WorldPack: Dynamic Frame Compression for Long-context Video World Modeling

Anonymous authors
Paper under double-blind review

Abstract

Video world models have attracted significant attention for their ability to produce high-fidelity future visual observations conditioned on past observations and navigation actions. However, achieving temporally and spatially consistent generation over long horizons remains an open challenge: existing approaches either compress past frames at fixed rates based on temporal proximity, discarding spatially critical information, or retrieve only a handful of relevant frames without increasing the total amount of retained history. In this paper, we propose *WorldPack*, a video world model that introduces spatially-aware compressed memory to address both limitations simultaneously. The key insight is that compression rates should not be uniform or temporally determined, but should instead be dynamically allocated based on 3D spatial relevance to the current viewpoint. WorldPack achieves this through two tightly coupled mechanisms: *trajectory packing*, which fits substantially more historical frames into a fixed-length context through hierarchical frame compression, and *geometric selection*, which leverages camera pose information and field-of-view overlap to assign lower compression to spatially important frames and higher compression to less relevant ones. Together, these mechanisms expand the effective context from 4 to 22 frames with only 16% additional inference time, while preserving the most informative frames for spatial reasoning with high fidelity. We evaluate WorldPack on LoopNav, a Minecraft benchmark for long-horizon spatial consistency, and conduct comprehensive experiments on the RECON, real-world navigation dataset, across multiple evaluation protocols. WorldPack consistently outperforms strong baselines—including Oasis, Mineworld, DIAMOND, NWM—with particularly pronounced gains in spatial reasoning tasks that require recall of distant observations.

1 Introduction

Video world models, i.e., neural world simulators based on video generation models, have recently attracted significant attention for their ability to produce high-fidelity future visual observations conditioned on past observations and navigation actions (Brooks et al., 2024; Ball et al., 2025; World Labs, 2025a; Hafner et al., 2025). By predicting and generating future visual observations from past observations and agent actions, these models hold the potential to serve as alternatives to conventional simulation environments. Their applications span a wide range of domains, such as robotic simulation (Bar et al., 2024; Hu et al., 2025; Zhu et al., 2025; Mao et al., 2025a; Chen et al., 2025), autonomous driving (Hu et al., 2023; Russell et al., 2025; Wang et al., 2023; Zhao et al., 2024; Gao et al., 2024a), and AI-driven content generation in game engines (Alonso et al., 2024; Valevski et al., 2024; Bruce et al., 2024).

Recent work has begun to address these challenges from two complementary but disconnected angles. On the one hand, FramePack (Zhang & Agrawala, 2025) compresses past frames at varying rates to fit more history into a fixed context, but it determines compression rates based on temporal proximity, assigning the highest fidelity to the most recent frames regardless of their spatial relevance. This is suboptimal for world modeling, where a temporally distant frame may share substantial 3D overlap with the current viewpoint and thus be critical for consistent generation. On the other hand, spatial memory retrieval methods (Xiao et al., 2025; Yu et al., 2025a) select past frames based on field-of-view overlap or 3D co-visibility, but they

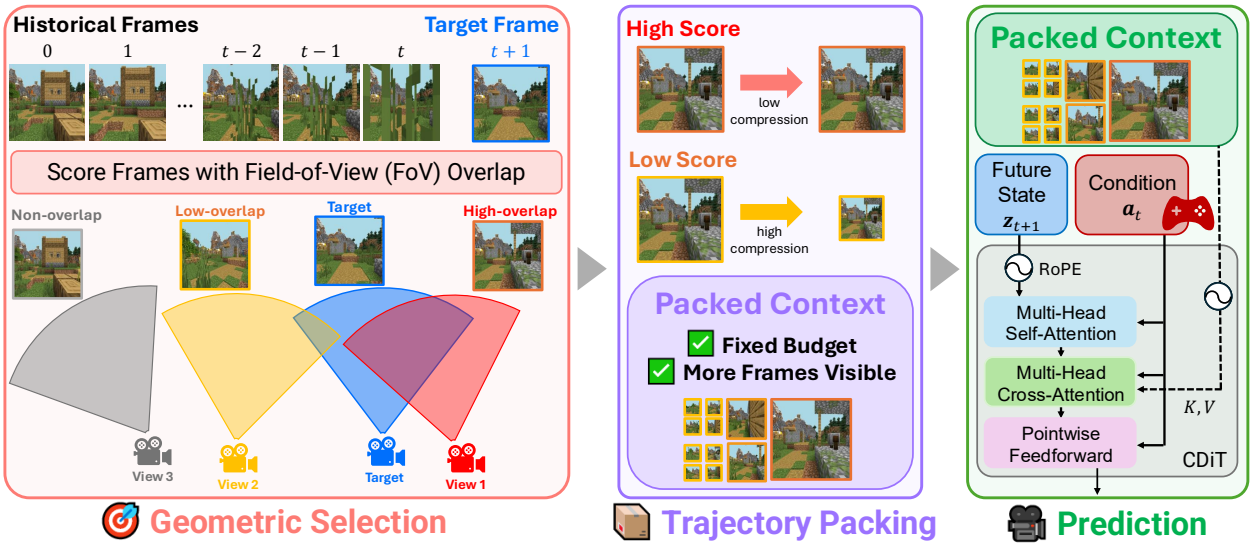

Figure 1: WorldPack consists of (1) geometric selection: dynamic allocation of compression rate based on camera pose information, (2) trajectory packing: packing the trajectory into the context, and (3) CDiT with RoPE-based timestep embedding.

operate within a fixed context window, replacing less relevant frames entirely rather than retaining them at reduced resolution. This binary keep-or-discard strategy limits the total amount of historical information available to the model.

In this paper, we propose WorldPack, a video world model that bridges these two directions by introducing spatially-aware compressed memory. Rather than treating frame selection and frame compression as independent problems, WorldPack unifies them: it packs many historical frames into a fixed-length context while dynamically allocating compression rates based on 3D spatial relevance. Frames that strongly overlap with the current observation are preserved at high resolution, while less relevant frames are aggressively compressed but still retained, ensuring that no historical information is entirely discarded. This design enables the model to reason over substantially longer horizons without incurring a proportional increase in computational cost.

We build WorldPack on a conditional diffusion transformer (CDiT) (Bar et al., 2024) backbone with RoPE-based (Su et al., 2023) temporal embeddings, and evaluate it on LoopNav (Lian et al., 2025), a Minecraft benchmark for long-horizon spatial consistency, across both spatial memory retrieval and spatial reasoning tasks. We further conduct comprehensive experiments on the RECON dataset (Shah et al., 2021) under multiple protocols to demonstrate effectiveness on real-world data. Through detailed ablation studies, we reproduce the two most closely related approaches within our own backbone as controlled baselines: (i) the temporal-proximity packing of FramePack (Zhang & Agrawala, 2025), and (ii) the spatial retrieval mechanism of WorldMem (Xiao et al., 2025) and Context-as-Memory (Yu et al., 2025a). Comparing WorldPack against (i) isolates the contribution of geometric selection, and against (ii) isolates the contribution of trajectory packing; together these reveal that the two mechanisms address complementary bottlenecks—expanding the amount of available history and determining how that history is compressed.

2 Related Work

Video World Models. Recent advances in video diffusion models have enabled photorealistic, high-resolution video generation, positioning them as "general-purpose world simulators" capable of producing diverse scenes with plausible dynamics from text (Ho et al., 2022c;b; Brooks et al., 2024; Google DeepMind, 2024; Kang et al., 2024; Bansal et al., 2024; Chefer et al., 2025; Wu et al., 2025; Oshima et al., 2025). Building

on this progress, video world models have attracted significant attention for their ability to generate high-fidelity future visual observations conditioned on past scene sequences and navigation actions (Ball et al., 2025; World Labs, 2025a; Mao et al., 2025c;b; Hong et al., 2025; HunyuanWorld, 2025; Hafner et al., 2025). Their applications span a wide range of domains, such as game engines (Valevski et al., 2024; Decart et al., 2024; Guo et al., 2025; Bruce et al., 2024), autonomous driving (Hu et al., 2023; Russell et al., 2025; Wang et al., 2023; Zhao et al., 2024; Gao et al., 2024a; Hu et al., 2024; Guo et al., 2024), and robotics (Bar et al., 2024; Zhu et al., 2025; Hu et al., 2025; Mao et al., 2025a; Chen et al., 2025). These applications underscore the importance of maintaining long-term temporal and spatial consistency, particularly in decision-making tasks such as driving and navigation.

However, achieving such coherence remains an unresolved challenge, even for state-of-the-art models, due to the prohibitively high computational costs required to process a long sequence of observations in the model context (Decart et al., 2024; Guo et al., 2025). Recent studies (Yu et al., 2025a; Xiao et al., 2025; World Labs, 2025b) propose spatial retrieval mechanisms that select past frames based on overlapping fields of view, improving spatial consistency by ensuring that relevant observations are included in the context. However, these methods operate within a fixed context window: they choose which frames to include, but cannot increase the total number of frames accessible to the model. As a result, the trade-off between spatial relevance and historical coverage remains unresolved; a spatially relevant frame from the distant past may be included only at the cost of discarding other potentially useful observations.

Long-Context Video Generation. In video generation, extensive research has focused on extending fixed-length generation horizons to long-term rollouts. Representative directions include temporal super-resolution with coarse-to-fine processing (Ho et al., 2022b; Yin et al., 2023), as well as architectural advances aimed at capturing long-range dependencies (Gu et al., 2021; Gu & Dao, 2023; Oshima et al., 2024; Gao et al., 2024b). While these methods enable the generation of longer videos, they ultimately remain constrained by fixed-length outputs. One of the major research directions toward overcoming this limitation is autoregressive long-term video generation. These approaches generate videos sequentially conditioned on recent frames (He et al., 2022; Henschel et al., 2024; Po et al., 2025b; Jin et al., 2024; Kodaira et al., 2025; Yang et al., 2025; Gao et al., 2025; Qiu et al., 2025), and include inference-time techniques that adapt pretrained models to longer rollouts without retraining (Qiu et al., 2023; Kim et al., 2024), as well as few-step model distillation methods (Yin et al., 2025).

However, autoregressive long-term video generation suffers from error accumulation and memory forgetting as the rollout length increases (Wang et al., 2025). To mitigate error accumulation, various stabilization methods have been explored, including combining next-token prediction with full-sequence diffusion (Chen et al., 2024; Ruhe et al., 2024; Song et al., 2025), and training models to correct drift by directly conditioning on their own generated frames during autoregressive rollouts (Huang et al., 2025; Shin et al., 2025; Cui et al., 2025; Po et al., 2025a; Yu et al., 2025b). Recently, Zhang & Agrawala (2025) proposed compressing past frames at varying rates when injecting them into the context, retaining long histories while reducing the impact of accumulated drift. However, their compression schedule is determined by temporal proximity—recent frames receive the highest fidelity while earlier frames are progressively compressed—which is suboptimal for world modeling, where spatial relevance rather than temporal recency determines which frames are most informative. In this work, we transfer such context compression techniques to the setting of video world modeling and, critically, replacing the temporal compression schedule with a spatially-aware one that dynamically allocates compression rates based on 3D co-visibility between past and current viewpoints.

3 Preliminaries

We begin by extending latent diffusion models (Rombach et al., 2022) to the temporal domain, formulating video diffusion models (He et al., 2022; Ho et al., 2022a). Given a sequence of frames $\mathbf{x}_{0:T} = (\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_T)$, we first encode frames into latent representations $\mathbf{z}_{0:T} = (\mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_T)$ using a pretrained VAE (Kingma & Welling, 2013), i.e., $\mathbf{z}_i = \mathrm{Enc}(\mathbf{x}_i)$. In this setting, all latent frames share the same noise level k, and the reverse diffusion process restores the clean sequence by iteratively denoising:

$$p_\theta(\mathbf{z}_{0:T}^{k-1} \mid \mathbf{z}_{0:T}^{k}) = \mathcal{N}\big(\mathbf{z}_{0:T}^{k-1}; \mu_\theta(\mathbf{z}_{0:T}^{k}, k), \sigma_k^2 I\big)\,, \tag{1}$$

where $\mathbf{z}_{0:T}^{k}$ denotes the noisy latent sequence at noise level k. This full-sequence formulation provides global guidance across frames, but constrains the sequence length to that used during training and lacks flexibility for long-horizon rollouts.

To overcome this limitation, we adopt an autoregressive formulation. Instead of generating the entire sequence jointly, the model conditions on the most recent m latent frames to predict the next one:

$$p_\theta(\mathbf{z}_{t+1} \mid \mathbf{z}_{t-m+1:t}), \tag{2}$$

where generation proceeds sequentially. This setup naturally extends video length beyond the training horizon and supports long-term coherent generation.

Finally, to obtain an interactive video world model, we further introduce action sequences into the formulation. Given past latent states $\mathbf{z}_{t-m:t}$ and the current action $\mathbf{a}_t$, we learn a stochastic transition model F_θ:

$$\mathbf{z}_{t+1} \sim F_\theta(\mathbf{z}_{t+1} \mid \mathbf{z}_{t-m:t}, \mathbf{a}_t). \tag{3}$$

This formulation approximates the environment dynamics $p(\mathbf{z}_{t+1} \mid \mathbf{z}_{\leq t}, \mathbf{a}_{\leq t})$, while operating in the compressed latent space. The predicted next state can then be decoded back into pixel space for visualization, enabling action-conditioned video generation and long-term world simulation.

4 WorldPack

The design of WorldPack is motivated by a gap between two existing approaches to long-horizon conditioning. FramePack (Zhang & Agrawala, 2025) compresses past frames at varying rates to expand the effective context, but determines compression rates by temporal proximity, a poor proxy for spatial relevance in world modeling. Spatial memory retrieval (Xiao et al., 2025; Yu et al., 2025a) selects the most relevant frames based on 3D co-visibility, but is constrained to a fixed context window and discards all non-selected frames. WorldPack unifies these ideas: it packs a large number of frames into the context while allocating compression rates based on spatial relevance (like memory retrieval), so that all historical frames are retained at a fidelity proportional to their importance.

4.1 Video World Modeling with Conditional Diffusion Transformer

Following Section 3, we design F_θ as a probabilistic mapping to simulate stochastic environments. To this end, we employ CDiT (Bar et al., 2024), which is a temporally autoregressive transformer model, and where efficient CDiT blocks are applied N times over the input sequence (Figure 1). Unlike a standard Transformer that applies self-attention across all tokens, CDiT restricts self-attention to the tokens of the denoised target frame and incorporates cross-attention over past frames, allowing efficient learning. This cross-attention contextualizes the representation through skip connections, and conditioning on input actions is incorporated. While a standard DiT (Peebles & Xie, 2023) can be directly applied, its computational complexity scales quadratically with context length, i.e., $O(m^2n^2d)$ for n tokens per frame, m frames, and token dimension d. In contrast, CDiT is dominated by the cross-attention complexity $O(mn^2d)$, which scales linearly with context length, enabling the use of longer contexts.

In addition, our model must integrate memory contexts located at arbitrary temporal distances from the current timestep. To achieve this, we adopt Rotary Position Embeddings (RoPE) (Su et al., 2023) as a position-aware design. RoPE enables consistent temporal representations regardless of variable context length, providing stable embeddings even for memory frames selected at arbitrary distances. This allows memory-aware inference over sequences with long-term dependencies.

4.2 Spatially-Aware Compressed Memory

Previous video world models are constrained by a fixed context length, preventing them from incorporating long-term history. While they remain sensitive to recent observations, predicting scenes that depend on events further in the past is challenging. This limitation causes errors to accumulate during rollouts, leading generated trajectories to gradually diverge from the original world (Decart et al., 2024; Guo et al., 2025).

Algorithm 1 WorldPack: Spatially-Aware Compressed Memory

Require: Historical frames $\{z_0, \ldots, z_t\}$, camera poses $\{p_0, \ldots, p_t\}$, current action a_t, context budget L_{pack}, uncompressed slots S
Ensure: Predicted next frame z_{t+1}
▷ Spatial importance scoring
for each historical frame z_i **do**
  $s_i \leftarrow \text{FoVOverlap}(p_i, p_t)$
end for
▷ Geometric selection
Sort frames by s_i in descending order
Assign top-S frames to uncompressed slots ($d_i = 0$)
Assign remaining frames with d_i proportional to rank
▷ Trajectory packing
for each frame z^i with priority d_i **do**
  Encode at resolution $\ell_i = L_f / \lambda^{d_i}$
end for
Concatenate into packed context $\mathbf{z}_{\text{ctx}}$
▷ Conditional generation via CDiT
$z_{t+1} \sim F_\theta(z_{t+1} \mid \mathbf{z}_{\text{ctx}}, a_t)$
return z_{t+1}

To overcome this, we propose a spatially-aware compressed memory that combines hierarchical frame compression (i.e., trajectory packing) with 3D-guided rate allocation (i.e., geometric selection) into a single mechanism. Past frames are encoded at different resolutions depending on their spatial importance: frames that share a large field-of-view overlap with the current viewpoints are preserved at high resolution, while spatially less relevant frames are compressed and stored at lower resolution.

Trajectory packing. Let a sequence of frames selected from the historical trajectory be $z^0, z^1, \ldots, z^N$, where N is the number of frames maintained in the context window. After the Transformer patchifying process, each frame z^i is assigned an effective context length ℓ_i determined by:

$$\ell_i = \frac{L_f}{\lambda^{d_i}}, \tag{4}$$

where L_f is the base context length for high-resolution frames, $\lambda > 1$ controls compression intensity, and d_i is the priority index of frame z^i. A lower d_i indicates higher priority, resulting in more tokens and higher visual fidelity. The total packed context length is:

$$L_{\text{pack}} = \sum_{i=0}^{S-1} L_f + \sum_{i=S}^{N} \ell_i, \tag{5}$$

where S denotes the number of uncompressed slots reserved for the most critical observations.

Geometric selection. The key question is how to assign d_i. FramePack (Zhang & Agrawala, 2025) assigns it by temporal recency, but in world modeling an agent revisiting a previously observed location needs high-fidelity access to those earlier observations regardless of how many timesteps have elapsed. We instead score each historical frame by how strongly it overlaps the current view in 3D space.

Each camera pose p induces a truncated viewing frustum $V(p) \subset \mathbb{R}^3$ given the field-of-view angle and a near/far depth range. We define the field-of-view overlap of a historical frame i as

$$o_i \;=\; \frac{\text{Vol}\big(V(p_i) \cap V(p_t)\big)}{\text{Vol}\big(V(p_t)\big)} \;\in\; [0, 1], \tag{6}$$

estimated by Monte Carlo sampling of M points in $V(p_t)$ and counting those also inside $V(p_i)$. To break ties among frames with similar overlap, where temporally closer frames tend to have less pose drift and fewer

compounding generation artifacts, we add a mild temporal penalty:

$$s_i \;=\; w_o\, o_i \;-\; w_t\, \Delta t_i, \qquad \Delta t_i = t - i, \quad w_o \gg w_t > 0. \tag{7}$$

Frames are then sorted by s_i and assigned compression: the top-S go to the uncompressed slots ($d_i = 0$), and d_i grows with rank for the rest. The contrast with temporal-proximity packing is that spatially critical frames are preserved at high resolution regardless of Δt_i.

Implementation details. We use three compression ratios $2^0, 2^2, 2^4$ ($\lambda = 2$ with $d_i \in \{0, 1, 2\}$), giving per-frame context lengths of 2, 4, 16 tokens. The packed context holds $S = 2$ uncompressed frames, 4 frames at ratio 2^2, and 16 frames at ratio 2^4, totaling $1{+}1{+}4{+}16 = 22$ historical frames. By Eq. 5 the packed length is $2L_f + (4/2^2)L_f + (16/2^4)L_f = 4L_f$, matching the budget of the 4-frame baseline while exposing 5.5× more frames to the model. Each compression ratio uses an independent input projection layer, initialized by interpolating the pretrained patchify layer of the base model (kernel size $(4, 4)$).

Algorithm 1 summarizes the complete procedure.

5 Evaluation on Spatial Consistency

We primarily focus on evaluating video world models' ability to retain long-term spatial memory. For this purpose, we leverage LoopNav (Lian et al., 2025), a benchmark constructed in Minecraft environments. LoopNav is designed for loop-style navigation tasks, in which the agent explores a portion of the environment and then returns to an earlier location. This design provides a precise and targeted method for testing whether a model can recall and reconstruct previously observed scenes, making LoopNav a distinctive benchmark for evaluating spatial memory.

Spatial Memory Retrieval Task (ABA). The most basic setting of LoopNav is the A→B→A trajectory (Figure 2; **Left**). In this case, the segment from A to B acts as the exploration phase, supplying contextual observations to the model. The return path from B to A constitutes the reconstruction phase, during which the model must demonstrate spatial consistency in regenerating observations from earlier locations. Because the ground-truth sequence has already been observed, this scenario is best viewed as a spatial retrieval task that explicitly probes whether the model can reproduce information embedded in the context.

Spatial Reasoning Task (ABCA). Here, A→B→C forms the exploration phase, while C→A is evaluated as the reconstruction phase (Figure 2; **Right**). Unlike an A→B→A loop, this task challenges the model to rely on accumulated spatial memory to reconstruct the environment along an extended path, potentially across areas observed from different viewpoints or at earlier time steps. This setup is closely related to a spatial reasoning task, where success requires leveraging contextual knowledge to generate coherent future observations rather than simply retrieving frames.

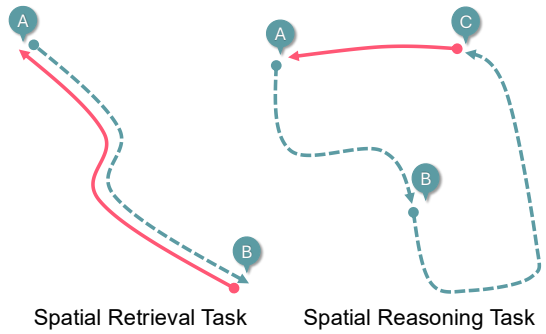

Figure 2: Illustration of the two LoopNav benchmark tasks. (**Left**) Spatial Memory Retrieval Task: the agent explores along A→B (blue path) and must reconstruct earlier observations on the return path B→A (red path). (**Right**) Spatial Reasoning Task: the agent explores along A→B→C (blue path) and must reconstruct the environment on the longer return path C→A (red path), requiring reasoning across accumulated spatial memory.

Metrics. For evaluation, we use LPIPS (Zhang et al., 2018) to assess semantic-level perceptual fidelity, SSIM (Wang et al., 2004) to evaluate low-level structural alignment, and Fréchet Video Distance (FVD) (Unterthiner et al., 2019) to evaluate video synthesis quality. We further employ DreamSim (Fu et al., 2023), which measures perceptual similarity based on deep feature representations, and PSNR to capture pixel-level reconstruction quality. Since no single metric fully reflects semantic accuracy or long-term spatial coherence, we complement these quantitative results with qualitative inspection by human observers.

Table 1: Model performance on tasks of varying type and difficulty. ABA denotes the spatial memory retrieval tasks, and ABCA denotes the spatial reasoning tasks. The navigation range (5, 15, 30, 50) indicates the size of the area within which the agent is required to move. SSIM (↑) evaluates better structural consistency, while LPIPS (↓) reflects perceptual fidelity, and FVD (↓) measures temporal video quality. We refer to baseline evaluation results from Lian et al. (2025).

Nav. Range	Model	Context	Frames	SSIM ↑		LPIPS ↓		FVD ↓	
				ABA	ABCA	ABA	ABCA	ABA	ABCA
5	Oasis	32	32	0.36	0.34	0.76	0.82	2615	2583
	Mineworld	15	15	0.31	0.32	0.73	0.72	2089	1914
	DIAMOND	4	4	0.40	**0.37**	0.75	0.79	3353	3336
	NWM	4	4	0.33	0.31	0.64	0.67	1950	2240
	WorldPack (ours)	4	22	**0.41**	0.35	**0.51**	**0.58**	**1510**	**1449**
15	Oasis	32	32	0.37	0.38	0.82	0.81	2516	3146
	Mineworld	15	15	0.34	0.32	0.74	0.74	2367	2009
	DIAMOND	4	4	0.38	0.39	0.78	0.79	3691	3302
	NWM	4	4	0.30	0.33	0.67	0.65	2132	2338
	WorldPack (ours)	4	22	**0.38**	**0.41**	**0.55**	**0.54**	**1448**	**1339**
30	Oasis	32	32	0.33	**0.35**	0.86	0.85	3131	3199
	Mineworld	15	15	0.33	0.28	0.77	0.77	2316	2094
	DIAMOND	4	4	**0.37**	**0.35**	0.81	0.81	3708	3473
	NWM	4	4	0.32	0.30	0.69	0.71	1893	2437
	WorldPack (ours)	4	22	0.32	0.34	**0.63**	**0.60**	**1777**	**1618**
50	Oasis	32	32	0.36	0.36	0.86	0.83	3334	3162
	Mineworld	15	15	0.31	0.32	0.78	0.75	2077	2144
	DIAMOND	4	4	**0.37**	**0.38**	0.83	0.81	3249	2994
	NWM	4	4	0.28	0.33	0.72	0.65	2715	1537
	WorldPack (ours)	4	22	0.36	0.36	**0.57**	**0.59**	**2004**	**1440**

6 Experiments

6.1 Baselines

Oasis (Decart et al., 2024) is a world model that employs a ViT (Dosovitskiy et al., 2020) as a spatial autoencoder and a DiT (Peebles & Xie, 2023) as the latent diffusion backbone, trained with Diffusion Forcing (Chen et al., 2024). It generates frames autoregressively with user-controllable conditioning, and the publicly available Oasis-500M model is evaluated with a context length of 32. Mineworld (Guo et al., 2025) is an interactive world model based on a pure Transformer architecture that generates new scenes from paired game frames and actions, with its pretrained checkpoint evaluated at a context length of 15. DIAMOND (Alonso et al., 2024) is a diffusion-based world model built upon a UNet architecture (Ronneberger et al., 2015), generating frames conditioned on past observations and actions, and evaluated with a context length of 4. NWM (Bar et al., 2024) is a controllable video generation model that predicts future observations conditioned on navigation actions, leveraging CDiT with a context length of 4.

6.2 Results

In the multi-step rollout generation (Table 1 and Table 2), WorldPack, despite the shortest context length, outperforms the baselines – Oasis, Mineworld, DIAMOND, and NWM – in SSIM and LPIPS, and also surpasses NWM in PSNR, DreamSim, and FVD. However, the SSIM results were not decisively superior, remaining only partially competitive. This tendency can be explained by the inherent limitations of distortion-based metrics, which favor spatially averaged or blurred predictions that minimize pixel-wise differences at the expense of perceptual fidelity (Blau & Michaeli, 2018). Indeed, Lian et al. (2025) also reported that SSIM exhibits only a weak correlation with perceptual quality in visualizations.

Table 2: Evaluation of models on spatial memory (ABA) and reasoning (ABCA) tasks under different navigation ranges. PSNR (↑) reflects pixel-level reconstruction accuracy, DreamSim (↓) captures perceptual similarity based on deep features.

Nav. Range	Model	PSNR ↑		DreamSim ↓	
		ABA	ABCA	ABA	ABCA
5	NWM	12.1	11.5	0.34	0.36
	WorldPack (ours)	**13.2**	**12.1**	**0.28**	**0.34**
15	NWM	10.7	11.8	0.44	0.37
	WorldPack (ours)	**12.8**	**12.6**	**0.32**	**0.31**
30	NWM	10.7	10.0	0.46	0.47
	WorldPack (ours)	**11.3**	**11.4**	**0.42**	**0.38**
50	NWM	9.4	10.3	0.52	0.47
	WorldPack (ours)	**11.9**	**11.6**	**0.35**	**0.37**

Collectively, these results demonstrate consistent improvements across both the ABA and ABCA tasks, as evidenced by quantitative metrics across all navigation ranges. In particular, the proposed compressed memory mechanism plays a crucial role in balancing high context efficiency with long-term spatial consistency. Accommodating more frames than uncompressed baselines allows the essential frames for world modeling to remain accessible even under the shortest context-length constraints.

6.3 Ablation Study

Table 3: Ablation Study of WorldPack on ABA-5 in LoopNav. Each baseline ablates one component: Nearest Frame Packing applies trajectory packing without geometric selection (TP only), while Memory Retrieval applies geometric selection without trajectory packing (GS only). Thus, the gap between Nearest Frame Packing and WorldPack isolates the effect of geometric selection (GS), and the gap between Memory Retrieval and WorldPack isolates the effect of trajectory packing (TP). Collectively, these results suggest that both components are vital for robust world modeling.

Method	**TP**	**GS**	**DreamSim ↓**	**LPIPS ↓**	**PSNR ↑**	**SSIM ↑**	**FVD ↓**
Baseline	✗	✗	0.44	0.60	10.7	0.37	2030
Nearest Frame Packing	✓	✗	0.40	0.60	11.4	0.34	1683
Memory Retrieval	✗	✓	0.36	0.56	12.0	0.38	1694
WorldPack (ours)	✓	✓	**0.32**	**0.55**	**12.8**	**0.38**	**1510**

To evaluate the individual contributions of trajectory packing and geometric selection, we conducted an ablation study comparing the following four configurations. For a fair comparison, all settings are constrained to a fixed context size of four frames.

- **Baseline:** Following the standard approach (Bar et al., 2024), the four most recent frames are used directly as the context.
- **Nearest Frame Packing:** Following the protocol in FramePack (Zhang & Agrawala, 2025), the 22 most recent frames are compressed into a 4-frame context, with compression rates determined by temporal proximity.
- **Memory Retrieval:** Following the spatial memory retrieval mechanism of WorldMem (Xiao et al., 2025) and Context-as-Memory (Yu et al., 2025a), the context consists of the four frames with the highest FoV-based spatial similarity scores. This setting serves as a direct comparison with these retrieval-based methods.

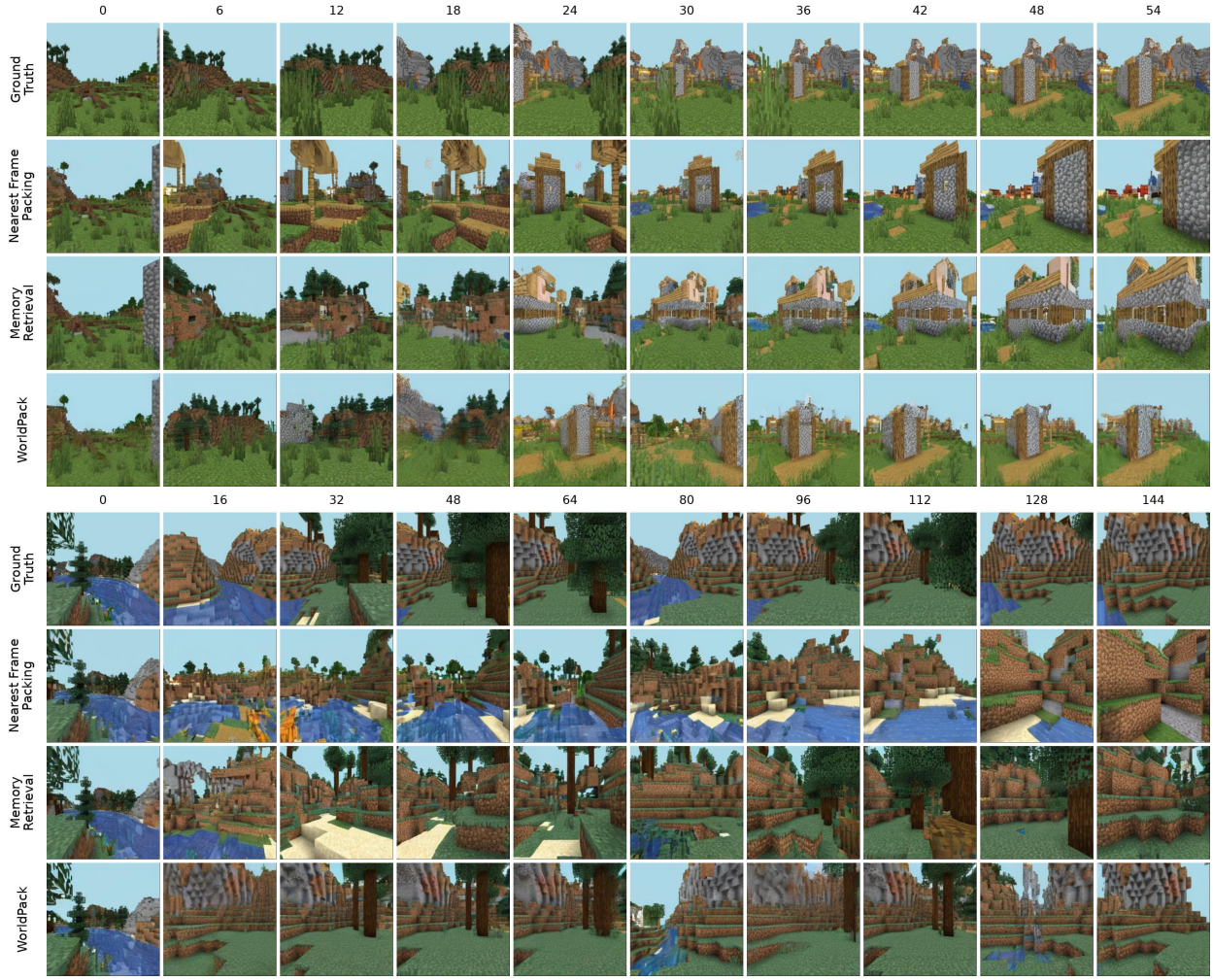

Figure 3: Qualitative comparison of rollouts. We compare Ground Truth, Nearest Frame Packing, Memory Retrieval, and WorldPack. WorldPack preserves the environment's spatial structure more consistently than the other variants.

- **WorldPack (ours):** The 22 frames are compressed into a 4-frame context, where compression rates are determined based on spatial similarity scores.

First, the results of the ablation study for ABA-5 in LoopNav are presented in Table 3. The comparison between Nearest Frame Packing and WorldPack demonstrates the effectiveness of geometric selection, which adaptively determines compression rates based on 3D-aware importance rather than employing a fixed rate. Furthermore, since the Memory Retrieval setting reproduces the selection mechanism of WorldMem (Xiao et al., 2025) and Context-as-Memory (Yu et al., 2025a), its comparison with WorldPack serves as a direct evaluation against these retrieval-based methods and highlights the efficacy of trajectory packing, which enables the handling of larger frame sizes without increasing context length by compressing past frame information. Collectively, these results suggest that both components are vital for robust world modeling. Figure 3 qualitatively compares the three packing/retrieval variants against the ground truth: WorldPack preserves the environment's spatial structure more faithfully than Nearest Frame Packing and Memory Retrieval across the displayed frames.

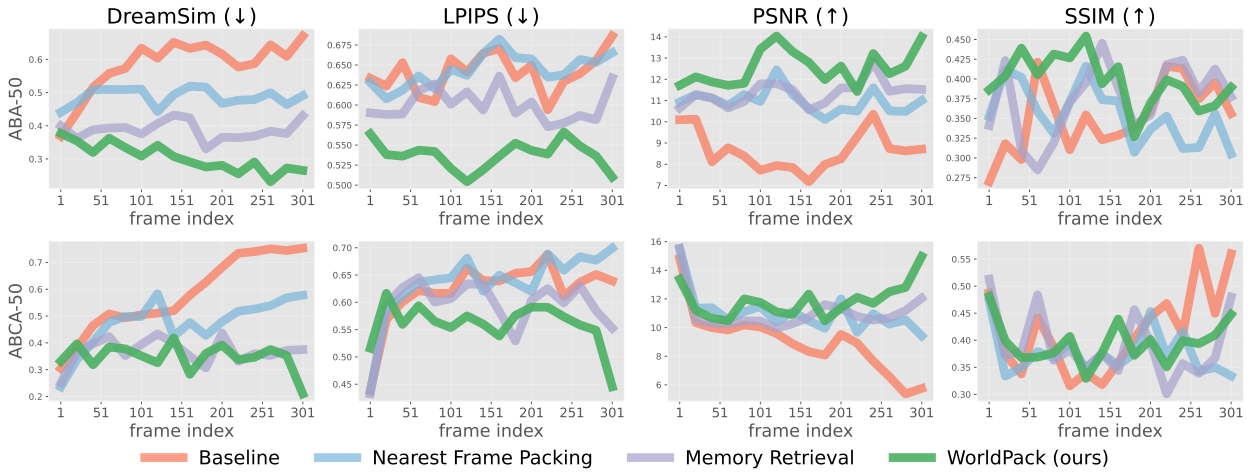

Figure 4: Prediction performance on the terminal frames of ABCA trajectories with different navigation ranges. **Top**: 301 frames rollout in ABA-50. **Bottom**: 301 frames rollout in ABCA-50. WorldPack not only accesses task-relevant information based on 3D spatial cues but also retains a significantly larger number of frames within the context through frame compression. Consequently, the model can effectively correct the generation by fully leveraging past observations, thereby minimizing quality degradation.

Next, for a more detailed analysis, Figure 4 illustrates the transitions of each metric throughout a 301-frame rollout for the LoopNav ABA-50 and ABCA-50 tasks. Nearest Frame Packing sometimes shows performance improvements during the initial stages of the rollout, as it can maintain a larger context and allow for longer access to past observations (e.g., in ABCA-50). However, as generations progress, past observations are eventually evicted from the context window, leading to a gradual degradation in generation quality. Memory Retrieval, which corresponds to the retrieval mechanism of WorldMem (Xiao et al., 2025) and Context-as-Memory (Yu et al., 2025a), can extract past information essential for prediction based on 3D spatial proximity scoring. While this helps mitigate the divergence in generation quality to some extent, its effectiveness is limited by the fixed context length, which restricts the total number of frames the model can handle simultaneously. In contrast, WorldPack not only accesses task-relevant information based on 3D spatial cues but also retains a significantly larger number of frames within the context through frame compression. Consequently, the model can effectively correct the generation by fully leveraging past observations, thereby minimizing quality degradation. This advantage is particularly evident in the latter segments of ABCA-50, a spatial reasoning task, where WorldPack demonstrates significant performance gains. In such spatial reasoning tasks, the importance of past observations for accurate prediction is maximized during the latter stages of the rollout (see Figure 2; **Right**). While WorldPack successfully leverages this information to improve generation, other methods fail to recover quality, either because they cannot access past observations or lack sufficient context capacity to retain them.

6.4 Experiments with Real-World Data

To verify the practical usefulness of WorldPack beyond simulator environments such as Minecraft, we conducted experiments using real-world data. Specifically, we evaluated our method on the RECON dataset (Shah et al., 2021), one of the most commonly used datasets in prior video-generation world-model studies (Shah et al., 2022; Sridhar et al., 2024; Bar et al., 2024). In our experiments, we used the first 80 frames as context and generated the subsequent frames. The quantitative results are shown in Table 4. These results demonstrate that WorldPack achieves strong generative performance even on real-world data, confirming its effectiveness beyond simulated environments.

Table 4: Evaluation on RECON dataset, real-world generation performance, including DreamSim (↓), LPIPS (↓), PSNR (↑), and SSIM (↑).

Model	DreamSim ↓	LPIPS ↓	PSNR ↑	SSIM ↑	FVD ↓
Baseline	0.31	0.53	11.7	0.30	822
WorldPack	**0.18**	**0.45**	**13.3**	**0.40**	**694**

Table 5: Inference time and memory usage comparison.

Model	Frames	Inference Time (1-step, sec)	Memory Usage (GB)
Baseline	4	0.255	22.7
WorldPack	22	0.296	25.4

6.5 Analysis of Computational Efficiency

We present the single-step inference time and memory costs for the diffusion model in Table 5. Compared to the baseline, WorldPack extends the visible length of past frames from 4 to 22 frames, a 5.5× increase in context length. Notably, despite this substantial expansion, the computational overhead remains remarkably low. The inference time increases by only approximately 16%, which is a marginal increase given the additional temporal information processed. Furthermore, memory consumption demonstrates excellent scalability: while handling over five times as many frames, the memory footprint increases by only about 12%. These experimental results corroborate WorldPack's ability to maintain high computational efficiency and scalability even when dealing with long-range trajectory dependencies.

7 Discussion and Limitation

In this study, we focused on memory management for world modeling and employed a 3D scoring mechanism based on camera poses—a widely adopted approach in existing literature—to determine frame importance (HunyuanWorld, 2025; Yu et al., 2025a; Xiao et al., 2025). However, it has been noted that such scoring methods, which rely heavily on 3D information, may underperform in complex environments with occlusion (Xiao et al., 2025). Consequently, exploring more robust scoring metrics that can overcome these constraints will be crucial for achieving more sophisticated and reliable world modeling in the future. In addition, we primarily focused on the simulation capabilities of video world models and therefore evaluated their scene-generation performance. As a future direction, we believe that exploring policy learning and planning with video world models (Alonso et al., 2024) will further deepen the discussion on the utility of spatial memory capabilities.

8 Conclusion

We introduced WorldPack, a video world model that achieves long-horizon spatial consistency through spatially-aware compressed memory. By unifying trajectory packing with geometric selection, WorldPack retains substantially more historical context than prior methods while preserving high fidelity for the frames most relevant to spatial reasoning. Experiments on LoopNav and RECON reveal that simply expanding context length is less effective than intelligently compressing a larger history with spatial guidance and spatially adaptive compression rates, which provide clear benefits over both temporal-proximity-based packing and fixed-context spatial retrieval, with the advantage growing over the rollout horizon. Additionally, a controlled comparison that reproduces WorldMem's retrieval mechanism within our backbone confirms that the contribution lies not in spatial scoring itself but in using it to control compression rates across a larger set of frames.

Broader Impact Statement

This work studies memory mechanisms for video world models. Potential positive impacts include more efficient simulation and planning systems. Potential risks include the misuse of increasingly realistic interactive video generation systems to create deceptive or harmful content. Our work does not introduce new data collection or human-subject experiments, and our experiments are conducted on public navigation benchmarks. We encourage future deployments to incorporate provenance, access control, and safety evaluation.

References

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *arXiv preprint arXiv:2405.12399*, 2024.

Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, Christos Kaplanis, Marjorie Limont, Matt McGill, Yanko Oliveira, Jack Parker-Holder, Frank Perbet, Guy Scully, Jeremy Shar, Stephen Spencer, Omer Tov, Ruben Villegas, Emma Wang, Jessica Yung, Cip Baetu, Jordi Berbel, David Bridson, Jake Bruce, Gavin Buttimore, Sarah Chakera, Bilva Chandra, Paul Collins, Alex Cullum, Bogdan Damoc, Vibha Dasagi, Maxime Gazeau, Charles Gbadamosi, Woohyun Han, Ed Hirst, Ashyana Kachra, Lucie Kerley, Kristian Kjems, Eva Knoepfel, Vika Koriakin, Jessica Lo, Cong Lu, Zeb Mehring, Alex Moufarek, Henna Nandwani, Valeria Oliveira, Fabio Pardo, Jane Park, Andrew Pierson, Ben Poole, Helen Ran, Tim Salimans, Manuel Sanchez, Igor Saprykin, Amy Shen, Sailesh Sidhwani, Duncan Smith, Joe Stanton, Hamish Tomlinson, Dimple Vijaykumar, Luyu Wang, Piers Wingfield, Nat Wong, Keyang Xu, Christopher Yew, Nick Young, Vadim Zubov, Douglas Eck, Dumitru Erhan, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Raia Hadsell, Aäron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 3: A new frontier for world models. `https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/`, 2025. DeepMind blog post.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models, 2024. URL `https://arxiv.org/abs/2412.03572`.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6228–6237, 2018. doi: 10.1109/CVPR.2018.00652.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

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. *arXiv preprint arXiv:2402.15391*, 2024.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2018. URL `https://arxiv.org/abs/1705.07750`.

Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. VideoJAM: Joint appearance-motion representations for enhanced motion generation in video models. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=yMJcHWcb2Z`.

Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion, 2024. URL `https://arxiv.org/abs/2407.01392`.

Boyuan Chen, Tianyuan Zhang, Haoran Geng, Kiwhan Song, William T. Freeman, Jitendra Malik, Russ Tedrake, Vincent Sitzmann, and Yilun Du. Large video planner, 2025. URL http://arxiv.org/abs/2512.15840.

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.

Decart, Julian Quevedo, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. 2024. URL https://oasis-model.github.io/.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50742–50768, 2023.

Jianxiong Gao, Zhaoxi Chen, Xian Liu, Junhao Zhuang, Chengming Xu, Jianfeng Feng, Yu Qiao, Yanwei Fu, Chenyang Si, and Ziwei Liu. Longvie 2: Multimodal controllable ultra-long video world model, 2025. URL https://arxiv.org/abs/2512.13604.

Ruiqi Gao*, Aleksander Holynski*, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole*. Cat3d: Create anything in 3d with multi-view diffusion models. *Advances in Neural Information Processing Systems*, 2024.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability, 2024a.

Yu Gao, Jiancheng Huang, Xiaopeng Sun, Zequn Jie, Yujie Zhong, and Lin Ma. Matten: Video generation with mamba-attention, 2024b. URL https://arxiv.org/abs/2405.03025.

Google DeepMind. Veo 2, 2024. URL https://deepmind.google/technologies/veo/veo-2/.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. Mineworld: a real-time and open-source interactive world model on minecraft. *arXiv preprint arXiv:2504.08388*, 2025.

Xi Guo, Chenjing Ding, Haoxuan Dou, Xin Zhang, Weixuan Tang, and Wei Wu. Infinitydrive: Breaking time limits in driving world models, 2024. URL https://arxiv.org/abs/2412.01522.

Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models, 2025. URL https://arxiv.org/abs/2509.24527.

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.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022.

Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv:2210.02303*, 2022b.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022c.

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

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving, 2023.

Xiaotao Hu, Wei Yin, Mingkai Jia, Junyuan Deng, Xiaoyang Guo, Qian Zhang, Xiaoxiao Long, and Ping Tan. Drivingworld: Constructingworld model for autonomous driving via video gpt. *arXiv preprint arXiv:2412.19505*, 2024.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=c0dhw1du33.

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.

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

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

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model?: A physical law perspective. *arXiv preprint arXiv:2406.16860*, 2024.

Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. Fifo-diffusion: Generating infinite videos from text without training. *NeurIPS*, 2024.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Akio Kodaira, Tingbo Hou, Ji Hou, Masayoshi Tomizuka, and Yue Zhao. Streamdit: Real-time streaming text-to-video generation, 2025. URL https://arxiv.org/abs/2507.03745.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Kewei Lian, Shaofei Cai, Yilun Du, and Yitao Liang. Toward memory-aided world models: Benchmarking via spatial consistency, 2025. URL https://arxiv.org/abs/2505.22976.

Jiageng Mao, Sicheng He, Hao-Ning Wu, Yang You, Shuyang Sun, Zhicheng Wang, Yanan Bao, Huizhong Chen, Leonidas Guibas, Vitor Guizilini, Howard Zhou, and Yue Wang. Robot learning from a physical world model, 2025a. URL https://arxiv.org/abs/2511.07416.

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

Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An interactive world generation model. *arXiv preprint arXiv:2507.17744*, 2025c.

Yuta Oshima, Shohei Taniguchi, Masahiro Suzuki, and Yutaka Matsuo. SSM meets video diffusion models: Efficient video generation with structured state spaces. In *5th Workshop on practical ML for limited/low resource settings*, 2024. URL https://openreview.net/forum?id=jzbeme6FdW.

Yuta Oshima, Masahiro Suzuki, Yutaka Matsuo., and Hiroki Furuta. Inference-time text-to-video alignment with diffusion latent beam search, 2025. arXiv preprint arXiv:2501.19252.

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

Ryan Po, Eric Ryan Chan, Changan Chen, and Gordon Wetzstein. Bagger: Backwards aggregation for mitigating drift in autoregressive video diffusion models, 2025a. URL https://arxiv.org/abs/2512.12080.

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

Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling, 2023.

Haonan Qiu, Shikun Liu, Zijian Zhou, Zhaochong An, Weiming Ren, Zhiheng Liu, Jonas Schult, Sen He, Shoufa Chen, Yuren Cong, Tao Xiang, Ziwei Liu, and Juan-Manuel Perez-Rua. Histream: Efficient high-resolution video generation via redundancy-eliminated streaming. *arXiv preprint arXiv:2512.21338*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling diffusion models, 2024. URL https://arxiv.org/abs/2402.09470.

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

Dhruv Shah, Benjamin Eysenbach, Nicholas Rhinehart, and Sergey Levine. Rapid Exploration for Open-World Navigation with Latent Goal Models. In *5th Annual Conference on Robot Learning*, 2021. URL https://openreview.net/forum?id=d_SWJhyKfVw.

Dhruv Shah, Ajay Sridhar, Arjun Bhorkar, Noriaki Hirose, and Sergey Levine. GNM: A General Navigation Model to Drive Any Robot. In *arXiV*, 2022. URL https://arxiv.org/abs/2210.03370.

Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. Motionstream: Real-time video generation with interactive motion controls. *arXiv preprint arXiv:2511.01266*, 2025.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL https://arxiv.org/abs/1409.1556.

Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 19313–19325. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/a11ce019e96a4c60832eadd755a17a58-Paper.pdf.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion, 2025. URL https://arxiv.org/abs/2502.06764.

Ajay Sridhar, Dhruv Shah, Catherine Glossop, and Sergey Levine. Nomad: Goal masked diffusion policies for navigation and exploration. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 63–70, 2024. doi: 10.1109/ICRA57147.2024.10610665.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2019.

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

Jing Wang, Fengzhuo Zhang, Xiaoli Li, Vincent Y. F. Tan, Tianyu Pang, Chao Du, Aixin Sun, and Zhuoran Yang. Error analyses of auto-regressive video diffusion models: A unified framework, 2025. URL https://arxiv.org/abs/2503.10704.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

World Labs. Generating worlds. https://www.worldlabs.ai/blog/generating-worlds, 2025a. Product site.

World Labs. Rtfm: A real-time frame model. https://www.worldlabs.ai/blog/rtfm, 2025b. Company blog post.

Ziyi Wu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Ashkan Mirzaei, Igor Gilitschenski, Sergey Tulyakov, and Aliaksandr Siarohin. DenseDPO: Fine-grained temporal preference optimization for video diffusion models. *NeurIPS*, 2025.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory, 2025. URL https://arxiv.org/abs/2504.12369.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, and Song Hanand Yukang Chen. Longlive: Real-time interactive long video generation. 2025.

Shengming Yin et al. Nuwa-xl: Diffusion over diffusion for extremely long video generation. *arXiv preprint arXiv:2303.12346*, 2023.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. 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. *arXiv preprint arXiv:2506.03141*, 2025a.

Zhengyang Yu, Akio Hayakawa, Masato Ishii, Qingtao Yu, Takashi Shibuya, Jing Zhang, and Yuki Mitsufuji. Autorefiner: Improving autoregressive video diffusion models via reflective refinement over the stochastic sampling path, 2025b. URL https://arxiv.org/abs/2512.11203.

Lvmin Zhang and Maneesh Agrawala. Packing input frame contexts in next-frame prediction models for video generation. *Arxiv*, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. *arXiv preprint arXiv:2403.06845*, 2024.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

Appendix

A The Use of Large Language Models

In this paper, we mainly used LLMs to polish writing and propose paraphrases.

B Evaluation Metrics

To rigorously assess the semantic consistency of the world model outputs, we employ Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) and DreamSim (Fu et al., 2023). These metrics evaluate perceptual similarity based on deep features extracted from neural networks. Specifically, LPIPS utilizes image classification models (e.g., AlexNet (Krizhevsky et al., 2012) or VGG (Simonyan & Zisserman, 2015)) as its backbone to capture human perception of structural differences. Following Lian et al. (2025), we used VGG as a backbone.

Additionally, we use Peak Signal-to-Noise Ratio (PSNR) to quantify pixel-level quality by measuring the ratio of the maximum pixel value to the error; higher values indicate better quality.

To evaluate video synthesis quality, we use Fréchet Video Distance (FVD) (Unterthiner et al., 2019). FVD uses I3D (Carreira & Zisserman, 2018) as its backbone to compare the feature distributions of real and generated videos, with lower scores indicating higher visual quality.

C Further Ablation Study

Encoding Spatial Information Helps World Modeling. We investigate the impact of encoding spatial information on world modeling. Following Sitzmann et al. (2021); Xiao et al. (2025), we adopt Plücker embedding to convert 5D poses $p \in \mathbb{R}^5$ into dense positional features $PE(p) \in \mathbb{R}^{h \times w \times 6}$, consistent with recent works (He et al., 2024; Gao* et al., 2024). As shown in Table 6, removing the camera pose embedding (`w/o Camera Pose Embedding`) results in performance degradation across key metrics, including DreamSim and LPIPS. These results confirm that explicitly injecting spatial information via camera poses is highly effective for enhancing the understanding of 3D structures and improving prediction accuracy in memory-based world modeling.

Too Much Compression Collapses World Modeling. Next, we examine the effect of compression rates in the tokenizer on model performance. While our main method employs a frame-wise tokenizer with packing limited to the spatial dimension, this ablation study investigates configurations that incorporate temporal compression (Table 7).

First, we observed that compressing only the temporal dimension (`+ Temporal Compression`) improves performance compared to the baseline. This improvement is likely due to temporal compression, which allows the model to handle longer frame sequences within the same token budget, enabling the world model to leverage a broader range of past information. However, when further spatial compression (`+ Nearest Frame Packing`) or spatio-temporal compression (`+ Temporal Packing`) is applied, the performance deteriorates. These findings suggest that excessive compression leads to significant information loss, which outweighs the benefits of an extended context length. This confirms a critical trade-off between representation density and context length in effective world modeling.

Table 6: Ablation for encoding spatial information.

Method	**DreamSim ↓**	**LPIPS ↓**	**PSNR ↑**	**SSIM ↑**	**FVD ↓**
Baseline	0.44	0.60	10.7	0.37	2030
Memory Retrieval	0.36	0.56	12.0	0.38	1694
w/o Camera Pose Embedding	0.38	0.58	11.44	0.37	2067

Table 7: Ablation for compression rate and world modeling performance

Method	**Context**	**Frames**	**DreamSim ↓**	**LPIPS ↓**	**PSNR ↑**	**SSIM ↑**	**FVD ↓**
Baseline	4	4	0.44	0.60	10.7	0.37	2030
+ Temporal Compression	4	16	0.36	0.57	11.5	0.36	1847
+ Nearest Frame Packing	4	88	0.37	0.59	11.4	0.36	1714
+ Temporal Packing	4	296	0.42	0.61	10.8	0.32	1899

D Prediction Performance for Rollout

We describe LoopNav rollout results for ABA-{5, 15} and ABCA-{5, 15} in Figure 5, and for ABA-{30, 50} and ABCA-{30, 50} in Figure 6.

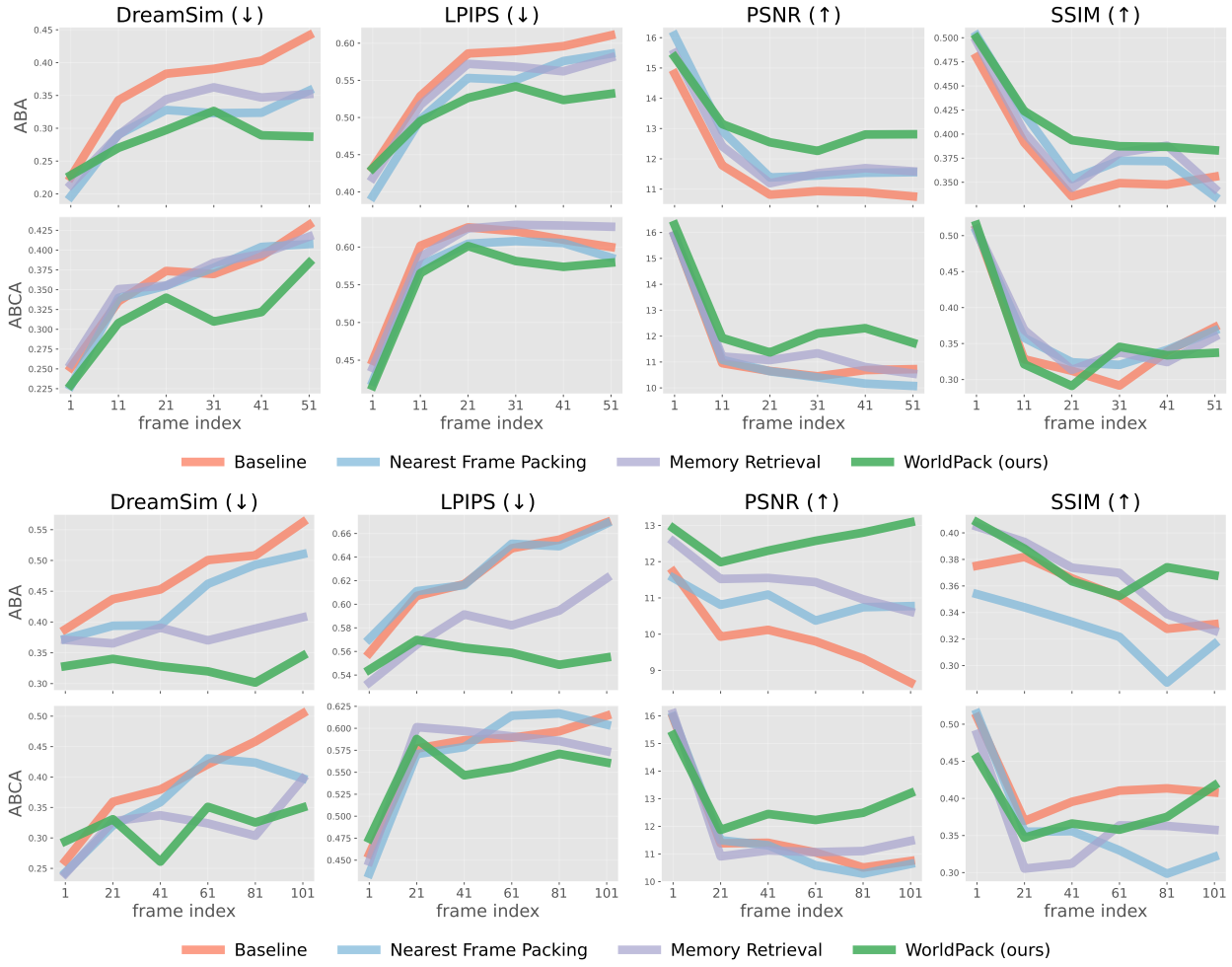

Figure 5: Prediction performance on the terminal frames of ABCA trajectories with different navigation ranges. **Top**: last 51 frames in ABA-5 and ABCA-5. **Bottom**: last 101 frames in ABA-15 and ABCA-15. WorldPack not only accesses task-relevant information based on 3D spatial cues but also retains a significantly larger number of frames within the context through frame compression. Consequently, the model can effectively correct the generation by fully leveraging past observations, thereby minimizing quality degradation.

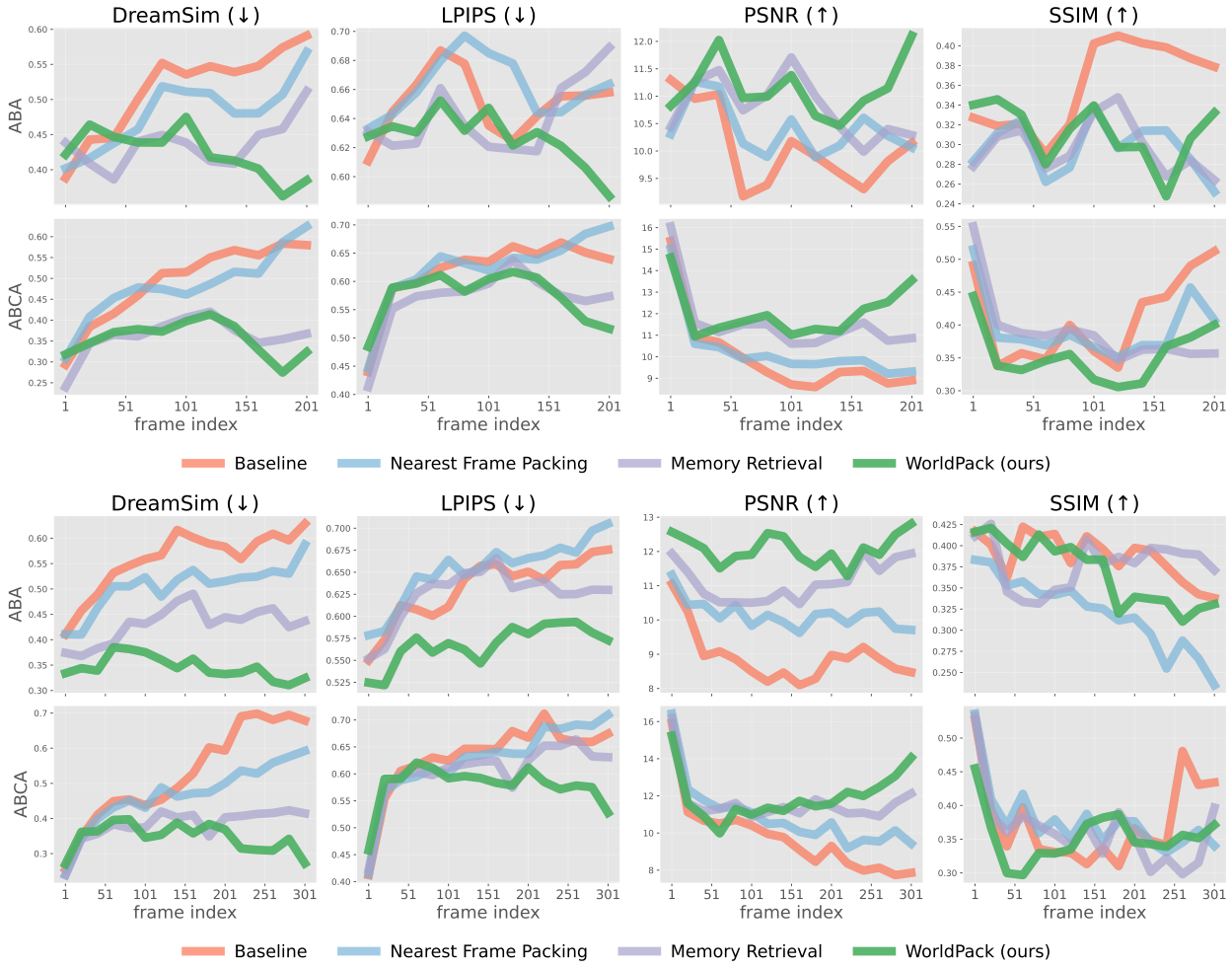

Figure 6: Prediction performance on the terminal frames of ABCA trajectories with different navigation ranges. **Top**: last 201 frames in ABA-30 and ABCA-30. **Bottom**: last 301 frames in ABA-50 and ABCA-50. WorldPack not only accesses task-relevant information based on 3D spatial cues but also retains a significantly larger number of frames within the context through frame compression. Consequently, the model can effectively correct the generation by fully leveraging past observations, thereby minimizing quality degradation.