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

Yuta Oshima[1] Yusuke Iwasawa[1] Masahiro Suzuki[1] Yutaka Matsuo[1] Hiroki Furuta[2,†]

[1]The University of Tokyo [2]Google DeepMind

yuta.oshima@weblab.t.u-tokyo.ac.jp

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. Temporally- and spatially-consistent, long-term world modeling has been a long-standing problem, unresolved even for recent state-of-the-art models, due to the prohibitively high computational costs of long-context inputs. In this paper, we propose WorldPack, a video world model with efficient compressed memory. This compression method allows the model to handle more frames without increasing the number of context tokens. The compressed memory consists of two key components: trajectory packing, which enables the model to handle a significantly larger number of frames while maintaining a constant token length, and dynamic compression, which adjusts compression rates based on camera poses to incorporate 3D spatial information into memory management. Together, these mechanisms ensure consistent rollouts even in later stages, where reliable spatial reasoning is crucial. Our performance is evaluated using LoopNav, a Minecraft benchmark specialized in long-term consistency, and RECON, a real-world navigation dataset. We verify that WorldPack notably outperforms strong state-of-the-art models across both domains.

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 action [2, 6, 26, 73]. 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 [4, 11, 36, 46, 85], autonomous driving [19, 34, 59, 71, 84], and AI-driven content generation in game engines [1, 7, 69].

Despite this promise, achieving temporally and spatially consistent world modeling over long horizons remains a long-standing challenge, even with recent state-of-the-art video world models [13, 24]. This difficulty stems from the prohibitively high computational cost of processing long-context inputs, which limits existing models to relatively short temporal windows [1, 4]. As a result, previously observed information is easily discarded, leading to inconsistencies in spatial layouts and object arrangements over time. For instance, an object visible in one view may abruptly vanish or shift position when the perspective changes, undermining the reliability of such models as world simulators.

In this paper, we propose *WorldPack*, a long-context-aware video world model that achieves efficient compressed memory while maintaining high generation quality. Despite operating with relatively short context lengths, WorldPack substantially improves long-term spatial consistency. The compressed memory consists of two key components: *trajectory packing*, which enables the model to handle a significantly larger number of frames while maintaining a constant token length, and *dynamic compression*, which adjusts compression rates based on camera poses, including both position and orientation, to incorporate 3D spatial information into the memory management process. Together, these mechanisms ensure consistent rollouts even in later stages, where reliable spatial reasoning is crucial. We adopt the conditional diffusion transformer (CDiT) [4] as a base backbone architecture and incorporate RoPE-based temporal embeddings [67], enabling effective utilization of memories regardless of their temporal distance from a target scene.

Our experiments evaluate WorldPack on LoopNav [45], a benchmark designed to assess long-horizon temporal- and spatial-consistency in a Minecraft-based environment. On

[†]Work done as an advisory role only.

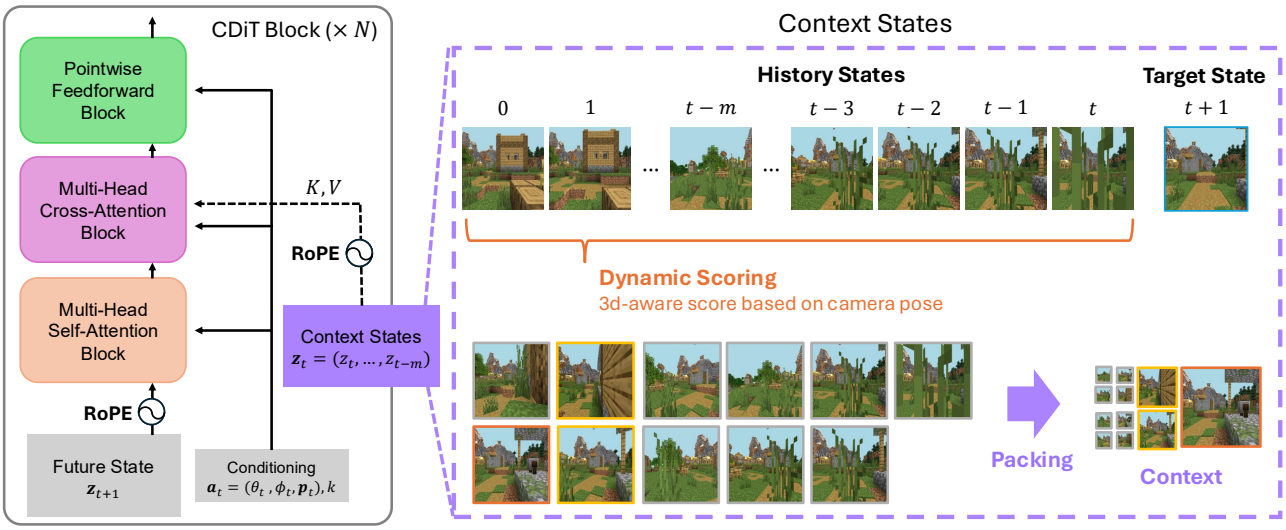

Figure 1. WorldPack consists of (1) CDiT with RoPE-based timestep embedding, (2) packing the trajectory into the context, and (3) dynamic allocation of compression rate based on camera pose information.

both the spatial memory retrieval task, which measures the ability to recall past observations, and the spatial reasoning task, which evaluates consistency under long-horizon rollouts, WorldPack demonstrates superior scene prediction performance. Furthermore, we extend our evaluation to the RECON dataset [60] to demonstrate that WorldPack's effectiveness is not limited to simulated environments. Notably, it substantially outperforms strong state-of-the-art baselines such as Oasis, MineWorld [24], Diamond [1], and NWM [4], as validated across multiple quality metrics, including SSIM [72], LPIPS [83], FVD [68], PSNR, and DreamSim [15].

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 simulator" capable of producing diverse scenes with plausible dynamics from text [3, 6, 9, 21, 31, 32, 40, 50, 75]. 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 [2, 26, 33, 38, 47, 48, 73]. Their applications span a wide range of domains, such as game engines [7, 13, 24, 69], autonomous driving [19, 25, 34, 35, 59, 71, 84], and robotics [4, 11, 36, 46, 85]. 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 [13, 24]. Recent studies [74, 76, 80] propose spatial retrieval mechanisms that select past frames based on overlapping fields of view, but the context window remains fixed, limiting the amount of information that can be incorporated.

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 [31, 78], as well as architectural advances aimed at capturing long-range dependencies [20, 22, 23, 49]. 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 [17, 28, 29, 39, 43, 53, 55, 77], and include inference-time techniques that adapt pretrained models to longer rollouts without retraining [41, 54], as well as few-step model distillation methods [79].

However, autoregressive long-term video generation suffers from error accumulation and memory forgetting as the rollout length increases [70]. To mitigate error accumulation, various stabilization methods have been explored, including combining next-token prediction with full-sequence diffusion [10, 58, 65], and training models to correct drift by directly conditioning on their own generated frames during autoregressive rollouts [12, 37, 52, 62, 81]. Recently, Zhang and Agrawala [82] proposed compressing past frames at varying rates when injecting them into the context, aiming to retain long histories while reducing the impact of drift ac-

cumulated through autoregressive generation. However, the compression schedule is primarily designed based on temporal proximity (e.g., prioritizing recent frames), which is suboptimal for tasks such as video world modeling, where recalling past states should depend on view-dependent relevance rather than temporal distance alone. In this work, we transfer such context compression techniques for long-context generation to the setting of video world modeling.

3. Preliminaries

We begin by extending latent diffusion models [56] to the temporal domain, formulating video diffusion models [28, 30]. Given a sequence of frames $\mathbf{x}_{0:T} = (\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_T)$, we first encode frames into latent representations $\mathbf{z}_{0:T} = (\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_T)$ using a pretrained VAE [42], 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 $p(\mathbf{z}_{t+1} \mid \mathbf{z}_{\le t}, \mathbf{a}_{\le t})$, while operating in the compressed latent space. Predicted next state can then be decoded back to pixel space for visualization, enabling action-conditioned video generation and long-term world simulation.

4. WorldPack

WorldPack adopts a conditional diffusion transformer (CDiT) [4] as the backbone for history and action conditioning and incorporates RoPE-based temporal embeddings [67], allowing effective use of memories regardless of temporal distance (Section 4.1). The compressed memory combines *trajectory packing* to efficiently handle a larger number of frames (Section 4.2 and *dynamic compression* to retain spatially important frames for consistent long-horizon rollout (Section 4.3).

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 [4], 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 [51] 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) [67] 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. Packing Trajectory into Context

Previous video world models have been constrained by a fixed context length, which prevented them from incorporating long-term history. As a result, while they remained sensitive to recent observations, it was challenging to predict scenes that depend on events further in the past. This limitation caused errors to accumulate during rollouts, leading the generated trajectories to gradually diverge from the original world [13, 24].

To overcome this issue, we propose trajectory packing. Trajectory packing enables efficient utilization of long-term history within a fixed-length context by hierarchically compressing and allocating trajectories. Specifically, past frames are encoded at different resolutions depending on their importance: more important frames are preserved at high resolution, while older frames are compressed and stored at lower resolution.

Formally, let a sequence of important frames selected from the historical trajectory be $\mathbf{z}^0, \mathbf{z}^1, \ldots, \mathbf{z}^N$, where $\mathbf{z}^0$ represents the most important observation. Here, N represents the number of consecutive past frames maintained in the context window. After the Transformer patchifying process, each frame $\mathbf{z}^i$ is assigned an effective context length ℓ^i, determined by:

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

where L_f denotes the base context length for high-resolution frames and $\lambda > 1$ is a hyperparameter controlling the compression intensity. where L_f is the base context length for the most recent frame, $\lambda > 1$ controls how aggressively frames are compressed, and d_i represents the priority index of the frame $\mathbf{z}^i$, determined by its 3D spatial relevance and importance to the current prediction. A lower d_i indicates a higher priority, resulting in a larger allocation of tokens and higher visual fidelity. For example, $\lambda = 2, i = 2$ corresponds to a 4×4 patchify kernel, while $i = 4$ corresponds to an 8×8 kernel. The total packed context length is then given by:

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

where S denotes the number of "uncompressed slots" reserved for the most critical observations. This design ensures that high-resolution details are preserved for immediate past observations, while more distant or redundant historical frames are progressively compressed. Consequently, the model can reason over vast temporal scales by fitting significantly more frames into the same token budget.

In practice, we represent frames more efficiently by applying geometric compression [82]. Specifically, we set compression ratios of 2^0, 2^2, and 2^4, which correspond to context lengths of 2, 4, and 16, respectively, and train across a total of 22 context lengths. Furthermore, to account for distributional differences across compression levels, we assign independent input projection layers for each compression ratio, rather than sharing a single projection. These layers are initialized by interpolating from the pretrained patchify layer of the base model with a kernel size of $(4, 4)$. As a result, the model achieves generalized temporal representations that can handle memory contexts selected from arbitrary historical contexts.

4.3. Dynamic Compression

While existing packing studies predominantly rely on fixed compression rates, such as prioritizing the most recent frames [82], our approach moves beyond this uniform approach. We leverage camera pose information and utilize Field-of-View (FOV) overlap–based scoring [16, 76, 80] to dynamically determine the compression ratio for each frame, ensuring that information preservation is based on 3d world information rather than mere temporal proximity.

We first consider FOV overlap as a measure of how strongly a past frame is related to the current observation in three-dimensional space, quantifying the extent to which the two frames observe common regions of the environment. Unlike temporal distance–based criteria, FOV overlap directly captures spatial relevance and viewpoint relationships. Based on this score, we dynamically adjust the compression ratio for each frame, assigning lower compression rates to more important conditioning frames and higher compression to less important ones. As a result, frames that are temporally distant but spatially critical (i.e., those that share strong three-dimensional correspondence with the current viewpoint) are often overly compressed, leading to the loss of important contextual information. By incorporating 3D co-visibility through FOV overlap, our approach mitigates this issue and enables more faithful preservation of spatially salient frames. Overall, dynamically controlling compression based on FOV overlap allows us to make more effective use of spatially important frames, achieving a better spatial consistency in long-term rollout.

5. Evaluation on Spatial Consistency

We primarily focus on evaluating the ability of video world models to retain long-term spatial memory. For this purpose, we leverage LoopNav [45], a benchmark constructed in Minecraft environments. LoopNav is designed for loop-style navigation tasks, where the agent explores a portion of the environment and then returns to an earlier location within it. 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, explicitly probing 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 suc-

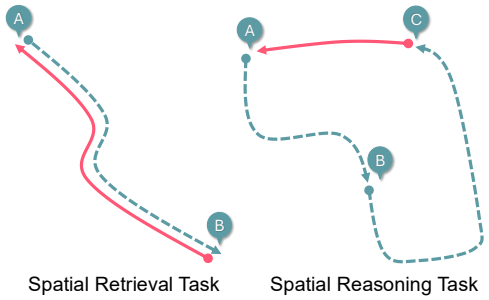

Figure 2. Illustration of the two LoopNav benchmark tasks. (**Left**) Spatial Memory Retrieval Task: the agent explores along A→B (blue path) and must reconstruct earlier observations on the return path B→A (red path). (**Right**) Spatial Reasoning Task: the agent explores along A→B→C (blue path) and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

cess requires leveraging contextual knowledge to generate coherent future observations rather than simply retrieving frames.

Metrics. For evaluation, we use LPIPS [83] to assess semantic-level perceptual fidelity, SSIM [72] to evaluate low-level structural alignment, and Fréchet Video Distance (FVD) [68] to evaluate video synthesis quality. We further employ DreamSim [15], 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.

6. Experiments

6.1. Baselines

Oasis [13] is a world model that employs a ViT [14] as a spatial autoencoder and a DiT [51] as the latent diffusion backbone, trained with Diffusion Forcing [10]. It generates frames autoregressively with user-controllable conditioning, and the publicly available Oasis-500M model is evaluated with a context length of 32. Mineworld [24] 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 [1] is a diffusion-based world model built upon a UNet architecture [57], generating frames conditioned on past observations and actions, and evaluated with a context length of 4. NWM [4] is a controllable video generation model that predicts future observations conditioned on navigation actions, leveraging CDiT with a context length of 4.

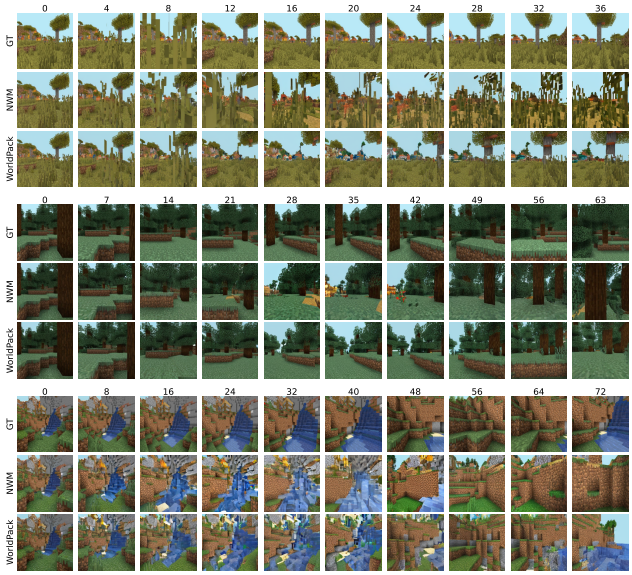

Figure 3. Visualization of rollouts. We compare ground truth (GT), NWM [4], and WorldPack. WorldPack can predict more similar states than NWM, especially in the latter part of the rollouts.

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 [5]. Indeed, Lian et al. [45] also reported that SSIM exhibits only a weak correlation with perceptual quality in visualizations. In addition, qualitative evaluations confirmed that WorldPack maintains long-term consistency, showing only minor deviations from the ground truth even when rollouts are extended (Figure 3).

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

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. [45].

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**

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**

6.3. Ablation Study

To evaluate the individual contributions of trajectory packing and dynamic compression, 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 [4], the four most recent frames are used directly as the context.
- **Nearest Frame Packing:** Following the protocol in Zhang and Agrawala [82], the 22 most recent frames are compressed into a 4-frame context, with compression rates determined by temporal proximity.
- **Memory Retrieval:** Following the protocol in Xiao et al. [76], Yu et al. [80], the context consists of the four frames that exhibit the highest spatial similarity scores based on FoV.
- **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 dynamic compression, which adaptively determines compression rates based on 3D-aware importance rather than employing a fixed rate. ‘ Furthermore, the comparison between Memory Retrieval and WorldPack highlights the efficacy of trajectory packing, which enables

Table 3. Ablation Study of WorldPack on ABA-5 in LoopNav. The comparison between Nearest Frame Packing and WorldPack demonstrates the effectiveness of dynamic compression (DC), which adaptively determines compression rates based on 3D-aware importance rather than employing a fixed rate. Furthermore, the comparison between Memory Retrieval and WorldPack highlights the efficacy of trajectory packing (TP), 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.

Method	TP	DC	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**

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.

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 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 [60], one of the most commonly used datasets in prior video-generation world-model studies [4, 61, 66]. 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.

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 5.5x 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 [38, 76, 80]. However, it has been noted that such scoring methods, which rely heavily on 3D information, may underperform in complex environments with occlusion [76]. 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 be-

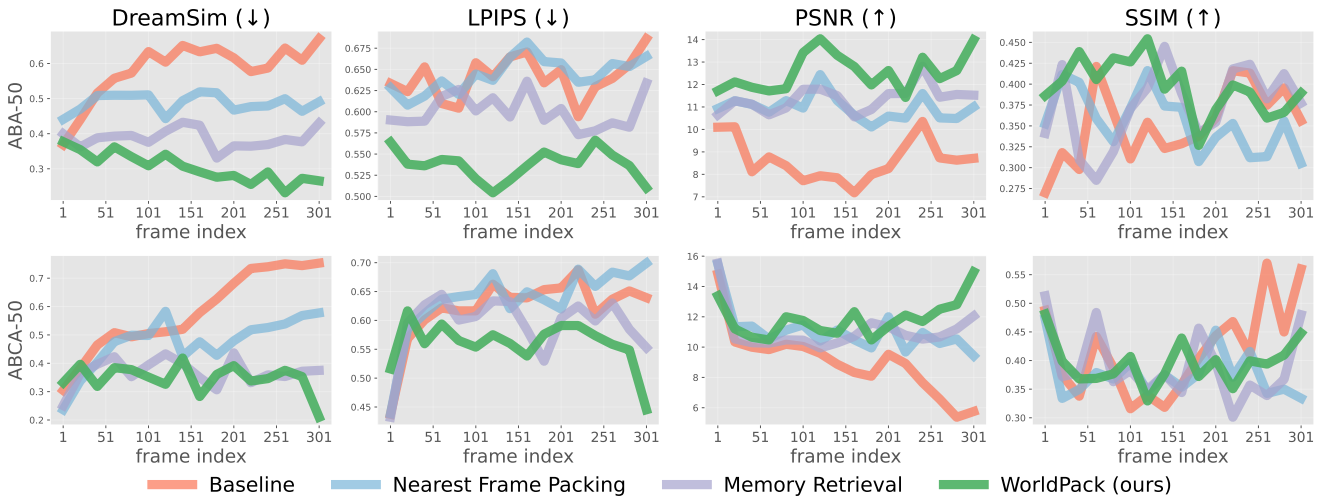

Figure 4. Prediction performance on the terminal frames of ABCA trajectories with different navigation ranges. **Top**: 301 frames rollout in ABA-50. **Bottom**: 301 frames rollout in ABCA-50. WorldPack not only accesses task-relevant information based on 3D spatial cues but also retains a significantly 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.

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

lieve that exploring policy learning and planning with video world models [1] will further deepen the discussion on the utility of spatial memory capabilities.

8. Conclusion

In this paper, we introduce WorldPack, a long-context-aware video world model through context compression. The memory retrieval module facilitates scene generation by selectively using non-recent contextual spatial information. Trajectory packing enables the retention of long-term information without increasing computational costs by compressing past observations. Dynamic compression adaptively allocates compression rates based on spatial similarity scores, prioritizing the most informative frames for world modeling. We hope that this study will further promote the handling of long-context memory in video world models.

References

[1] 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. 1, 2, 5, 8

[2] 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. 1, 2

[3] 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. 2

[4] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models, 2024. 1, 2, 3, 5, 6, 7

[5] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018. 5

[6] 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. 1, 2

[7] 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. 1, 2

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2018. 12

[9] 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. 2

[10] Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion, 2024. 2, 5

[11] Boyuan Chen, Tianyuan Zhang, Haoran Geng, Kiwhan Song, William T. Freeman, Jitendra Malik, Russ Tedrake, Vincent Sitzmann, and Yilun Du. Large video planner, 2025. 1, 2

[12] 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. 2

[13] Decart, Julian Quevedo, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. 2024. 1, 2, 3, 5

[14] 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. 5

[15] 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*, pages 50742–50768, 2023. 2, 5, 12

[16] Xiao Fu, Shitao Tang, Min Shi, Xian Liu, Jinwei Gu, Ming-Yu Liu, Dahua Lin, and Chen-Hsuan Lin. Plenoptic video generation, 2026. 4

[17] 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. 2

[18] 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. 13

[19] 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, 2024. 1, 2

[20] Yu Gao, Jiancheng Huang, Xiaopeng Sun, Zequn Jie, Yujie Zhong, and Lin Ma. Matten: Video generation with mamba-attention, 2024. 2

[21] Google DeepMind. Veo 2, 2024. 2

[22] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2

[23] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 2

[24] 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. 1, 2, 3, 5

[25] Xi Guo, Chenjing Ding, Haoxuan Dou, Xin Zhang, Weixuan Tang, and Wei Wu. Infinitydrive: Breaking time limits in driving world models, 2024. 2

[26] Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models, 2025. 1, 2

[27] 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. 13

[28] 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. 2, 3

[29] 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. 2

[30] 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*, 2022. 3

[31] 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*, 2022. 2

[32] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 2

[33] 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. 2

[34] 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. 1, 2

[35] 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. 2

[36] 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. 1, 2

[37] 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. 2

[38] Team HunyuanWorld. Hy-world 1.5: A systematic framework for interactive world modeling with real-time latency and geometric consistency. *arXiv preprint*, 2025. 2, 7

[39] 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. 2

[40] 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. 2

[41] Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. Fifo-diffusion: Generating infinite videos from text without training. *NeurIPS*, 2024. 2

[42] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[43] Akio Kodaira, Tingbo Hou, Ji Hou, Masayoshi Tomizuka, and Yue Zhao. Streamdit: Real-time streaming text-to-video generation, 2025. 2

[44] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2012. 12

[45] Kewei Lian, Shaofei Cai, Yilun Du, and Yitao Liang. Toward memory-aided world models: Benchmarking via spatial consistency, 2025. 1, 4, 5, 6, 12

[46] 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, 2025. 1, 2

[47] Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, Tong He, Jiangmiao Pang, Yu Qiao, and Kaipeng Zhang. Yume-1.5: A text-controlled interactive world generation model. *arXiv preprint arXiv:2512.22096*, 2025. 2

[48] 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*, 2025. 2

[49] 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. 2

[50] Y. Oshima, M. Suzuki, Y. Matsuo, and H. Furuta. Inference-time text-to-video alignment with diffusion latent beam search, 2025. arXiv preprint arXiv:2501.19252. 2

[51] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. 3, 5

[52] Ryan Po, Eric Ryan Chan, Changan Chen, and Gordon Wetzstein. Bagger: Backwards aggregation for mitigating drift in autoregressive video diffusion models, 2025. 2

[53] Ryan Po, Yotam Nitzan, Richard Zhang, Berlin Chen, Tri Dao, Eli Shechtman, Gordon Wetzstein, and Xun Huang. Long-context state-space video world models, 2025. 2

[54] 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. 2

[55] 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. 2

[56] 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. 3

[57] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 5

[58] David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling diffusion models, 2024. 2

[59] 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. 1, 2

[60] 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. 2, 7

[61] Dhruv Shah, Ajay Sridhar, Arjun Bhorkar, Noriaki Hirose, and Sergey Levine. GNM: A General Navigation Model to Drive Any Robot. In *arXiV*, 2022. 7

[62] 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. 2

[63] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 12

[64] Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. In *Advances in Neural Information Processing Systems*, pages 19313–19325. Curran Associates, Inc., 2021. 13

[65] Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion, 2025. 2

[66] 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)*, pages 63–70, 2024. 7

[67] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. 1, 3

[68] 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. 2, 5, 12

[69] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024. 1, 2

[70] 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. 2

[71] 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. 1, 2

[72] 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. 2, 5

[73] World Labs. Generating worlds. https://www.worldlabs.ai/blog/generating-worlds, 2025. Product site. 1, 2

[74] World Labs. Rtfm: A real-time frame model. https://www.worldlabs.ai/blog/rtfm, 2025. Company blog post. 2

[75] 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. 2

[76] Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory, 2025. 2, 4, 6, 7, 13

[77] 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. 2

[78] Shengming Yin et al. Nuwa-xl: Diffusion over diffusion for extremely long video generation. *arXiv preprint arXiv:2303.12346*, 2023. 2

[79] 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. 2

[80] 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*, 2025. 2, 4, 6, 7

[81] 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, 2025. 2

[82] Lvmin Zhang and Maneesh Agrawala. Packing input frame contexts in next-frame prediction models for video generation. *Arxiv*, 2025. 2, 4, 6

[83] 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. 2, 5, 12

[84] 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. 1, 2

[85] 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. 1, 2

Appendix

A. The Use of Large Language Models

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

B. Impact Statement

WorldPack advances the development of reliable ”world simulators” by improving long-term spatial consistency and fidelity in video world models. It addresses the long-standing issue of landscapes or objects shifting or disappearing when returning to a previously seen location, achieved through its unique trajectory packing and memory retrieval mechanisms. This allows for the creation of highly consistent simulations that more closely mimic the real world.

This technology is expected to be highly effective in fields requiring extreme safety, such as autonomous driving and robotics. Notably, it achieves high computational efficiency. This optimized approach provides a robust foundation for agents to learn complex navigation tasks in environments that accurately reflect physical reality.

At the same time, high-fidelity video generation carries inherent ethical risks, such as the potential to spread misinformation through deepfakes or to create biased simulations based on training data. While these risks must be carefully managed, the progress in physical consistency is a vital step toward AI truly understanding our world. It is hoped that this technology will continue to develop in a positive direction, leading to safer, more sophisticated simulations.

C. Evaluation Metrics

To rigorously assess the semantic consistency of the world model outputs, we employ Learned Perceptual Image Patch Similarity (LPIPS) [83] and DreamSim [15]. These metrics evaluate perceptual similarity based on deep features extracted from neural networks. Specifically, LPIPS utilizes image classification models (e.g., AlexNet [44] or VGG [63]) as its backbone to capture human perception of structural differences. Following Lian et al. [45], 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) [68]. FVD uses I3D [8] as its backbone to compare the feature distributions of real and generated videos, with lower scores indicating higher visual quality.

D. Further Ablation Study

Encoding Spatial Information Helps World Modeling. We investigate the impact of encoding spatial information on world modeling. Following Sitzmann et al. [64], Xiao et al. [76], 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 [18, 27]. 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

E. 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.

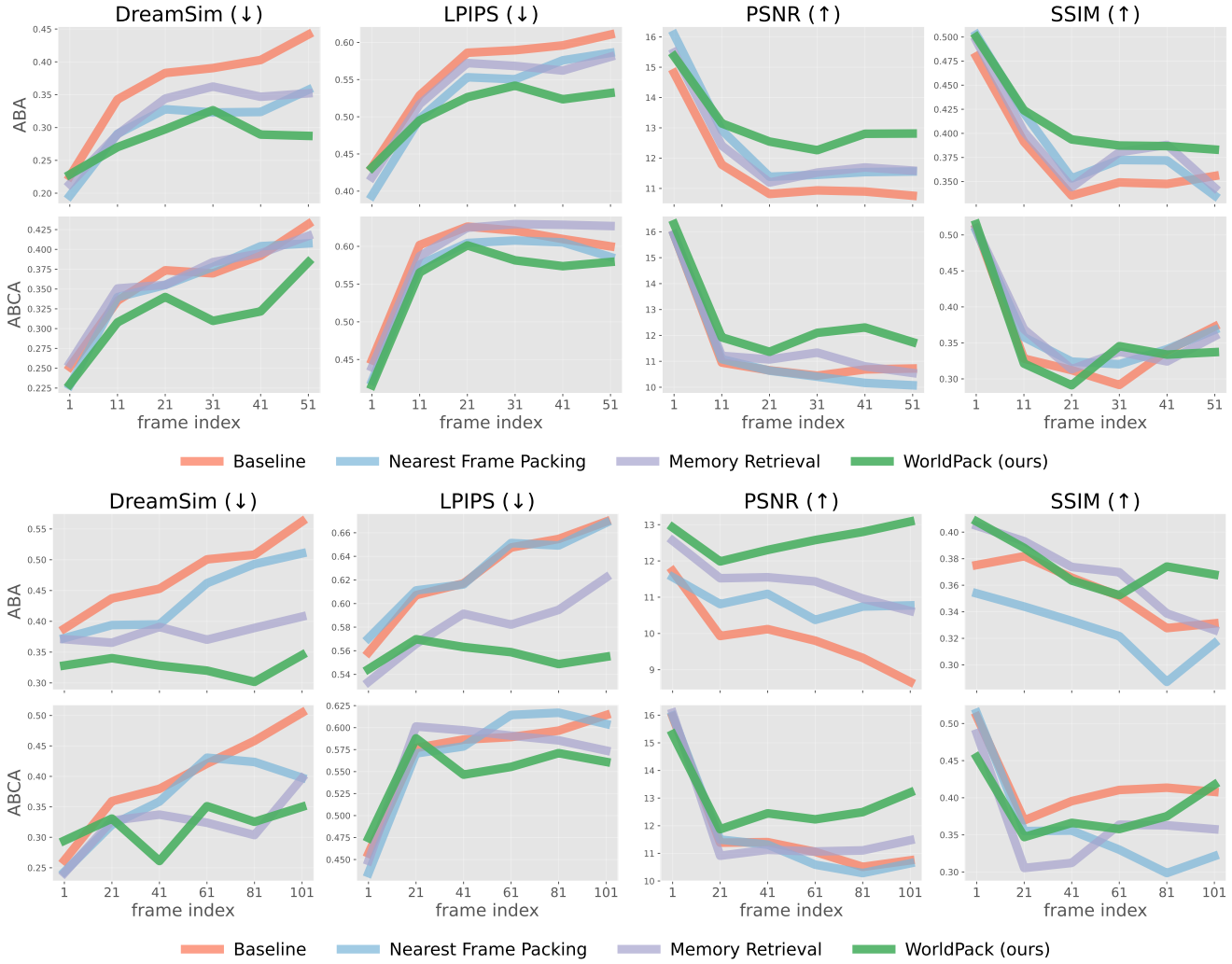

Figure 5. Prediction performance on the terminal frames of ABCA trajectories with different navigation ranges. **Top**: last 51 frames in ABA-5 and ABCA-5. **Bottom**: last 101 frames in ABA-15 and ABCA-15. WorldPack not only accesses task-relevant information based on 3D spatial cues but also retains 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.

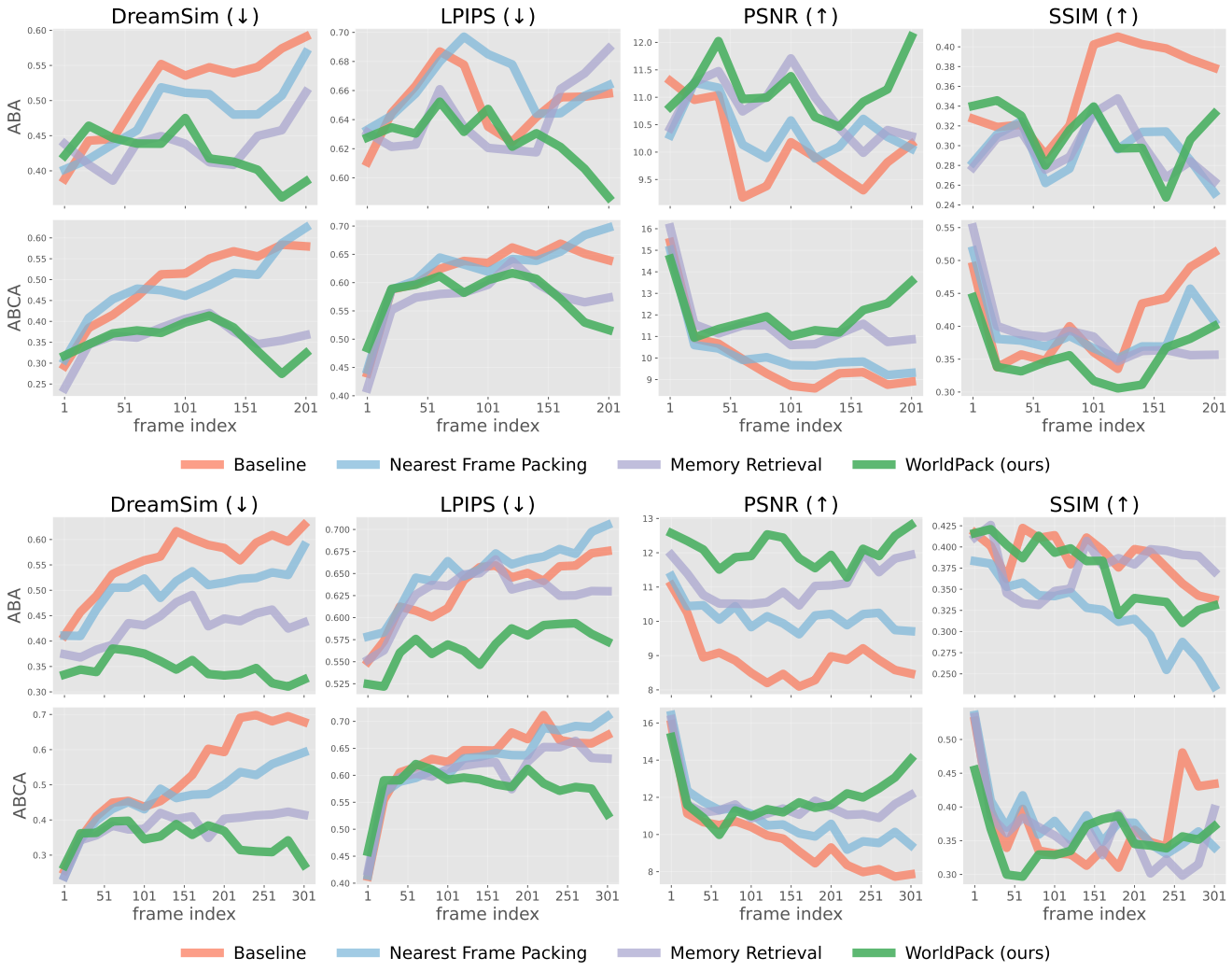

Figure 6. Prediction performance on the terminal frames of ABCA trajectories with different navigation ranges. **Top**: last 201 frames in ABA-30 and ABCA-30. **Bottom**: last 301 frames in ABA-50 and ABCA-50. WorldPack not only accesses task-relevant information based on 3D spatial cues but also retains 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.