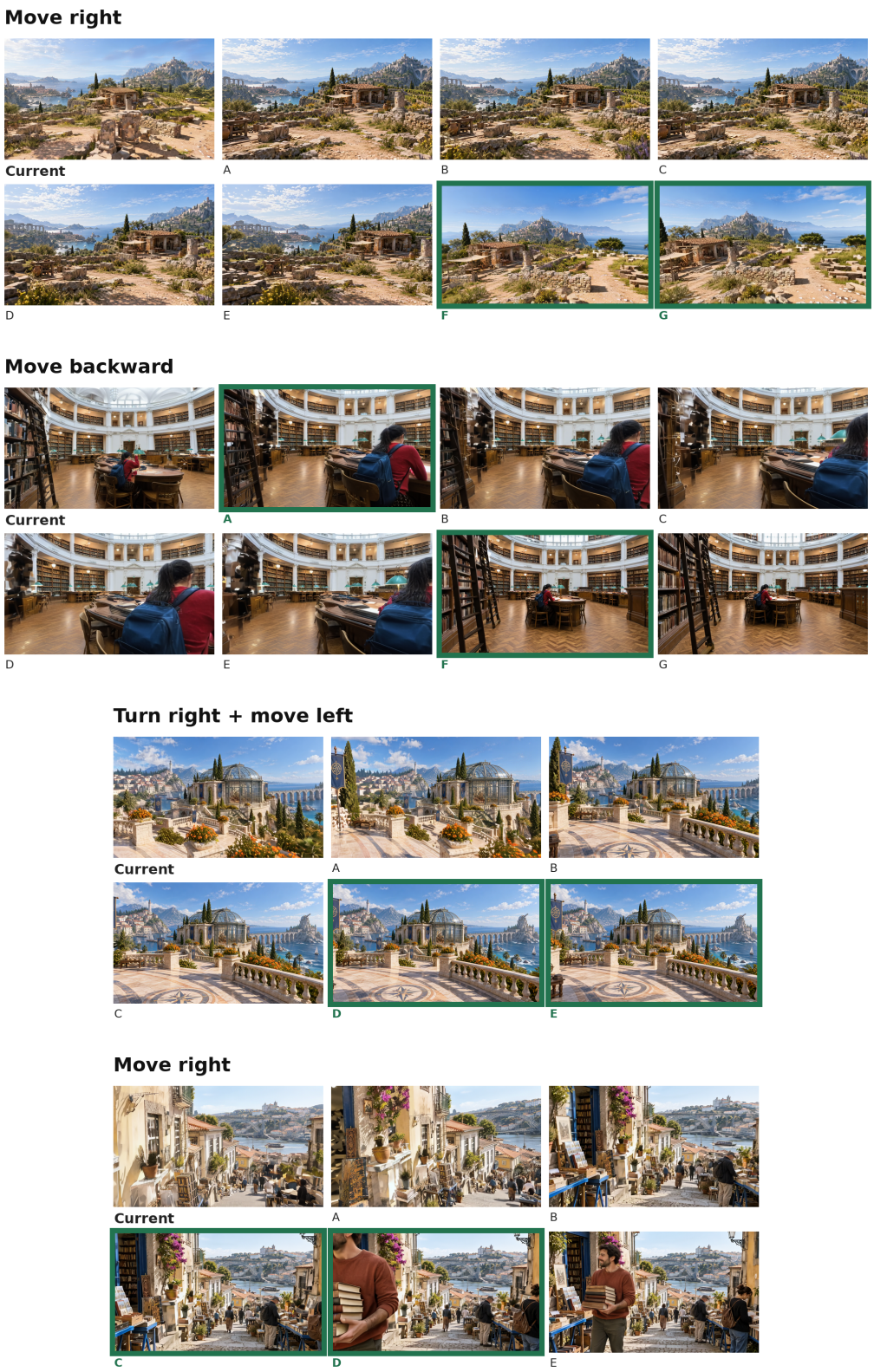

Figure 9: Additional action-aware retrieval examples. ‘Current’ denotes the current observation, and A–G denote historical candidates. Green borders mark the selected candidates.

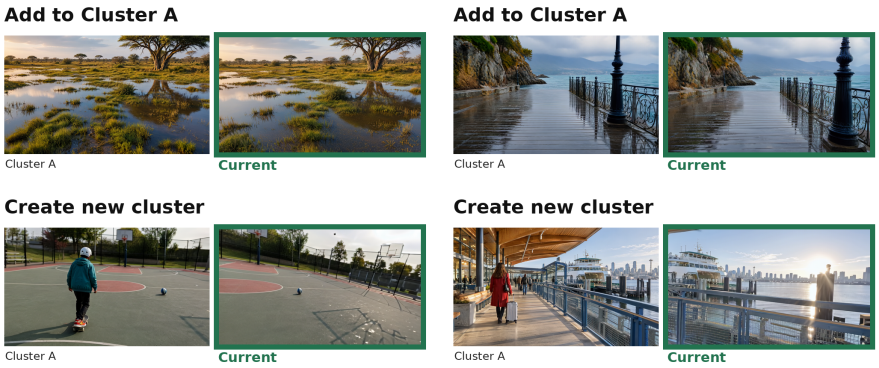

Figure 10: Spatial clustering: add to an existing cluster (top) or create a new cluster (bottom).

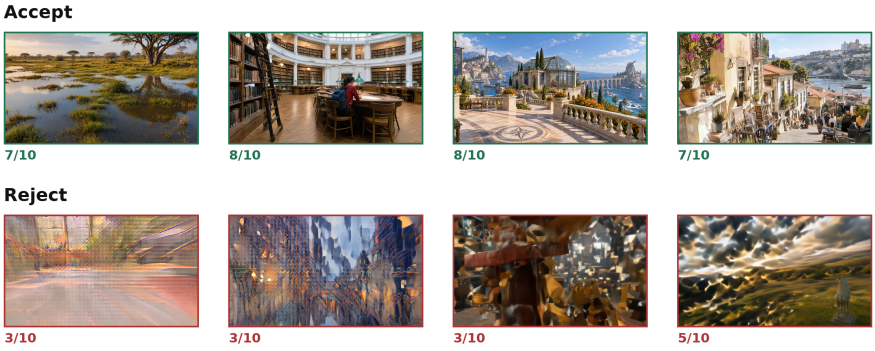

Figure 11: Reliability-aware filtering: accepted chunks (top) and rejected chunks (bottom).

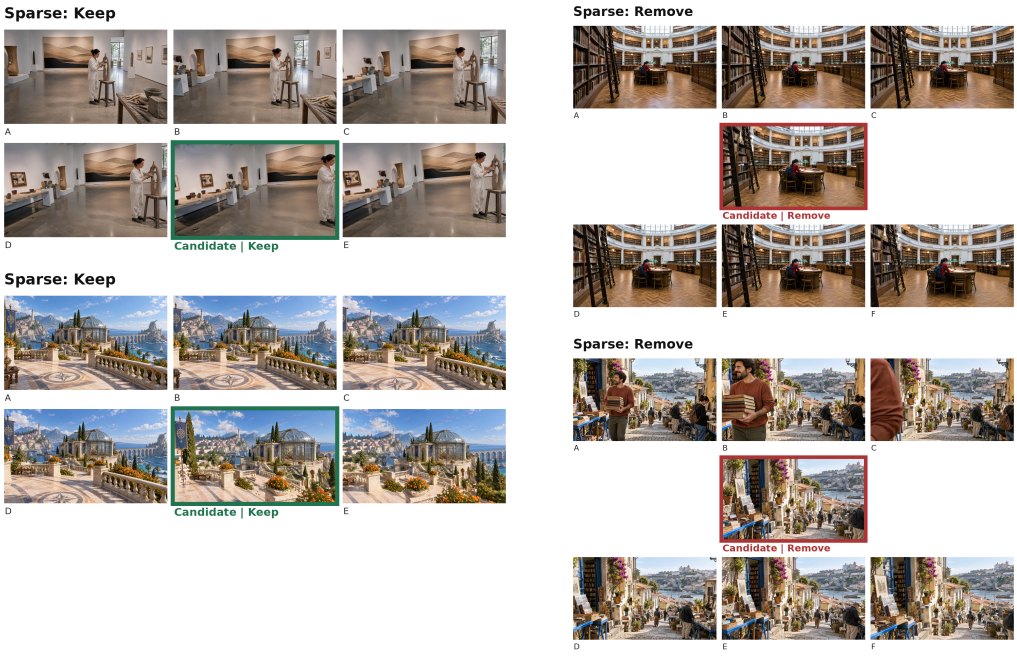

Figure 12: Within-cluster sparsification: KEEP decisions (left) and REMOVE decisions (right).

B Related Work

B.1 Video World Models

Classical world models learn predictive environment representations for planning and control (Ha & Schmidhuber, 2018; Hafner et al., 2023). Enabled by scalable Diffusion Transformer architectures (Peebles & Xie, 2023), video generation has advanced substantially (HaCohen et al., 2024; Yang et al., 2025c; Wan et al., 2025), accelerating the development of video world models for interactive simulation of evolving environments. Early video world models focused on controllable visual simulation in games and virtual environments (Bruce et al., 2024; Che et al., 2025; Zhang et al., 2025b). Subsequent work has moved toward persistent world simulation, combining streaming generation (Chen et al., 2024a; Huang et al., 2025b; Yuan et al., 2026) with interactive control (Yu et al., 2025b; Sun et al., 2025) to support long-horizon interaction in evolving environments (Li et al., 2025a; Hong et al., 2025; Wang et al., 2026a; Team et al., 2026b). Recent work has further extended video world modeling to multi-agent interaction (Savva et al., 2026; Hu et al., 2026; Liu et al., 2026a) and embodied simulation for robotics and autonomous systems (Pai et al., 2025; Agarwal et al., 2026; Ye et al., 2026; Li et al., 2026). Despite these advances, persistent world simulation remains challenging, as long-horizon generation requires distilling an ever-growing observation history into a sparse memory while retaining the information needed to preserve scene consistency. This motivates mechanisms for joint memory sparsification and consistency-aware retrieval.

B.2 Memory in Video Generation and World Models

Because world models generate autoregressively, their spatial memory grows with the interaction horizon, making selective retention critical for generation efficiency, scene consistency, and task success. Existing approaches manage historical context through compression, sparsification, and retrieval. Compression-based methods (Zhang et al., 2025a; Jin et al., 2025; Savov et al., 2025; Wu et al., 2026; Hong et al., 2025; Zhang et al., 2026b; Wei et al., 2026) condense an expanding context into a bounded representation, but may discard fine-grained visual evidence. Sparse-computation methods (Xi et al., 2025; Cai et al., 2023; Choromanski et al., 2020; Xia et al., 2025; Cai et al., 2025) reduce computation through sparsity, routing, or linear attention. Retrieval-based methods instead select a subset of the history. Field-of-view and geometry-based retrieval (Yu et al., 2025a; Xiao et al., 2025; Wu et al., 2025c; Ren et al., 2025; Li et al., 2025c; Huang et al., 2025a) exploit camera or scene geometry, but remain vulnerable to occlusion, dynamic or semantic changes, geometric errors, and queries requiring higher-level reasoning. Semantic retrieval (Ji et al., 2025; Wu et al., 2025d) captures high-level relevance but lacks explicit spatial grounding, making semantically similar observations from different locations or viewpoints difficult to distinguish. Some methods jointly address memory compression and implicit content querying (Zhang et al., 2026b) through an auxiliary encoder and the generative model itself, leaving memory interpretation largely coupled to the generator. Collectively, these limitations motivate a memory manager with spatial perception, semantic understanding, and long-horizon reasoning.

B.3 Multimodal intelligence

Spatial-memory management inherently requires spatial perception, semantic understanding, and long-horizon reasoning. Recent advances in multimodal intelligence provide a natural foundation for these requirements. MLLMs have acquired increasingly strong capabilities in long-video reasoning (Ren et al., 2024; Song et al., 2024; He et al., 2024). Meanwhile, VLMs have progressed from basic spatial perception toward reasoning over 3D and multi-view environments, including cognitive-map construction and continuous scene understanding (Chen et al., 2024b; Ma et al., 2024; 2025b; Yang et al., 2025a; Wang et al., 2025; Zhu et al., 2025; Yang et al., 2025b). Furthermore, recent models explore depth and multi-view inputs, geometry-aware representations, and spatially grounded learning strategies to improve spatial perception and reasoning (Daxberger et al., 2025; Wu et al., 2025b; Ma et al., 2025a; Li et al., 2025b; Shen et al., 2025; Zhao et al., 2026; Liu et al., 2026c). In parallel, unified multimodal models are beginning to dissolve the traditional boundary between understanding and generation (Lu et al., 2024; Jin et al., 2024; Wu et al., 2025a; Xie et al., 2025; Wu et al., 2025e; Deng et al., 2025). These advances point toward a forward-looking paradigm in which understanding is not merely an auxiliary capability applied after generation, but an integral component that actively organizes and governs the generative process.

C Additional Experimental Details

We implement the four spatial-memory operations on HY1.5 and Wan2.2 as follows.

Spatial clustering. We extract the middle frame from the newly generated chunk and from each candidate cluster prototype. The MLLM determines whether they depict the same spatial region. If a match exists, the new chunk is assigned to the corresponding cluster; otherwise, a new cluster is created.

Within-cluster sparsification. When the number of chunks in a cluster reaches a predefined threshold, we sample two frames from each chunk. The MLLM compares their spatial content to identify redundancy. An intermediate chunk is removed if it provides no additional spatial structure, object states, or occlusion information relative to its neighboring memories.

Action-aware retrieval. We first shortlist the seven historical chunks whose camera positions are closest to the current camera position. The MLLM then considers the spatial location depicted in recent observations and the current action to select historical chunks that provide relevant spatial information for the current generation. Both backbones retain the two most recent chunks as context. No additional chunk is retrieved if none of the candidates is suitable.

Reliability-aware filtering. The MLLM scores each newly generated chunk on a scale of 1–10 for visual reliability, with higher scores indicating greater reliability. Chunks scoring below 6 are discarded rather than added to the memory bank, reducing the accumulation of generation errors through subsequent memory reuse.

D GPT-5.6-sol Evaluation

D.1 Evaluation Protocol

Input construction. For each generated video, we sample one representative frame from every generated chunk and preserve their temporal order to form a long-horizon evaluation sequence. The corresponding action sequence is provided when judging camera-motion adherence. This chunk-level sampling retains the evolution of the generated scene while keeping the evaluation input tractable.

Blind multi-agent scoring. We use three independent GPT-5.6-sol evaluator agents. Method identities are hidden, and the candidate videos are presented in randomized order. All evaluators receive the same instruction and independently score the complete sequence. Their predictions are aggregated only after the three evaluations are completed, reducing the influence of ordering and individual evaluator variation.

Evaluation dimensions. The protocol assesses adjacent transition continuity, long-range identity preservation, cumulative geometric consistency, overall scene consistency, camera-motion adherence, structural integrity, artifact cleanliness, detail clarity, and overall visual quality. These dimensions jointly capture the consistency, stability, continuity, and visual reliability of long-horizon generation.

Reporting. The main paper reports Camera-Constrained Scene Consistency (the Camera-Gated Scene column below) and Overall Visual Quality. Tables 5 and 6 report all nine score dimensions for HY1.5 and Wan2.2. Bold marks the highest score in each column.

Table 5: Complete GPT-5.6-sol evaluation on HY1.5. All scores are on a 0–100 scale; higher is better.

Method	Adjacent transition	Long-range identity	Cumulative geometry	Overall scene consistency	Camera-gated scene	Structural integrity	Artifact cleanliness	Detail clarity	Overall visual quality
Base (FoV)	63.5462	51.8905	45.5081	53.2088	51.6053	55.9933	48.0538	53.7048	53.1538
FramePack	56.9633	42.8667	35.3834	44.5193	43.2800	49.8634	40.5100	47.8134	46.6474
Deep Forcing	44.2133	39.6633	28.2400	37.9464	37.6760	37.5733	25.1633	36.4701	33.6297
MoC	61.4733	49.6100	42.3133	50.7506	48.9196	54.4600	46.8767	53.3933	51.9717
VMem	66.0933	55.7800	48.6767	56.5863	54.4727	59.2067	51.3800	57.4433	56.5060
MemFlow	56.7033	41.1300	34.8067	43.4457	43.3160	48.9834	39.9500	45.8667	45.6500
SMI (ours)	**75.3666**	**68.4500**	**62.2167**	**68.6197**	**67.0200**	**68.3000**	**60.6034**	**65.1833**	**65.3677**

Table 6: Complete GPT-5.6-sol evaluation on Wan2.2. All scores are on a 0–100 scale; higher is better.

Method	Adjacent transition	Long-range identity	Cumulative geometry	Overall scene consistency	Camera-gated scene	Structural integrity	Artifact cleanliness	Detail clarity	Overall visual quality
Base (FoV)	47.8335	31.0978	29.1101	34.8049	34.6524	39.6402	34.8039	37.8779	37.8363
FramePack	51.6066	30.6701	28.8400	35.4547	34.7984	39.7833	34.9867	38.0200	37.9917
Deep Forcing	**61.6470**	44.8030	43.2600	**48.6450**	46.8560	50.8130	**49.6300**	52.1630	50.7300
MoC	59.8633	**45.9233**	43.0867	48.2164	45.5013	48.9634	47.7200	**52.3930**	49.2763
VMem	58.9533	41.0967	37.2434	44.6153	43.8800	45.6400	43.5534	47.3770	45.3613
MemFlow	54.3633	36.1333	31.5034	39.5520	38.8487	40.2234	39.1734	42.1770	40.2990
SMI (ours)	59.2771	42.8278	**43.5601**	47.1430	**46.9309**	**52.8471**	48.9702	51.8770	**51.4909**

E User Study

We compare Base, VMem, FramePack, and SMI on 15 scenes, with videos lasting 45–60 seconds under matched initial observations and action sequences. Twenty participants evaluate scene consistency and generation stability. As shown in Table 7, SMI leads all compared methods in both dimensions, receiving best-choice preference shares of 52.50% for scene consistency and 57.83% for generation stability.

	Base	VMem	FramePack	SMI (ours)
Scene consistency (%)	11.00	27.00	7.50	**52.50**
Generation stability (%)	11.33	27.83	2.00	**57.83**

Table 7: Best-choice preference shares (%) on 15 scenes. Bold indicates the highest share in each row.

F Baseline Implementation Details

Evaluation settings. We compare Base, FramePack, Deep Forcing, Mixture of Contexts (MoC), VMem, MemFlow, and SMI on HY1.5 and Wan2.2, keeping the backbone weights unchanged. Within each backbone, all methods share the same initial observations, action sequences, random seeds, context budget, and generation settings. We retain the two most recent chunks as local context and retrieve at most two additional historical chunks for HY1.5 and three for Wan2.2.

Base. The original backbone retains the local context and retrieves earlier chunks according to field-of-view overlap, up to the backbone-specific cap. It does not use an auxiliary memory manager.

FramePack. We adapt FramePack's temporal hierarchy to the chunk interface. One initial anchor is retained, and older observations are selected by temporal distance using the $4\times$, $2\times$, and $1\times$ tiers. The selected observations follow the native context path; no FramePack-specific encoder or weight is introduced.

Mixture of Contexts (MoC). We adapt the MoC routing idea as a global chunk router at inference. Each chunk contributes a compact block-0 attention descriptor. At each step, the router ranks remote

chunks by query-key similarity and combines the highest-scoring chunks with the two local chunks, subject to the backbone-specific cap.

Deep Forcing. Deep Forcing uses a token-level cache with a 10-frame deep-sink prefix and a 4-frame recent query window. When the cache is full, non-sink historical tokens are retained according to attention scores; temporal RoPE positions are compensated after reorganization. We use the strict token-cache implementation and do not use the separate random-selection or compression-free variants.

VMem. Our VMem implementation builds a surfel map from geometry estimated by LoGeR (Zhang et al., 2026a) and scores historical chunks by their contribution to the target-view rendering. It combines the two local chunks with at most two additional chunks on HY1.5 or three on Wan2.2, and then re-encodes the selected chunks through the native context path.

MemFlow. We adapt MemFlow's semantic-retrieval component to the native chunk interface. Every second chunk contributes its first decoded RGB frame as a candidate, and the last decoded frame of the preceding chunk is used as the query. We rank candidates by CLIP (Radford et al., 2021) image-image cosine similarity, exclude the two local chunks, and retrieve at most two remote chunks on HY1.5 or three on Wan2.2.

G Operation-Oriented Data Construction

Collection and manual annotation. We first collect approximately one thousand videos, each longer than one minute, from actual world-model inference and manually remove low-quality videos. We then manually annotate 354 samples in total across the four operations, providing reliable supervision examples for the subsequent annotation stage.

Teacher annotation and quality control. Using the manually annotated samples as few-shot demonstrations, GPT-5.6-sol and Gemini-3.1 Pro annotate the remaining data. Claude Fable 5 then examines the annotations produced by the two teacher MLLMs and flags anomalous outputs. Finally, we further review and clean the flagged samples to obtain the complete training dataset.

Operation-specific supervision. The resulting samples are organized according to the four atomic operations. Reliability-aware filtering predicts the visual reliability of a newly generated chunk. Spatial clustering determines whether a new chunk matches an existing spatial prototype. Within-cluster sparsification determines whether a candidate chunk is redundant given the remaining memories in the same cluster. Action-aware retrieval selects historical chunks that can support generation under the current context and planned action. Each sample pairs its operation-specific multimodal input $\mathcal{I}_n$ with a structured target output Y_n.

Dataset composition. Because HY1.5 and Wan2.2 have different generation characteristics, we construct a separate supervision dataset for each backbone. Table 8 reports the number of training samples for every operation. The HY1.5 training set contains 63,911 samples in total, including 24,623 samples for action-aware retrieval (Coverage), accounting for approximately 38.53% of the training set.

Table 8: Composition of the operation-oriented training datasets.

Operation	HY1.5	Wan2.2
Spatial clustering	12,652	12,538
Within-cluster sparsification	7,156	7,135
Action-aware retrieval	24,623	6,852
Reliability-aware filtering	19,480	4,411
Total	63,911	30,936

G.1 Operation-Wise Label Composition

We further report the supervision-label composition of each atomic operation. All statistics below are computed over the effective training occurrences used for optimization.

Table 9: Label composition of the spatial clustering, within-cluster sparsification, and reliability-aware filtering training data.

Operation	Label	HY1.5	Wan2.2
Spatial clustering	Belongs	9,041 (71.46%)	8,972 (71.56%)
Spatial clustering	Does not belong	3,611 (28.54%)	3,566 (28.44%)
Within-cluster sparsification	Keep	2,045 (28.58%)	2,051 (28.75%)
Within-cluster sparsification	Remove	5,111 (71.42%)	5,084 (71.25%)
Reliability-aware filtering	Filtered (≤ 6)	4,783 (24.55%)	1,711 (38.79%)
Reliability-aware filtering	Retained (> 6)	14,697 (75.45%)	2,700 (61.21%)

Spatial clustering. Each training sample compares a newly generated chunk with one candidate spatial prototype. A *belongs* label indicates that the two observations describe the same local spatial region, whereas *does not belong* indicates that they should not be assigned to the same cluster. HY1.5 contains 9,041 belongs samples and 3,611 does-not-belong samples; Wan2.2 contains 8,972 and 3,566 samples, respectively.

Within-cluster sparsification. Each sample evaluates a candidate chunk relative to the remaining memories in the same cluster. A *keep* label indicates that the chunk retains useful spatial evidence, while *remove* indicates that it is redundant. The HY1.5 dataset contains 2,045 keep samples and 5,111 remove samples, whereas Wan2.2 contains 2,051 and 5,084 samples, respectively.

Reliability-aware filtering. The visual-quality task assigns a reliability score from 1 to 10 to a newly generated chunk. Following the filtering rule used by SMI, chunks with scores no greater than 6 are labeled as filtered, while chunks with scores greater than 6 are retained. Under this binary grouping, 24.55% of HY1.5 samples and 38.79% of Wan2.2 samples are filtered. The remaining 75.45% and 61.21%, respectively, are retained. These statistics characterize the supervision distribution rather than the runtime filtering rate.

Action-aware retrieval. Coverage supervision conditions memory selection on the planned camera motion. Table 10 reports the motion distribution over 24,623 HY1.5 and 6,852 Wan2.2 training occurrences. The percentages therefore describe the effective exposure seen during training, including motion balancing. For HY1.5, 15 and 9 legacy records with non-standard motion strings are merged into the `tilt_down` and `tilt_up` categories, respectively.

H Prompts for the Four Atomic Operations

We use four operation-specific prompts for reliability-aware filtering, spatial clustering, within-cluster sparsification, and action-aware retrieval. The templates below preserve the decision rules used at runtime; implementation paths, hashes, and dynamically populated image and geometry fields are omitted.

Reliability-Aware Filtering

Role and input. Evaluate one frame from a newly generated chunk; when multiple frames are provided, judge the overall visual reliability of the chunk.
Task. Assign a visual-quality score from 1 to 10 indicating whether the generated chunk is reliable enough to serve as a future memory reference. Judge only visible corruption, blur, structural artifacts, unstable geometry, distorted layout, broken regions, and overall reference reliability. An unusual scene, object, or viewpoint should not be penalized unless it contains a visible quality failure.
Scale. Scores 1–2 denote an unusable result; 3–4 denote obvious artifacts or instability; 5 denotes a marginal reference; 6–7 denote a mostly reliable reference with minor defects; 8–9 denote a clean and stable reference; and 10 denotes an excellent reference without meaningful defects.
Output. Return strict JSON containing only the integer field `visual_quality_score`.

Table 10: Motion composition of the action-aware retrieval training data.

Motion	HY1.5	Wan2.2
mostly_static	0.75%	2.01%
move_backward	9.36%	7.41%
move_forward	12.50%	7.68%
move_left	10.20%	7.59%
move_right	9.81%	7.41%
tilt_down	2.70%	4.25%
tilt_up	2.47%	4.44%
tilt_down+move_backward	1.87%	2.85%
tilt_down+move_forward	1.91%	3.12%
tilt_down+move_left	0.75%	1.69%
tilt_down+move_right	0.75%	1.59%
tilt_up+move_backward	0.75%	1.72%
tilt_up+move_forward	1.87%	2.47%
tilt_up+move_left	0.75%	1.71%
tilt_up+move_right	1.87%	1.88%
turn_left	8.72%	5.65%
turn_right	8.57%	6.67%
turn_left+move_backward	1.87%	3.30%
turn_left+move_forward	1.23%	3.65%
turn_left+move_left	1.87%	3.30%
turn_left+move_right	1.62%	4.28%
turn_left+tilt_down	0.19%	0.03%
turn_left+tilt_up	0.19%	0.07%
turn_right+move_backward	1.87%	3.50%
turn_right+move_forward	1.22%	3.59%
turn_right+move_left	1.44%	4.67%
turn_right+move_right	1.87%	3.44%
turn_right+tilt_down	0.19%	0.03%
Unclassified motion metadata	10.86%	0.00%

Action-Aware Retrieval

Input. The final frame of the immediately preceding chunk, the planned camera motion, and candidate historical final frames with their labels, chunk identifiers, and precomputed visual-quality scores.
Task. Judge which candidates can support the expected next view through viewpoint coverage, visible-boundary continuity, spatial support, wider context, or recent dynamic-state continuity. Evaluate every candidate independently.
Motion grouping. Assign every candidate exactly once to `supports`, `neutral`, or `opposes`. Motion alignment and coverage usefulness are separate judgments; an opposing candidate may remain useful when it provides strong spatial support, although an equally useful supporting or neutral candidate receives higher priority. Forward or nearly static motion favors high-overlap stable references; backward or widening motion favors complementary broad coverage; and lateral or turning motion favors support on the corresponding side of the expected next view.
Acceptance and selection. Put every useful reference in `acceptable`. Then choose at most two labels from this set for `selected`, prioritizing the strongest next-view support and complementary coverage while avoiding near duplicates. When no candidate is useful, both lists are empty. Visual-quality scores are reliability hints rather than the sole selection criterion.
Output. Return strict JSON containing only the three motion groups, `acceptable`, and `selected`.

Spatial Clustering

Input. The middle frame of the new chunk, the middle frame of one existing class prototype, and `prototype_fov_overlap` in $[0, 1]$ as a geometric hint.
Task. Decide whether the new chunk and the prototype belong to the same local viewpoint range. This is only the clustering decision; no sparsification or memory-admission decision is made here. Semantic similarity alone is insufficient: the two views must share concrete visible area, compatible local layout, and viewpoint continuity.
Labels and rules. Assign `view_overlap_score` from 1 to 6, where 1 indicates no compatible local overlap and 6 indicates very strong spatial alignment. Set `match=false` when the score is at most 2 and `match=true` when it is at least 3. The FoV overlap is used as a hint rather than the sole criterion.
Output. Return strict JSON containing only `match` and `view_overlap_score`.

Within-Cluster Sparsification

Input. One middle frame from every remaining chunk in the same cluster, the middle frame of the candidate chunk, and `coverage_by_remaining_chunks` in $[0, 1]$ as a geometric hint.
Task. Remove the candidate only when the remaining chunks already cover its useful spatial information and it contributes no useful information beyond them. Otherwise, keep it.
Labels. Predict `spatial_coverage` from {none, weak, medium, strong, full} and `unique_useful_information` from {none, small, clear}. Unique information may include an unrepresented viewpoint, boundary, layout, transition, occlusion, dynamic state, or local spatial relation that is useful for future consistency.
Decision rules. Output `remove` only when coverage is strong or full and unique useful information is none or small. Keep the candidate when coverage is none, weak, or medium, or whenever unique useful information is clear. The numeric coverage hint supports the decision but never determines it alone.
Output. Return strict JSON containing only `action`, `spatial_coverage`, and `unique_useful_information`.

I Limitations and Future Work

Limitations. SMI has several limitations. First, the current framework is not trained end to end with the world model and therefore does not directly improve the intrinsic generation capability of the underlying backbone. Its performance remains bounded by the visual quality, controllability, and stability of the base generator. Second, because the understanding model and the generation model operate with different representations, the MLLM-based memory controller introduces additional computation, particularly during the early stages of a rollout before memory sparsification yields substantial savings. Third, the current task setting does not explicitly address long-horizon streaming understanding and may therefore remain insufficient for memory tasks involving complex long-term dependencies or logical reasoning.

Future work. Based on these limitations, several directions can be explored. First, generation, understanding, and memory management could be jointly learned within a shared representation space. Such a formulation may not only reduce computational overhead and enable end-to-end optimization, but also allow understanding and generation to mutually reinforce each other, thereby further enhancing the world model's modeling and memory capabilities. Second, reinforcement learning could be introduced into long-horizon world models with unified understanding and generation to further strengthen the understanding model's ability to manage long-term memory. Third, higher-quality and more comprehensive datasets could be constructed to enable understanding models to learn broader memory-management capabilities.

J Additional Qualitative Comparisons

We provide additional qualitative comparisons on HY1.5 (Figures 13–16) and Wan2.2 (Figures 17–18). SMI achieves comprehensive improvements in generation stability and spatial consistency.

J.1 HY1.5

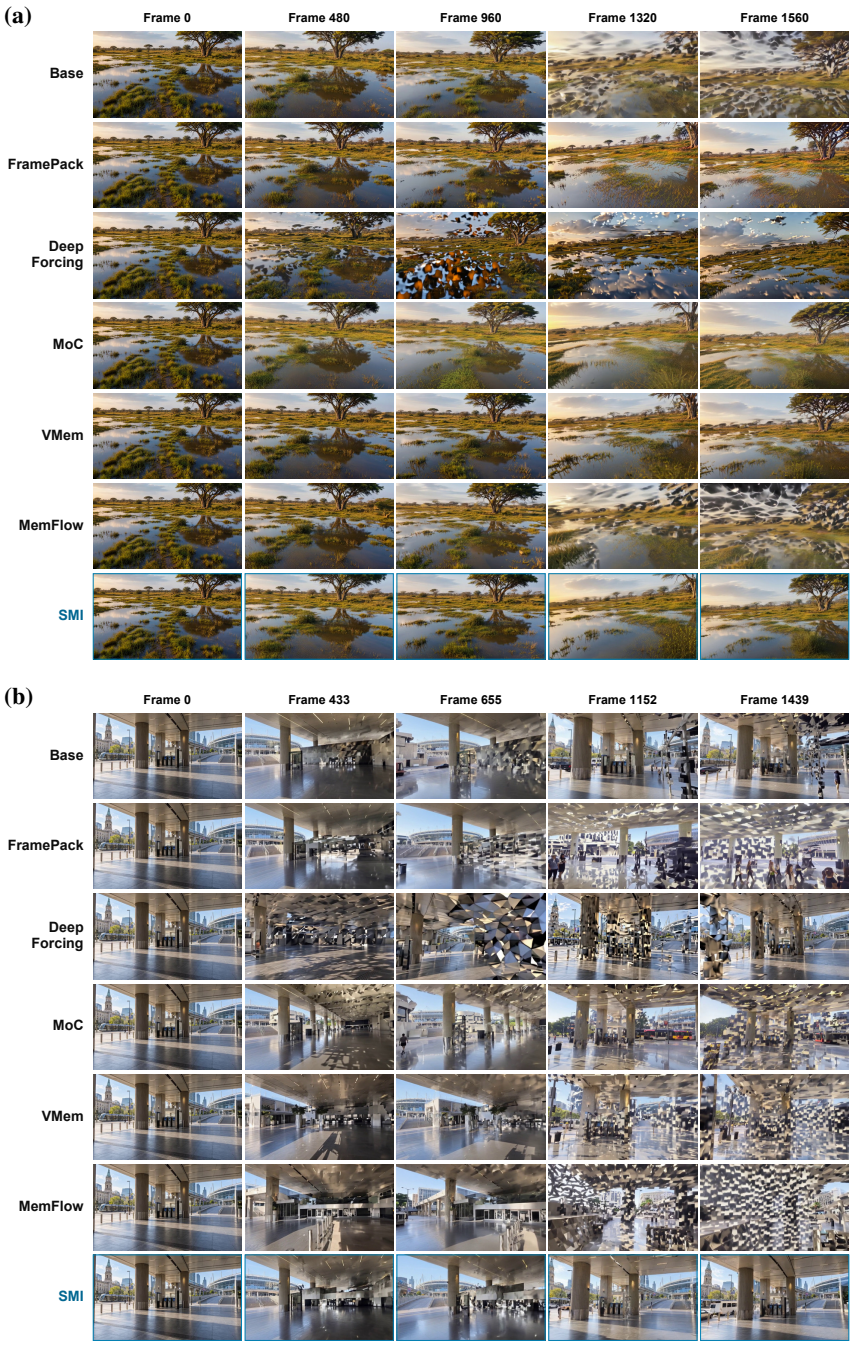

Figure 13: Additional qualitative results on HY1.5.

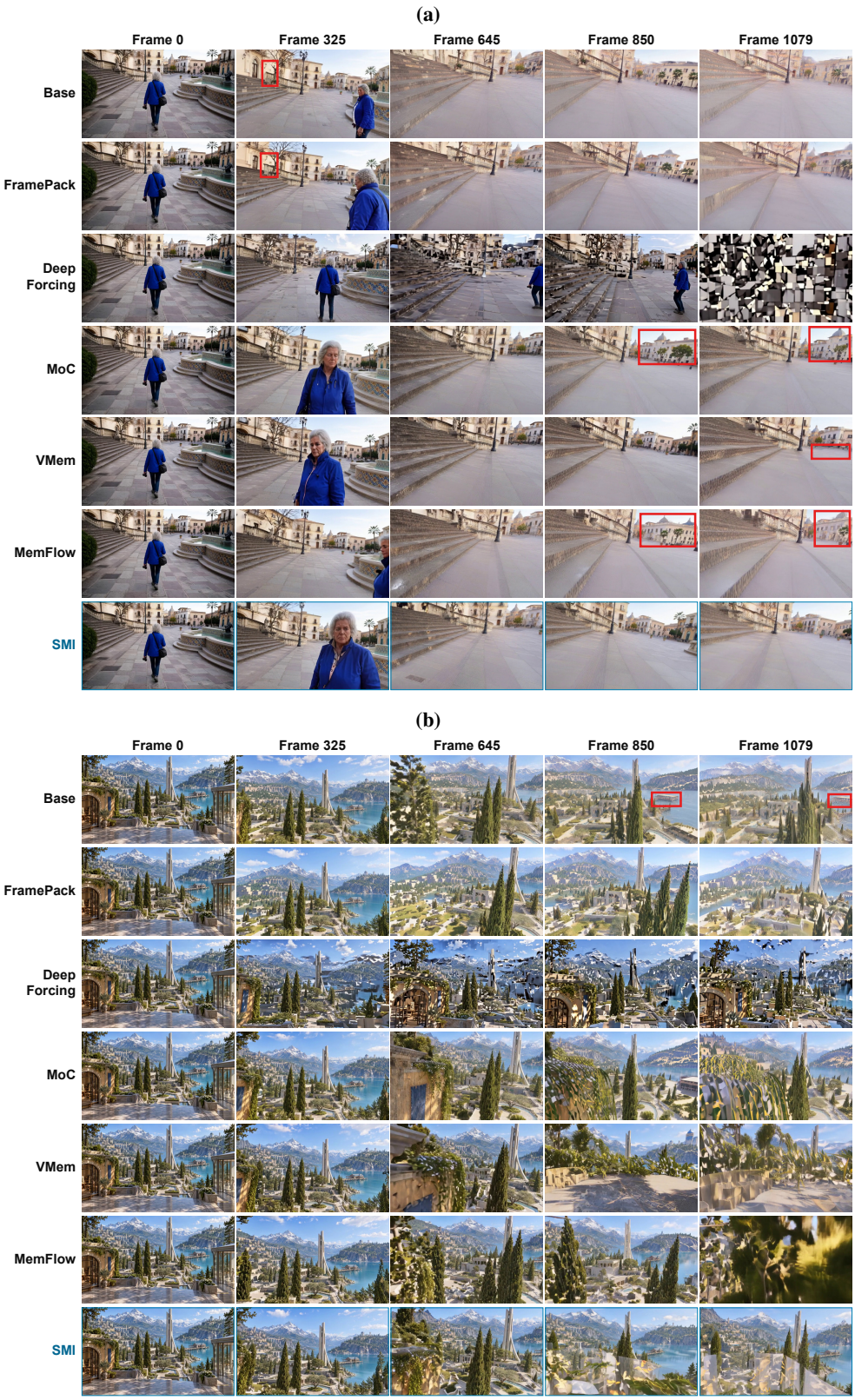

Figure 14: More qualitative results on HY1.5. Red boxes mark spatial inconsistencies.



Figure 15: More qualitative results on HY1.5. Red boxes mark spatial inconsistencies.

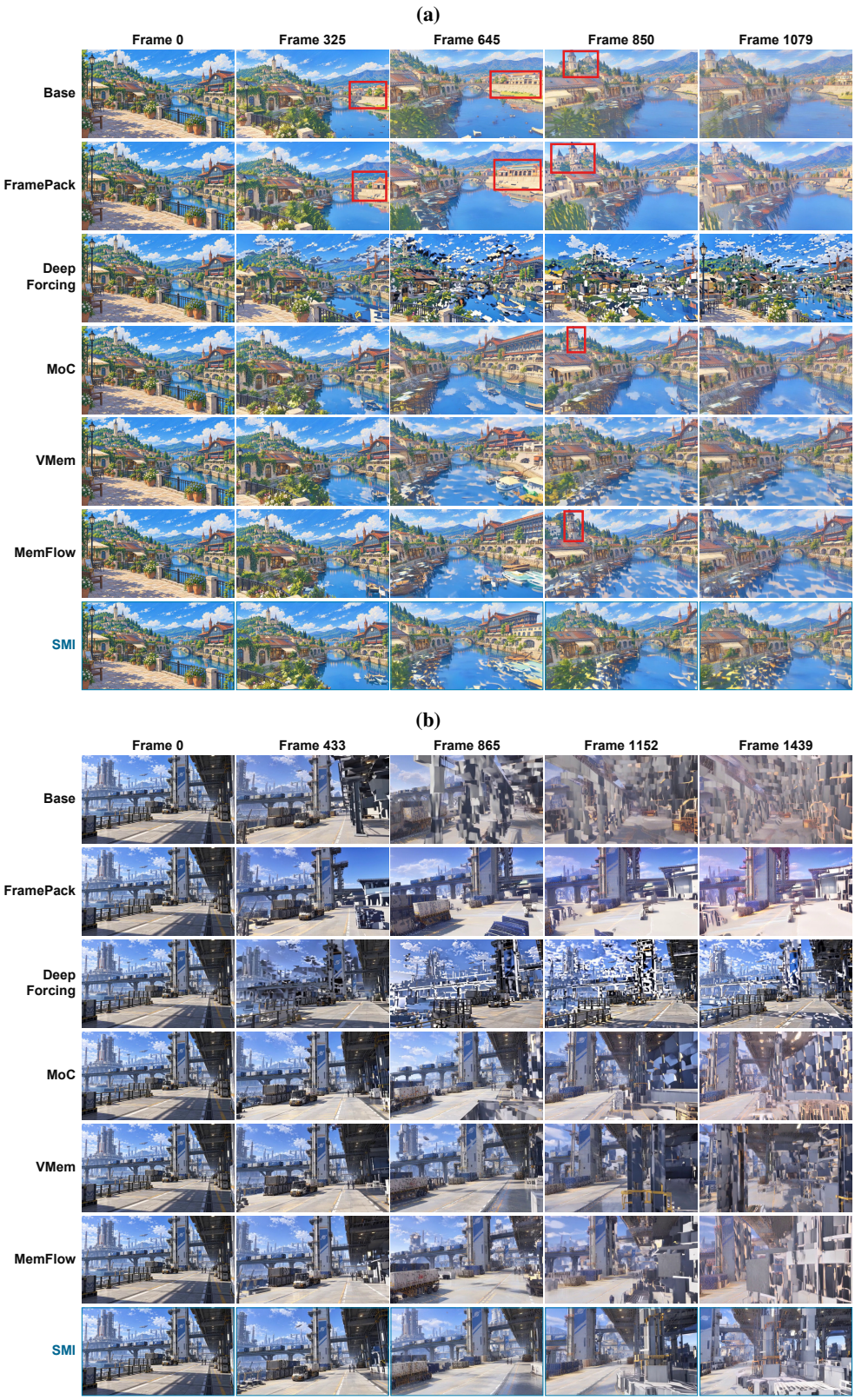

Figure 16: More qualitative results on HY1.5. Red boxes mark spatial inconsistencies.

J.2 WAN2.2

All seven methods use the same frame indices within each case. Red boxes mark visible inconsistencies in the final baseline frames.

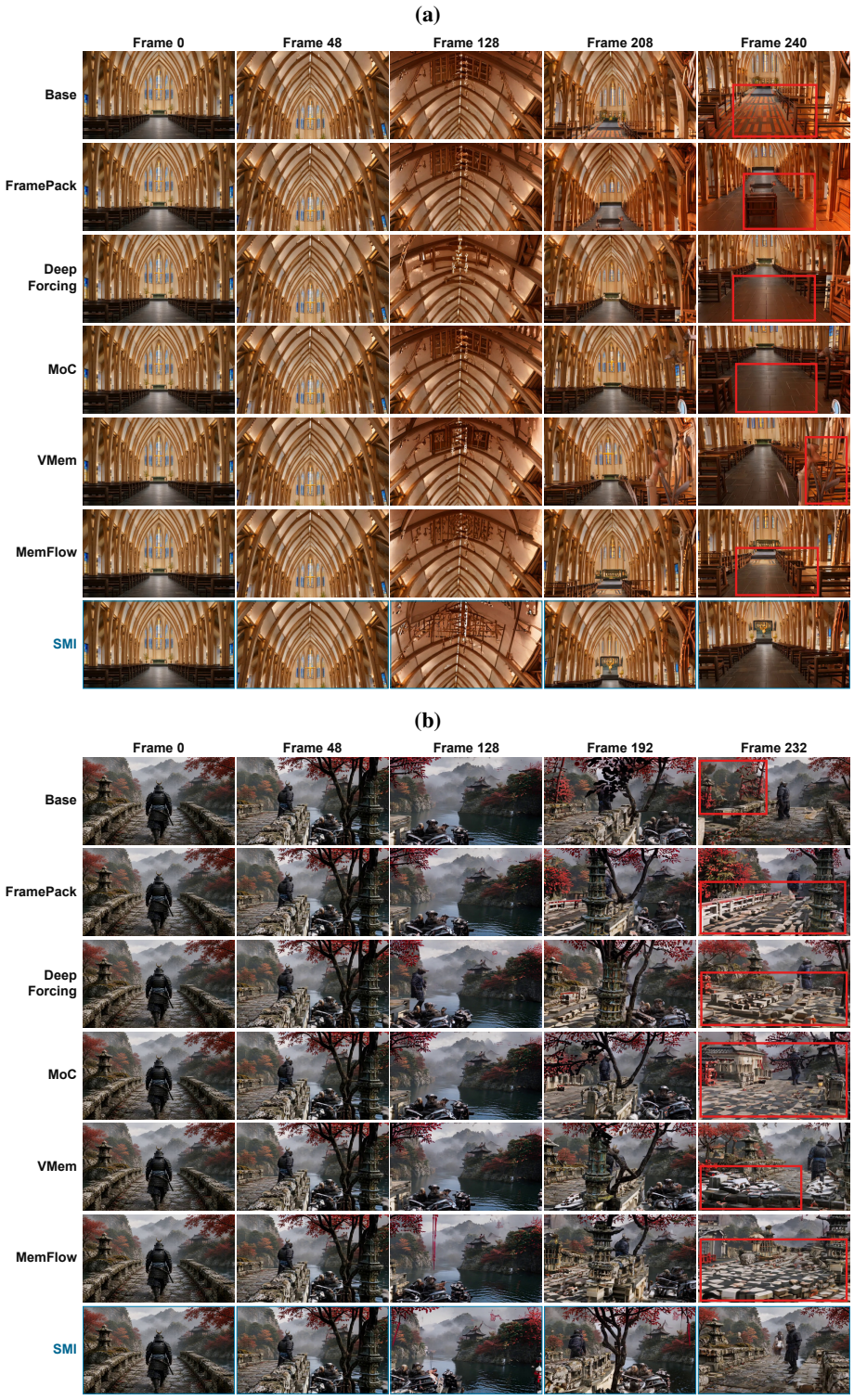

Figure 17: Additional qualitative results on Wan2.2.



Figure 18: More qualitative results on Wan2.2.