Memorize When Needed: Decoupled Memory Control for Spatially Consistent Long-Horizon Video Generation

Yanjun Guo[1,2,⋆], **Zhengqiang Zhang**[1,2,⋆], **Pengfei Wang**[1,⋆],
Xinyue Liang[1], **Zhiyuan Ma**[1], **Lei Zhang**[1,2,†]
[1]The Hong Kong Polytechnic University [2]OPPO Research Institute
yanjunn.guo@connect.polyu.hk cslzhang@comp.polyu.edu.hk
[⋆]Equal contribution [†]Corresponding author

Abstract

Spatially consistent long-horizon video generation aims to maintain temporal and spatial consistency along predefined camera trajectories. Existing methods mostly entangle memory modeling with video generation, leading to inconsistent content during scene revisits and diminished generative capacity when exploring novel regions, even when trained on extensive annotated data. To address these limitations, we propose a decoupled framework that separates memory conditioning from generation. Our approach significantly reduces training costs while simultaneously enhancing spatial consistency and preserving the generative capacity for novel scene exploration. Specifically, we employ a lightweight, independent memory branch to learn precise spatial consistency from historical observations. We first introduce a hybrid memory representation to capture complementary temporal and spatial cues from generated frames, then leverage a per-frame cross-attention mechanism to ensure each frame is conditioned exclusively on the most spatially relevant historical information, which is injected into the generative model to ensure spatial consistency. When generating new scenes, a camera-aware gating mechanism is proposed to mediate the interaction between memory and generation modules, enabling memory conditioning only when meaningful historical references exist. Compared with the existing method, our method is highly data-efficient, yet the experiments demonstrate that our approach achieves state-of-the-art performance in terms of both visual quality and spatial consistency. Project page can be found at `https://iguoyanjun.github.io/memorize-when-needed-project-page/`.

1 Introduction

Recent state-of-the-art video generation models [15, 30, 6, 28, 29, 50, 59] have achieved impressive spatio-temporal coherence within short-term sequences. However, extending such fidelity to long-horizon synthesis remains a fundamental challenge [14, 46, 28]. This limitation is particularly obvious when scene revisit is required, where the model must maintain consistency with previously seen environments [4]. Due to the inherent memory bottleneck imposed by finite context windows, existing models [8, 41, 20] often fail to perceive distant historical observations, ultimately leading to inconsistent content and visual discontinuities.

To overcome context limitations, many recent methods have introduced memory retrieval mechanisms [45, 60, 57, 31]. By leveraging camera trajectories as geometric cues, these approaches dynamically fetch relevant historical frames to guide the synthesis, aiming to maintain visual consistency

[†]This work is supported by the PolyU-OPPO Joint Innovative Research Center.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

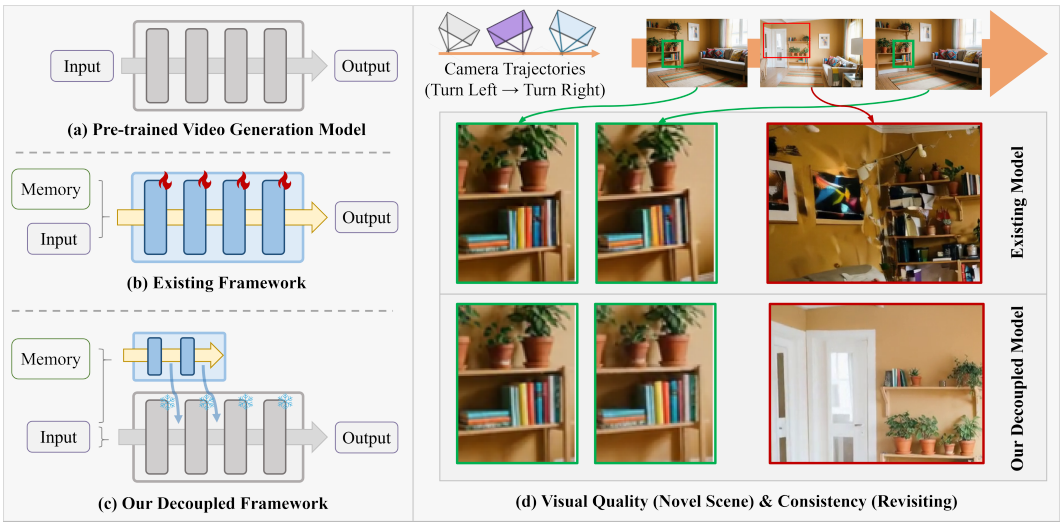

Figure 1: **Comparison of model architecture of long-horizon video generation methods.** (a) A standard pre-trained video generation model. (b) Existing methods typically entangle generation and memory modeling within a unified fine-tuning framework. (c) In contrast, our model decouples memory modeling in a lightweight branch while keeping the pre-trained backbone frozen. (d) Our design enables realistic texture synthesis in novel scenes (e.g., the realistic door in second frame) while ensuring consistency in revisited locations (e.g., the consistent wooden shelf in the green box), outperforming existing methods that suffer from structural distortions (see red box).

during revisiting. For example, WorldMem [57] and Context-as-Memory [60] concurrently propose FOV-overlap scoring for memory selection, a practical strategy adopted by Hunyuanworld1.5 [45]. Despite the potential of retrieval-augmented memory, such an architecture entanglement introduces a compromise between memory adherence and generative quality, inevitably leading to suboptimal long-term consistency and degraded visual fidelity (e.g., the structural distortions shown in the red box of Fig. 1(d)), even after extensive fine-tuning.

We address these limitations with a decoupled memory framework that separates memory modeling from the generative process (see Fig. 1(c) vs. (b)). By freezing the video backbone and employing a lightweight auxiliary branch for memory injection, we eliminate architectural entanglement, preserving the rich priors in pretrained generation models and enabling consistent long-horizon synthesis without sacrificing the quality of novel scenes. As shown in Fig. 1(d), this design enables consistent long-horizon synthesis (e.g., the shelf) without sacrificing the visual fidelity of novel scenes (e.g., the door). To adaptively modulate the memory injection, we introduce a camera-aware gating mechanism that activates memory conditioning only when informative references exist, effectively enabling the exploration of unseen regions. Our design balances the dual demands of long-horizon generation: enforcing spatial consistency during revisits while leveraging the backbone's generative priors for novel exploration. Moreover, this decoupled paradigm also reduces training overhead — the memory branch can be optimized on generic videos using simple data augmentations, obviating the need for costly, manually annotated datasets.

In the memory control branch, we first retrieve relevant historical frames using a Field-of-View (FOV)-guided retrieving strategy, following [57]. We feed these frames into a video encoder to extract hybrid memory representations: continuous temporal tokens capture motion patterns and scene dynamics, while fine-grained spatial tokens provide sharp, static references for the current viewpoint. This combination yields complementary contextual and geometric cues. Subsequently, to achieve frame-level visual alignment, we employ per-frame cross-attention blocks. These blocks allow the generative tokens to selectively query only the most spatially aligned historical frame from the hybrid representation via attention masking. Our design serves a dual purpose: ensuring precise control by filtering out irrelevant spatial noise, and reducing computational overhead by avoiding redundant attention across the entire history buffer. Finally, the aligned memory features are injected into the frozen backbone via the aforementioned camera-aware gating mechanism. Our experiments demonstrate the effectiveness of our decoupled memory control design, which is not only much

more data-efficient than existing methods, but also achieves state-of-the-art video quality and spatial consistency, especially for scene revisits.

In summary, our main contributions are as follows:

- We propose a decoupled memory control method with a camera-aware gating mechanism. This design effectively balances the preservation of pretrained generative priors for novel scene exploration and the effective memory injection for long-horizon video consistency.
- We extract hybrid memory representations from historical frames and employ a masked per-frame cross-attention module to achieve precise frame-level alignment. This approach leverages both motion patterns and spatial details for accurate context retrieval.
- Experiments demonstrate that our method achieves state-of-the-art performance in terms of visual quality and spatial consistency, especially for revisited scenes. Notably, our method is highly data-efficient, capable of learning robust memory control from generic videos using simple augmentations.

2 Related Work

Video Generation and Controllability. Video generation has evolved from early UNet-based diffusion models [10, 5] to large-scale latent diffusion transformers [13, 15, 30, 6, 28, 29, 35, 50, 59]. Benefiting from advances in model architectures [21, 37, 39] and large-scale video data curation pipelines [30, 50], foundation text/image-to-video models exhibit zero-shot generalization with consistent physical dynamics and notable 3D consistency over short temporal ranges [54]. Recent studies have also investigated whether video models can reliably capture spatio-temporal structure beyond perceptual realism, exploring verifiable spatio-temporal reasoning [68] and robust multimodal grounding [53]. To enable viewpoint control, recent efforts inject camera poses into pretrained video diffusion models via conditioning modules, using camera parameters [52], dense Plücker ray embeddings [3, 17], or learnable pose tokens [16, 32, 36, 42, 43]; other approaches use synthetic game-engine data with discrete action commands [4, 12, 19, 44, 49, 61] or enforce 3D consistency through depth- and pose-based frame warping [7, 38, 62, 63].

Long-Term Context and Memory Modeling. Generating long videos that extend beyond the native temporal window of pretrained models is a challenging issue, given the quadratic attention complexity. Training-free methods [34, 65, 56] extend pretrained models through noise rescheduling, temporal-frequency rebalancing, or sparse attention. Diffusion Forcing [8] and History-Guidance [41] condition denoising on previous frames with decayed noise, and are scaled by SkyReels-V2 [9] and Magi-1 [48]. Meanwhile, recent multi-shot video generation methods [1] explicitly model cross-shot dependencies through fixed-window attention or keyframe conditioning to improve narrative consistency across shots. Other works distill bidirectional diffusion into causal generators [11, 24, 27, 58] to reduce autoregressive error accumulation, while [20, 64] augment pretrained models with memory modules and generate long videos iteratively, achieving high visual quality.

Beyond temporal coherence, long-horizon video generation also requires maintaining spatial consistency when revisiting previously observed regions. Existing methods use either explicit 3D reconstruction [7, 38, 62, 63, 23, 55, 51, 2] or implicit context conditioning [31, 57, 60, 22, 45]. Explicit methods provide strong geometric guidance but heavily depend on reconstruction quality, introducing additional computational overhead. Implicit methods retrieve spatially relevant history from memory banks, e.g., through Field-of-View overlap scoring [57, 60], offering a practical solution for maintaining spatial coherence. However, these methods entangle memory processing with the generation backbone (e.g., via input concatenation or modified attention), leading to high training costs, inflexible conditioning, and over-reliance on history that reduces diversity in novel regions.

3 Method

In this section, we first formulate the problem of long-horizon consistent video generation. We then present our decoupled framework, detailing the camera-aware gating mechanism, the architectures of the memory control branch (including hybrid memory representation and per-frame cross-attention blocks), and the tailored training strategies designed to handle complex exploration trajectories.

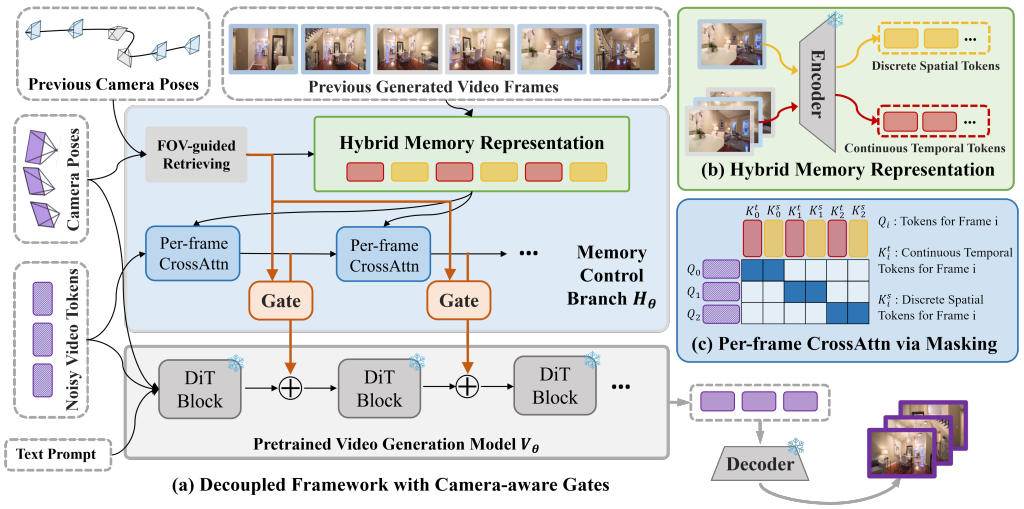

Figure 2: **Overview of our decoupled framework.** (a) We disentangle memory control (blue branch) from the generative backbone (gray branch), employing a camera-aware gating mechanism (orange part) to adaptively modulate memory injection. (b) The hybrid memory representation (green part) extracts complementary spatio-temporal cues from history frames. (c) Per-frame cross-attention blocks (blue part) enforce frame-level alignment, where each latent token attends solely to its spatially corresponding historical frame, ensuring both precise control and computational efficiency.

3.1 Preliminaries

Long-horizon video generation aims to synthesize video sequences that maintain long-term spatial consistency along a predefined camera trajectory. Given an initial image I_0, a text prompt $\mathcal{T}$, and a sequence of camera poses $\{P_t\}_{t=1}^{T}$, the goal is to generate a sequence of frames $\{I_t\}_{t=1}^{T}$ that follow the specified viewpoint changes while preserving scene consistency. In latent diffusion-based I2V models, the initial image I_0 conditions DiT-based iterative denoising of video latents, which are finally decoded into video sequences by a 3D VAE.

A core challenge is how to maintain spatial consistency, particularly when the camera revisits previously observed locations. Prior works [60, 31, 18] typically adopt a segment-based iterative approach, dividing the sequence into K consecutive segments $\{S_k\}_{k=1}^{K}$. In this pipeline, each segment is generated by utilizing previously generated content as conditioning context:

$$S_k = V_\theta(I_{\text{pre}}, \mathcal{P}_{S_k}, \mathcal{M}_k, \mathcal{T}), \quad \{I_t\}_{t=1}^{T} = \text{Concat}(\mathcal{D}(S_1), \ldots, \mathcal{D}(S_K)), \tag{1}$$

where I_{pre} denotes the last frame of the segment S_{k-1} ($S_0 = \{I_0\}$), $\mathcal{P}_{S_k}$ is the set of camera poses corresponding to segment S_k, and $\mathcal{M}_k = \{(I_0, (\mathcal{D}(S_1), P_1), \ldots,$
$(\mathcal{D}(S_{k-1}), P_k)\}$ represents the accumulated visual memory and camera poses from the initial image and all preceding segments. $\mathcal{D}$ is the video decoder.

3.2 Motivation and Framework Overview

Existing methods generally entangle memory modeling with video generation within a unified framework. The model V_θ is burdened with the dual task of creating new content and enforcing consistency constraints simultaneously. This often leads to a trade-off, compromising either the perceptual quality of generated frames or the long-term spatial consistency. To address this limitation, we propose to decouple these objectives into separate streams. Specifically, we retain V_θ for high-fidelity video generation while introducing an additional lightweight memory control branch H_θ dedicated to processing historical memory. The generation process is thus reformulated as:

$$S_k = V_\theta(I_{\text{pre}}, \mathcal{P}_{S_k}, H_\theta(\mathcal{M}_k, \mathcal{P}_{S_k}), \mathcal{T}), \tag{2}$$

where $H_\theta(\cdot)$ extracts compact, consistency-aware features from the raw memory $\mathcal{M}_k$, allowing V_θ to focus primarily on synthesizing novel content aligned with the text prompt and camera trajectory.

Our architecture is illustrated in Fig. 2(a), which consists of a frozen video generation backbone V_θ (the gray branch) dedicated to high-quality synthesis, and a lightweight memory control branch H_θ

(the blue branch) designed to extract relevant historical context. In this decoupled framework, the inherent conflict between maintaining long-term consistency and generating diverse novel content is resolved by isolating their respective optimization goals. The backbone focuses solely on visual fidelity and are free from the constraints of historical alignment, while the memory branch specializes in context extraction.

3.3 Camera-aware Gating Mechanism

The core design principle of our framework is to inject historical information *only when needed*.Indiscriminately enforcing memory constraints can impede novel content generation in previously unseen regions and introduce artifacts in regions with unreliable memory retrieval. To strike a balance between consistency and creativity, we introduce an on-the-fly camera-aware gating mechanism (the orange part in Fig. 2(a)) that dynamically modulates the influence of the memory branch based on the exploration trajectory. Unlike previous works that must interact with $\mathcal{M}_k$ in a segment-wise manner, our approach regulates the interaction between the two streams (V_θ and H_θ) in a frame-wise manner via a gating mechanism. This provides more precise control for video generation while minimizing unintended artifacts.

Specifically, given camera poses P_{S_k} for the current segment and the corresponding history $\mathcal{M}_k$, we first obtain the geometric relevance scores s_i that incorporate both FOV overlap and camera distance between each current pose $P_i \in P_{S_k}$ and all historical poses. The geometric score s_i provides the initial gating decision, which is further refined by translation-distance and temporal-redundancy filtering, yielding the final binary gate $g_i \in \{0, 1\}$ for each frame. Details are provided in Appendix. Let $\mathbf{f}_l^i$ denote the intermediate features of V_θ at layer l for frame I_i, and $\mathbf{h}_l^i$ represent the corresponding features from H_θ. The feature injection is formulated as:

$$\mathbf{f}_l^i = \mathbf{f}_l^i + g_i \cdot \mathbf{h}_l^i, \tag{3}$$

where g_i serves as a **hard gate**: it enables memory injection only for frames with reliable historical correspondences, which act as anchors for the segment. For novel or unreliably matched frames, the memory connection is disabled, allowing the backbone to generate content without unnecessary or erroneous memory constraints while maintaining consistency through segment-level temporal coherence. This mechanism effectively harmonizes generative capability with consistency constraints.

3.4 Memory Control Branch

As illustrated in Fig. 2(a), our memory control branch first identifies relevant historical frames and segments via an FOV-guided retrieval strategy (similar to that in [57]). Subsequently, it constructs a hybrid memory representation through per-frame cross-attention blocks to aggregate pertinent information. The details are presented as follows.

Hybrid Memory Representation. The efficacy of the memory control branch hinges on the quality and reliability of features retrieved from historical frames. To enhance robustness to the imperfect matches introduced by FOV-based retrieval, as depicted in Fig. 2(b), we propose a hybrid memory representation that combines complementary chunk-level and frame-level cues from the retrieved history. One is continuous temporal memory, which encodes a short window of consecutive historical frames to aggregate multi-frame evidence K^t around the retrieved frames, thereby improving robustness to single-frame retrieval errors. The other is discrete spatial memory to provide a sharp reference for the current viewpoint. We obtain those discrete spatial tokens K^s by inputting individual retrieved frames. Together, the chunk-level memory offers stable multi-frame context, while the frame-level memory preserves precise appearance details.

Per-frame Cross-Attention Blocks. To extract relevant information for current latents, we employ them as queries to retrieve useful context from historical tokens (K^s and K^t) within cross-attention blocks, where camera poses of queries and keys are injected as additional positional embeddings. To precisely control the interaction between those historical tokens and current video latents, we apply a mask strategy to impose a strict constraint where each latent interacts only with the tokens from its corresponding history frame, as shown in Fig. 2 (c). This localized attention mechanism not only reduces computational overhead but also alleviates optimization difficulties.

3.5 Training Strategies

A key advantage of our decoupled architecture is that it streamlines the optimization of the memory control branch. Specifically, it allows us to augment existing real-world videos to construct revisiting and exploration data, circumventing the need for expensive 3D dataset rendering. We synthesize training sequences by strategically sampling and reordering video frames.

Synthesizing Pseudo-Loops for Memory Training. To train the model on revisiting events using standard videos, we synthesize sequences by reorganizing frame orders to simulate closed-loop trajectories (e.g., creating forward-backward loops). A naive approach — using the same frame as both history and target — would allow the model to learn a trivial identity mapping. To avoid this, we apply a temporal stride strategy: we use a frame at time I_t as the history reference but require the model to generate its neighbor at $I_{t+\delta}$. This controlled perturbation between the memory reference and the target view mimics the imperfect matches produced by FOV-based retrieval and encourages the memory branch to learn robust content-level correspondence from geometrical cues.

History Dropout for Novel Scene Exploration. To ensure the model can seamlessly transition between revisiting known regions and exploring new ones, we introduce a history dropout strategy. During training, we randomly mask out the historical reference frames for certain segments. This compels the model to dynamically adapt: when history is unavailable (masked), it must rely on the backbone's generative priors to synthesize novel content; when history is present, it utilizes the memory for consistency. This simple regularization effectively guides the model to activate memory only when valid references exist.

4 Experiment

This section presents extensive experiments to validate our method. We begin by outlining the experimental setup and protocols in Sec. 4.1, followed by a comparative analysis against existing methods in Sec. 4.2. Sec. 4.3 further evaluates the robustness of our method on out-of-distribution data and challenging trajectory patterns. Ablation studies are provided in Sec. 4.3, where we analyze the effects of the hybrid memory representation and camera-aware gating mechanism.

4.1 Experimental Setup

Training Settings. We adopt two representative image-to-video (I2V) diffusion models as our generative backbones: Wan2.1-14B-I2V [50] for high-quality generation and CogVideoX-5B-I2V [59] for practical efficiency. Following AC3D [3], we utilize Plücker coordinates for camera pose representation and incorporate an additional module for precise camera control. During memory branch training, we freeze the parameters of both the video backbone and the camera control module, optimizing only the separate memory control branch. We use the Adam optimizer with a learning rate of 1×10^{-4}. The effective batch size is set to 16, and the models are trained on NVIDIA A800 GPUs for 10K iterations. More training details are provided in the **Appendix**.

Training and Testing Data. For training, we utilize the RealEstate10K [67] dataset to synthesize video sequences that simulate revisiting and exploration scenarios (details in Sec. 3.5). The data is processed into 49-frame video segments at 480p resolution. For testing, following the protocol in WorldPlay [45] and VMem [31], we sample trajectories from RealEstate10K and extend them with a return path along the initial trajectory for consistency evaluation.

Furthermore, to assess generalization, we collect 200 additional test scenes sourced from both the internet and AI models, which cover indoor and outdoor environments and span diverse content and visual styles. We also design a suite of challenging trajectory patterns to rigorously evaluate consistency under diverse conditions. (a) *Repeated Revisiting*: where the camera rotates forth and back to the same viewpoint multiple times, creating a repetitive pattern that tests the model's stability against error accumulation; (b) *Random Loop Insertion*: introducing return loops of varying lengths at random positions along a trajectory; and (c) *Spatially-Offset Return*: where the camera returns via a slightly shifted path instead of exact retracing, testing robustness to viewpoint deviations.

Evaluation Metrics. We evaluate our method from two primary perspectives: long-term consistency in revisiting scenarios and generation quality on novel scenes. For *long-term consistency*, following [31, 45, 60], we employ PSNR, SSIM, LPIPS as quantitative metrics. Specifically, we utilize

Table 1: Long-term consistency and video quality comparisons on RealEstate10K. **Red**: best, blue: second best in each column.

Method	Revisiting							Exploration			
	long-term consistency			video quality				video quality			
	PSNR↑	SSIM↑	LPIPS↓	AQ(%)↑	IQ(%)↑	MS(%)↑	DD(%)↑	AQ(%)↑	IQ(%)↑	MS(%)↑	DD(%)↑
AC3D[3]	13.13	0.52	0.52	51.37	64.71	99.29	95.24	**49.88**	63.84	99.21	92.86
DFoT[41]	14.93	0.48	0.34	43.00	66.80	98.58	35.71	42.20	65.71	98.67	35.71
SEVA[66]	14.80	0.60	0.47	42.78	61.67	98.40	88.10	44.44	64.62	98.28	88.10
VMem[31]	15.37	0.65	0.37	42.34	59.80	97.66	78.57	42.38	59.37	98.04	80.95
WorldPlay[45]	16.31	0.64	0.36	**51.51**	70.45	99.06	92.86	49.24	70.57	99.04	92.86
WorldPlay-d[45]	15.09	0.63	0.31	51.34	71.43	**99.32**	88.10	49.00	70.12	99.32	90.48
Ours-Cog	20.23	0.69	0.23	50.64	63.53	99.17	92.86	49.20	63.49	**99.42**	95.24
Ours-Wan	**21.85**	**0.71**	**0.20**	51.24	**72.27**	98.86	**97.62**	49.55	**72.48**	98.88	**95.56**



Figure 3: Visual comparison. Our method generates clearly structured staircases in unseen regions (pink boxes) and faithfully reproduces fine-grained details such as the two chairs when revisiting the original scene (red boxes).

cyclic camera trajectories that revisit previously observed viewpoints, comparing each generated frame on the return path with its counterpart from the initial pass. Higher PSNR/SSIM scores and lower LPIPS scores indicate superior consistency. For *generation quality*, we assess the synthesized videos across both spatial and temporal dimensions, following VBench [25]. We use Aesthetic Quality (AQ)[40] and Image Quality (IQ)[26] to measure image quality, and adopt Motion Smoothness (MS)[33] and Dynamic Degree (DD)[47] to estimate motion magnitude.

Compared Methods. We conduct comprehensive comparisons with existing methods, categorized into two groups. (1) *Models without memory*: AC3D[3] (re-implemented on CogVideoX-5B-I2V), DFoT[41], and SEVA[66]. These methods support camera-controlled generation, but lack explicit mechanisms for maintaining cross-clip consistency. (2) *Models with memory*: VMem[31] and WorldPlay [45]. For WorldPlay, we evaluate both its bidirectional attention variant (WorldPlay) and the final distilled version (WorldPlay-d).

Table 2: Comparison of training data, parameters, and computational cost.

Method	Training Data	Parameters			FLOPs (TFLOPs)		
		Backbone	Memory	Total	Backbone	Memory	Total
VMem[31]	-	1.26B	0	1.26B	67.46	44.96	112.44
WorldPlay [45]	320K	8.56B	0	8.56B	923.03	3065.00	3988.03
Ours-Cog	14K	5.60B	128.42M	5.73B	440.81	2.97	443.78
Ours-Wan	14K	14.75B	1371.38M	16.12B	1240.22	71.62	1311.84

Table 3: Out-of-distribution quantitative comparison. **Red**: best, blue: second best.

Method	**Revisiting**							**Exploration**			
	long-term consistency			video quality				video quality			
	PSNR↑	SSIM↑	LPIPS↓	AQ(%)↑	IQ(%)↑	MS(%)↑	DD(%)↑	AQ(%)↑	IQ(%)↑	MS(%)↑	DD(%)↑
AC3D[3]	13.74	0.48	0.49	55.94	70.40	98.62	**97.46**	57.20	73.25	99.11	97.58
DFoT[41]	14.80	0.47	0.32	47.44	71.30	98.34	29.29	47.76	72.04	98.54	24.40
SEVA[66]	12.31	0.48	0.45	50.10	73.30	98.50	90.44	54.18	73.96	98.44	97.36
VMem[31]	15.43	0.49	0.27	54.95	69.12	97.45	60.98	55.69	70.11	97.97	60.98
WorldPlay[45]	16.54	0.71	0.27	58.99	**75.81**	98.47	92.68	57.71	73.81	98.76	90.46
WorldPlay-d[45]	16.57	0.69	0.26	58.44	75.70	98.81	93.00	57.69	74.22	98.94	95.35
Ours-Cog	20.79	0.72	0.23	57.96	74.38	**98.91**	95.00	57.89	75.83	**99.12**	97.50
Ours-Wan	**21.35**	**0.73**	**0.22**	**60.09**	75.77	98.84	94.50	**59.83**	**78.29**	98.78	**97.70**

4.2 Main Results

Quantitative Results. As shown in Table 1, our method (Ours-Wan) achieves state-of-the-art performance in both consistency and generation quality metrics. Methods without explicit memory modeling — such as AC3D[3], DFoT[41], and SEVA[66] — struggle to maintain consistency over extended trajectories, resulting in inferior PSNR, SSIM, and LPIPS scores in the *Revisiting* phase. Incorporating memory modeling notably improves consistency. In specific, WorldPlay [45] achieves improved values of 16.31 (PSNR), 0.64 (SSIM), and 0.36 (LPIPS). Our decoupled framework surpasses these competitors by a significant margin, achieving superior results of 21.85 (PSNR), 0.71 (SSIM), and 0.20 (LPIPS), respectively. This demonstrates that our separate memory control branch provides significantly more precise memory retrieval and faithful scene reproduction. Regarding video quality, while WorldPlay achieves a marginally higher AQ (51.51%) compared to ours (51.24%), our method outperforms all competitors in both Image Quality (IQ, 72.27%) and Dynamic Degree (DD, 97.62%). This indicates that our approach effectively preserves high-fidelity generative capabilities and temporal coherence even in complex revisiting scenarios, without sacrificing visual details.

We also evaluate all methods on the *Exploration* columns in Table 1 to assess their generative capacity on novel scenes. Our method (CogVideoX-based) achieves AQ on par with AC3D [3], demonstrating that the decoupled design preserves the generative prior of the backbone and ensures stable generation capability during exploration. Ours-Wan achieves the highest IQ (72.48%) and DD (95.56%) in the exploration phase with only a slight performance drop in AQ (1.69 points) compared to WorldPlay. This demonstrates the effectiveness of our dynamic gating mechanism, which successfully balances retrieving memory for consistency and generating novel content for exploration.

Qualitative Results. We provide qualitative comparisons in Figure 3, where the camera first rotates forward into unseen regions and then moves back to the starting viewpoint. Methods AC3D [3], DFoT [41] and SEVA [66] lack explicit cross-clip spatial memory, hallucinating mismatched content when the camera rotates back. Although VMem [31] incorporates memory through image-conditioned generation, it suffers from severe error accumulation over extended sequences, synthesizing unreasonable leather texture during exploration. WorldPlay [45] produces noticeable artifacts when the camera moves into unseen regions. In contrast, our method synthesizes coherent novel content, such as clearly structured staircases and a bathroom with an open door. When the camera rotates back, our method faithfully reproduces the original living room scene, including fine-grained details such as the two chairs beside the dining table, demonstrating the capability of our decoupled framework in both novel scene exploration and revisited scene reconstruction.

Computational and Data Efficiency. Table 2 compares the extra computational overhead, extra parameters, and training data for methods with memory mechanisms. Although WorldPlay [45]

Table 4: Quantitative comparison across three challenging revisiting trajectories. IQ: Imaging Quality. **Red (bold)**: best; blue: second best.

Method	Repeated Revisiting			Random Loop Insertion			Spatially-Offset Return		
	PSNR↑	LPIPS↓	IQ(%)↑	PSNR↑	LPIPS↓	IQ(%)↑	PSNR↑	LPIPS↓	IQ(%)↑
AC3D [3]	13.16	0.47	55.09	16.49	0.32	63.58	10.50	0.56	65.38
DFoT[41]	14.39	0.39	66.44	20.62	**0.13**	70.39	10.41	0.55	67.24
SEVA[66]	13.26	0.51	53.53	18.99	0.20	**75.38**	9.71	0.71	68.04
VMem[31]	18.91	0.30	59.71	19.03	0.25	65.50	**27.34**	**0.15**	61.87
WorldPlay [45]	19.24	0.19	73.17	17.40	0.30	72.23	13.95	0.48	68.96
WorldPlay-d [45]	17.09	0.26	**73.85**	17.61	0.53	71.98	11.79	0.70	67.72
Ours-Cog	21.92	0.15	59.38	19.53	0.22	63.88	14.35	0.44	64.69
Ours-Wan	**22.03**	**0.14**	71.48	**21.76**	0.16	72.33	16.38	0.29	**73.27**

Table 5: Ablation studies on hybird memory representation.

Method	PSNR↑	SSIM↑	LPIPS↓	IQ↑	MS↑
Only K^t	16.93	0.59	0.29	73.13%	98.98%
Only K^s	20.78	0.69	0.23	72.18%	97.55%
Ours	20.79	0.72	0.23	74.38%	98.91%

introduces no additional parameters, it results in high computational costs due to the quadratic complexity of self-attention on the extended sequence. As shown in Table 2, the memory operations in WorldPlay [45] require 3,065 TFLOPs, which is approximately $3.3\times$ its backbone (923.03 TFLOPs) and accounts for the majority of the total inference cost. Our method adopts a separate, lightweight branch for memory modeling. This design adds 128.42M parameters to the CogVideoX [59] backbone (about 2.2% of the total), but significantly reduces computational load. Specifically, the memory branch in our method requires only 2.97 TFLOPs, which is over $1,000\times$ lower than that of WorldPlay [45]. As a result, the total inference cost of our method (443.78 TFLOPs) is close to the backbone-only cost. More Comparisons on memory and latency are provided in the **Appendix**.

In terms of training data efficiency, our method needs only **14K** training samples, while WorldPlay [45] requires **320K** samples, about $23\times$ higher than ours. Overall, our approach achieves effective memory modeling with substantially lower computational and data costs than prior works.

4.3 Robustness Analysis

To further evaluate the robustness and adaptability of our decoupled framework, we conduct additional experiments on out-of-distribution (OOD) data and more challenging camera trajectories.

OOD Generalization. Table 3 evaluates the generalization on unseen distributions. Our methods achieve the top two results across all long-term consistency metrics while maintaining competitive video quality in both Revisiting and Exploration. It also achieves the second-best IQ in Revisiting and leading IQ and DD in Exploration. Compared to WorldPlay [45], our method shows a significant leap (+4.81 PSNR, -0.05 LPIPS), validating that our decoupled memory mechanism enables robust generalization to novel domains and styles. We provide more visual analysis in the **Appendix**.

Complex Trajectories. As discussed in Sec. 4.1, we evaluate performance under three challenging trajectory settings (Table 4), where our models continue to excel in both generation quality and spatial consistency. Specifically, Ours-Wan achieves state-of-the-art consistency in *Repeated Revisiting* (22.03 PSNR), and excels in *Random Loop Insertion* with 21.76 PSNR. For the *Spatially-Offset Return* path, we evaluate consistency by computing metrics between the first and last frames. Although VMem [31] achieves strong first-last frame consistency, it struggles to maintain high visual quality throughout the long-range generation process. Our method achieves a balance between video generation quality and spatial consistency, attaining 16.38 PSNR while maintaining an IQ of 73.27%.

4.4 Ablation Studies

We conduct ablation studies using our CogVideoX-based variant, i.e., Ours-Cog, selected for its computational efficiency and smaller parameter count.

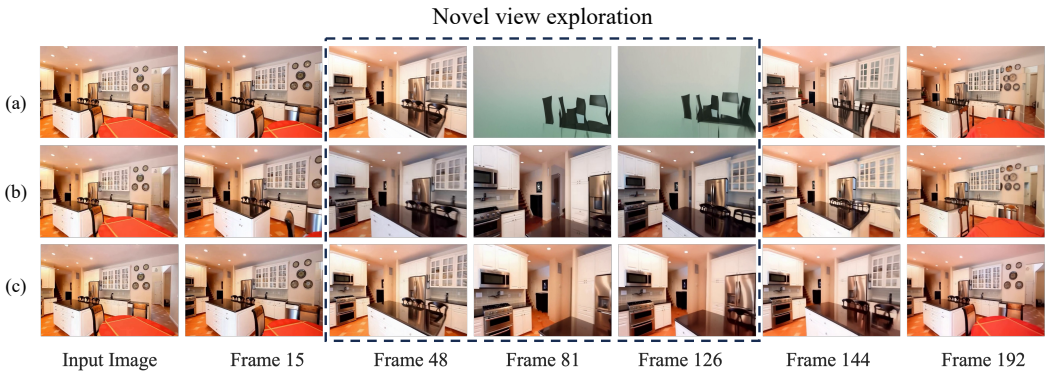

Figure 4: **Ablation Study** on camera-aware gating mechanism and history dropout strategies. (a) Baseline without gating fails on novel viewpoints. (b) Adding the camera-aware gate improves realism in novel views but causes inconsistency. (c) Full model with history dropout achieves both realistic and temporally consistent results.

Hybrid Memory Representation. Table 5 presents the ablation study on the two components of our hybrid memory representation using out-of-distribution (OOD) data. When using only the temporal memory representation (K^t), the generated videos achieve the highest MS score of 98.98%, outperforming the spatial-only variant (K^s, 97.55%). This suggests that K^t is particularly effective at capturing motion consistency across frames. In contrast, compared with the temporal-only variant, the "Only K^s" configuration substantially improves the spatial quality metrics, increasing PSNR from 16.93 to 20.78 and SSIM from 0.59 to 0.69, while reducing LPIPS from 0.29 to 0.23. It also achieves a slightly higher IQ score of 72.18%. Combining K^t and K^s achieves the best overall performance, with the highest PSNR, SSIM, and IQ, confirming their complementary benefits for temporal coherence and spatial fidelity.

Camera-aware Gating Mechanism. We evaluate the camera-aware gating mechanism and history dropout strategy in Figure 4. The test sequence comprises a historical reference (Frame 15), a novel view exploration phase (Frames 48–126), and a revisiting phase (Frames 144–192). The baseline without gating (Figure 4 (a)) maintains consistency during revisiting but suffers from texture artifacts in novel views. Incorporating the camera-aware gate (Figure 4 (b)) effectively filters irrelevant history during exploration, significantly improving texture realism. However, this selective injection leads to temporal inconsistencies between the exploration and revisiting phases. Finally, our history dropout strategy remedies this by enforcing robustness against missing historical context. As shown in Figure 4 (c), the full model achieves both high-fidelity synthesis for novel views and temporal coherence for revisited regions.

5 Conclusion

We propose a decoupled memory-conditioned framework for spatially consistent long-horizon video generation. By separating memory modeling from the generation backbone, our method improves scene-revisit consistency while preserving generative capacity for novel regions with reduced training cost. We introduce a lightweight memory branch with hybrid temporal-spatial memory representation, per-frame cross-attention for spatially relevant conditioning, and camera-aware gating to activate memory only when valid historical references exist. Experiments show that our approach achieves state-of-the-art visual quality and spatial consistency across diverse camera trajectories while significantly reducing training data requirements.

Limitations. Following prior work, our method relies on FOV-based memory retrieval, where both the accuracy of FOV matching and camera pose control can affect the consistency and fidelity of generated revisited scenes. Moreover, as our framework builds upon existing video generation models, our performance upper bound is also constrained by the capability of the underlying base model.

Acknowledgments and Disclosure of Funding

This work is supported by the PolyU-OPPO Joint Innovative Research Center.

References

[1] Zhaochong An, Menglin Jia, Haonan Qiu, Zijian Zhou, Xiaoke Huang, Zhiheng Liu, Weiming Ren, Kumara Kahatapitiya, Ding Liu, Sen He, Chenyang Zhang, Tao Xiang, Fanny Yang, Serge J. Belongie, and Tian Xie. Onestory: Coherent multi-shot video generation with adaptive memory. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026, Denver, CO, USA, June 3-7, 2026*, pages 16173–16184, .

[2] Zhaochong An, Orest Kupyn, Théo Uscidda, Andrea Colaco, Karan Ahuja, Serge J. Belongie, Mar González-Franco, and Marta Tintore Gazulla. VGGRPO: towards world-consistent video generation with 4d latent reward. In *Computer Vision - ECCV 2026 - 19th European Conference, Malmö, Sweden, September 8-12, 2026*, pages 305–322, .

[3] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22875–22889, 2025.

[4] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15791–15801, 2025.

[5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[6] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

[7] Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–12, 2025.

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

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

[10] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7310–7320, 2024.

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

[12] Decart, Julian Quevedo, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer, 2024.

[13] Google Deepmind. Veo3 video model, 2025. `https://deepmind.google/models/veo/`.

[14] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27713–27724, 2025.

[15] Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.

[16] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

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

[18] Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13416–13426, 2025.

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

[20] Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2568–2577, 2025.

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[22] 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.

[23] 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*, 2025.

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

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

[26] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021.

[27] Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. Fifo-diffusion: Generating infinite videos from text without training. *Advances in Neural Information Processing Systems*, 37:89834–89868, 2024.

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

[29] Kuaishou. Kling video model, 2024. `https://klingai.com/global/`.

[30] 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, 2025.

[31] 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*, pages 25690–25699, 2025.

[32] Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. *arXiv preprint arXiv:2507.10496*, 2025.

[33] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023.

[34] Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. Freelong: Training-free long video generation with spectralblend temporal attention. *Advances in Neural Information Processing Systems*, 37:131434–131455, 2024.

[35] Minimax. Hailuo video model, 2024. `https://hailuoai.video`.

[36] Takeru Miyato, Bernhard Jaeger, Max Welling, and Andreas Geiger. Gta: A geometry-aware attention mechanism for multi-view transformers. *arXiv preprint arXiv:2310.10375*, 2023.

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

[38] 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*, pages 6121–6132, 2025.

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[40] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

[41] Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. *arXiv preprint arXiv:2502.06764*, 2025.

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

[43] Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 2024.

[44] Wenqiang Sun, Fangyun Wei, Jinjing Zhao, Xi Chen, Zilong Chen, Hongyang Zhang, Jun Zhang, and Yan Lu. From virtual games to real-world play. *arXiv preprint arXiv:2506.18901*, 2025.

[45] 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.

[46] HunyuanWorld Team, Zhenwei Wang, Yuhao Liu, Junta Wu, Zixiao Gu, Haoyuan Wang, Xuhui Zuo, Tianyu Huang, Wenhuan Li, Sheng Zhang, et al. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv preprint arXiv:2507.21809*, 2025.

[47] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.

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

[49] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024.

[50] 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.

[51] Pengfei Wang, Liyi Chen, Zhiyuan Ma, Yanjun Guo, Guowen Zhang, and Lei Zhang. One2scene: Geometric consistent explorable 3d scene generation from a single image. *arXiv preprint arXiv:2602.19766*, 2026.

[52] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[53] Xiaofei Wen, Wenjie Jacky Mo, Xingyu Fu, Rui Cai, Tinghui Zhu, Wendi Li, Yanan Xie, Muhao Chen, and Peng Qi. When vision speaks for sound. *CoRR*, abs/2605.16403.

[54] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

[55] 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. *arXiv preprint arXiv:2602.02393*, 2026.

[56] 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.

[57] Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*, 2025.

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

[59] 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. *arXiv preprint arXiv:2408.06072*, 2024.

[60] 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*, pages 1–11, 2025.

[61] Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11590–11599, 2025.

[62] Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 100–111, 2025.

[63] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.

[64] Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. *arXiv preprint arXiv:2504.12626*, 2025.

[65] Min Zhao, Guande He, Yixiao Chen, Hongzhou Zhu, Chongxuan Li, and Jun Zhu. Riflex: A free lunch for length extrapolation in video diffusion transformers. *arXiv preprint arXiv:2502.15894*, 2025.

[66] Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12405–12414, 2025.

[67] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 37, 2018.

[68] Tinghui Zhu, Sheng Zhang, James Y. Huang, Selena Song, Xiaofei Wen, Yuankai Li, Hoifung Poon, and Muhao Chen. Video models can reason with verifiable rewards. abs/2605.15458.

Appendix

In this appendix, we provide the following materials:

- Section A: The camera-aware gating algorithm (referring to Sec. 3.3 in the main paper);
- Section B: More training and efficiency details, including training settings, FLOPs calculation, GPU memory, and wall-clock latency (referring to Sec. 4.1, Sec. 4.2 and Table 2 in the main paper);
- Section C: More qualitative results (referring to Sec. 4.3 in the main paper).
- Section D: Broader impact of our work.

A Algorithm of Camera-aware Gating

This section presents in detail our camera-aware gating mechanism, including the computation of the geometric relevance score s_i and the filtering of unmatched frames based on spatial and temporal constraints. The goal is to determine when to enable memory interaction. We first enable all the gates and then deactivate the interaction in three steps, as described in the following.

Step 1: Deactivate Gate by Geometric Relevance Scoring

For each target pose, we calculate its geometric alignment with historical poses by computing the field-of-view (FOV) overlap [57] and a normalized translation-distance penalty to derive a relevance score for each historical frame. The frame with the highest score is selected as the initial candidate and this score is set as the geometric relevance score s_i for the target frame. A high s_i suggests strong alignment with known content, while a low s_i suggests exploration of new regions. If s_i is below a threshold, we deactivate the corresponding gate due to insufficient alignment with historical memory.

Step 2: Deactivate Gate by Translation Distance

We also deactivate the gate for candidates with large translation distances, even if their relevance scores are high. This step is crucial to avoid overestimating s_i in scenarios such as forward-moving trajectories, where significant FOV overlap with earlier frames might occur. Introducing a translation distance constraint helps prevent such false positives and enhances gating accuracy.

Step 3: Deactivate Gate due to Temporal Redundancy

To prevent over-constraining the generation process, we apply temporal redundancy filtering across adjacent target frames. Injecting highly similar historical contexts into consecutive frames can weaken the temporal dynamics. In general, if a frame within a brief interval is grounded on historical memory, its neighboring frames will maintain spatial consistency. This filtering reduces the frequency of memory usage, balancing spatial coherence with temporal dynamics.

We summarize the whole algorithm of camera-aware gating in Algorithm 1. Frames with high s_i (which remain active after the filtering steps) are treated as revisits to known regions, thus enabling memory injection, while frames with low s_i (or those deactivated by distance/temporal constraints) are considered novel and generated without memory interaction.

B Experimental Details

B.1 Training Details

Our training follows a two-stage procedure. First, we train the camera control module following AC3D [3]. Specifically, camera pose conditioning is injected only at the low-frequency denoising steps, where global structure is determined. We observe that restricting camera supervision to these early denoising steps accelerates convergence and leads to more accurate camera pose control. In the second stage, we freeze both the video backbone and the camera control module, and train the memory control branch independently. During the first 20% of training iterations in this stage, the camera control module is temporarily disabled and the memory branch is trained in isolation, allowing it to first learn stable memory representations before interacting with the camera control signal.

Algorithm 1 Camera-Aware Gating via Geometric Relevance Scoring

Input:
- Target poses $P_{\text{tgt}} = \{P_t\}_{t=1}^{N}$, where $P_t = (R_t, \mathbf{t}_t) \in \text{SE}(3)$.
- Historical poses $P_{\text{hist}} = \{P_r\}_{r=1}^{F}$, where $P_r = (R_r, \mathbf{t}_r) \in \text{SE}(3)$.

Output:
- Geometric relevance scores $S = \{s_t\}_{t=1}^{N}$.
- Binary gates $G = \{g_t\}_{t=1}^{N}$, where $g_t \in \{0, 1\}$.

begin

For each target pose index $t = 1, \ldots, N$:

Initialize current gate:

$$g_t \leftarrow 1$$

Compute FOV overlap with all historical poses [57]:

$$\rho_r^{(t)} \leftarrow \text{OverlapComputation}(P_t, P_r), \qquad r = 1, \ldots, F$$

Compute normalized translation distances:

$$d_r^{(t)} \leftarrow \text{Norm}(\|\mathbf{t}_t - \mathbf{t}_r\|_2), \qquad r = 1, \ldots, F$$

Compute geometric relevance score:

$$c_r^{(t)} \leftarrow \rho_r^{(t)} - \lambda_d d_r^{(t)}, \qquad r = 1, \ldots, F$$

Select the highest-scoring historical pose:

$$r_t^* \leftarrow \arg\max_r c_r^{(t)}, \qquad s_t \leftarrow c_{r_t^*}^{(t)}$$

△ **Step 1: Deactivate Gate by Geometric Relevance Scoring**

$$\text{if } s_t < \tau_c, \qquad g_t \leftarrow 0$$

△ **Step 2: Deactivate Gate by Translation Distance**

$$\text{if } d_{r_t^*}^{(t)} > \tau_d, \qquad g_t \leftarrow 0$$

For each target pose index $t = 2, \ldots, N$:

△ **Step 3: Deactivate Gate due to Temporal Redundancy**

$$\text{if } g_t = 1 \wedge g_{t-1} = 1 \wedge |r_t^* - r_{t-1}^*| < \tau_{\text{temp}}, \qquad g_t \leftarrow 0$$

Return S **and** G**.**

end

B.2 FLOPs Estimation Details

We detail the FLOPs estimation protocol used in Table 2 of the main paper. For fair comparison, all methods are evaluated under the same setting: a single denoising step with 61 input frames. Backbone FLOPs refer to the computation cost of the pretrained generation model itself, while memory FLOPs refer to the additional cost introduced by the memory mechanism. The total FLOPs are reported as the sum of backbone FLOPs and memory FLOPs.

For VMEM [31], the backbone FLOPs are calculated by running SEVA [66] (historical frames=1) to generate 61 frames under its iterative inference paradigm. The memory FLOPs refer to the extra cost arising from its memory mechanism, which retains more historical frames(=4) and reduces the newly generated frames per iteration, thus increasing the total number of iterations required.

For WorldPlay[45], Ours-Wan, and Ours-Cog, the backbone FLOPs are calculated by running the pretrained video generation models (HunyuanVideo 1.5 [28], Wan 2.1 [50], and CogVideoX [59]), respectively, to generate 61 frames with camera pose conditioning. For WorldPlay[45], the memory FLOPs arise from concatenating retrieved memory frames into the input sequence, which increases quadratically the attention cost of the backbone. For our methods, the memory FLOPs arise from the computation of the lightweight memory branch, which operates separately from the backbone and introduces negligible overhead.

B.3 Peak Memory and Wall-Clock Latency

We also report the peak GPU memory and wall-clock latency on an NVIDIA A800 GPU in TableA1. The results distinguish the backbone-only cost from the total cost with memory modeling, directly showing the additional practical overhead.

Table A1: Peak GPU memory and wall-clock latency under the same inference setting. Latency is measured per denoising step.

Method	Peak Memory (GB)			Wall-Clock Latency (s)		
	Backbone	Total	Increase	Backbone	Total	Increase
WorldPlay	21.57	27.03	+5.46 (+25.3%)	11.18	45.47	+34.29 (+306.7%)
Ours-Cog	26.02	31.26	+5.24 (+20.1%)	5.78	5.97	+0.20 (+3.5%)
Ours-Wan	41.30	48.89	+7.60 (+18.4%)	9.29	10.81	+1.52 (+16.4%)

C More Qualitative Results

More results on RealEstate10K and OOD scenes. We provide more qualitative examples on RealEstate10K [67] in Figure A1 and additional OOD results in Figure A2, including scenes with various artistic styles in indoor and outdoor environments. Our method not only synthesizes reasonable novel contents during scene exploration, but also preserves scene consistency when the camera revisits previously observed views. Although trained only on RealEstate10K, our decoupled design preserves the generative capability of the pretrained video model, enabling strong generalization to diverse scenes beyond the training distribution.

Additional comparisons with memory-based baselines. Figure A3 provides two additional comparison examples with memory-based baselines, including VMem [31] and WorldPlay [45], on OOD data and RealEstate10K. As highlighted by the red boxes, prior methods tend to introduce visible artifacts when synthesizing novel contents in unseen regions. In contrast, our method generates cleaner and more coherent results, demonstrating superior robustness in scene exploration and revisiting.

D Broader Impact

This work improves spatial consistency and visual realism during scene revisiting in long-horizon video generation, and can be applied to embodied intelligence, world simulation, game content generation, and virtual environment construction.

However, more coherent and realistic long-video generation capabilities may also be misused to create misleading synthetic content, such as fabricated scene walkthroughs or surveillance-style videos. Since the proposed method can reduce artifacts such as spatial inconsistency, the generated content may become more difficult to identify as synthetic videos.

Potential mitigation measures include synthetic-content watermarking, provenance tracking, sensitive-data filtering, controlled release of model weights, and clear labeling of generated content. This work is primarily intended as foundational research but not direct deployment in high-risk scenarios.

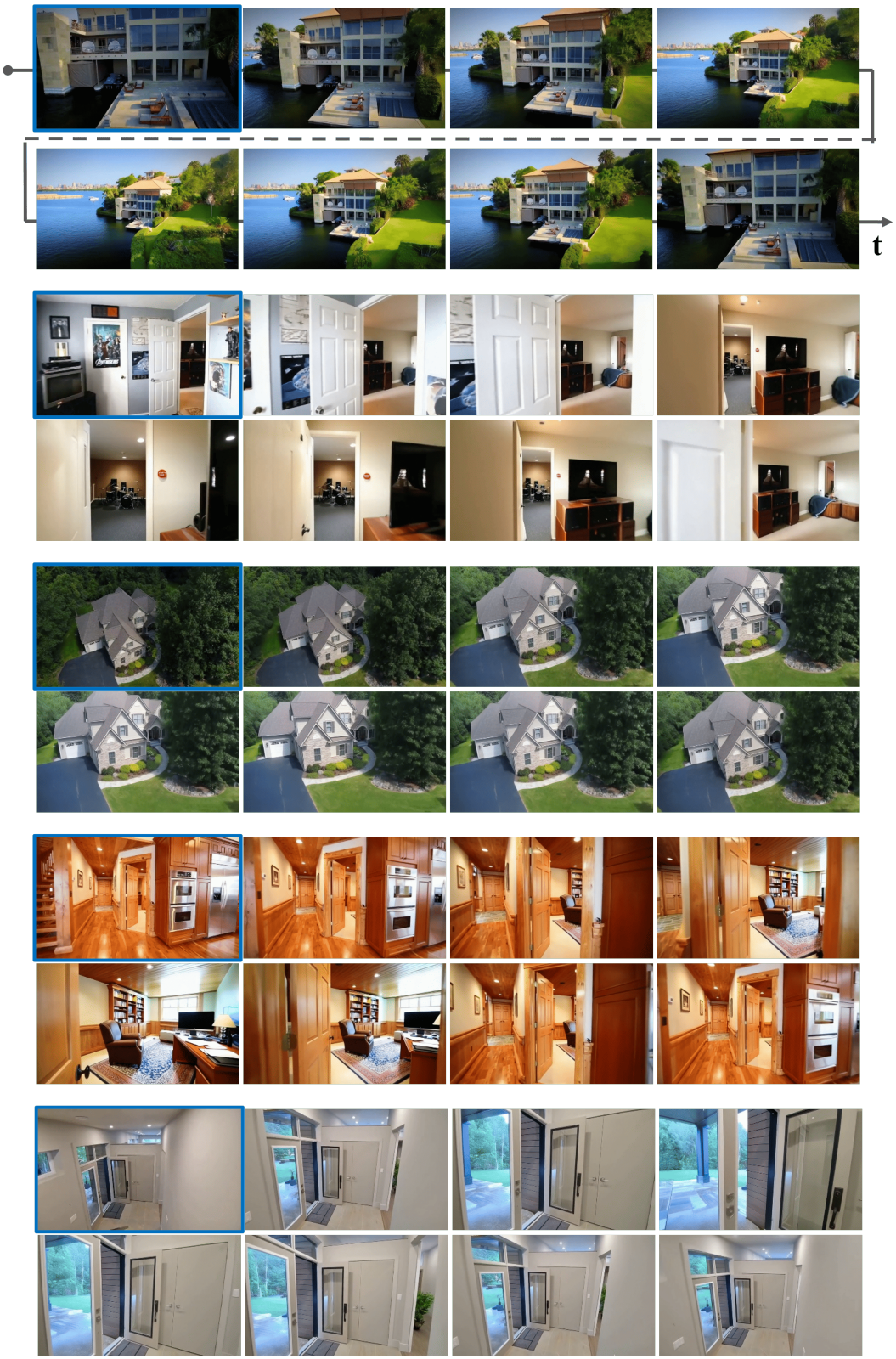

Figure A1: **Qualitative results on the RealEstate10K dataset [67].** The boxed frame marks the starting frame, with following sequence from the first row to the second.

.

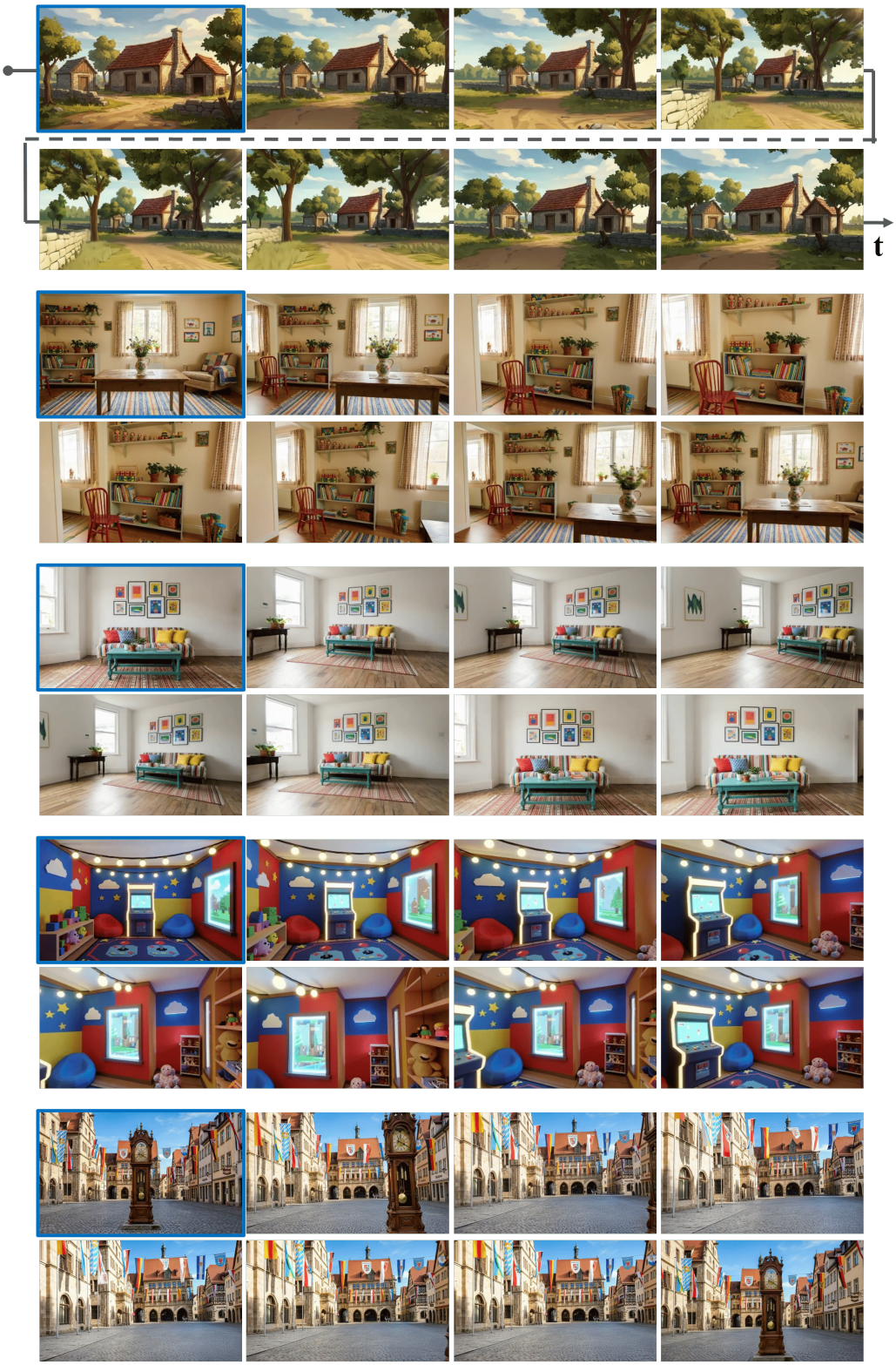

Figure A2: **Qualitative results on out-of-distribution (OOD) scenes**, with diverse styles in both indoor and outdoor environments. The boxed frame marks the starting frame, with the following sequence from the first row to the second.



Figure A3: **Comparison with memory-based baselines.** Top: an OOD example. Bottom: a RealEstate10K [67] example. boxed frame marks the starting frame, with the following sequence from the first row to the second. Red boxes mark artifacts of VMem [31] and WorldPlay [45] when generating novel contents in unseen regions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, the main claims in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations in the final section (after the conclusion).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include formal theoretical results, theorems, or proofs. The motivation of the work comes from practical observations, and the effectiveness of the proposed method is validated through extensive experiments.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, the paper fully discloses all necessary information to reproduce the main experimental results, supporting the main claims and conclusions.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All training data are publicly available. We will release the codes and data if the paper is accepted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, the paper provides all necessary details, including hyperparameters, evaluation metrics and *etc.* to fully understand the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We detail the type of compute resources in Experiment.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, the research in the paper fully conforms to the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, we analyze the potential social impact in appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, we clearly indicate the baseline methods and testing data used in the paper. Their licenses permit use with academic scope.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Code and dataset will be all released if the paper is accepted.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowd sourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowd sourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Our core method focuses on video generation and does not involve LLMs as part.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.