RecallWorld: Episodic Memory for Persistent Embodied Video World Models

Sarosij Bose[1,2] Gitika Karumuri[1] Chenqi Li[1,3] Allen Tu[1,4] Srikumar Sastry[1,5]
Mellon Zhang[1,6] Avinash Ravichandran[1] Alexander Schwing[1] Konstantinos Karydis[2] Amit K. Roy Chowdhury[2]
[1]Amazon [2]University of California, Riverside [3]University of Oxford
[4]University of Maryland, College Park [5]Washington University in St. Louis [6]Georgia Institute of Technology

https://recallworld.github.io/

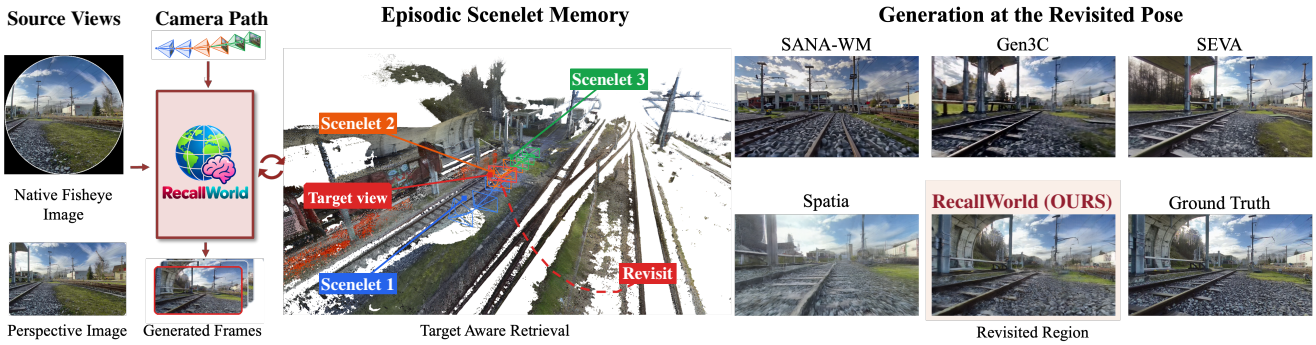

Figure 1. **RecallWorld** consolidates perspective and fisheye views into episodic Scenelets that condition the world model to generate faithful revisits along the target camera trajectory. Retrieving relevant Scenelets on revisits preserves scene structure and appearance that prior methods often lose over long-horizon generation. When the camera revisits a pose in GrandTour [44], prior methods [37, 66, 67, 70] alter scene identity, whereas episodic memory preserves the observed railway structure and appearance.

Abstract

Video world models generate plausible novel views but often lack persistent memory, failing to preserve a place's geometry and appearance when an agent revisits a scene. A central challenge is how previous views should be effectively used for creating memory. Reconstructing them independently underutilizes geometric constraints between nearby views, while repeatedly reconstructing a growing history can propagate later prediction errors into earlier geometry. We introduce RecallWorld, a persistent embodied video world model that jointly reconstructs short multi-view episodes into Scenelets while keeping geometry independent across episodes. Starting from a single source image, camera-aware ray lifting places perspective and native fisheye images in a common world frame, retaining wide-field content for perspective generation. Each generated segment extends the Scenelet memory and conditions subsequent generation. Coverage-based retrieval selects complementary Scenelets from the full history, while target-view visibility is resolved geometrically before appearance fusion, providing consistent memory conditioning for guiding subsequent generation. Across four datasets, RecallWorld improves revisit PSNR over the strongest evaluated baseline by 2.6 dB on partial revisits and 3.9 dB on full cycles, demonstrating substantially stronger preservation of scene identity across exploration and return.

1. Introduction

Video world models generate future visual observations conditioned on actions or camera motion, supporting robotic interaction and manipulation [50, 53, 68], autonomous driving [11, 39, 65], and synthetic training-data generation [21, 33, 56]. However, for embodied exploration, plausible next frames are not enough: the model must preserve a place's geometry and appearance after its observations leave the recent context as an agent may later approach the same location from a new direction. Persistent video generation therefore requires long-term access to earlier observations without reconstructing or conditioning on the full history.

No prior video world model combines these capabilities. Camera-controlled generators produce locally coherent video [2, 14, 38, 43, 70] but do not guarantee consistent revisits. Memory-based methods retrieve earlier views or

project stored 3D content [28, 37, 51, 52], but *leave the granularity at which memory should be consolidated unresolved*. Independent per-view reconstruction discards constraints among nearby observations, whereas repeatedly reconstructing the growing history lets later generated content revise earlier geometry. We design persistent memory by reconstructing nearby views jointly as *Scenelets*, keeping episode geometry independent, and retrieving target-relevant content within a fixed conditioning budget.

We present **RecallWorld**, a persistent embodied video world model with episodic Scenelet memory. Firstly, camera-aware ray lifting unifies perspective and native fisheye observations in a common world frame [36], retaining surfaces outside the perspective image used for generation. Next, each completed segment forms a Scenelet: overlapping views are reconstructed jointly, and geometry remains independent across episodes. Then, coverage-based retrieval utilizes complementary Scenelets under a fixed conditioning budget, and target-view geometry resolves visibility before appearance fusion. This design separates local reconstruction from long-term retention while keeping earlier Scenelets accessible. Figure 2 illustrates our pipeline.

We evaluate RecallWorld on real robot-navigation trajectories from GrandTour [44] and M2DGR [58], synthetic trajectories from TartanGround [35], and real scene captures from DL3DV [30]. On exploration-and-return trajectories starting from the same source image, RecallWorld's generated revisits preserve the structure and appearance of previously observed places, whereas camera-controlled and global-memory baselines alter scene content. Our ablation studies show that Scenelets improve revisit fidelity over independent views as memory and reduce peak reconstruction memory relative to global memory. Natively using fisheye images further expands the surface coverage retained in memory. We also measure camera adherence and video coherence throughout each rollout. Together, these results show that RecallWorld preserves place identity across exploration and partial or full return cycles.

Our main contributions are:

- **Episodic Scenelet memory.** We show that how observations are grouped into memory is important for persistent video generation. We introduce episodic Scenelets, which jointly reconstruct nearby views while keeping geometry independent across different episodes.
- **Persistent Retrieval and visibility-aware conditioning.** We retrieve Scenelets from the full history under a fixed memory budget and use their geometry to determine which surfaces are visible in the target view before combining appearance for generation.
- **Camera general persistent representation.** We use camera-aware ray lifting to retain perspective and native fisheye observations in a common spatial representation, preserving wide-field surfaces for later perspective generation without changing the Scenelet pipeline.

Table 1. **Memory and conditioning in video world models.** **RecallWorld** combines persistent 3D memory, navigation video generation, and local 3D units with independent multi-view reconstructions. **RecallWorld-PF** extends this design by storing perspective and native fisheye observations in a unified memory.

Method	3D Memory	Navigation Video Gen.	Local 3D Units	Unified Memory
SANA-WM [70]	✗	✗	✗	✗
NWM [2]	✗	✓	✗	✗
Gen3C [37]	✓	✗	✗	✗
VMem [28]	✓	✗	✗	✗
SPMem [51]	✓	✗	✗	✗
Spatia [66]	✓	✗	✗	✗
Lyra 2.0 [40]	✓	✗	✗	✗
RecallWorld	✓	✓	✓	✓$_{\text{PF}}$

2. Related Work

Camera Conditioned Video Generation. Video diffusion models synthesize temporally coherent sequences [3, 16]. Video world models predict future observations from actions or camera motion. Action-conditioned models support interactive navigation [2, 4, 50], while pose-conditioned diffusion and explicit 3D guidance enable broader camera control [5, 13, 14]. SANA-WM generates minute-scale videos along continuous camera trajectories [70], SEVA synthesizes image-conditioned novel views [67], and GEN3C uses a 3D cache for world-consistent camera control [37]. Camera control, however, does not ensure a consistent revisit once a place leaves the model's recent context.

Feed Forward Reconstruction. Beyond camera control, feed-forward models recover geometry and correspondences from images [26, 46], while generative reconstruction jointly models geometry and appearance to ground synthesized views in observed structure [6, 42, 64]. Most generative models assume pinhole imagery, and rectifying wide field-of-view inputs discards peripheral coverage [9]. UniK3D instead estimates rays and radial range for perspective, fisheye, and panoramic cameras [36]. Fisheye3R and Cam3R generalize feed-forward reconstruction [9, 12], while ray-based generators extend camera control across lens geometries [18, 23, 61]. RecallWorld uses π^3X [47] for joint multi-view reconstruction, Depth Anything 3 [29] for perspective initialization, and UniK3D [36] for native fisheye initialization. Together, these components provide camera-aware geometry for episodic memory.

Memory Aware Generation. For persistence over long horizons, video models retain spatial evidence as retrieved images indexed by pose, time or field-of-view overlap [52, 59], projected 3D caches or RGB-D features [15, 69], and

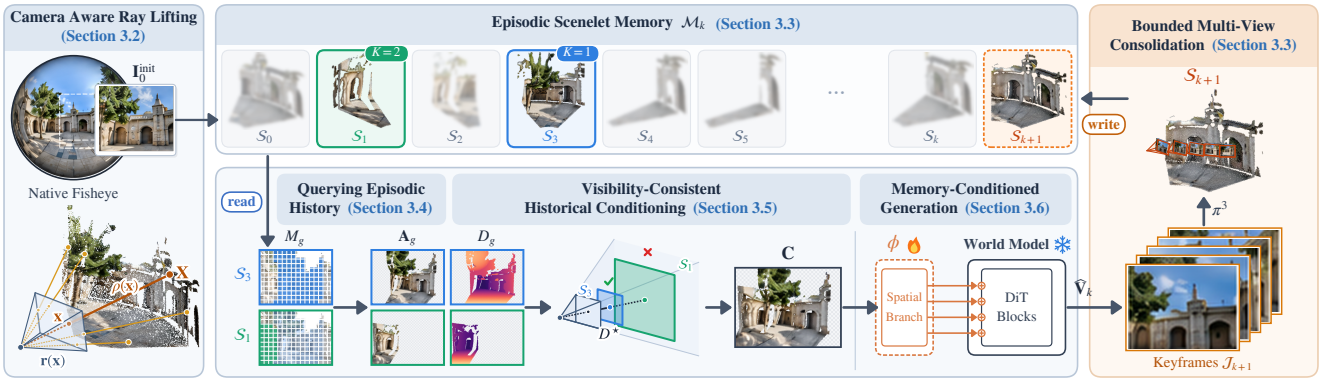

Figure 2. **Overview of RecallWorld.** Ray lifting unifies perspective and native fisheye inputs in 3D; consolidation forms episodic Scenelet memory $\mathcal{M}_k$. Complementary retrieval, target-view projection and visibility resolution form $\mathbf{C}_k$, which conditions SANA-WM [70] via a spatial-conditioning branch. Generated keyframes are jointly reconstructed into $\mathcal{S}_{k+1}$ for continued memory-aware generation.

evolving global or streaming maps [7, 22, 27, 60]. Memory organization then governs which observations share a reconstruction and remain available for future queries. VMem indexes earlier views with surfels [28], SPMem combines geometric support with recent and episodic images [51], and Spatia maintains an updatable point-cloud memory [66]. AnchorWeave retrieves complementary frame-associated geometry by marginal coverage [49]. RecallWorld follows this retrieval family while changing the persistent memory unit: nearby views are jointly reconstructed before retention, whereas separate episodes retain independent geometry and are reconciled after target-view projection. Table 1 compares these components.

3. RecallWorld

A persistent video world model must preserve observed geometry and appearance after a place leaves its recent context. **RecallWorld** generates video autoregressively and consolidates each completed segment into an independent **Scenelet**. Nearby keyframes share a joint reconstruction, while stored Scenelets remain fixed, preventing later prediction errors from rewriting earlier geometry. Camera-aware ray lifting places perspective and fisheye observations in a common world-frame representation. Before generating each segment, RecallWorld retrieves a complementary set of Scenelets covering the target views, resolves visibility among their projected surfaces, and conditions the generator on the resulting appearance. This separates local reconstruction from long-term retention while keeping active conditioning bounded. Figure 2 illustrates our framework.

3.1. Persistent Autoregressive Generation

RecallWorld generates video in consecutive segments from a single source image and target camera poses. The source image $\mathbf{I}_0^{\text{init}}$ forms the initial Scenelet $\mathcal{S}_0$, giving $\mathcal{M}_0 = \{\mathcal{S}_0\}$. For segment $k > 0$, let $\mathbf{I}_k^{\text{init}}$ denote its starting image, $\mathcal{Q}_k$ the target camera poses, and $\mathcal{M}_k = \{\mathcal{S}_0, \ldots, \mathcal{S}_{k-1}\}$ the set of Scenelets accumulated so far. Projecting the accumulated Scenelets into the target views $\mathcal{Q}_k$ and resolving visibility produces the bounded, target-aligned condition $\mathbf{C}_k$. We augment the pretrained SANA-WM [70] backbone θ with a learned ControlNet-style [62] spatial-conditioning branch ϕ. Let $\widehat{\mathbf{V}}_k$ denote the generated segment and $\mathcal{S}_{k-1}$ its reconstructed Scenelet ($\mathcal{S}_0$ from source image). The memory-conditioned generator $\mathcal{F}_{\theta,\phi}$ operates causally:

$$\begin{aligned} \widehat{\mathbf{V}}_k &= \mathcal{F}_{\theta,\phi}\big(\mathbf{I}_k^{\text{init}}, \mathcal{Q}_k; \mathbf{C}_k\big), \\ \mathbf{I}_k^{\text{init}} &= \text{last}(\widehat{\mathbf{V}}_{k-1}), \qquad \mathcal{M}_k = \mathcal{M}_{k-1} \cup \{\mathcal{S}_{k-1}\}. \end{aligned} \tag{1}$$

The pretrained decoder produces $\widehat{\mathbf{V}}_k$ for autoregressive feedback and Scenelet construction. Only $\mathbf{I}_0^{\text{init}}$ and $\mathcal{S}_0$ come directly from the source observation. After initialization, each segment starts from a generated image and appends a Scenelet built from generated frames, so prediction errors can influence later segments.

3.2. Camera Aware Ray Lifting

Persistent memory must retain the geometry and appearance captured across the camera's full field of view, rather than being limited to the narrower perspective view used for generation. RecallWorld lifts each source pixel along its viewing ray using the corresponding radial range, mapping it to a 3D point in the shared world space. Perspective and fisheye images can therefore be represented in the same world space, including surfaces outside the rectified perspective view.

For an observation $\mathbf{I}_i$, let $\mathcal{C}_i$ denote its calibrated pinhole or Kannala–Brandt fisheye camera model [25] and $\mathbf{T}_i = (\mathbf{R}_i, \mathbf{t}_i)$ its camera-to-world pose, where $\mathbf{R}_i \in \text{SO}(3)$ and $\mathbf{t}_i \in \mathbb{R}^3$. Generated keyframes take their poses from the target trajectory. At image coordinate $\mathbf{x} \in \mathbb{R}^2$, calibrated unprojection gives the unit viewing ray $\mathbf{r}_i(\mathbf{x})$, and radial range $\rho_i(\mathbf{x})$ locates the corresponding world-space point:

$$\begin{aligned} \mathbf{r}_i(\mathbf{x}) &= \pi_{\mathcal{C}_i}^{-1}(\mathbf{x}), \qquad \|\mathbf{r}_i(\mathbf{x})\|_2 = 1, \\ \mathbf{X}_i(\mathbf{x}) &= \mathbf{R}_i[\rho_i(\mathbf{x})\mathbf{r}_i(\mathbf{x})] + \mathbf{t}_i. \end{aligned} \tag{2}$$

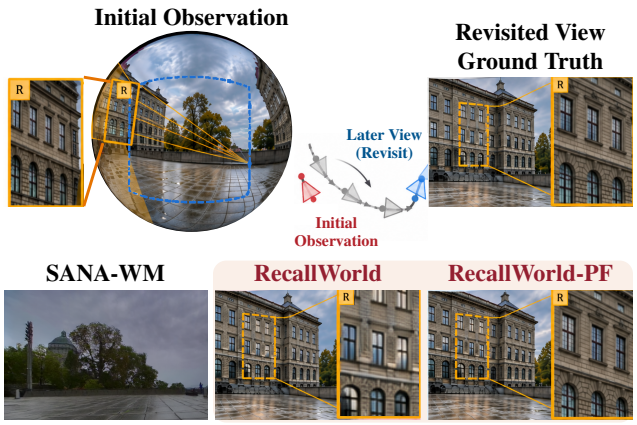

Figure 3. **Wide-field memory on GrandTour [44].** Region R lies outside the rectified view, marked by the dashed box in the initial observation. From the rectified view, SANA-WM [70] has no memory of R while RecallWorld recalls it. RecallWorld-PF initializes memory from the full fisheye view and retains R sharply.

Unlike optical depth, radial range remains defined for fisheye rays at or beyond 90° from the optical axis.

We introduce two RecallWorld variants that differ in source initialization and initial memory coverage but share all subsequent processing, including the use of generated perspective frames for later Scenelets. **RecallWorld** uses the perspective source image to initialize both generation and $\mathcal{S}_0$. Depth Anything 3 [29] estimates its depth, which we convert to radial range. **RecallWorld-PF** instead initializes $\mathcal{S}_0$ from a calibrated native fisheye image and generation from its perspective resampling $\mathbf{I}_0^{\text{init}}$. UniK3D [36] estimates the native image's radial range directly along calibrated rays. Figure 3 shows how RecallWorld-PF retains surfaces outside the initial perspective view and uses them to condition a later revisit.

3.3. Episodic Scenelet Memory

Beyond locating observed surfaces in 3D, persistent memory depends on consolidation granularity applied to memory, which determines how nearby views share a reconstruction. Per-view reconstruction discards constraints among nearby frames, whereas repeatedly reconstructing the growing history allows newly generated content to revise earlier geometry. RecallWorld avoids both extremes by grouping each segment into a bounded Scenelet whose views share a local reconstruction that remains independent across episodes.

For each Scenelet $\mathcal{S}_g$ with $g \geq 1$, $\mathcal{J}_g$ indexes five evenly spaced keyframes from the corresponding completed segment. π^3X [47] jointly reconstructs the keyframes using their intrinsics and target poses. A similarity transform then aligns the predicted cameras to those poses, using Kabsch [24] for rotation and the camera centers for scale and translation. The resulting depths are converted to radial range and lifted (see Equation 2), ensuring that stored geometry matches the generator's camera conditioning. For each $i \in \mathcal{J}_g$, $\mathcal{S}_g$ stores episode-specific point map $\mathbf{X}_i^{(g)}$, appearance features $\mathbf{F}_i = \mathcal{E}(\mathbf{I}_i)$ from frozen video encoder $\mathcal{E}$, and validity mask $W_i^{(g)}$ for valid non-sky geometry:

$$\mathcal{S}_g = \left\{ (\mathbf{X}_i^{(g)}, \mathbf{F}_i, W_i^{(g)}) \mid i \in \mathcal{J}_g \right\}. \tag{3}$$

Each tuple remains tied to view i, with $\mathbf{X}_i^{(g)}$ and $W_i^{(g)}$ specific to episode g. A boundary frame appearing in adjacent segments therefore receives a separate record in each Scenelet. The source image forms the single-view Scenelet $\mathcal{S}_0$, and new Scenelets are appended without revising existing geometry.

3.4. Target-Aligned Scenelet Retrieval

To keep conditioning bounded as Scenelet memory grows with the trajectory [54, 57], RecallWorld retains every Scenelet but activates at most K for each target segment. Let Ω be the set of spatiotemporal locations p in the target segment's latent grid; we omit the segment index k for clarity. For each candidate $\mathcal{S}_g \in \mathcal{M}_k$, projection into the target poses produces

$$\Pi(\mathcal{S}_g, \mathcal{Q}_k) = (\mathbf{A}_g, D_g, B_g). \tag{4}$$

At each p, the nearest valid surface among the Scenelet's keyframes determines projected appearance $\mathbf{A}_g(p)$ and target-camera depth $D_g(p)$. Binary support $B_g(p) \in \{0, 1\}$ indicates whether such a surface exists.

Ranking Scenelets individually by pose distance, recency, or view overlap [52, 59] can select redundant candidates covering the same target region. Following [49, 59], RecallWorld instead scores each candidate by the additional target locations it covers beyond the current selection. Retrieval begins with selected index set $\mathcal{G} = \varnothing$ and uncovered set $\mathcal{U} = \Omega$. At each step, it selects

$$g^\star = \underset{\substack{\mathcal{S}_g \in \mathcal{M}_k \\ g \notin \mathcal{G}}}{\arg\max} \sum_{p \in \mathcal{U}} B_g(p). \tag{5}$$

After selecting $g^\star$, we add it to $\mathcal{G}$ and remove each location it supports from $\mathcal{U}$. The procedure stops after K Scenelets or when no candidate adds coverage. Because set coverage is monotone submodular, greedy selection achieves at least $1 - 1/e$ of the optimal coverage with K Scenelets [32]. Earlier Scenelets can therefore be selected when they cover regions missed by later ones.

3.5. Visibility-Resolved Conditioning

Coverage-based retrieval determines which Scenelets support a target location, but not which projected surface is visible. Independently reconstructed Scenelets may project different surfaces onto the same target location, such as a

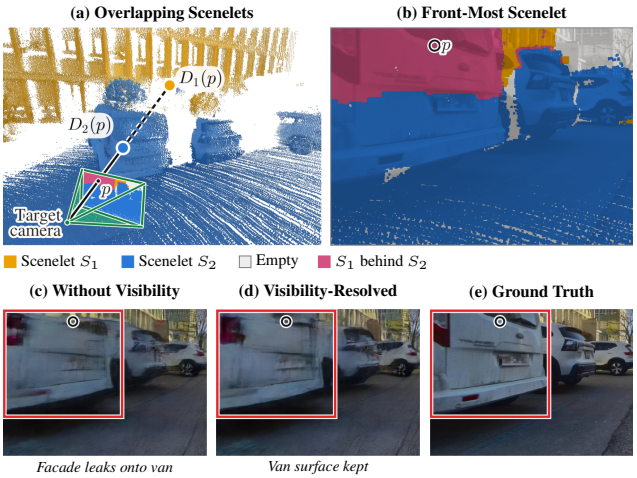

Figure 4. **Visibility resolution across Scenelets on GrandTour [44].** (a) Independently reconstructed Scenelets project conflicting surfaces into the target view. (b) Nearest depth selects the frontmost surface. (c) Fusion without visibility resolution contaminates the generated revisit. (d) RecallWorld resolves the conflict before fusion and better matches (e) ground truth.

wall and the building behind it. Averaging their appearance features would introduce occluded content, so RecallWorld resolves visibility geometrically before fusion.

At a supported location p, let $D_{\min}(p) > 0$ denote the nearest projected depth among selected Scenelets with $B_g(p) = 1$. Given a relative depth tolerance $\delta \geq 0$, define the visibility-compatible set

$$\mathcal{V}(p) = \{g \in \mathcal{G} : B_g(p) = 1,\ D_g(p) \leq (1+\delta) D_{\min}(p)\}. \quad (6)$$

For each $g \in \mathcal{V}(p)$, binary weight $w_g(p)$ indicates whether its projected appearance passes validity checks. If at least one appearance is valid, the fused condition is

$$\mathbf{C}(p) = \frac{\sum_{g \in \mathcal{V}(p)} w_g(p)\, \mathbf{A}_g(p)}{\sum_{g \in \mathcal{V}(p)} w_g(p)}. \quad (7)$$

Depth-compatible projections are treated as the same surface, and valid observations receive equal weight. A binary mask $M(p)$ indicates whether any valid projected appearance remains, and $\mathbf{C}(p) = \mathbf{0}$ wherever $M(p) = 0$. Because visibility depends only on geometry, a front surface without valid appearance still occludes farther surfaces, leaving its location unconditioned rather than exposing hidden content. Across the target grid, $\mathbf{C}$ and M form the segment condition $\mathbf{C}_k$ for the spatial branch. Figure 4 illustrates this geometry-first rule and its effect on the generated revisit.

3.6. Training and Inference

Memory-conditioned training. The spatial branch copies a subset of backbone blocks, embeds $\mathbf{C}_k$, and injects its outputs into the backbone through zero-initialized projections. During this stage, only ϕ is trained; the backbone, video encoder, and pretrained decoder remain fixed. We train the branch with the conditional flow-matching objective used by SANA-WM [31, 70]. For a recorded training segment with latent $\mathbf{z} = \mathcal{E}(\mathbf{V})$ and noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the noisy latent and target velocity are

$$\mathbf{z}_{\boldsymbol{\tau}} = (\mathbf{1} - \boldsymbol{\tau}) \odot \mathbf{z} + \boldsymbol{\tau} \odot \boldsymbol{\epsilon}, \qquad \mathbf{u} = \boldsymbol{\epsilon} - \mathbf{z}, \quad (8)$$

where $\boldsymbol{\tau}$ holds one noise level per latent frame and $\odot$ is elementwise multiplication. The first latent frame encodes the starting image and remains clean ($\tau_0 = 0$); the remaining frames share one sampled noise level in $[0, 1]$. Let m be a binary mask over the latent tensor that excludes the first frame. The branch minimizes

$$\mathcal{L}_{\mathrm{FM}}(\phi) = \mathbb{E}\left[\frac{1}{\|m\|_1} \left\| m \odot (\mathbf{v}_{\theta,\phi} - \mathbf{u}) \right\|_2^2\right], \quad (9)$$

where $\mathbf{v}_{\theta,\phi}$ is the predicted velocity at $(\mathbf{z}_{\boldsymbol{\tau}}, \boldsymbol{\tau}; \mathbf{I}_k^{\mathrm{init}}, \mathcal{Q}_k, \mathbf{C}_k)$, and $\|m\|_1$ counts the supervised latent entries. During training, $\mathbf{C}_k$ is formed from preceding recorded frames. At inference, it is formed from the source observation and prior model predictions.

Decoder refinement. The pretrained decoder loses fine scene detail that the latents retain. Holding the encoder and generator fixed, we fine-tune a copy of the decoder, $\mathcal{D}_\psi$, for higher-fidelity reconstruction. For a recorded video $\mathbf{V}$ with L frames and N RGB values, let $\widetilde{\mathbf{V}} = \mathcal{D}_\psi(\mathcal{E}(\mathbf{V}))$ denote its reconstruction. The decoder objective is

$$\mathcal{L}_{\mathrm{dec}}(\psi) = \tfrac{1}{N} \left\| \widetilde{\mathbf{V}} - \mathbf{V} \right\|_1 + \tfrac{1}{L} \sum_{\ell=1}^{L} \mathcal{L}_{\mathrm{LPIPS}}\left(\widetilde{\mathbf{V}}_\ell, \mathbf{V}_\ell\right), \quad (10)$$

where ℓ indexes frames and $\mathcal{L}_{\mathrm{LPIPS}}$ is the LPIPS perceptual loss [63]. Table 5 shows that refinement improves both reconstruction and generation fidelity on the same latents.

Autoregressive inference. RecallWorld retrieves up to K Scenelets, resolves visibility, and generates each segment according to the causal update in Equation 1. The pretrained decoder supplies the next starting image and the keyframes used for Scenelet construction. The refined decoder produces only the reported output video.

4. Experiments

RecallWorld is based on the SANA-WM [70] backbone, generating 129-frame clips at 16 fps. Five evenly spaced keyframes from each completed segment are jointly reconstructed with π^3X [47]. Retrieval activates at most $K = 4$ Scenelets. During memory-conditioned training, only the copied spatial-conditioning blocks and their output projections are optimized; the backbone, video encoder, and pretrained decoder remain fixed. The pretrained decoder supplies autoregressive feedback and keyframes for Scenelet construction. The encoder provides appearance features, and the refined decoder produces the reported RGB outputs. We train on 8 H100 GPUs.

Datasets and trajectories. We construct exploration-and-return trajectories from four datasets: GrandTour [44] and

Table 2. **Persistent embodied video generation.** We compare camera-controlled and persistent-memory models on 32-second exploration-and-return trajectories from GrandTour [44], M2DGR [58], TartanGround [35] and DL3DV [30]. Metrics cover revisit fidelity, camera adherence and video consistency. Partial revisit averages high-, moderate- and low-overlap cases. Best in **bold**, second best underlined.

	Partial Revisit			Full Revisit (Cycle)			Camera Adherence			Video Consistency		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	TransErr↓	RotErr↓	ATE↓	WS-3D↑	VB-BG↑	MEt3R↓
Camera-controlled generation												
SANA-WM [70]	12.82	0.366	0.710	13.96	0.325	0.618	1.18%	1.43°	3.00%	0.468	0.909	0.040
GEN3C [37]	13.22	0.320	0.732	14.06	0.383	0.622	2.01%	2.03°	7.06%	0.628	0.915	0.044
SEVA [67]	13.67	0.377	0.694	13.94	0.350	0.609	**1.10%**	**1.12°**	**1.88%**	0.635	**0.930**	**0.028**
Persistent-memory generation												
VMem [28]	12.31	0.240	0.886	12.47	0.249	0.841	4.28%	8.89°	11.94%	0.407	0.806	0.080
SPMem [51]	13.74	0.313	0.782	13.31	0.346	0.708	4.00%	7.57°	8.04%	0.448	0.853	0.073
Spatia [66]	14.54	0.379	0.728	15.46	0.402	0.616	3.32%	6.61°	8.15%	0.417	0.901	0.107
Lyra 2.0 [40]	13.88	0.370	0.754	12.94	0.359	0.731	1.65%	2.60°	6.34%	0.552	0.886	0.058
RecallWorld	**17.11**	**0.541**	**0.394**	**19.34**	**0.557**	**0.327**	1.24%	2.43°	2.81%	**0.685**	0.914	0.035

M2DGR [58] provide real robot navigation videos, TartanGround [35] provides synthetic navigation videos, and DL3DV [30] provides real scene videos. The collection includes full revisits to earlier camera poses, following loop-closure evaluations [28, 55], and partial revisits that reobserve previously seen regions from different viewpoints. Cycle trajectories retrace the outgoing path in reverse; partial revisits span *high*, *moderate*, and *low overlap* with earlier views. The main camera-pose comparison uses 32-second trajectories and common target poses. Scores are averaged by clip.

Baselines and configurations. Camera-controlled generation is represented by SANA-WM [70], SEVA [67] and GEN3C [37]; persistent-memory models include VMem [28], SPMem [51], Spatia [66] and Lyra 2.0 [40]. Table 3 reports a separate comparison with Navigation World Models (NWM) [2] on the 32-second navigation subset. We map target camera poses to NWM's native planar-translation and yaw actions, repeat the source image across its context slots, and score at 4 Hz. Unless stated otherwise, RecallWorld denotes the perspective variant. External baselines retain their native encoders and decoders.

Metrics. Revisit fidelity uses PSNR [17], SSIM [48], and LPIPS [63] against recorded target views. Following GEAR [55], we align ViPE [19] poses once per sequence with a similarity transform [45] and report relative translation and rotation errors (TransErr and RotErr). TransErr uses one-second intervals; ATE [41] measures accumulated trajectory deviation. Both translation errors are normalized by target path length and reported as percentages, while RotErr is reported in degrees. WorldScore 3D consistency [8] and VBench background consistency [20] assess spatial and temporal consistency. MEt3R [1] measures cross-view feature disagreement after 3D alignment. The navigation-command comparison additionally reports DreamSim [10] and Scene Revisit Consistency (SRC) [69], which measures cosine similarity between DINOv2 [34] features from first-visit and return frames at the same camera pose.

Table 3. **Navigation video generation.** We compare RecallWorld with NWM [2] on the 32-second navigation subset. NWM uses camera-derived movement actions.

Method	**PSNR↑**	**LPIPS↓**	**DreamSim↓**	**SRC↑**
NWM [2]	12.17	0.651	0.669	0.554
RecallWorld	**17.48**	**0.363**	**0.106**	**0.880**

5. Results

RecallWorld achieves the best reported revisit fidelity across all three metrics for both partial and full revisits in Table 2. Against the strongest baseline for each metric, PSNR improves by 2.6 dB on partial revisits and 3.9 dB on full cycles, while LPIPS decreases by 43.2% and 46.3%, respectively. RecallWorld also attains the highest WorldScore 3D consistency and the second-lowest ATE. SEVA [67] leads in camera adherence, background consistency and MEt3R but has substantially worse revisit fidelity. Better trajectory adherence is therefore not sufficient for faithful revisits. The fidelity advantage also carries over to the separate navigation subset in Table 3, where RecallWorld outperforms NWM [2] on all four metrics, raising SRC from 0.554 to 0.880.

Figure 5 compares how faithfully each method recovers observed scene content across partial and full revisits. SANA-WM [70] generates a plausible street view but replaces the observed foreground van. SEVA [67] likewise loses observed content despite its stronger camera adherence. Spatia [66] and Lyra 2.0 [40] alter scene structure despite using persistent memory, showing that retaining history does not by itself ensure faithful returns. By contrast, RecallWorld more closely

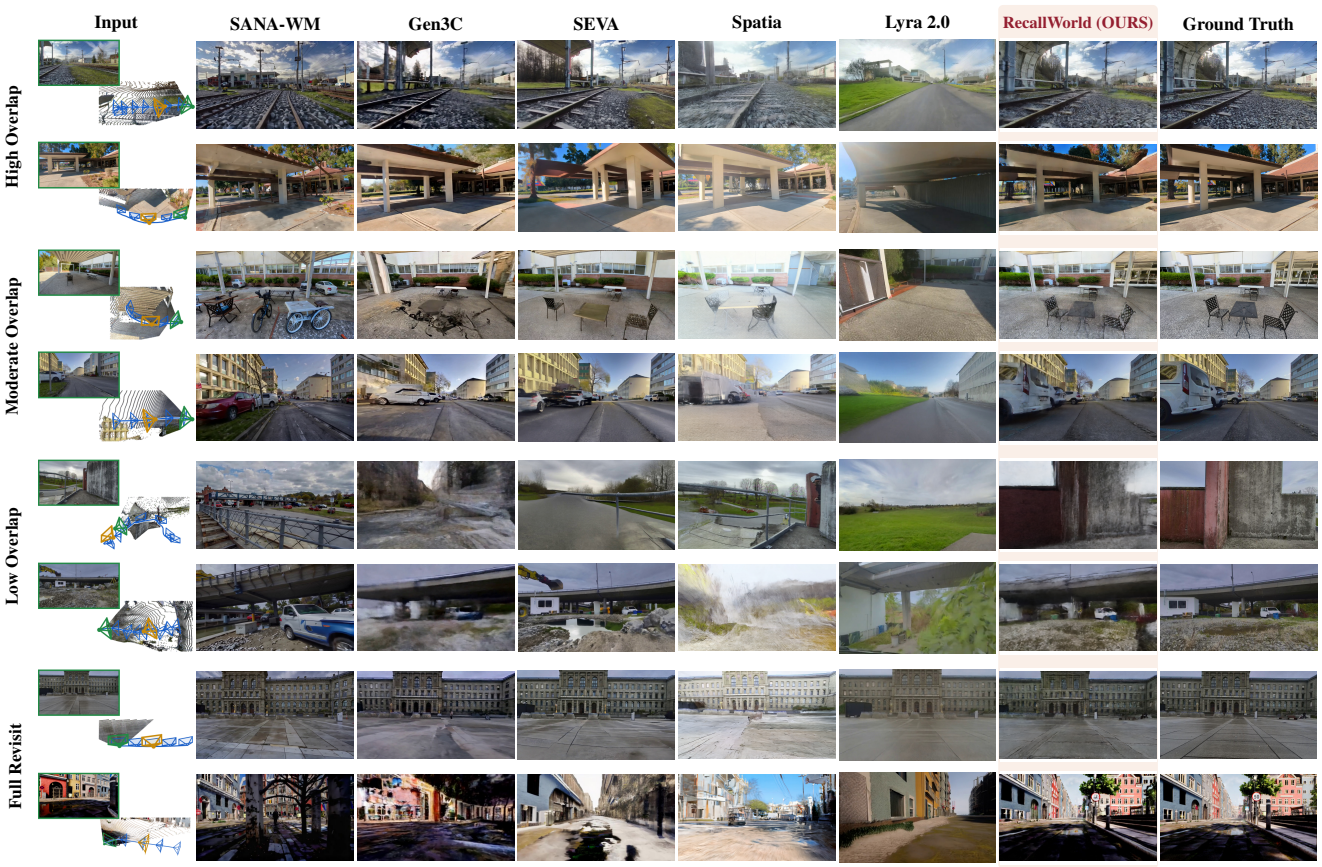

Figure 5. **Revisits on GrandTour [44], DL3DV [30] and TartanGround [35].** Rows are grouped into high-, moderate- and low-overlap partial revisits, followed by full retraced cycles. The input column shows the source image and camera path: blue marks the path, green the source pose, and gold the revisit pose. The remaining columns compare generated revisits with recorded target views.

Table 4. **Memory organization and visibility on GrandTour [44].** Revisit fidelity with memory reconstructed per keyframe, over the full history or jointly per segment as Scenelets. RecallWorld adds visibility resolution before Scenelet fusion.

Memory Organization	Partial Revisit		Full Revisit (Cycle)	
	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
Independent views	16.41	0.365	16.27	0.255
Global memory	17.02	0.344	17.80	0.241
Scenelets w/o visibility	20.72	0.339	20.27	0.238
RecallWorld	**20.82**	**0.322**	**20.84**	**0.237**

matches the recorded target views, preserving railway structures, street layouts and building facades across both revisit types. Its advantage is more faithful recovery of observed place identity, not visual plausibility alone.

5.1. Memory and Decoder Ablations

Local multi-view consolidation accounts for most of the PSNR gain in Table 4. Even without visibility resolution, Scenelets improve PSNR over global memory by 3.7 dB on partial revisits and 2.5 dB on full cycles. Visibility resolution further improves both PSNR and LPIPS. The complete model reduces LPIPS by 11.8% and 7.1%, respectively, relative to independent-view memory. In Figure 7, global memory distorts scene structure and independent-view memory propagates hallucinations. In Figure 4, visibility resolution prevents the occluded facade's appearance from leaking onto the foreground van.

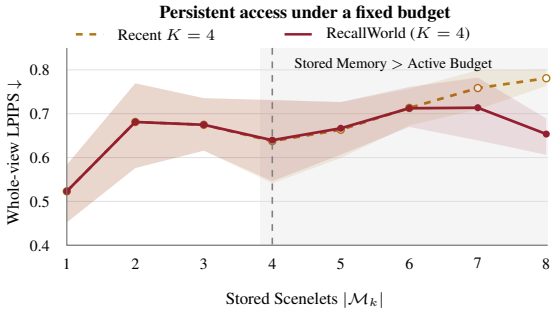

Figure 6. **Long-horizon fidelity with bounded Scenelet retrieval.** Whole-view LPIPS on 64 second GrandTour [44] cycles as stored memory grows under a $K = 4$ Scenelet retrieval budget.

Complementary retrieval achieves lower whole-view LPIPS than recent selection at the longest history in

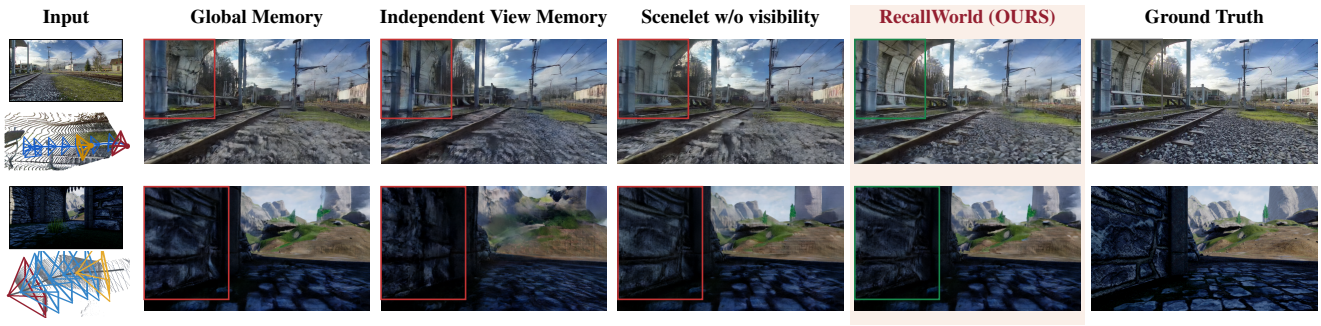

Figure 7. **Memory ablations on GrandTour [44] and TartanGround [35].** RecallWorld more closely matches the recorded target views, preserving local structure and detail better than global memory, independent-view memory, and Scenelets without visibility resolution.

Table 5. **Decoder refinement.** Pretrained and refined decoders are applied to the same latents, from encoded recorded videos for reconstruction and from whole RecallWorld rollouts for generation.

Dataset	Decoder	Reconstruction		Generation	
		PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
GrandTour [44]	Pretrained	19.16	0.301	13.93	0.605
	+ Refinement	**25.29**	**0.196**	**14.22**	**0.563**
DL3DV [30]	Pretrained	19.23	0.208	13.43	0.574
	+ Refinement	**22.92**	**0.188**	**13.78**	**0.568**

Figure 6. Both use at most $K = 4$ active Scenelets on 64-second GrandTour [44] cycles. The strategies coincide while stored memory fits within this budget. As history grows, recent selection excludes earlier episodes and LPIPS rises, whereas complementary retrieval recalls target-relevant Scenelets. Scenelets lower peak memory versus global reconstruction by 19.7%.

With latents held fixed, decoder refinement raises reconstruction PSNR by 6.1 dB on GrandTour and 3.7 dB on DL3DV [30] in Table 5. Generation PSNR improves by 0.29 dB and 0.35 dB, respectively. GrandTour generation LPIPS improves from 0.605 to 0.563. Generation fidelity drops by only $\sim 1\%$ even without our fine-tuned decoder. The pretrained decoder still supplies autoregressive feedback and memory keyframes, leaving rollout history and Scenelet construction unchanged.

5.2. Fisheye Memory

Native fisheye memory retains observed surfaces lost to perspective rectification. In Table 6, RecallWorld-PF increases projected coverage from 52.67% to 63.27%, a 20.1% relative increase, and lowers both peripheral and whole-view LPIPS. With encoding and range estimates held fixed, expanding the fisheye field of view from 90° to its native 120° [44] adds 2.58 percentage points of coverage and improves whole-view LPIPS by 1.4%. Figures 3 and 8 show sharper revisited structure and coherent street layouts. Wide-field memory improves perspective revisits without requiring fisheye generation.

Table 6. **Fisheye memory on GrandTour [44].** Coverage and revisit fidelity with perspective and native fisheye memory. Peripheral scores cover regions outside the initial perspective view.

Method	Memory		Peripheral		Whole
	FoV	Coverage↑	PSNR↑	LPIPS↓	LPIPS↓
RecallWorld	90°	52.67%	11.33	0.760	0.664
RecallWorld-PF	90°	60.69%	11.68	0.711	0.663
RecallWorld-PF	120°	**63.27%**	**11.93**	**0.687**	**0.654**



Figure 8. **Generation from Fisheye image on GrandTour [44].** Existing world models [70] cannot account for fisheye distortions, leading to severely distorted generations and blank periphery, whereas RecallWorld retains structures beyond the initial 90° FoV.

6. Conclusion

We introduced **RecallWorld**, a persistent embodied video world model that separates local reconstruction from long-term retention. Episodic Scenelets jointly reconstruct nearby views while keeping geometry independent across episodes, so later predictions do not rewrite earlier reconstructions. Complementary retrieval and geometry-first visibility resolution recall relevant history within a fixed conditioning budget. Camera-aware ray lifting unifies perspective and native fisheye observations in memory, retaining wide-field content for later perspective generation. Across four benchmarks, RecallWorld improves revisit PSNR over the strongest evaluated baseline by 2.6 dB on partial revisits and 3.9 dB on full cycles, demonstrating that it more faithfully preserves place identity across exploration and return.

Limitations and future work. Long-horizon autoregressive generation can compound prediction errors, motivating uncertainty-aware consolidation and cross-episode verification of Scenelet memory. Extending static Scenelets with time-indexed geometry and appearance could track moving actors, object changes, and illumination across re-

visits while preserving enduring scene structure. Native fisheye generation could further carry wide-field coverage from retained observations into generated views. Together, these directions would extend persistent world models from remembering where an agent has been to modeling how places evolve.

LLM Usage Statement

GPT 5.6 Sol assisted with implementing code, running experiments and drawing diagrams. The authors designed the method and experiments, wrote the manuscript, reviewed all results, and take full responsibility for the work.

References

[1] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. MEt3R: Measuring multi-view consistency in generated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6034–6044, 2025.

[2] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation World Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15791–15801, 2025.

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

[4] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C.Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando De Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *Proceedings of the 41st International Conference on Machine Learning*, pages 4603–4623, 2024.

[5] Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3C: Unifying precisely 3D-enhanced camera and human motion controls for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–12, 2025.

[6] Ata Çelen, Jaewoo Jung, Federico Tombari, Marc Pollefeys, Sunghwan Hong, Michael Niemeyer, and Daniel Barath. GenRec: Knowing where to reconstruct and where to generate. *arXiv preprint arXiv:2608.17832*, 2026.

[7] Yutian Chen, Shi Guo, Renbiao Jin, Tianshuo Yang, Xin Cai, Yawen Luo, Mingxin Yang, Mulin Yu, Linning Xu, and Tianfan Xue. Anyrecon: Arbitrary-view 3d reconstruction with video diffusion model. *arXiv preprint arXiv:2604.19747*, 2026.

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

[9] Ruxiao Duan, Erin Hong, Dongxu Zhao, Eric Turner, Alex Wong, and Yunwen Zhou. Fisheye3R: Adapting unified 3D feed-forward foundation models to fisheye lenses. In *Computer Vision – ECCV 2026*, pages 1–19, 2026.

[10] 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.

[11] 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. In *Advances in Neural Information Processing Systems*, pages 91560–91596, 2024.

[12] Namitha Guruprasad, Abhay Yadav, Cheng Peng, and Rama Chellappa. CAM3R: Camera-agnostic model for 3D reconstruction. *arXiv preprint arXiv:2603.22631*, 2026.

[13] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for video diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[14] Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. CameraCtrl II: Dynamic Scene Exploration via Camera-controlled Video Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13416–13426, 2025.

[15] Yuze He, Bin Tan, Zelin Gao, Yujun Shen, Yong-Jin Liu, Nan Xue, and Yinghao Xu. Persistent planar memory for video world models. *arXiv preprint arXiv:2610.23795*, 2026.

[16] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, pages 8633–8646, 2022.

[17] Alain Horé and Djemel Ziou. Image quality metrics: PSNR vs. SSIM. In *International Conference on Pattern Recognition*, pages 2366–2369, 2010.

[18] Boni Hu, Xiong Wei, Haoming Huang, Yong Huang, Chenbo Wang, Yi Yang, Jiancheng Wang, Ruicheng Zhu, Zhimin Yang, Guanglai Liu, Qiaowan Jin, Dongzhuo Wang, Haiwei Kuang, Jiajun Fan, Yue Wu, Jiaxin Wei, Hao Sun, Feihong Yan, Yuyao Zhou, Wei Bi, Kaixuan Wang, Zichao Guo, and Xiaozhi Chen. ZYT-World: A real-time controllable world model for closed-loop autonomous-driving simulation. *arXiv preprint arXiv:2609.21712*, 2026.

[19] 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-Taixé, and Sanja Fidler. ViPE: Video pose engine for 3D geometric perception. *arXiv preprint arXiv:2508.10934*, 2025.

[20] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21807–21818, 2024.

[21] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, Loïc Magne, Ajay Mandlekar, Avnish Narayan, You Liang Tan, Guanzhi Wang, Jing Wang, Qi Wang, Yinzhen Xu, Xiaohui Zeng, Kaiyuan Zheng, Ruijie Zheng, Ming-Yu Liu, Luke Zettlemoyer, Dieter Fox, Jan Kautz, Scott Reed, Yuke Zhu, and Linxi Fan. DreamGen: Unlocking generalization in robot learning through video world models. In *9th Annual Conference on Robot Learning*, 2025.

[22] Haian Jin, Rundi Wu, Tianyuan Zhang, Ruiqi Gao, Jonathan T. Barron, Noah Snavely, and Aleksander Hołyński. ZipMap: Linear-time stateful 3d reconstruction via test-time training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21748–21759, 2026.

[23] Seonghyun Jin, Youngmin Kim, Sunwoo Park, and Jong Chul Ye. CRePE: Curved ray expectation positional encoding for unified-camera-controlled video generation. *arXiv preprint arXiv:2605.12938*, 2026.

[24] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5): 922–923, 1976.

[25] J. Kannala and S. S. Brandt. A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8):1335–1340, 2006.

[26] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with MASt3R. In *Computer Vision – ECCV 2024*, pages 71–91, 2024.

[27] Moyang Li, Zihan Zhu, Wei Zhang, Marc Pollefeys, and Daniel Barath. CLoSeR: Closing the loop for long-context streaming reconstruction. *arXiv preprint arXiv:2610.01927*, 2026.

[28] Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. VMem: Consistent Interactive Video Scene Generation with Surfel-Indexed View Memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 25690–25699, 2025.

[29] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *The Fourteenth International Conference on Learning Representations*, 2026.

[30] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. DL3DV-10K: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22160–22169, 2024.

[31] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[32] George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14: 265–294, 1978.

[33] NVIDIA. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

[34] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[35] 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. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 20524–20531, 2025.

[36] Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniK3D: Universal camera monocular 3D estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1028–1039, 2025.

[37] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-Informed World-Consistent Video Generation with Precise Camera Control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6121–6132, 2025.

[38] Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, Jie Liu, Yansong Cheng, Yao Yao, Jiayi Zhu, Yihao Meng, Kecheng Zheng, Qingyan Bai, Jingye Chen, Zehong Shen, Yue Yu, Xing Zhu, Yujun Shen, and Hao Ouyang. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

[39] 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. *arXiv preprint arXiv:2503.20523*, 2025.

[40] Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, Sanja Fidler, Jiahui Huang, Huan Ling, Jun Gao, and Xuanchi Ren. Lyra 2.0: Explorable generative 3D worlds. *arXiv preprint arXiv:2604.13036*, 2026.

[41] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 573–580, 2012.

[42] Stanislaw Szymanowicz, Jason Y. Zhang, Pratul Srinivasan, Ruiqi Gao, Arthur Brussee, Aleksander Holynski, Ricardo Martin-Brualla, Jonathan T. Barron, and Philipp Henzler. Bolt3D: Generating 3d scenes in seconds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 24846–24857, 2025.

[43] 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. *arXiv preprint arXiv:2503.20314*, 2025.

[44] Turcan Tuna, Jonas Frey, Frank Fu, Katharine Patterson, Tianao Xu, Maurice Fallon, Cesar Cadena, and Marco Hutter. Grandtour: A legged robotics dataset in the wild for multi-modal perception and state estimation. *arXiv preprint arXiv:2602.18164*, 2026.

[45] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991.

[46] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, 2024.

[47] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

[48] 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.

[49] Zun Wang, Han Lin, Jaehong Yoon, Jaemin Cho, Yue Zhang, and Mohit Bansal. AnchorWeave: World-consistent video generation with retrieved local spatial memories. In *Computer Vision – ECCV 2026*, pages 203–224, 2026.

[50] Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. iVideoGPT: Interactive VideoGPTs are scalable world models. In *Advances in Neural Information Processing Systems*, pages 68082–68119, 2024.

[51] Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video World Models with Long-term Spatial Memory. In *Advances in Neural Information Processing Systems*, pages 49371–49393, 2025.

[52] Zeqi Xiao, Yushi LAN, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term consistent world simulation with memory. In *Advances in Neural Information Processing Systems*, pages 49632–49652, 2025.

[53] Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Leslie Pack Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *The Twelfth International Conference on Learning Representations*, 2024.

[54] Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. NeoVerse: Enhancing 4D world model with in-the-wild monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 40340–40351, 2026.

[55] Zesong Yang, Weikai Chen, Liyuan Cui, Lutao Jiang, Runze Zhang, Yingda Yin, Xiaoyang Huang, Kai Yan, Keyang Luo, Wangguandong Zheng, Xin Wang, Hujun Bao, and Zhaopeng Cui. Geometry as address: Routing attention to visual memory for long-horizon camera-controlled video generation. *arXiv preprint arXiv:2609.34722*, 2026.

[56] Junjie Ye, Rong Xue, Basile Van Hoorick, Runhao Li, Harshitha Rajaprakash, Pavel Tokmakov, Muhammad Zubair Irshad, Vitor Guizilini, and Yue Wang. RoboDream: Compositional world models for scalable robot data synthesis. *arXiv preprint arXiv:2606.02577*, 2026.

[57] Jung Yi, Minjae Kim, Paul Hyunbin Cho, Wooseok Jang, Sangdoo Yun, and Seungryong Kim. WorldKV: Efficient world memory with world retrieval and compression. *arXiv preprint arXiv:2605.22718*, 2026.

[58] Jie Yin, Ang Li, Tao Li, Wenxian Yu, and Danping Zou. M2dgr: A multi-sensor and multi-scenario slam dataset for ground robots. *IEEE Robotics and Automation Letters*, 7(2): 2266–2273, 2022.

[59] Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as Memory: Scene-Consistent Interactive Long Video Generation with Memory Retrieval. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–11, 2025.

[60] Wangbo Yu, Kunhao Liu, Wenbo Hu, Shenghai Yuan, Chaoran Feng, Haiyang Zhou, Yukun Huang, Yiran Wang, Wang Zhao, Yingmin Luo, and Ying Shan. WorldCrafter: Consistent video world model with implicit 3D-aware memory. *arXiv preprint arXiv:2609.24984*, 2026.

[61] Cheng Zhang, Boying Li, Meng Wei, Yan-Pei Cao, Camilo Gambardella, Dinh Phung, and Jianfei Cai. Unified camera positional encoding for controlled video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 38027–38037, 2026.

[62] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023.

[63] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[64] Yisu Zhang, Chenjie Cao, Tengfei Wang, Xuhui Zuo, Junta Wu, Jianke Zhu, and Chunchao Guo. WorldStereo: Bridging camera-guided video generation and scene reconstruction via 3d geometric memories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 40327–40339, 2026.

[65] 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(10):10412–10420, 2025.

[66] Jinjing Zhao, Fangyun Wei, Zhening Liu, Hongyang Zhang, Chang Xu, and Yan Lu. Spatia: Video generation with updatable spatial memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4245–4257, 2026.

[67] Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable Virtual Camera: Generative View Synthesis with Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12405–12414, 2025.

[68] Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. RoboDreamer: Learning compositional world models for robot imagination. In *Proceedings of the 41st International Conference on Machine Learning*, pages 61885–61896, 2024.

[69] Siyuan Zhou, Yilun Du, Yuncong Yang, Lei Han, Peihao Chen, Dit-Yan Yeung, and Chuang Gan. Learning 3D Persistent Embodied World Models. In *Advances in Neural Information Processing Systems*, pages 104807–104829, 2025.

[70] Haoyi Zhu, Haozhe Liu, Yuyang Zhao, Tian Ye, Junsong Chen, Jincheng Yu, Tong He, Song Han, and Enze Xie. SANA-WM: Efficient minute-scale world modeling with hybrid linear diffusion transformer. *arXiv preprint arXiv:2605.15178*, 2026.