WorldScape: An Efficient Real-Time Interactive World Model with Universal Control

WorldScape Team

Abstract. Interactive world models enable training AI agents within an unlimited curriculum of rich simulation environments. We present WorldScape, an autoregressive video diffusion model capable of real-time, streaming prediction of how visual observations evolve given navigation or manipulation actions. The generated observations implicitly construct a world that exhibits physical consistency and memory. WorldScape is powered by four key components: (1) For **interactivity**, we design a unified interaction-aware conditioning scheme that enables visual control over both navigation and manipulation. (2) For **3D spatial consistency**, we introduce spatial-aware consistency training incorporating 3D Gaussian Splatting–based supervision to inject geometric priors into generation. (3) For **real-time capability**, we adopt an asymmetric distillation strategy that converts the bidirectional diffusion backbone into a fast autoregressive model, achieving a generation speed of 16 FPS on a single H100 GPU. (4) For **memory**, we develop a geometry-aware KV cache optimization with hierarchical memory management to maintain long-range spatial coherence in streaming generation. Experiments demonstrate that WorldScape achieves balanced state-of-the-art performance across visual quality, interactivity, memory, and real-time capability, maintaining superior overall competence compared to existing models. Codes and demos can be found at: `https://worldscape.io`.

Date: Apr 03, 2026
Webpage: https://worldscape.io/

1 Introduction

Recent advances in world models, such as Cosmos [1] and V-JEPA [2], have demonstrated remarkable predictive capabilities through video-based future generation [3]. However, these models operate in an open-loop setting, producing single-pass forecasts without incorporating action feedback and thus severely limiting their applicability [4, 5]. We aim to develop a more general interactive world model that performs closed-loop, streaming prediction: continuously taking control actions as input and generating the evolution of both the environment and the agent's influence upon it [6, 7]. This allows the world models to function directly as a world simulator, enabling the training of robots and virtual agents entirely within boundless simulated environments, without the need for costly physical setups [8, 9]. To fulfill this vision, an interactive world model must exhibit four essential properties [10, 11]: interactivity, 3D spatial consistency, real-time capability, and memory.

1) Interactivity. We expect a general interactive world model is capable of understanding and responding to general forms of action inputs—ranging from spatial navigation to object manipulation—thus enabling bidirectional communication between agent and environment. This requires the generation of causally consistent environment responses conditioned on action sequences, reflected directly in the predicted video frames. However, most existing models remain restricted to language-based interaction [12, 13], while a few support only narrow types of non-linguistic control, such as spatial movement [14, 15] or object manipulation [16]. A unified representation and training framework for universal interactivity is still absent.

2) 3D spatial consistency. Each predicted frame should conform to consistent 3D geometry and maintain structural plausibility [17, 18]. The generated content is expected to respect spatial continuity and physical constraints, ensuring that objects, depth, and camera perspectives align within a coherent three-dimensional scene. Existing models [19, 20] often exhibit spatial collapse, with abrupt and implausible geometry changes between adjacent frames, as well as spatial drift due to the absence of explicit 3D priors. These issues typically arise from training objectives that rely solely on

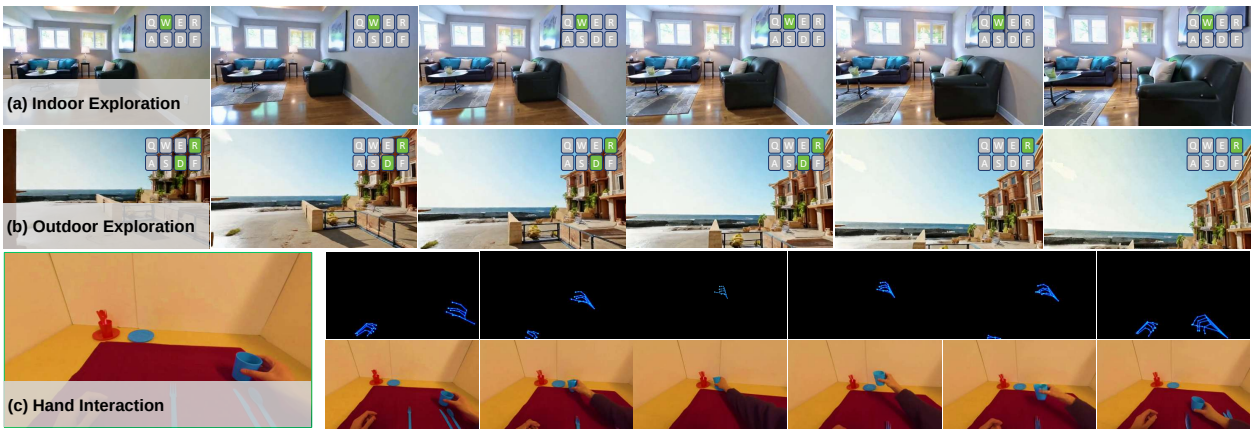

Figure 1: Introduction of our interactive world model. We use WASD to represent movement forward, left, backward, and right, and QERF to represent camera rotation to the left, right, up, and down. Manipulation interaction can be realized by specifying the hand poses.

pixel-level reconstruction.

3) Real-time capability. Beyond generation speed, real-time capability demands temporal alignment between action and response, ensuring that agents and humans perceive the passage of time in synchrony with the physical world [10, 14, 21]. Such temporal coherence allows a world model not only to produce visually plausible scenes but also to replicate realistic dynamics. For online decision-making, this enables latency-free action planning based on the most recent state; for immersive human–AI interaction, it sustains the sense of presence and naturalness. Yet, state-of-the-art models [4, 5, 22] with high visual fidelity typically require minutes to generate a single video, far from real-time usage. Diffusion-based models, though visually superior, depend on multi-step denoising, making the trade-off between quality and latency a persistent challenge [23].

4) Memory. In interactive world modeling, memory ensures that content generated over time remains part of a single, continuous world [24, 25]. When revisiting previously encountered regions, the model should recall and reproduce their corresponding scene states, maintaining temporal and semantic coherence across interactions. Effective memory are crucial for training agents that act and learn through continuous, open-ended experiences. However, how to efficiently propagate geometric knowledge in streaming generation remains underexplored.

To address these challenges, we propose **WorldScape**, an autoregressive diffusion-based world modeling framework that achieves interactive, spatially consistent, real-time, and memory-aware video generation, as shown in Figure 1. We start from a full-sequence diffusion foundation model and **first** introduce a unified interaction-aware conditioning scheme that enables visual generation control for both spatial navigation and hand manipulation, achieving universal interactive capability. To enhance 3D spatial consistency in video generation, we **then** propose a spatial-aware consistency training framework that jointly optimizes flow matching with depth and rendering supervision derived from 3D Gaussian Splatting, effectively integrating geometric priors into the generative process. **Furthermore**, to enable real-time autoregressive generation, we design a distillation architecture that converts the trained interactive and 3D-consistent full-sequence model into a fast causal diffusion model, using sliding-window self forcing to address both latency and error accumulation. **Finally**, to preserve long-range spatial memory during streaming interaction, we develop a memory-aware KV cache optimization framework with geometry-guided cache selection and hierarchical memory management, maintaining 3D coherence under bounded memory. **Taken together, our contributions lie in building an interactive world model that simultaneously achieves (1) unified navigation–manipulation control, (2) 3D spatial consistency, (3) real-time interaction of 16 FPS on a single H100 GPU, and (4) memory-preserving streaming generation.**

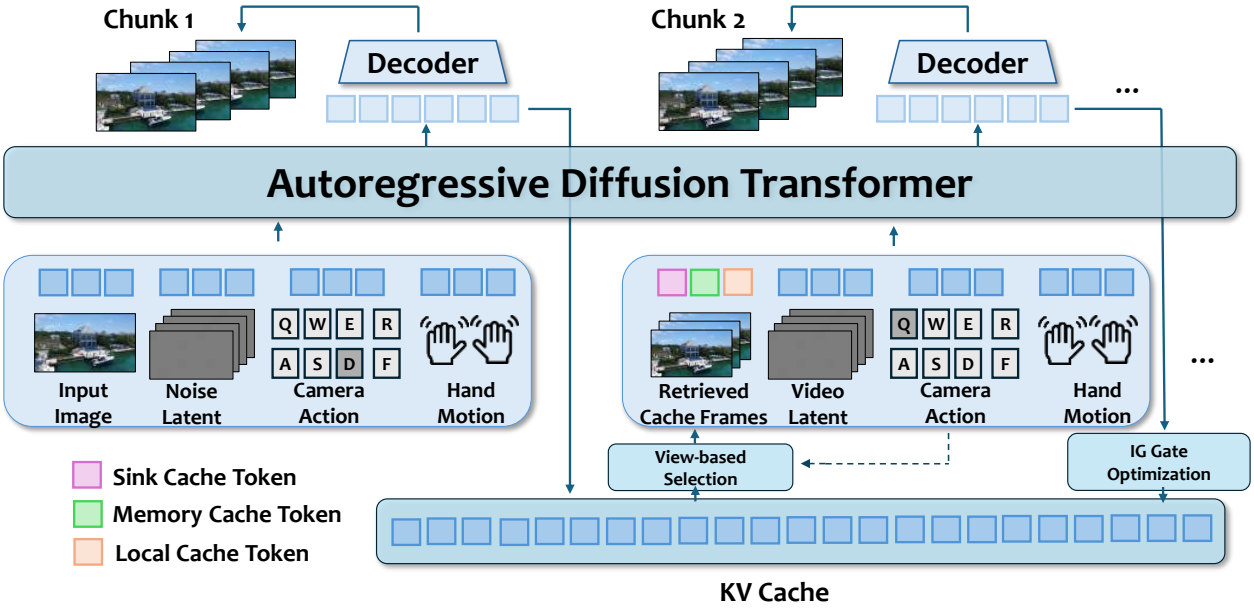

Figure 2: Model overview. WorldScape adopts a chunk-level autoregressive video generation architecture, where the denoising process of each chunk is conditioned on the camera pose inputs, hand motion inputs, and the KV cache of past denoising results.

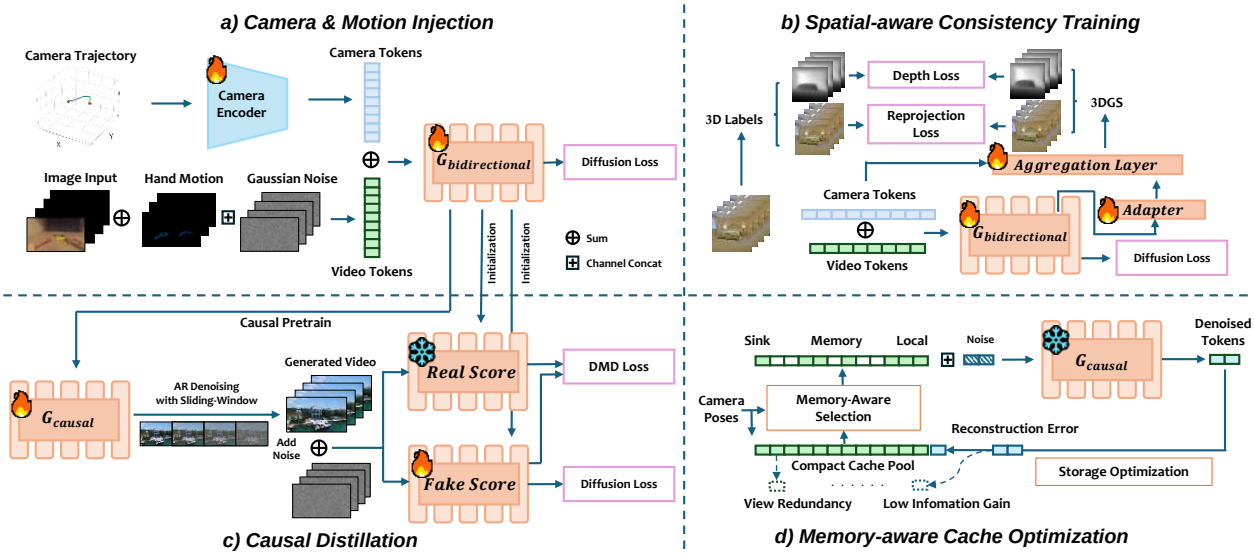

Figure 3: Framework of WorldScape. **a)** Given camera trajectory and pose video of hand motion, WorldScape trains a bidirectional teacher model backbone along with a camera encoder to enable controllable video generation. **b)** In Spatial-aware Consistency Training, WorldScape trains a feature aggregator together with an adaptor that projects DiT latents into the input space of the aggregator. The aggregator predicts depth and 3D information to compute two consistency losses, thereby training the model to maintain 3D consistency during video generation. **c)** WorldScape employs a Self Forcing-style distillation to transfer the controllability and spatial consistency of the bidirectional teacher DiT into an autoregressive few-step student model. **d)** During inference, WorldScape maintains the KV cache based on camera poses and information gain, enabling autoregressive generation with long-horizon memory.

2 WorldScape

2.1 Controllable Video Generation

Camera Control Injection. To incorporate camera trajectories as control conditions during video generation, following CameraCtrl [26], we represent camera trajectories using Plücker embeddings [27]. For the i-th frame in a video, its

Plücker embedding can be expressed as $\mathbf{P}_i \in \mathbb{R}^{6\times h\times w}$, where h and w are the height and width of the frame. As presented in Figure 3a, the camera encoder employs a lightweight adapter that takes per-pixel Plücker coordinates as input and uses pixel-unshuffle followed by convolutional residual blocks to produce spatially-aligned control features that are added to the DiT hidden states directly.

Hand Motion Control Injection. To further enhance controllability, we inject hand motion control capabilities into the model. We convert hand motions into a pose video and concatenate it after the first-frame image along the frame dimension to form a complete control video. The control video is then concatenated with the noisy video latent along the channel dimension and fed into the DiT model for denoising, enabling hand motion–controlled video generation.

2.2 Spatial-Aware Consistency Training

Multi-task Framework. Motivated by the potential of 3D geometric priors to enhance spatial stability [17], we adopt a multi-task objective that jointly optimizes flow matching and 3D signals from 3D Gaussian Splatting [28]. This approach organically infuses spatial structural cues into the generative process: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{fm}} + \alpha\mathcal{L}_{\text{depth}} + \beta\mathcal{L}_{\text{render}}$. As shown in Figure 3b, the denoised latent derived from the DiT's predicted velocity field is projected via a lightweight convolutional adapter into a ViT-based 3DGS aggregator [29]. This reconstructs the scene representation, rendering depth and RGB images for supervision.

Objective Formulation. The primary loss $\mathcal{L}_{\text{fm}}$ ensures generative fidelity by minimizing the error between predicted and target velocity fields:

$$\mathcal{L}_{\text{fm}} = \mathbb{E}_{t,x_1,x_t}\left[\|\hat{v}_\theta(x_t,t) - v_t\|_2^2\right].$$

To explicitly enforce 3D consistency, we incorporate $\mathcal{L}_{\text{depth}}$ and $\mathcal{L}_{\text{render}}$, which supervise geometry and appearance by comparing rendered depth maps $\hat{D}_i$ and RGB images $\hat{F}_i$ against ground truth (GT) across N viewpoints:

$$\mathcal{L}_{\text{depth}} = \frac{1}{N}\sum_{i=1}^{N}\|\hat{D}_i - D_i\|_2^2, \quad \mathcal{L}_{\text{render}} = \frac{1}{N}\sum_{i=1}^{N}\|\hat{F}_i - F_i\|_2^2.$$

2.3 Causal Distillation

To enable real-time streaming of video generation with camera control, in this section, we distill the slow teacher model into a fast causal video diffusion model using the pipeline shown in Figure 3c.

Causal Initialization. Following the initialization protocol of CausVid [30], we initialize the student model with the weights of the camera-guided teacher diffusion model and adapt it to a causal attention architecture and a few-step trajectory by regressing on ODE solution pairs sampled from the teacher.

Self Forcing–Style Distillation We start with the training paradigm proposed in Self Forcing [31] for model distillation. The key challenge is that strict chunk-level causality in Self-Forcing leads to motion pauses when camera control is introduced. To address this, similar to Rolling Forcing [32], we relax causality during denoising by allowing chunks at different noise levels to attend to each other via a sliding window, while preserving causal consistency through a monotonic KV cache.

We first define a video latent divided into L chunks as $\{z_t^i\}_{i=1}^L$, where t is the timestep. At the w-th window iteration ($w = 1,\ldots,L+K-1$), the generator processes chunks indexed by $\mathcal{W}_w = \{\max(1, w-K+1),\ldots,\min(L,w)\}$. Within each window, chunks are assigned heterogeneous timesteps forming a gradient from clean to noisy: the j-th chunk in the window (relative index) is assigned timestep t_{K-j+1}, where $\{t_k\}_{k=1}^K$ denotes the denoising schedule in descending order. Intuitively, each window contains up to K consecutive chunks, where earlier chunks are closer to clean states while later ones remain noisier. For example, when $K = 3$ and $w = 4$, the window processes chunks $\{2,3,4\}$ with timesteps $\{t_3, t_2, t_1\}$, respectively. Formally, the attention context for chunk z^i at window w is: $C_i^w = \{z^j_{t(j,w)}\}_{j\in\mathcal{W}_w} \cup \{z_0^j\}_{j<\min(\mathcal{W}_w)}$, where $t(j,w) = t_{K-(j-\min(\mathcal{W}_w))}$ assigns progressively noisier timesteps within the window, and $\{z_0^j\}_{j<\min(\mathcal{W}_w)}$ denotes the cached clean key-value features from previously completed chunks. The generator G_θ produces all chunks within the window jointly: $\hat{z}_0^i = G_\theta(z^i_{t(i,w)} \mid C_i^w)$ for $i \in \mathcal{W}_w$. After each window iteration, each chunk advances one denoising step. Only the leftmost chunk of each window (i.e., the chunk closest to completion) updates the KV cache with its clean-state features upon reaching t_K. This design ensures that only

temporally finalized content contributes to future attention, preserving causal consistency despite overlapping denoising windows. This rolling pipeline enables mutual refinement across chunks at different denoising stages through attention, while maintaining causal consistency via the monotonically advancing cache.

After autoregressively generating all L chunks, we obtain the complete video $\hat{z}_0 = \{z_0^1, \ldots, z_0^L\}$. We apply the DMD objective [33] to minimize the reverse KL divergence: $\mathcal{L}_{\text{DMD}} = \mathbb{E}_t \left[D_{\text{KL}}(p_t^{\text{gen}} \| p_t^{\text{data}}) \right]$. The gradient becomes:

$$\nabla_\theta \mathcal{L}_{\text{DMD}} \approx -\mathbb{E}_{t,\hat{z}_0} \left[\begin{pmatrix} s_{\text{real}}(\Psi(\hat{z}_0, t), t, c_i, c_m) \\ - s_{\text{fake}}(\Psi(\hat{z}_0, t), t, c_i, c_m) \end{pmatrix} \cdot \frac{\partial \hat{z}_0}{\partial \theta} \right],$$

where c_i is the control video condition, c_m is the camera trajectory condition and $\Psi(\hat{z}_0, t)$ denotes the forward diffusion operator. The real score s_{real} is computed using the frozen bidirectional teacher, and the fake score uses a trainable critic. This distills the teacher's camera control capability into the causal student with minimal inference cost.

2.4 Memory-aware KV Cache Optimization

To prevent unbounded KV cache growth without sacrificing spatial memory, we propose Memory-aware Cache Optimization (MACO). By leveraging camera trajectory priors for long-term 3D consistency, MACO manages the cache via a three-level hierarchy: a permanent sink anchor, an optimized global memory pool, and a local sliding window. At step t, the attention context is concatenated as:

$$\mathbf{K}/\mathbf{V}_t = [\mathbf{K}/\mathbf{V}_{\text{sink}} \oplus \mathbf{K}/\mathbf{V}_{\text{mem}} \oplus \mathbf{K}/\mathbf{V}_{\text{local}}].$$

As shown in Figure 3a, this hierarchical strategy achieves sublinear memory complexity while maintaining high generative fidelity through three following mechanisms:

View-guided Cache Selection. Each KV cache entry is associated with a camera state $\mathbf{C} = (\mathbf{v}, \mathbf{t})$, where $\mathbf{v}$ denotes the normalized view direction and $\mathbf{t}$ represents the camera position. Given a query pose i and a candidate pose j in the memory pool $\mathcal{M}_{\text{kv}}$, we define a geometry-aware similarity score: $S(i, j) = \mathbf{v}_i \cdot \mathbf{v}_j - \alpha \cdot \|\mathbf{t}_i - \mathbf{t}_j\|_2$. The first term measures Field-of-View alignment via cosine similarity, while the second penalizes distant viewpoints. This formulation prioritizes proximal views with aligned orientations, which are more likely to contain relevant scene content for the current query.

MDL-based Inbound Deduplication. To reduce redundancy in continuous video trajectories, we introduce an information-theoretic gating mechanism inspired by the Minimum Description Length (MDL) principle. We aim to maintain a memory pool $\mathcal{M}_{\text{kv}}$ that minimizes global redundancy, formulated as the sum of pairwise mutual information:

$$\min_{\mathcal{M}_{\text{kv}}} \sum_{\mathbf{k}_i, \mathbf{k}_j \in \mathcal{M}_{\text{kv}}, i \neq j} I(\mathbf{k}_i; \mathbf{k}_j) \quad \text{s.t.} \quad |\mathcal{M}_{\text{kv}}| \leq N_{\text{budget}}.$$

Since global optimization is intractable in autoregressive generation, we employ a greedy approximation by evaluating the *surprise* of a candidate key $\mathbf{k}_{\text{old}}$ exiting the local window. We quantify this reconstructibility using the maximum cosine similarity against the current pool:

$$S_{\text{mdl}} = \max_{\mathbf{k}_i \in \mathcal{M}_{\text{kv}}} \frac{\mathbf{k}_{\text{old}} \cdot \mathbf{k}_i}{\|\mathbf{k}_{\text{old}}\|_2 \|\mathbf{k}_i\|_2}.$$

S_{mdl} serves as a proxy for the *negative reconstruction error*. Entries with high S_{mdl} are discarded as redundant, while those with low similarity ($S_{\text{mdl}} < \tau_{\text{mdl}}$) are preserved as they represent novel content or high-frequency details.

Intra-trajectory Global Pruning. When the memory pool exceeds the budget N_{budget}, we perform global pruning to enforce spatial diversity. Leveraging camera trajectory priors, we apply Farthest Point Sampling (FPS) to select a representative subset $P \subset \mathcal{M}_{\text{kv}}$. The distance metric $d(i, j)$ incorporates both translation and rotation dynamics: $d(i, j) = \|\mathbf{t}_i - \mathbf{t}_j\|_2 + \lambda \|\mathbf{v}_i - \mathbf{v}_j\|_2$, where λ is a balancing weight. This strategy retains geometrically distinct viewpoints critical for consistent 3D representation while effectively pruning redundancies.

2.5 Training Recipe

We first fine-tune a pretrained video generation model on datasets consisting of video–camera trajectory pairs, resulting in an image-to-video (I2V) base model with fundamental camera trajectory following capability. We then perform spatial-aware consistency training on this base model using extra datasets with manually annotated 3D labels, which improves the 3D spatial coherence of the generated videos. Building upon this enhanced model, we further apply LoRA-based fine-tuning on a dataset composed of quadruplets, including video, text, camera trajectory, and hand motion, to enable hand motion–conditioned video generation. In addition, we employ Self Forcing with sliding window for step distillation, resulting in a final few-step causal generator.

3 Experiments

Table 1: Quantitative comparison on spatial navigation. We conduct quantitative comparisons with multiple baselines from four aspects: visual quality, interactivity, memory consistency, and efficiency. Best results are **bold**, and second best are underlined.

Models	Visual Quality				Interactivity	Memory	Realtime	
	Imaging Quality ↑	Motion Smoothness ↑	Subject Consistency ↑	Background Consistency ↑	Trajectory Accuracy ↑	Memory Symmetry ↑	FPS ↑	Pixel Throughput ↑
Large Scale Models (≥ 5B)								
RealCam-I2V	0.598	0.987	0.847	0.915	0.580	0.802	0.97	0.445
AC3D	0.451	**0.992**	0.878	**0.923**	0.595	**0.909**	0.11	0.038
YUME 1.5	0.591	0.980	0.830	0.917	**0.714**	0.509	0.27	0.243
HY-World 1.5	**0.656**	**0.992**	**0.932**	**0.923**	0.699	0.841	**1.12**	**0.447**
Small Scale Models (< 5B)								
Caml2V	0.503	**0.989**	0.799	0.908	0.691	0.376	1.34	0.220
CameraCtrl	0.451	0.981	0.735	0.880	0.679	0.414	5.02	0.329
MotionCtrl	0.458	0.975	0.723	0.875	0.662	0.305	5.13	0.336
Astra	0.535	0.981	0.782	0.885	0.608	0.440	0.41	0.164
Matrix-Game 2.0	0.495	0.985	0.727	0.884	0.671	0.308	**8.09**	1.823
WorldScape	**0.685**	0.986	**0.891**	**0.923**	**0.717**	**0.686**	6.27	**2.504**

Table 2: Quantitative comparison on WorldScore benchmark. We conduct quantitative comparisons with multiple baselines on the WorldScore benchmark, including the top eight independently developed models listed on the leaderboard[1] ranked by WorldScore-Static score. Best results are **bold**, and second best are underlined.

Models	WorldScore-Static	Camera Control	Object Control	Content Alignment	3D Consistency	Photometric Consistency	Style Consistency	Subjective Quality
CogVideoX-I2V	62.15	38.27	40.07	36.73	86.21	88.12	83.22	62.44
LucidDreamer	70.40	88.93	41.18	**75.00**	**90.37**	90.20	48.10	58.99
FlashWorld	70.85	84.43	50.28	56.54	85.87	86.72	79.36	52.75
WonderWorld	72.69	**92.98**	51.76	71.25	86.87	85.56	70.57	49.81
Voyager	77.62	85.95	66.92	68.92	81.56	85.99	84.89	**71.09**
TeleWorld	78.23	76.58	74.44	73.20	87.35	88.82	85.59	61.66
FantasyWorld-1.0	80.45	81.45	**87.90**	66.94	84.62	**94.07**	86.69	61.46
WorldScape	**80.76**	74.90	86.93	71.49	87.65	89.51	**90.63**	64.20

3.1 Experimental Setup

Training Details. We train our model based on Wan 2.1 (1.3B), with a generation resolution of 832 × 480. To equip the model with basic camera-trajectory-following capability, we preprocess videos from RealEstate10K [34], obtaining 20K clips for initial training. For spatial-aware consistency and memory enhancement, we train the model on a hybrid dataset

[1]The same results can also be found at the online leaderboard: https://huggingface.co/spaces/Howieeeee/WorldScore_Leaderboard/

of approximately 8K samples, over 80% of which contain revisiting or looped trajectories. Specifically, half of the samples come from RE10K and are temporally reversed to synthesize closed-loop camera paths, while the other half is drawn from the Context-as-Memory dataset [35], which naturally includes revisiting sequences or is similarly processed via temporal reversal. Camera poses are annotated using the VIPE pipeline [36], and depth labels are automatically generated using WorldMirror [29].

To further enrich the training distribution and strengthen the model's world-modeling capability, we curate a high-quality and diverse UE5 dataset that spans a broad spectrum of scene types, including urban, rural, post-apocalyptic, and stylized fantasy environments. The collected camera trajectories cover the major motion categories such as pan, move, pull, push, and orbit, providing diverse viewpoint transitions. At the final stage, we use 70K UE5 clips combined with 20K RE10K samples to extend training and improve overall performance under complex scenes and diverse trajectory patterns.

For hand-motion–conditioned generation, we perform LoRA-based finetuning on 200K samples from the EgoDex [37] dataset. And for causal distillation, 30K samples are drawn from each training stage to supervise the autoregressive student model with a bidirectional teacher.

Metrics. We evaluate WorldScape from multiple aspects, including visual quality, trajectory following capability, memory consistency and hand motion control capability. For visual quality, we adopt VBench [38] to assess generated videos in terms of imaging quality, subject consistency, background consistency, and motion smoothness. To evaluate interactivity, we introduce Trajectory Accuracy, which measures how well a model follows given camera trajectories by comparing the motion directions of generated and target trajectories. For memory and spatial consistency, we propose Memory Symmetry, which evaluates a model's ability to preserve scene structure when revisiting previously observed regions along mirrored trajectories. The complete formulation of these two metrics are detailed in Appendix A.1.1. For hand motion control, we use FID-VID, FVD, and image-level FID, which together capture frame-level appearance quality, temporal coherence, and overall visual fidelity.

Evaluation Datasets. We evaluate WorldScape's navigation capability on four independent datasets. For the in-distribution setting, we randomly sample 200 test frames from RealEstate10K. For out-of-distribution (OOD) evaluation, we sample 200 instances each from Tartanground [39] and Tartanair [40]. For each initial frame, we generate four videos conditioned on synthesized camera trajectories. To assess overall world-model performance, we evaluate WorldScape on the WorldScore benchmark [3], which measures multi-perspective consistency, style preservation, controllability, and scene-level stability. For hand motion following, we select 100 samples from the EgoDex test set and evaluate both image- and video-level fidelity.

3.2 Quantitative Evaluations

We compare our method with RealCam-I2V [41], AC3D [42], YUME 1.5, HY-World 1.5, CamI2V [43], CameraCtrl, MotionCtrl [44], Astra and Matrix-Game 2.0. For hand motion control, we choose baselines including Wan 2.2, HunyuanVideo [45], Cosmos-Predict 2.5, MimicMotion [46] and MagicDance [47]. For WorldScore benchmark evaluation, we choose the top 8 independently developed models in the leaderboard in terms of WorldScore-Static score, including CogVideoX-I2V [19], LucidDreamer [48], FlashWorld [49], WonderWorld [50], Voyager [51], TeleWorld [52] and FantasyWorld [53].

Visual Quality. As reported in Table 1, despite its faster generation speed, WorldScape achieves the best performance among models with fewer than 5B parameters on imaging quality, subject consistency, and background consistency, while remaining highly competitive in terms of motion smoothness. Remarkably, its imaging quality and background consistency are on par with, or even surpass, those of all larger-scale baseline models, demonstrating state-of-the-art or highly competitive visual quality.

Interactivity. As shown in Table 1, WorldScape achieves the best trajectory accuracy among all models with fewer than 5B parameters, outperforming other small-scale baselines by a clear margin. Compared with CamI2V, CameraCtrl and MotionCtrl, WorldScape demonstrates significantly improved adherence to input camera trajectories, indicating more precise and stable camera control. Notably, its trajectory accuracy is also competitive with larger-scale models, closely approaching or surpassing several models with over 5B parameters. These results suggest that WorldScape effectively learns fine-grained camera motion control, enabling accurate following of complex and long-horizon trajectories, which is critical for interactive and controllable video generation scenarios.

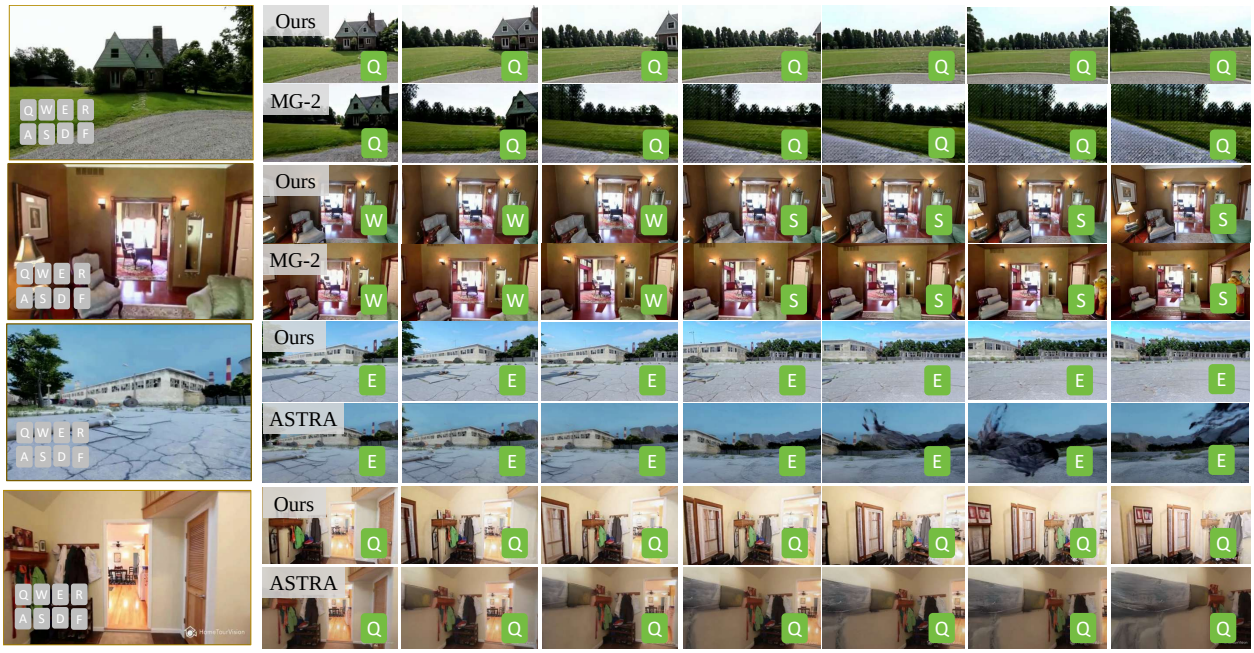

Figure 4: Qualitative comparisons on spatial navigation. Under diverse environments and lighting conditions, the proposed WorldScape maintains better spatial consistency and memory during interaction, enabling improved streaming visual generation.

Memory. As shown in Table 1, WorldScape achieves the highest memory symmetry score among all models with fewer than 5B parameters, significantly outperforming other baselines. This indicates that WorldScape is more effective at preserving scene structure and spatial coherence when revisiting previously observed regions over long temporal horizons. These results highlight its strong long-range memory consistency, which is crucial for interactive and exploration-oriented video generation.

Efficiency. We compare the real-time performance of WorldScape with baseline methods on a single A800 (80GB) GPU. As shown in Table 1, WorldScape achieves the second-highest FPS among all methods, trailing only Matrix-Game 2.0. This is primarily because Matrix-Game 2.0 generates videos at a significantly lower resolution than WorldScape. Consequently, in terms of pixel throughput, WorldScape outperforms all baseline models.

Hand Motion Control. As shown in Table 3, WorldScape consistently outperforms baseline methods on FVD and FID, obtaining the best results among both general video generation and pose-guided baselines. We note that FID-VID primarily measures frame-level appearance alignment and is less sensitive to long-range temporal coherence. In contrast, FVD better reflects the temporal consistency and motion realism of generated videos, which is crucial for hand motion control. The superior FVD score indicates that WorldScape generates more temporally coherent and stable hand motions, while the best image-level FID further demonstrates its strong visual fidelity. Overall, these results validate the effectiveness of WorldScape in producing high-quality and temporally consistent hand-controlled videos.

Broad Capability Assessment. As shown in Table 2, WorldScape achieves the highest WorldScore-Static score among all baselines. For the three consistency-oriented metrics, WorldScape ranks first in style consistency, second in 3D consistency, and third in photometric consistency—highlighting the effectiveness of our spatial-aware training and UE5-enhanced data in suppressing spatial drift and improving cross-view stability. In terms of prompt-following ability, WorldScape maintains strong performance on Object Control and Content Alignment, demonstrating reliable intention sensitivity and visual alignment. While the model shows less improvement on Camera Control and Subjective Quality, this is likely attributable to both its relatively small parameter count and the inherent difficulty of obtaining highly accurate camera annotations. Overall, WorldScape exhibits strong and well-balanced world-modeling capabilities, outperforming existing models on comprehensive metrics and demonstrating robust generalization across diverse environments.

Table 3: Quantitative comparison with on hand motion control. Best results are **bold**, and second best are underlined.

Model	Video		Image	FPS ↑
	FID-VID ↓	FVD ↓	FID ↓	
Large Scale Models (≥ 5B)				
Wan2.2-TI2V-5B	129.14	1421.37	207.77	0.23
HunyuanVideo-1.5	22.56	530.17	55.58	0.05
Cosmos-Predict 2.5	**14.47**	612.84	52.11	0.11
Small Scale Models (< 5B)				
MimicMotion	25.86	601.20	47.84	0.15
MagicDance	68.48	1552.93	93.26	0.11
WorldScape	28.88	**373.79**	**45.64**	**6.27**

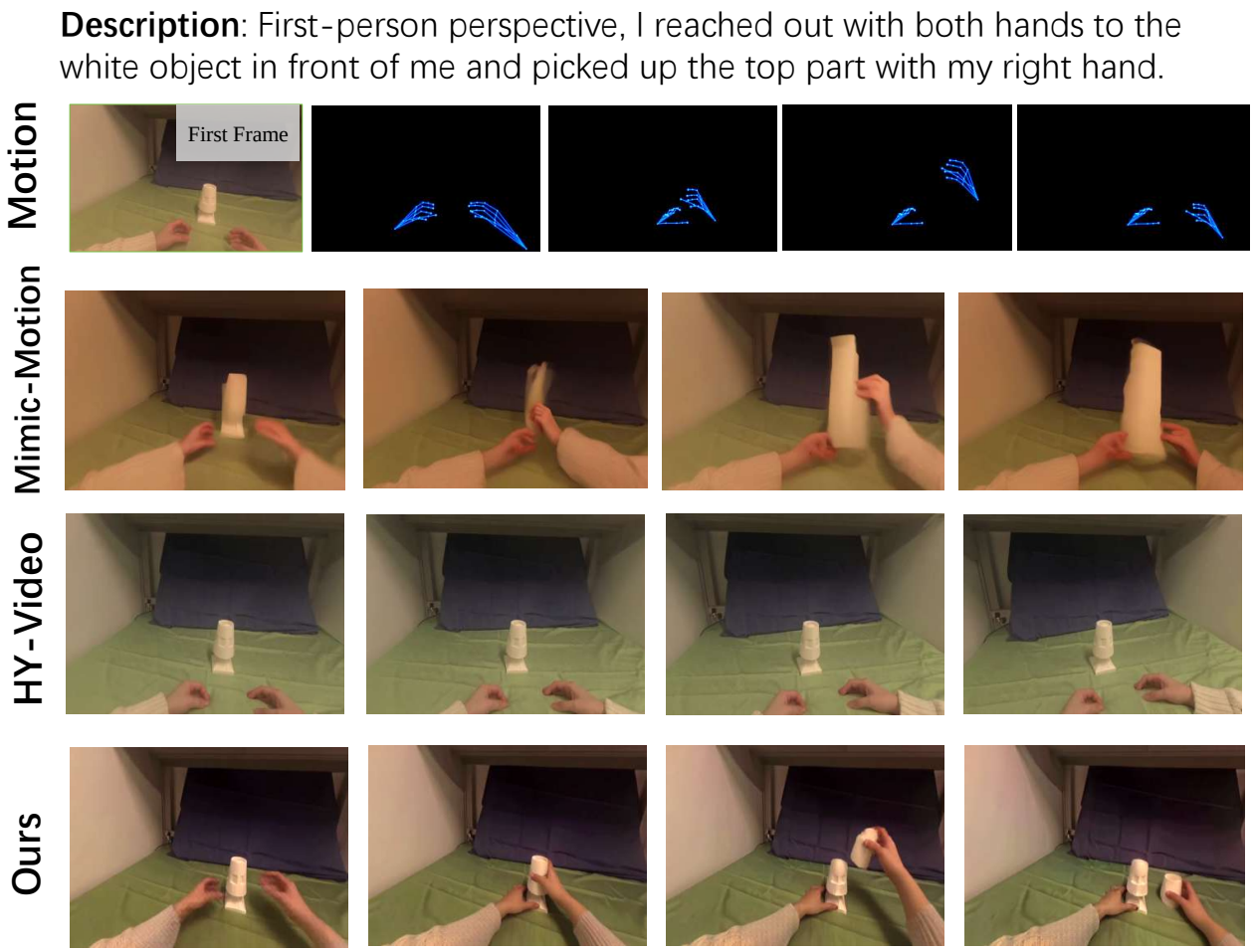

Figure 5: Qualitative comparisons on manipulation. Compared to existing models, the proposed WorldScape effectively performs pick-and-place tasks on objects based on the given hand motion.

3.3 Ablation Studies

Spatial-aware Consistency Training. We conduct an ablation study by comparing WorldScape with and without Spatial-aware Consistency Training, as shown in Table 4. The proposed training strategy consistently improves memory symmetry, multiview consistency, and subject consistency, while maintaining comparable background consistency. These results demonstrate the effectiveness of spatial-aware consistency training in enforcing stable 3D structures and long-term memory. Qualitative comparisons in Figure 6 further show that our method alleviates shape distortions and

Table 4: Ablation of 3D consistent training.

	Memory Symmetry	Multiview Consistency	Subject Consistency	Background Consistency
w/o	0.866	0.331	0.936	**0.929**
w	**0.919**	**0.406**	**0.955**	**0.929**

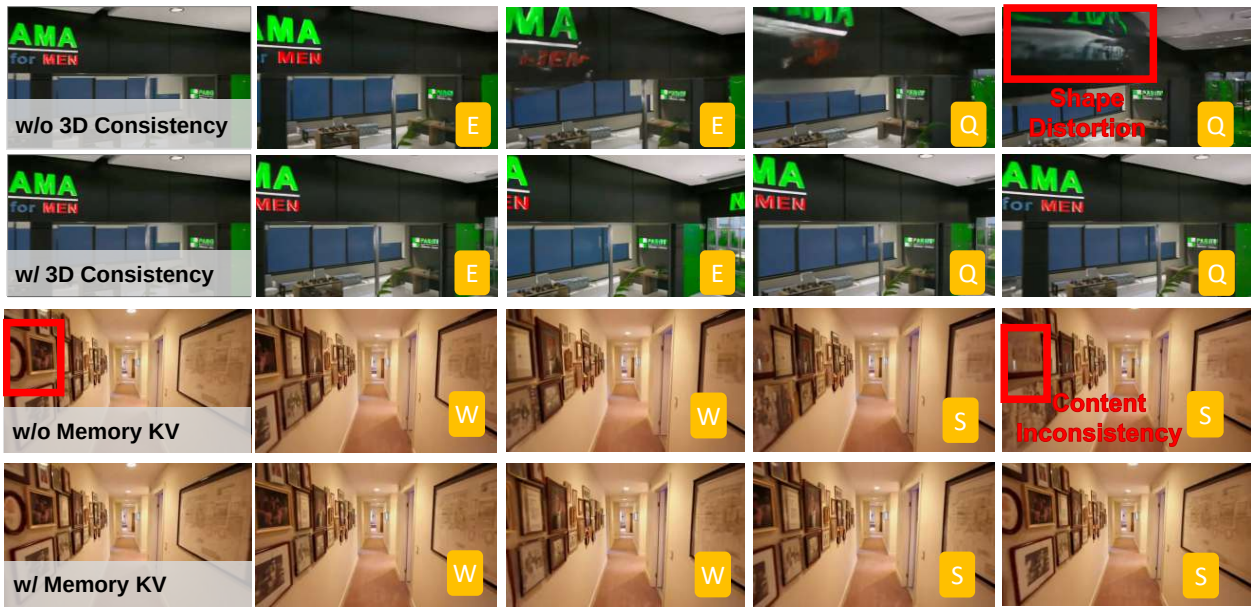

Figure 6: Comparison of results with and without spatial-aware consistency training and memory-aware KV cache. Without 3D consistency, videos exhibit shape distortion and spatial inconsistency under camera motion (red boxes line 1). With 3D consistency, spatial structures remain stable across viewpoints. Without memory, object deformation and structural inconsistency occur over time (red boxes in line 3), while memory-aware KV cache preserves long-term geometric and appearance consistency.

Table 5: Ablation of memory-aware KV cache optimization.

	Imaging Quality	Memory Symmetry	Subject Consistency	Background Consistency	FPS
w/o	0.720	0.536	0.890	0.893	**6.82**
w	**0.728**	**0.570**	**0.892**	**0.896**	6.27

preserves visual consistency under camera rotations.

Memory-aware KV Cache Optimization. We conduct an ablation study to evaluate the effect of the proposed Memory-aware KV Cache Optimization in long-term autoregressive generation. With the cache size kept constant, we compare our method with a naive rolling KV cache that only retains the most recent frames. Both approaches are evaluated on long sequences of 145 frames.

As shown in Table 5, memory-aware KV Cache optimization consistently improves imaging quality, memory symmetry, and both subject and background consistency, demonstrating its effectiveness in preserving long-term spatial information. While the optimization incurs a modest reduction in FPS due to additional memory-related computation, the performance remains comparable and the gain in long-term consistency significantly outweighs the computational overhead.

Causal Distillation. We ablate our causal distillation, as shown in Table 6. By distilling a bidirectional teacher into a 4-step autoregressive student, WorldScape achieves approximately 8.7× generation speedup. Meanwhile, the degradation in generation quality is minimal, demonstrating the effectiveness of autoregressive distillation.

Table 6: Ablation study of autoregressive distillation.

	Imaging Quality	Motion Smoothness	Subject Consistency	Background Consistency	FPS
w/o	0.621	**0.989**	0.886	**0.934**	0.72
w	**0.685**	0.986	**0.891**	0.923	**6.27**

3.4 Qualitative Results

Figures 4 and 5 present qualitative case studies of interactive video generation under spatial navigation and hand motion control, respectively, with comparisons to prior methods. Baseline models often suffer from spatial drift, structural distortion, or limited interaction fidelity, failing to maintain consistent scene layouts or complete object-centric manipulation tasks. In contrast, WorldScape produces temporally and spatially coherent generations, preserving stable geometry during long-horizon navigation and enabling task-consistent hand-object interactions.

4 Conclusion

In this paper, we construct WorldScape, an autoregressive diffusion–based world modeling framework that achieves interactive, real-time, and spatially consistent video generation. Through an universal interactive control action learning, enhanced self-forcing distillation framework, and 3D consistancy learning, WorldScape achieves state-of-the-art performance compared with existing world models. As for the future work, we plan to apply the world model into the downstream applications such as robotics, to further test WorldScape's performance.

A Appendix

A.1 Evaluation Details

A.1.1 Metrics.

Memory Symmetry. Memory Symmetry evaluates the visual consistency between the beginning and ending frames of a generated video, which is crucial for assessing temporal coherence in looping or reversible scenarios. Given a video with n frames $\{F_0, F_1, ..., F_{n-1}\}$, we compute the Mean Squared Error (MSE) between symmetric frame pairs (F_i, F_{n-1-i}) for $i \in [0, \lfloor n/2 \rfloor)$. Each MSE value is then mapped to a similarity score using the following transformation:

$$s_i = \exp\left(-k_1 \cdot \max(MSE_i - a, 0)^{k_2}\right),$$

where $k_1 = 0.0001$, $k_2 = 1.1$, and $a = 1000$ is an offset threshold below which perfect similarity is assigned. The final score is computed as a distance-weighted average, where frame pairs closer to the video boundaries receive higher weights:

$$\text{Memory Symmetry} = \frac{\sum_{i=0}^{m-1} s_i \cdot \exp(\alpha \cdot |i - \frac{n-1}{2}|)}{\sum_{i=0}^{m-1} \exp(\alpha \cdot |i - \frac{n-1}{2}|)},$$

where $m = \lfloor n/2 \rfloor$ and $\alpha = 0.1$. This weighting scheme emphasizes the first-last frame pair, which most directly reflects the "return-to-origin" property.

Trajectory Accuracy. Trajectory Accuracy measures the alignment between the generated camera trajectory and the target trajectory specified by the control signal. Given a generated trajectory and a target trajectory, both represented as sequences of 3×4 camera poses, we first compute the frame-wise derivatives $\Delta P_t = P_{t+1} - P_t$ for both trajectories. Each 12-dimensional derivative is converted to a 6-DoF vector $\mathbf{v} = [\mathbf{t}, \mathbf{r}] \in \mathbb{R}^6$, where $\mathbf{t}$ denotes translation and $\mathbf{r}$ represents Euler angles extracted from the rotation matrix. The 6-DoF cosine similarity at each frame is computed as:

$$\text{sim}t = \frac{1}{2}\left(\frac{|\mathbf{t}_t^{gen} \cdot \mathbf{t}_t^{tgt}|}{\|\mathbf{t}_t^{gen}\|\|\mathbf{t}_t^{tgt}\|} + \frac{|\mathbf{r}_t^{gen} \cdot \mathbf{r}_t^{tgt}|}{\|\mathbf{r}_t^{gen}\|\|\mathbf{r}_t^{tgt}\|}\right).$$

The raw accuracy is the average similarity across all frames. We then apply a threshold $\tau = 0.55$ to obtain the binary score:

$$\text{Trajectory Accuracy} = \mathbb{1}\left[\frac{1}{T-1}\sum_{t=1}^{T-1} \text{sim}t \geq \tau\right].$$

This metric evaluates whether the model correctly follows motion instructions by comparing instantaneous movement directions, independent of absolute trajectory scale.

A.1.2 Detailed Experimental Results

We report the average performance of WorldScape on both in-domain and OOD test sets in Table 1. Here we provide the detailed experimental results for each of the three test sets separately in the Appendix, as shown in Table 7, Table 8 and Table 9, respectively.

A.2 Visualization of Cache Optimizaiton Mechanism

Figure 7 visualizes the evolution of the proposed cache optimization strategy along a reversed orbit trajectory, using a local window size of 6 and a global memory budget of 4 with 2 selected. Each subplot corresponds to a generation step and shows how cache entries are distributed among the current query (yellow star), sink anchor (green), local window (blue), and global spatial memory (red star).

At early steps (e.g., Frame 7), the attention context mainly relies on the sink anchor and the local window, and only a minimal spatial memory is retained since nearby viewpoints are highly redundant. When the trajectory reverses

Table 7: Quantitative results on RealEstate10K dataset. We evaluate 200 initial frames generating 800 videos. Best results in each category are **bold**, and second best are underlined.

Models	Visual Quality				Interactivity	Memory
	Imaging Quality ↑	Motion Smoothness ↑	Subject Consistency ↑	Background Consistency ↑	Trajectory Accuracy ↑	Image Consistency ↑
Large Scale Models (≥ 5B)						
RealCam-I2V	0.639	0.989	0.862	0.920	0.531	0.814
AC3D	0.452	**0.992**	0.879	0.924	0.614	**0.912**
YUME 1.5	0.623	0.982	0.850	**0.926**	**0.762**	0.508
HY-World 1.5	**0.714**	**0.992**	**0.942**	0.922	0.709	0.790
Small Scale Models (< 5B)						
CamI2V	0.548	**0.990**	0.815	0.913	0.688	0.408
CameraCtrl	0.484	0.984	0.751	0.876	0.689	0.382
MotionCtrl	0.488	0.980	0.736	0.870	0.649	0.293
Astra	0.558	0.983	0.800	0.883	0.586	0.455
Matrix-Game 2.0	0.532	0.985	0.728	0.878	0.655	0.240
WorldScape	**0.689**	0.987	**0.896**	**0.931**	**0.792**	**0.672**

Table 8: Quantitative results on TartanGround dataset. We evaluate 100 initial frames generating 400 videos. Best results in each category are **bold**, and second best are underlined.

Models	Visual Quality				Interactivity	Memory
	Imaging Quality ↑	Motion Smoothness ↑	Subject Consistency ↑	Background Consistency ↑	Trajectory Accuracy ↑	Image Consistency ↑
Large Scale Models (≥ 5B)						
RealCam-I2V	0.613	0.985	0.863	0.918	0.605	0.771
AC3D	0.462	**0.992**	0.882	0.925	0.563	0.899
YUME 1.5	0.592	0.976	0.823	0.912	0.748	0.467
HY-World 1.5	**0.633**	**0.992**	**0.930**	**0.927**	**0.725**	**0.930**
Small Scale Models (< 5B)						
CamI2V	0.485	**0.987**	0.787	0.908	**0.754**	0.350
CameraCtrl	0.409	0.978	0.716	0.883	0.739	0.442
MotionCtrl	0.421	0.969	0.709	**0.881**	0.728	0.311
Astra	0.551	0.980	0.788	0.892	0.688	0.416
Matrix-Game 2.0	0.481	0.985	0.734	0.892	0.718	0.322
WorldScape	**0.717**	0.986	**0.893**	0.914	0.709	**0.659**

(Frame 17), the global memory begins to accumulate views that are spatially separated but temporally distant, allowing the current frame to retrieve relevant past observations via view-guided selection. In later steps (Frames 27 and 32), although the camera revisits previously seen regions, the composition of the global memory remains stable and bounded, retaining only representative viewpoints along the trajectory.

Figure 8 displays how cache pool changes during the whole autoregressive generation process on this trajectory. According to the color definitions in the legend, we can better understand when MDL actively admits or rejects incoming cache and how deduplication happens when the budget is exceeded. In this case, each column has at most 4 caches in pool and only two of them will be selected when generating.

For example, as the camera moves forward (Frame 12 in Figure 7), frames leaving the local window are selectively admitted into the global memory when they introduce geometrically distinct viewpoints, while redundant ones are

Table 9: Quantitative results on TartanAir dataset. We evaluate 100 initial frames generating 400 videos. Best results in each category are **bold**, and second best are underlined.

Models	Visual Quality				Interactivity	Memory
	Imaging Quality ↑	Motion Smoothness ↑	Subject Consistency ↑	Background Consistency ↑	Trajectory Accuracy ↑	Image Consistency ↑
Large Scale Models (≥ 5B)						
RealCam-I2V	0.502	0.986	0.802	0.903	**0.653**	0.810
AC3D	0.439	**0.992**	0.871	0.920	0.590	**0.913**
YUME 1.5	0.527	0.978	0.799	0.903	0.583	0.554
HY-World 1.5	**0.562**	0.991	**0.911**	**0.924**	0.650	0.854
Small Scale Models (< 5B)						
CamI2V	0.429	**0.989**	0.780	0.896	0.634	0.340
CameraCtrl	0.426	0.980	0.720	0.885	0.600	0.451
MotionCtrl	0.436	0.973	0.709	0.880	0.623	0.324
Astra	0.471	0.980	0.740	0.880	0.571	0.434
Matrix-Game 2.0	0.435	0.985	0.721	0.889	**0.656**	0.428
WorldScape	**0.646**	0.986	**0.878**	**0.917**	0.577	**0.738**

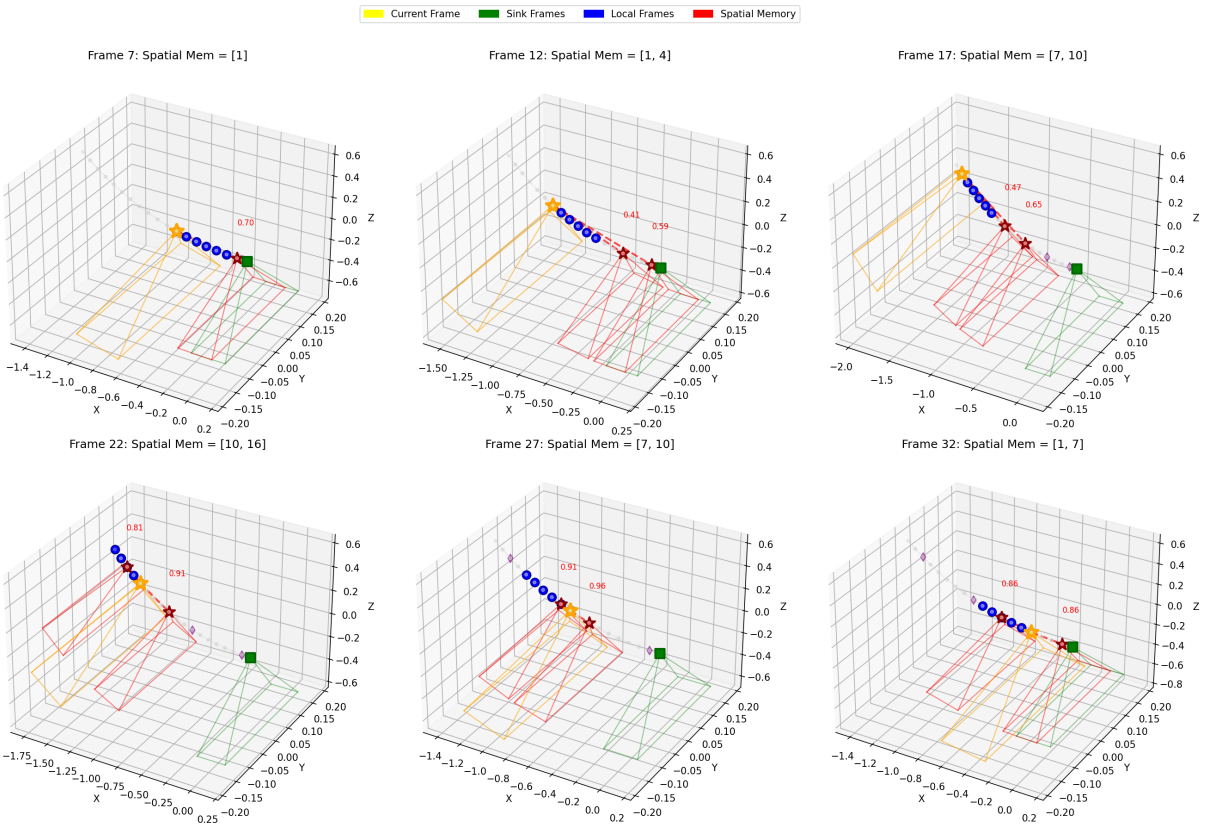

Figure 7: Case: a reversed orbit trajectory.

filtered by MDL-based deduplication. Once the memory budget is reached (Frame 22 in Figure 7), newly added entries trigger intra-trajectory global pruning, where geometrically redundant memories are replaced to preserve spatial diversity.

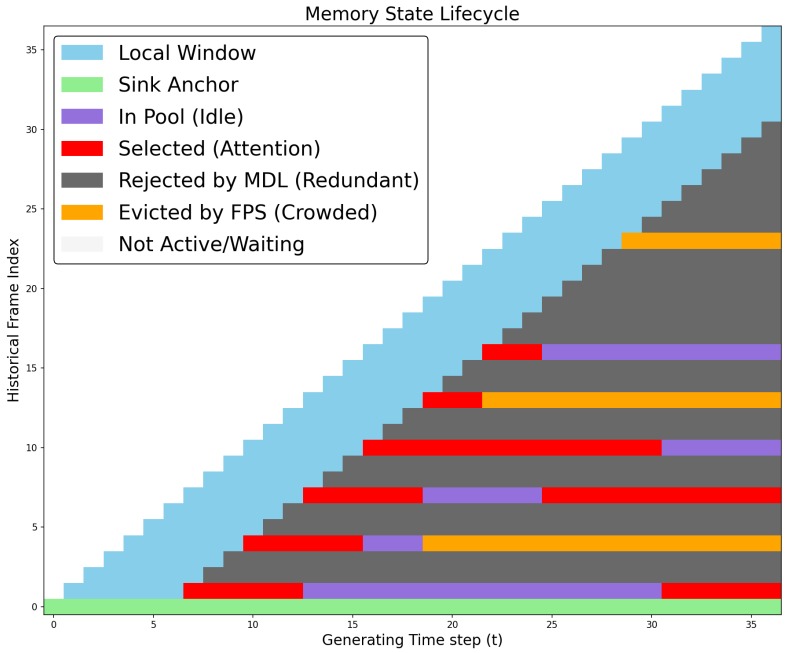

Figure 8: Cache statics during generation.

Overall, these figures demonstrate how the proposed strategy maintains long-term 3D consistency under a fixed KV cache budget by jointly leveraging geometric deduplication and trajectory-aware pruning.

A.3 Inference with Semantic-Aware Permutation

To further improve real-time inference, we integrate the Semantic-Aware Permutation (SAP) mechanism from SVG2 [54] into WorldScape, built upon the Wan2.1-Fun-1.3B-Control backbone. SAP exploits the inherent sparsity of attention by clustering tokens based on semantic similarity in the latent space and permuting them into contiguous layouts, enabling efficient block-wise sparse attention without padding overhead while preserving output equivalence. With SAP enabled, the inference speed improves from 6.27 FPS to 6.64 FPS under the same 81-frame generation setting, demonstrating additional gains in streaming efficiency without extra training.

A.4 More Qualitative Results

Figure 9 presents a series of case studies illustrating the model's performance in spatial navigation scenarios. The generated videos demonstrate the model's ability to maintain spatial coherence over extended trajectories while responding to users' navigation commands.

Figure 10 presents case studies of hand manipulation tasks. The generated sequences exhibit consistent hand-object contact and plausible object dynamics across time, indicating the model's ability to capture fine-grained interaction details in manipulation scenarios.

References

[1] NVIDIA, :, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura

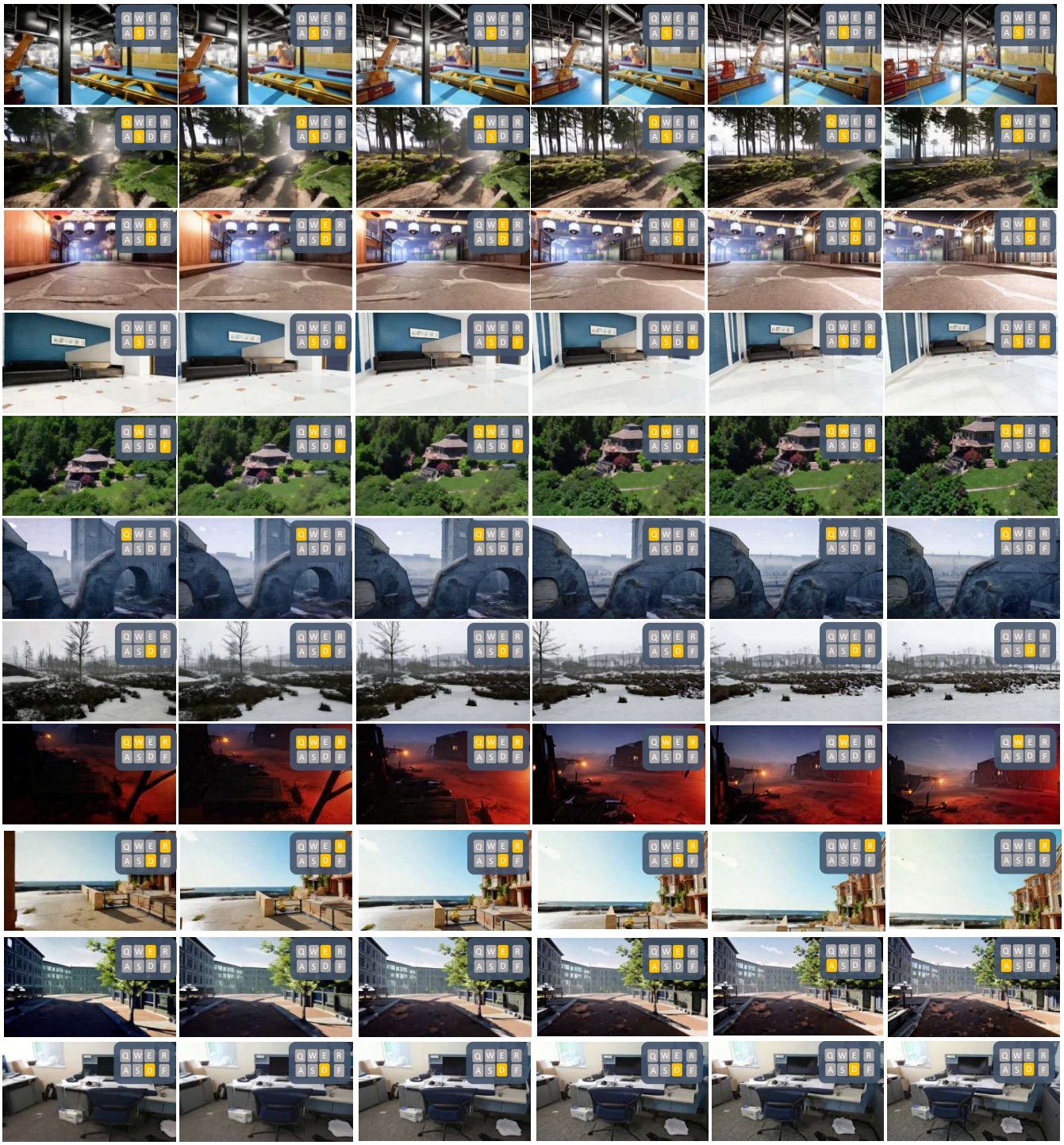

Figure 9: More case studies on spatial navigation.

Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei,

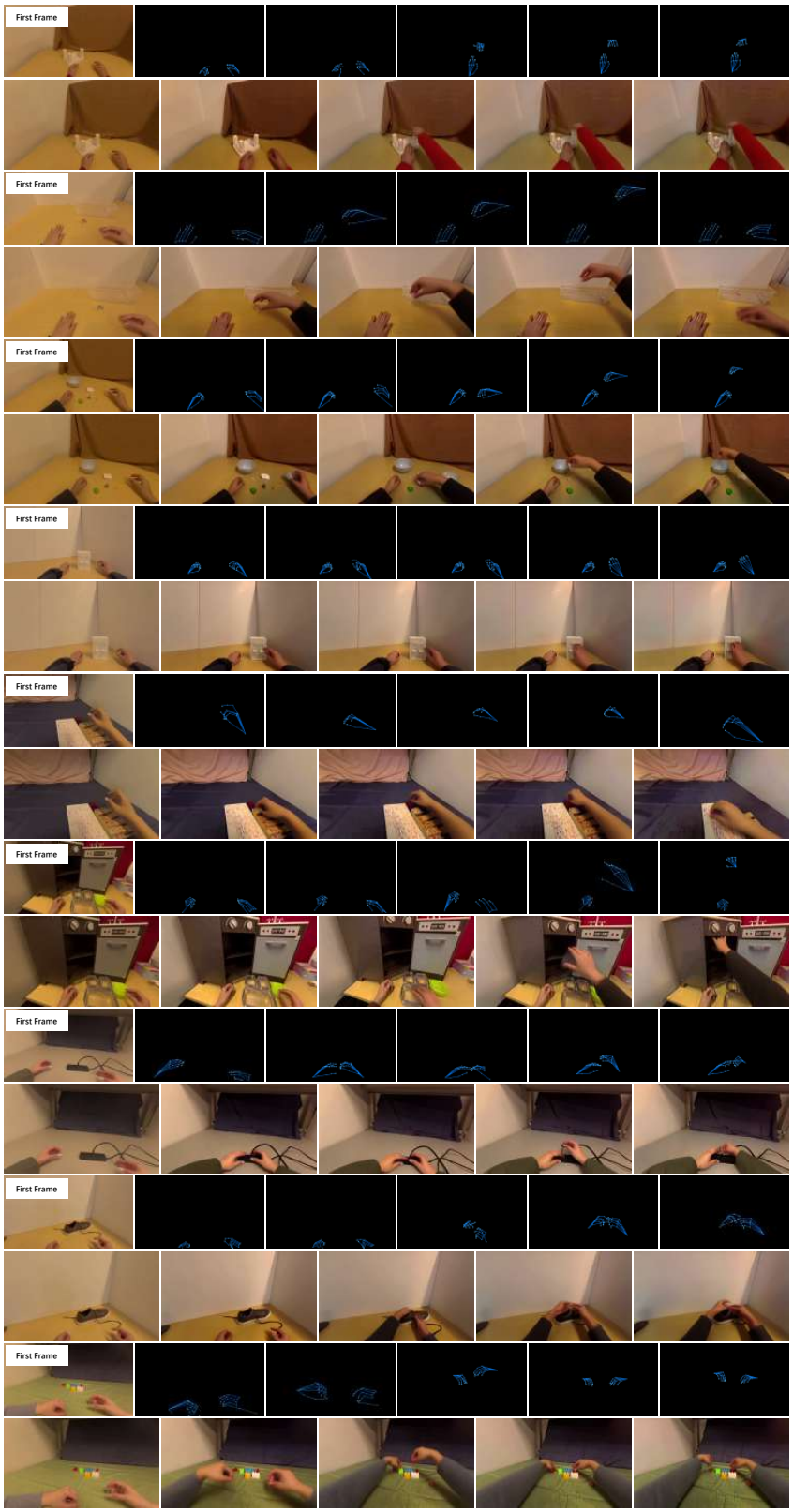

Figure 10: More case studies on hand motion control.

Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai, 2025. URL https://arxiv.org/abs/2501.03575.

[2] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985.

[3] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation, 2025. URL https://arxiv.org/abs/2504.00983.

[4] OpenAI. Sora: Video generation models as world simulators. https://openai.com/index/video-generation-models-as-world-simulators/, 2024. Accessed: 2026-01-27.

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

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

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

[8] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, pages 1–7, 2025.

[9] Gemini Robotics Team, Krzysztof Choromanski, Coline Devin, Yilun Du, Debidatta Dwibedi, Ruiqi Gao, Abhishek Jindal, Thomas Kipf, Sean Kirmani, Isabel Leal, Fangchen Liu, Anirudha Majumdar, Andrew Marmon, Carolina Parada, Yulia Rubanova, Dhruv Shah, Vikas Sindhwani, Jie Tan, Fei Xia, Ted Xiao, Sherry Yang, Wenhao Yu, and Allan Zhou. Evaluating gemini robotics policies in a veo world simulator, 2026. URL https://arxiv.org/abs/2512.10675.

[10] 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 Gharamani, Raia Hadsell, Aäron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 3: A new frontier for world models. 2025.

[11] Jiwen Yu, Yiran Qin, Haoxuan Che, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Hao Chen, and Xihui Liu. A survey of interactive generative video, 2025. URL https://arxiv.org/abs/2504.21853.

[12] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14749–14759, 2024.

[13] Baining Zhao, Rongze Tang, Mingyuan Jia, Ziyou Wang, Fanhang Man, Xin Zhang, Yu Shang, Weichen Zhang, Wei Wu, Chen Gao, et al. Airscape: An aerial generative world model with motion controllability. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12519–12528, 2025.

[14] Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Size Wu, Wei Li, Xuchen Song, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-game 2.0: An open-source real-time and streaming interactive world model, 2025. URL https://arxiv.org/abs/2508.13009.

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

[16] Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: A fine-grained world model for robot manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9834–9844, 2025.

[17] Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling, 2025. URL https://arxiv.org/abs/2507.07982.

[18] Anthony Chen, Wenzhao Zheng, Yida Wang, Xueyang Zhang, Kun Zhan, Peng Jia, Kurt Keutzer, and Shanghang Zhang. Geodrive: 3d geometry-informed driving world model with precise action control, 2025. URL https://arxiv.org/abs/2505.22421.

[19] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2025. URL https://arxiv.org/abs/2408.06072.

[20] 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, 2025. URL https://arxiv.org/abs/2504.08388.

[21] Xinle Cheng, Tianyu He, Jiayi Xu, Junliang Guo, Di He, and Jiang Bian. Playing with transformer at 30+ fps via next-frame diffusion, 2025. URL https://arxiv.org/abs/2506.01380.

[22] Google DeepMind and Google Research. Veo 3. https://aistudio.google.com/models/veo-3, 2026. Accessed: 2026-01-27.

[23] 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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22963–22974, 2025.

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

[25] Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory, 2025. URL https://arxiv.org/abs/2506.05284.

[26] 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.

[27] Vincent Sitzmann, Semon Rezchikov, William T. Freeman, Joshua B. Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. In *Proc. NeurIPS*, 2021.

[28] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering, 2023. URL https://arxiv.org/abs/2308.04079.

[29] Yifan Liu, Zhiyuan Min, Zhenwei Wang, Junta Wu, Tengfei Wang, Yixuan Yuan, Yawei Luo, and Chunchao Guo. Worldmirror: Universal 3d world reconstruction with any-prior prompting, 2025. URL https://arxiv.org/abs/2510.10726.

[30] 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. In *CVPR*, 2025.

[31] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion, 2025. URL https://arxiv.org/abs/2506.08009.

[32] Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.

[33] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, 2024.

[34] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 37, 2018. URL https://arxiv.org/abs/1805.09817.

[35] 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, 2025. URL https://arxiv.org/abs/2506.03141.

[36] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. Vipe: Video pose engine for 3d geometric perception. In *NVIDIA Research Whitepapers arXiv:2508.10934*, 2025.

[37] Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video, 2025. URL https://arxiv.org/abs/2505.11709.

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

[39] Manthan Patel, Fan Yang, Yuheng Qiu, Cesar Cadena, Sebastian Scherer, Marco Hutter, and Wenshan Wang. Tartanground: A large-scale dataset for ground robot perception and navigation, 2025. URL https://arxiv.org/abs/2505.10696.

[40] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam, 2020. URL https://arxiv.org/abs/2003.14338.

[41] Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, and Xi Li. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. *arXiv preprint arXiv:2502.10059*, 2025.

[42] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. *arXiv preprint arXiv:2411.18673*, 2024.

[43] Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024.

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

[45] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025. URL https://arxiv.org/abs/2412.03603.

[46] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance, 2025. URL https://arxiv.org/abs/2406.19680.

[47] Di Chang, Yichun Shi, Quankai Gao, Jessica Fu, Hongyi Xu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. *arXiv preprint arXiv:2311.12052*, 2023.

[48] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes, 2023. URL https://arxiv.org/abs/2311.13384.

[49] Xinyang Li, Tengfei Wang, Zixiao Gu, Shengchuan Zhang, Chunchao Guo, and Liujuan Cao. Flashworld: High-quality 3d scene generation within seconds, 2025. URL https://arxiv.org/abs/2510.13678.

[50] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T. Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image, 2025. URL https://arxiv.org/abs/2406.09394.

[51] Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson W. H. Lau, Wangmeng Zuo, and Chunchao Guo. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation, 2025. URL https://arxiv.org/abs/2506.04225.

[52] Xunzhi Xiang, Yabo Chen, Guiyu Zhang, Zhongyu Wang, Zhe Gao, Quanming Xiang, Gonghu Shang, Junqi Liu, Haibin Huang, Yang Gao, Chi Zhang, Qi Fan, and Xuelong Li. Macro-from-micro planning for high-quality and parallelized autoregressive long video generation, 2025. URL https://arxiv.org/abs/2508.03334.

[53] Yixiang Dai, Fan Jiang, Chiyu Wang, Mu Xu, and Yonggang Qi. Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction, 2025. URL https://arxiv.org/abs/2509.21657.

[54] Shuo Yang, Haocheng Xi, Yilong Zhao, Muyang Li, Jintao Zhang, Han Cai, Yujun Lin, Xiuyu Li, Chenfeng Xu, Kelly Peng, et al. Sparse videogen2: Accelerate video generation with sparse attention via semantic-aware permutation. *arXiv preprint arXiv:2505.18875*, 2025.