OmniPersist: Persistent Dynamic World Modeling via Coupled Panoramic Evolution

Zhilu Zhang[1] Xiaoyu Liu[2] Wenxi Liu[3] Shengfeng He[1]*
[1]Singapore Management University [2]Harbin Institute of Technology [3]Fuzhou University

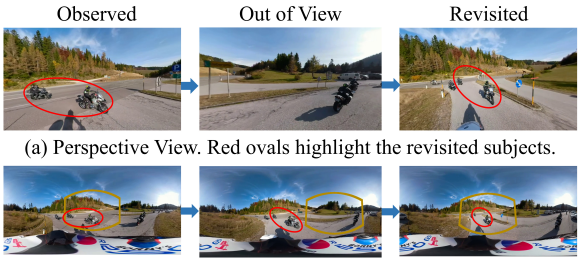
Project: https://omnipersist.github.io/

Abstract

Action-conditioned world models play a vital role in robotics and autonomous driving, yet existing video diffusion models struggle to maintain dynamic object persistence. When the camera turns away and returns, out-of-view moving entities fail to evolve plausibly in latent space, often mutating or vanishing entirely due to bounded perspective views. To resolve this issue, we propose OmniPersist, a unified framework that couples perspective video generation with panoramic evolution within a single-stage diffusion model, without relying on auxiliary models or initial panoramic inputs. To maintain efficiency, we model the panoramic stream at a lower frame rate, bridging both views across space and time via Geometric Anchoring (GeoAnch) for spatial correspondence and Temporal Compensation (TemComp) for temporal aggregation. OmniPersist is agnostic to camera-control architectures, supporting both latent-level and attention-level conditioning. Extensive experiments on a newly constructed dynamic benchmark across image-to-video and video continuation tasks demonstrate consistent gains in visual quality, camera controllability, revisit consistency, and dynamic persistence.

1. Introduction

World models [10, 34, 36, 47, 64] aim to predict future states from current observations conditioned on actions, playing an increasingly important role in robotics [20, 52] and autonomous driving [22, 44]. A prevalent paradigm generates future videos via diffusion models [3, 26, 39, 54], guided by textual prompts alongside camera trajectories or discrete motion signals, as demonstrated by LingBot-World [10, 36] and Matrix-Game 3.0 [47]. Despite demonstrating compelling visual fidelity through large-scale training, these models fundamentally struggle to maintain object persistence in dynamic scenes. When the viewpoint shifts away and subsequently returns, out-of-view objects fail to evolve plausibly in the latent space. As a result, their visual consistency and motion continuity break down, even causing returning entities to mutate or vanish entirely.



(a) Perspective View. Red ovals highlight the revisited subjects.

(b) Panoramic Evolution. Yellow boxes denote the perspective viewport.

Figure 1. **An example of dynamic subjects across a camera revisit.** While moving subjects temporarily exit the perspective view, the 360° panoramic view continuously captures their motion outside the perspective viewport. Best viewed when zoomed in.

We attribute this limitation to both training data biases and fundamental modeling bottlenecks. On the data side, current training data for dynamic scenes heavily favor continuous tracking, rarely featuring scenarios where moving entities exit and subsequently re-enter the camera's field of view (FoV). On the modeling side, vanilla diffusion models are inherently limited in preserving out-of-view dynamics without dedicated constraints and priors. Recent works have recognized this challenge and sought to mitigate it mainly from two perspectives. Some methods utilize retrieval frames [6] and compressed implicit contexts [56] as memory. Others attempt to predict dynamics of specific entity through auxiliary generative models [8] and LLMs [40]. However, the former often recall static appearance cues that fail to reflect continuous natural motion, while the latter introduce heavy system complexity and scale poorly to multi-entity scenarios. Furthermore, these methods are designed primarily for object consistency across temporal chunks. In fact, even within a single video chunk, combining intra-window bidirectional attention with advanced camera control approaches (*e.g.*, UCPE [60]) remains insufficient to

*Corresponding author.

sustain out-of-view dynamic evolution.

To address this problem, we propose modeling scene evolution in panoramic space, leveraging complete 360° spatial context to improve the persistence of dynamic world models. This paradigm offers two distinct advantages. First, conceptually, the omnidirectional nature of panoramic space ensures that when an object leaves the camera view, its continuous motion can still be explicitly predicted on a global canvas. Moreover, this formulation naturally supports multiple moving objects and inherently preserves background memory, as shown in Fig. 1. Second, practically, panoramic scenes can be represented as standard 2D frames via equirectangular projection (ERP), allowing existing video diffusion foundation models to adapt seamlessly to panoramic data while fully capitalizing on their strong pretrained priors.

Specifically, we propose **OmniPersist**, which jointly models the perspective camera view and its panoramic counterpart within a single video diffusion model. This unified framework achieves persistent dynamic world modeling in an end-to-end manner, without relying on auxiliary models or additional input. Since generating a dense panorama for every perspective frame introduces heavy computational overhead, we propose to design the panoramic stream to operate at a lower frame rate. To effectively bridge these two representations across space and time, we introduce two complementary modules, *i.e.*, *Geometric Anchoring (GeoAnch)* and *Temporal Compensation (TemComp)*. In particular, *GeoAnch* establishes explicit geometric correspondences between the perspective and panoramic views, enabling direct feature exchange at synchronized timestamps. Meanwhile, *TemComp* aggregates features across neighboring frames, effectively compensating for the temporal discrepancy caused by the differing frame rates. Concurrently, AlayaVista [33] (within the *past two months*) also explores panoramic representations, but relies on a multi-stage pipeline of panoramic initialization, state evolution, and perspective refinement. In contrast, OmniPersist avoids generating redundant full-rate panoramas, establishing a single-stage, unified framework for persistent dynamic world modeling.

To ensure broad applicability, OmniPersist is designed to be agnostic to underlying camera-control architectures. We integrate our framework with two representative paradigms, *i.e.*, latent-level trajectory injection via CameraCtrl [11] and attention-level pose conditioning via UCPE [60]. Furthermore, to benchmark dynamic persistence under complex camera movements, we construct a paired panoramic–perspective dataset featuring diverse trajectory patterns. We systematically evaluate OmniPersist under two settings, namely image-to-video (I2V) generation and video continuation. Extensive experiments across visual quality, camera controllability, revisit consistency, and dynamic persistence demonstrate that our coupled panoramic evolution yields consistent gains across both control mechanisms.

In summary, our main contributions are as follows:

- We propose **OmniPersist**, a single-stage unified framework that jointly generates perspective video and temporally sparse panoramic evolution to improve out-of-view dynamic persistence.
- We introduce *GeoAnch* and *TemComp* to couple dense perspective and sparse panoramic streams through synchronized geometric exchange and neighboring-time feature aggregation.
- We construct a paired panoramic–perspective dynamic dataset and demonstrate consistent gains in dynamic persistence and revisit consistency across diverse camera-control mechanisms.

2. Related Work

Camera-Controlled Video Generation. Camera trajectory control has become an essential capability for interactive world models [10, 34, 36, 47, 64]. Existing efforts encompass text-to-video [1, 46], image-to-video [21, 28, 63], and video re-shooting [2, 35, 57], and roughly divide into two paradigms. (1) *Camera-parameter conditioning* directly encodes viewpoints as extrinsic matrices [46], Plücker rays [1, 11], or relative positional embeddings [17, 55, 60]. These representations are used to modulate feature latents via auxiliary branches [1, 11] or inject geometric biases into attention mechanisms [17, 55, 60]. These methods operate strictly within the visible frustum, offering no spatial substrate to sustain the evolution of out-of-view dynamic entities. (2) *Explicit 3D reprojection* grounds camera trajectories by warping reconstructed 3D geometry into target image or latent spaces. For instance, ViewCrafter [58] and GEN3C [28] condition diffusion on rendered point clouds with iterative scene expansion or 3D caching. MosaicMem [59] retrieves 3D-lifted patches from spatial memory, whereas LSM-World [43] caches 3D latent tokens directly to bypass RGB rendering and re-encoding. However, these approaches rely heavily on external depth estimators and are predominantly restricted to static scenes. In contrast, **OmniPersist** bypasses 3D reconstruction and provides an omnidirectional canvas that naturally sustains unseen dynamic evolution. It acts as an orthogonal, plug-and-play spatial substrate, seamlessly integrating with the camera-parameter conditioning paradigm.

Persistent Dynamic World Modeling. World models simulate future states conditioned on current observations and actions. A prevalent paradigm generates future videos guided by text and camera trajectories [10, 36, 47, 64]. However, maintaining the evolving states of dynamic out-of-view entities remains an unresolved challenge. Existing approaches broadly follow three directions. (1) *Memory-based modeling*. HyDRA [6] and MemLearner [56] query

or compress historical observations to preserve generation consistency. ReMind [51] employs a node-structured curriculum that compels the model to retrieve relevant historical states. While historical memory provides informative anchors, these methods inherently infer unobserved evolution from past evidence, often struggling to sustain continuous visual transitions. (2) *Auxiliary entity tracking*. LiveWorld [8] uses multiple generative monitors to advance unseen entity states within a persistent scene, while WorldDirector [40] employs an LLM to predict object trajectories prior to visual generation. However, relying on external entity-level monitors or LLM planners introduces heavy system overhead, causes pipeline brittleness, and scales poorly to dense, multi-entity interactive scenarios. (3) *Panoramic world-state modeling*. In concurrent work, AlayaVista [33] first evolves panoramic latent states and then synthesizes perspective observations through viewport rendering and refinement stages. However, it relies on a separate panorama initialization stage and incurs high computational costs by predicting full-rate panoramas. In contrast, **OmniPersist** derives its panoramic conditions directly from the observed perspective prefix and predicts panoramic states at a lower temporal rate, establishing a single-stage joint diffusion framework.

Panoramic Video Generation. Panoramic video generation aims to synthesize temporally coherent 360° content, typically achieved by adapting pretrained perspective video diffusion models. Therein, perspective-to-panorama methods expand input videos while preserving their content and dynamics [19, 24, 48, 50], whereas text- and image-conditioned approaches synthesize omnidirectional scenes and motion [9, 42, 49, 50, 61]. Recent methods further extend panoramic synthesis to camera-controlled exploration [18, 32] and 3D world construction [7, 41, 53]. Technically, these methods address spherical distortion and boundary seams through ERP-aware designs [32, 48, 49], cross-view coherence via multi-view attention [9, 50], and exploration consistency through geometric guidance and spatial memory [7, 18, 41, 53]. Collectively, these advances establish the feasibility of using video diffusion models to evolve panoramic states. Unlike prior works targeting standalone panoramic synthesis, **OmniPersist** couples an efficient, low-rate panoramic stream with perspective rollouts to anchor out-of-view dynamic persistence.

3. Method

3.1. Overview

We study persistent dynamic world modeling where dynamic subjects may leave and subsequently re-enter the camera field of view. Given a textual prompt, an observed perspective prefix $\mathbf{X}_{0:K-1}$, and a T-frame camera trajectory $\mathbf{C}^{\mathrm{x}}$ with known intrinsics, the objective is to synthesize the remaining frames of the T-frame perspective video $\mathbf{X}$. Crucially, returning subjects must maintain identity consistency and plausible physical motion throughout the interval spent outside the viewport.

OmniPersist addresses this task by jointly generating $\mathbf{X}$ alongside a temporally sparse panoramic video $\mathbf{P}$. The panoramic stream captures scene evolution across all viewing directions, while the perspective stream adheres to the prescribed camera trajectory (see Fig. 1). Both streams share a single diffusion transformer backbone and exchange features via Geometric Anchoring (GeoAnch) and Temporal Compensation (TemComp), as illustrated in Fig. 2. Both streams rely solely on the perspective prefix, text prompt, and camera parameters, without requiring any auxiliary depth, panoramic images, or video inputs.

3.2. Coupled Panoramic Evolution

We jointly denoise the dense perspective stream and the panoramic stream using a shared pretrained video diffusion transformer [39]. Because predicting a full-rate panoramic stream incurs prohibitive computational overhead, we design the panoramic stream to operate at a lower frame rate. With a panoramic sampling interval s, panoramic frame $\mathbf{P}_j$ corresponds to perspective frame $\mathbf{X}_{sj}$ in a T-frame video, yielding $J = (T-1)/s + 1$ states over the same temporal horizon. These states can represent scene dynamics across all viewing directions around the camera, including regions outside the perspective viewport.

Joint Denoising of Perspective and Panoramic Streams. For each stream, the noisy latents and corresponding VAE-encoded [16] visual conditions are concatenated channel-wise and mapped into token sequences via view-specific input projections. The resulting token sequences from both perspective and panoramic streams are then concatenated along the sequence dimension. Joint spatiotemporal self-attention subsequently processes both streams under a shared text prompt and diffusion timestep, after which a shared prediction head outputs the flow fields to be split and reshaped back into their respective latent grids. Retaining per-view spatial RoPE [30], we assign temporal RoPE coordinates i and sj to perspective and panoramic latent frames i and j (corresponding to original frame timestamps $4i$ and $4sj$ under $4\times$ temporal VAE compression), aligning synchronized states.

Visual Conditions. We construct two visual conditions from the observed perspective prefix $\mathbf{X}_{0:K-1}$. The perspective condition comprises the VAE encoding of the prefix concatenated with an indicator mask denoting known frames. To construct the panoramic condition, the observed frames are reprojected onto equirectangular projection (ERP) coordinates using camera poses and intrinsics, encoded via the VAE, and concatenated with a spatial visibility mask. Unobserved future frames and unmapped ERP

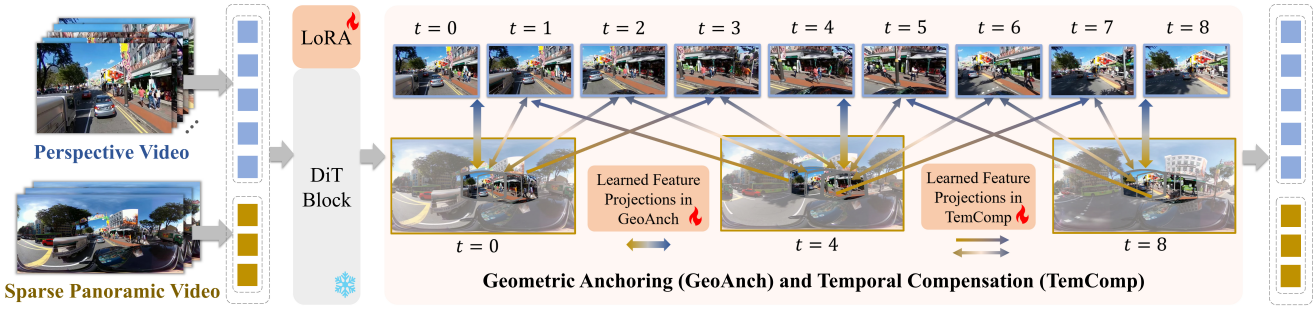


Figure 2. **Overview of the OmniPersist framework.** Tokens from the dense perspective stream and the temporally sparse panoramic stream are jointly processed by a shared, frozen Diffusion Transformer (DiT) block adapted via LoRA. The two streams coordinate across space and time via dedicated learned feature projections to exchange complementary context. GeoAnch transfers spatially aligned features at synchronized timestamps, while TemComp aggregates features across neighboring timestamps for intermediate observations. For illustration, we depict a temporal sampling interval of $s = 4$ in image-space indices.

regions are filled with zeros prior to VAE encoding. Crucially, both condition tensors are derived from the single perspective prefix, requiring no auxiliary visual inputs.

Camera Conditions. We align the initial panoramic coordinate frame with the first perspective view (centered at the ERP center), so both cameras share the same initial orientation. Representing camera-to-world poses as $\mathbf{C}_t^{\mathrm{x}}, \mathbf{C}_j^{\mathrm{p}} \in \mathbb{R}^{4\times4}$ and the pure-rotation viewport transformation from the perspective to the panoramic frame as $\mathbf{V}_t \in \mathbb{R}^{4\times4}$ (with $\mathbf{V}_0 = \mathbf{I}$), the panoramic pose is obtained as $\mathbf{C}_j^{\mathrm{p}} = \mathbf{C}_{sj}^{\mathrm{x}}\mathbf{V}_{sj}^{-1}$. This operation eliminates viewport rotation while preserving shared camera centers, yielding $J = (T-1)/s + 1$ panoramic poses when subsampling every s frames. During training, $\mathbf{V}_t$ is known from the viewport rotations used to render perspective clips from source panoramas. At inference, the panoramic camera retains the initial perspective orientation $\mathbf{R}_0^{\mathrm{x}} \in \mathbb{R}^{3\times3}$, setting the rotation component of $\mathbf{V}_t$ to $(\mathbf{R}_0^{\mathrm{x}})^\top\mathbf{R}_t^{\mathrm{x}}$. Consequently, all required poses are determined solely by the input perspective trajectory without additional camera inputs.

3.3. Geometric Anchoring

Because spatial RoPE [30] in diffusion transformers operates independently on separate image grids, it does not explicitly capture geometric correspondences between perspective and panoramic views. Although spherical positional embeddings, such as SpheRoPE [13], offer an alternative formulation, directly replacing pretrained spatial RoPE with spherical coordinates disrupts learned spatial priors and impedes joint optimization. We therefore introduce Geometric Anchoring (GeoAnch) to establish explicit cross-view correspondences while preserving the pretrained spatial encoding.

Hereafter, i and j index perspective and panoramic latent frames (corresponding to RGB frames $4i$ and $4j$), respectively. Let $\mathbf{q}$ denote the image-space patch center of a perspective token, and let $\mathbf{A}_i \in \mathbb{R}^{3\times3}$, $\mathbf{R}_i^{\mathrm{x}} \in \mathbb{R}^{3\times3}$, and $\mathbf{R}_j^{\mathrm{p}} \in \mathbb{R}^{3\times3}$ denote the corresponding camera intrinsics and rotations. At synchronized timestamps ($i = sj$), both views share an optical center, yielding the mapping:

$$\pi_{i\to j}(\mathbf{q}) = \Pi_{\mathrm{erp}}\left((\mathbf{R}_j^{\mathrm{p}})^\top\mathbf{R}_i^{\mathrm{x}}\mathbf{A}_i^{-1}\widetilde{\mathbf{q}}\right), \tag{1}$$

where $\widetilde{\mathbf{q}}$ is the homogeneous image coordinate and Π_{erp} projects a 3D ray onto the panoramic grid. Conversely, panoramic directions that fall within the perspective field of view are mapped back to the perspective grid using the inverse geometric transformation.

Based on these geometric correspondences, GeoAnch bilinearly samples aligned features from the other view, concatenates them with the current view's features, and estimates a residual update using a learned linear layer. Because the perspective camera observes only a sub-region of the visual sphere, updates to the panoramic stream are masked to the perspective field of view. In addition, feature sampling on the ERP grid employs horizontal circular wrapping to preserve spherical boundary continuity.

3.4. Temporal Compensation

Due to the temporal sparsity of the panoramic stream, intermediate perspective frames lack synchronized panoramic anchors. Temporal Compensation (TemComp) bridges this gap through bidirectional cross-temporal feature aggregation. For each intermediate perspective frame, TemComp performs interpolation between its two adjacent panoramic states. Conversely, each panoramic frame aggregates a local temporal window of perspective features centered at its synchronized timestamp to learn a correction relative to its synchronized geometric anchor.

Let $\mathbf{g}$ denote a query patch feature representation of the current frame. Following the ray projection in Eq. (1), we map each patch location of the current frame into candidate frames of the other view, bilinearly sampling features to obtain $\widetilde{\mathbf{h}}_k$ at candidate timestamp k. Candidate features are

aggregated as:

$$\bar{\mathbf{h}}(\mathbf{g}) = \sum_{k \in \mathcal{N}(\mathbf{g})} \alpha_k \widetilde{\mathbf{h}}_k, \quad \alpha_k \propto \exp\Big(\cos(\mathbf{g}, \widetilde{\mathbf{h}}_k)/\theta\Big), \tag{2}$$

where $\mathcal{N}(\mathbf{g})$ spans two bracketing panoramic frames for perspective queries, or up to $2\lfloor s/2 \rfloor + 1$ centered perspective frames for panoramic queries. The aggregation weights α_k are normalized to sum to one, and $\theta = 0.2$ serves as the temperature hyperparameter. The geometric mapping provides approximate directional correspondences to guide candidate sampling, while feature similarity suppresses incompatible evidence across time.

To update representations, TemComp incorporates a learned linear residual layer for each exchange direction, preserving base features while attenuating unreliable contributions. Specifically, synchronized perspective frames retain their direct GeoAnch updates, whereas intermediate perspective frames receive a residual computed from the aggregated representation $\bar{\mathbf{h}}$. TemComp is thus formulated as a conservative local feature correction around synchronized geometric anchors, enabling the model to selectively exploit neighboring temporal observations.

3.5. Training and Inference

Perspective frames are directly encoded using the pretrained VAE [16]. For panoramic targets, we adopt circular latent encoding [48], which applies horizontal circular padding to ERP frames prior to VAE encoding and subsequently crops the padded latent margins to restore the original grid dimensions. This provides a seamless latent representation across horizontal boundaries for diffusion training.

Flow matching and loss optimization operate directly on these latent spaces. For each view $v \in \{\mathrm{x}, \mathrm{p}\}$, we sample independent Gaussian noise $\boldsymbol{\epsilon}^v \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ at a shared noise level $\sigma \in [0, 1]$ to define the linear interpolation and velocity target:

$$\mathbf{z}_\sigma^v = (1-\sigma)\mathbf{z}^v + \sigma\boldsymbol{\epsilon}^v, \qquad \mathbf{u}^v = \boldsymbol{\epsilon}^v - \mathbf{z}^v. \tag{3}$$

The diffusion transformer jointly predicts the velocity fields $\hat{\mathbf{u}}^{\mathrm{x}}$ and $\hat{\mathbf{u}}^{\mathrm{p}}$. We train the model by minimizing the weighted objective:

$$\mathcal{L} = \mathbb{E}_{\sigma,\boldsymbol{\epsilon}}\left[\frac{w(\sigma)}{2}\left(\|\hat{\mathbf{u}}^{\mathrm{x}} - \mathbf{u}^{\mathrm{x}}\|^2 + \|\mathbf{W}_{\mathrm{p}} \odot (\hat{\mathbf{u}}^{\mathrm{p}} - \mathbf{u}^{\mathrm{p}})\|^2\right)\right], \tag{4}$$

where $w(\sigma)$ is the scheduler weight and $\mathbf{W}_{\mathrm{p}}$ denotes a normalized, latitude-dependent loss weight for the panoramic canvas. During training, the LoRA [14] parameters, input projections, camera conditioning modules, and cross-view interaction layers are optimized jointly, while the VAE and text encoder remain frozen.

During inference, the shared VAE can decode both streams. When panoramic visualization is not explicitly required, only the perspective stream is decoded, incurring zero additional decoding overhead.

4. Experiments

4.1. Experimental Setup

Dataset Construction. Unlike HyDRA [6], which constructs its training and evaluation sets within simulators (*i.e.*, Unreal Engine 5) that exhibit limited visual realism, we curate our dataset directly from real-world environments. We construct paired panoramic and perspective videos by simulating camera viewing trajectories over panoramic footage from the 360-1M dataset [38]. Following the filtering protocols in Argus [24] and UCPE [60], ∼125K high-quality video clips are selected. Each clip consists of 81 frames at 16 fps, with a perspective resolution of 832×480 and a panoramic resolution of 1024×512. We design four distinct camera trajectory types: object tracking, single revisit, double revisit, and a full camera rotation of ∼360°, covering both continuous subject following and leave-and-revisit patterns. Perspective viewpoints retain the camera-center trajectory of the original panoramic video while dynamically varying the viewing orientation. The evaluation set contains 300 videos whose source footage is strictly excluded from the training split. Details of the construction pipeline are provided in the *suppl*.

Implementation Details. Under the image-to-video (I2V) protocol, generation is conditioned on a single perspective frame ($K = 1$), whereas the video continuation setting conditions on 21 frames ($K = 21$). Both settings target the generation of 81-frame videos. OmniPersist is implemented based on the Wan2.2-TI2V-5B backbone [39]. We instantiate OmniPersist across two representative camera-control paradigms, applying rank-32 LoRA [14] to linear layers in the self-attention and feed-forward blocks. Under CameraCtrl [11], view-specific ray adapters inject the camera features of the two views into their respective input tokens. Under UCPE [60], camera-attention layers process both views according to their ray geometries under a unified coordinate and scale reference. GeoAnch and TemComp are embedded after each DiT Block. All OmniPersist variants adopt mixed-interval sampling, uniformly drawing $s \in \{1, 2, 4, 10\}$ and prefix length $K \in \{1, 21\}$ per example. Using AdamW [23] with a learning rate of 10^{-4} and total batch size of 8 on 8 NVIDIA RTX PRO 6000 GPUs, the UCPE- and CameraCtrl-based variants converge in 20K and 80K iterations, respectively.

Comparison Setup. We benchmark against world models developed through large-scale, dedicated pretraining, *i.e.*, DreamX-World [34], Matrix-World 3.0 [47], SANA-WM [64], and LingBot-World 2.0 [10]. We evaluate their released checkpoints configured with bidirectional attention. Beyond these, camera-conditioned video generation serves as a foundational paradigm for world models. To ensure a fair comparison, we train CameraCtrl [11] and UCPE [60] using the same dataset, training settings,

Table 1. **Image-to-video (I2V) comparison.** † indicates training on our dataset with the same Wan2.2-TI2V-5B backbone and LoRA settings. Unmarked methods use released checkpoints directly, with parameter counts of diffusion models in parentheses.

Methods	Overall Quality			Camera Controllability				BG Revisit Fidelity			Dynamic Subject Persistence			
	FID↓	FVD↓	FVD$_c$↓	Rot↓	Trans↓	CamMC↓	ATE↓	PSNR↑	SSIM↑	LPIPS↓	ID↑	R-ID↑	Vis.↑	Traj.↓
HyDRA (1.3B) [6]	48.51	857.4	763.8	35.949	0.707	1.152	0.805	11.09	0.3986	0.676	0.6725	0.1319	0.6376	0.4876
Matrix-World 3.0 (5B) [47]	41.18	688.6	654.9	28.952	0.663	1.008	0.748	9.74	0.3747	0.693	0.5570	0.1643	0.5739	0.6456
DreamX-World (5B) [34]	26.28	392.3	368.4	24.743	0.569	0.876	0.636	11.78	0.4412	0.638	0.6044	0.2075	0.5916	0.5621
SANA-WM (2.6B+17B) [64]	33.28	432.9	402.4	20.116	0.590	0.819	0.655	12.22	0.4488	0.644	0.6108	0.1278	0.6108	0.5262
LiveWorld (14B) [8]	24.53	335.4	334.4	16.252	0.439	0.655	0.492	12.46	0.4664	0.595	0.6035	0.2928	0.6229	0.5016
LingBot-World 2.0 (28B) [10]	22.64	383.8	389.4	13.560	0.422	0.587	0.470	12.42	0.4645	0.609	0.6964	0.3360	0.7303	0.3853
CameraCtrl† [11]	23.60	288.5	295.1	13.086	0.531	0.677	0.594	12.29	0.4415	0.618	0.5834	0.2325	0.6519	0.4753
[11]+OmniPersist ($s = 1$)	22.64	258.6	266.8	6.145	0.422	0.490	0.470	13.47	0.4749	0.575	0.6525	0.2653	0.7375	0.3779
[11]+OmniPersist ($s = 2$)	22.54	253.8	256.0	5.586	0.415	0.474	0.463	13.52	0.4749	0.576	0.6411	0.2933	0.7340	0.3759
[11]+OmniPersist ($s = 4$)	22.18	262.3	252.6	5.556	0.399	0.458	0.446	13.58	0.4754	0.573	0.6395	0.3355	0.7290	0.3830
[11]+OmniPersist ($s = 10$)	22.91	266.6	276.4	5.695	0.418	0.478	0.465	13.37	0.4731	0.580	0.6429	0.2840	0.7333	0.3783
UCPE† [60]	25.33	333.7	325.7	8.127	0.431	0.526	0.481	12.80	0.4696	0.580	0.6043	0.2097	0.6561	0.4897
[60]+OmniPersist ($s = 1$)	21.41	274.1	262.2	4.236	0.382	0.426	0.427	13.91	0.4890	0.543	0.6435	0.3183	0.7348	0.4003
[60]+OmniPersist ($s = 2$)	21.90	264.6	261.5	4.490	0.374	0.425	0.418	13.99	0.4898	0.543	0.6416	0.3839	0.7412	0.3828
[60]+OmniPersist ($s = 4$)	21.89	272.7	264.0	4.976	0.376	0.429	0.422	13.90	0.4872	0.546	0.6486	0.3694	0.7351	0.3791
[60]+OmniPersist ($s = 10$)	22.57	285.8	284.7	5.670	0.365	0.430	0.408	13.76	0.4825	0.553	0.6283	0.3116	0.7338	0.4090

Table 2. **Video continuation comparison.** † indicates training on our dataset with the same Wan2.2-TI2V-5B backbone and LoRA settings. Unmarked methods use released checkpoints directly, with parameter counts of diffusion models in parentheses.

Methods	Overall Quality			Camera Controllability				BG Revisit Fidelity			Dynamic Subject Persistence			
	FID↓	FVD↓	FVD$_c$↓	Rot↓	Trans↓	CamMC↓	ATE↓	PSNR↑	SSIM↑	LPIPS↓	ID↑	R-ID↑	Vis.↑	Traj.↓
HyDRA (1.3B) [6]	61.30	779.7	711.2	41.420	0.724	1.259	0.787	11.30	0.3955	0.674	0.3102	0.0850	0.3816	0.7474
Matrix-World 3.0 (5B) [47]	36.94	427.2	426.8	25.654	0.588	0.922	0.644	10.86	0.4040	0.666	0.3386	0.1353	0.4229	0.7710
SANA-WM (2.6B+17B) [64]	32.75	320.8	330.6	22.923	0.546	0.855	0.588	12.86	0.4578	0.606	0.4019	0.2039	0.5006	0.6568
LiveWorld (14B) [8]	25.45	286.7	295.0	12.912	0.425	0.605	0.461	13.26	0.4797	0.556	0.4325	0.1762	0.4824	0.6358
CameraCtrl† [11]	22.61	227.5	223.5	11.704	0.451	0.597	0.489	13.43	0.4701	0.579	0.4946	0.2494	0.6053	0.5161
[11]+OmniPersist ($s = 1$)	21.87	197.2	201.8	4.831	0.403	0.459	0.439	15.28	0.5245	0.511	0.5667	0.4093	0.7042	0.4117
[11]+OmniPersist ($s = 2$)	21.70	196.6	202.9	4.690	0.401	0.456	0.433	15.31	0.5263	0.514	0.5655	0.4062	0.6768	0.4177
[11]+OmniPersist ($s = 4$)	21.59	194.9	201.9	4.345	0.401	0.450	0.433	15.22	0.5238	0.515	0.5534	0.3771	0.6678	0.4289
[11]+OmniPersist ($s = 10$)	21.69	201.0	200.6	4.661	0.401	0.455	0.435	15.20	0.5221	0.518	0.5559	0.3492	0.6835	0.4287
UCPE† [60]	23.60	264.6	254.7	6.711	0.443	0.519	0.476	14.55	0.5136	0.523	0.4932	0.2778	0.6354	0.5255
[60]+OmniPersist ($s = 1$)	20.53	196.7	197.9	3.265	0.338	0.375	0.367	15.90	0.5446	0.478	0.5504	0.3669	0.6759	0.4280
[60]+OmniPersist ($s = 2$)	21.04	197.6	202.4	3.863	0.341	0.388	0.369	15.79	0.5406	0.484	0.5580	0.4667	0.6829	0.4173
[60]+OmniPersist ($s = 4$)	21.41	210.3	212.6	4.169	0.342	0.393	0.372	15.72	0.5385	0.484	0.5577	0.3849	0.6800	0.4245
[60]+OmniPersist ($s = 10$)	22.48	218.1	219.1	4.353	0.354	0.407	0.384	15.39	0.5234	0.496	0.5126	0.4045	0.6649	0.4618

Wan2.2-TI2V-5B backbone, and LoRA configuration as OmniPersist. Furthermore, we include comparisons with HyDRA [6] and LiveWorld [8], which are specifically designed for out-of-sight dynamic persistence. During inference, all diffusion models employ 50 denoising steps. Some models are evaluated solely on I2V, due to incompatibility with video continuation tasks.

Evaluation Protocol. We evaluate performance across 14 perspective-video metrics structured into four groups. (1) Overall visual quality is measured via FID [12], FVD [37], and center-crop FVD (FVD$_c$). (2) Camera controllability is evaluated via rotation (Rot) and translation (Trans) errors [11], CamMC [46], and absolute trajectory error (ATE) on ViPE-reconstructed [15] poses anchored at frame zero. Following UCPE [60], predicted and reference translations are independently normalized by their maximum trajectory radius. (3) Background revisit fidelity is evaluated using PSNR, SSIM [45], and LPIPS [62] between generated and ground-truth frames at camera-selected revisit timestamps, masked by intersected static backgrounds. (4) Dynamic persistence is assessed via foreground DINOv2 [25] feature similarity for overall subject identity (ID) and re-entry identity (R-ID), visibility F1-score (Vis.), and normalized subject trajectory error (Traj.). These metrics evaluate whether a dynamic subject successfully returns, retains its appearance, and follows the ground-truth trajectory (the latter being more informative for video continuation). Detailed metric implementations are provided in the *suppl*.



Figure 3. **Generation results on I2V across camera revisits.** From top to bottom, we show a frame before the camera turns away and two revisit frames after it returns to illustrate identity preservation and continued motion. Frames in each row share the same frame index. Prompt overview: *a video shows two men walking along a paved path next to a road.*

4.2. Quantitative and Qualitative Results

We adopt $s = 4$ as the primary setting for both quantitative and qualitative evaluations.

Quantitative Comparisons. As shown in Table 1, under the I2V task, OmniPersist consistently improves all 14 metrics over the respective baselines for both CameraCtrl [11] and UCPE [60]. Specifically, re-entry identity increases from 0.2325 to 0.3355 with CameraCtrl and from 0.2097 to 0.3694 with UCPE. Camera rotation and translation errors drop from 13.086 / 0.531 to 5.556 / 0.399 for CameraCtrl, and from 8.127 / 0.431 to 4.976 / 0.376 for UCPE. FVD under UCPE improves from 333.7 to 272.7. Notably, even when compared against larger released world models, including LiveWorld [8] (14B, which explicitly models multi-entity evolution by additional diffusion models) and LingBot-World 2.0 [10] (28B), the UCPE-based OmniPersist outperforms LiveWorld across all 14 metrics and surpasses LingBot-World 2.0 on 13 of the 14 metrics.

As reported in Table 2, under the video continuation task, both variants again surpass their respective baselines across all 14 metrics. In particular, re-entry identity increases from 0.2494 to 0.3771 with CameraCtrl and from 0.2778 to 0.3849 with UCPE. Subject trajectory error decreases from 0.5161 to 0.4289 with CameraCtrl and from 0.5255 to 0.4245 with UCPE. Moreover, CameraCtrl-based OmniPersist improves background revisit PSNR by 1.79 dB and camera rotation error by 7.359. Both variants also outperform HyDRA [6], LiveWorld [8], Matrix-World 3.0 [47], and SANA-WM [64] across all 14 metrics.

In summary, OmniPersist achieves consistent gains in overall visual quality, camera controllability, background revisit fidelity, and dynamic subject persistence across both generation tasks and camera control mechanisms.

Qualitative Comparisons. Figure 3 demonstrates that OmniPersist maintains strong consistency for both static backgrounds and dynamic subjects across a camera revisit. Specifically, the road layout and surrounding vegetation remain structurally coherent, while both pedestrians preserve their clothing details and recognizable appearances upon re-entry. Compared with perspective-only baselines, our results better preserve both subjects and their continued walking motion after the camera returns. More qualitative results and panoramic visualization are provided in the *suppl*.

4.3. Ablation Studies

We first isolate the contributions of panoramic evolution, GeoAnch, and TemComp using CameraCtrl [11] at $s = 4$. We then analyze the effect of panoramic sampling density and evaluate the computational overhead of our framework.

Effect of Joint Panoramic Evolution. As shown in Table 3, compared to the perspective-only baseline (w/o Pano), adding panoramic support (w/ Pano) improves re-entry identity (R-ID) from 0.2325 to 0.3256 for I2V and from 0.2494 to 0.3167 for video continuation. However, visual quality degrades inconsistently without explicit cross-view alignment: I2V FID worsens from 23.60 to 26.16, while video continuation FVD increases from 227.5 to 246.0. When explicit spatial and temporal coupling are incorporated, these scores improve substantially to 22.18 and 194.9, respectively. These results indicate that although the panoramic stream supplies essential information about returning subjects, naive joint modeling alone is insufficient to maintain generation quality.

Effect of GeoAnch and TemComp. Adding GeoAnch improves all four camera-control and three background-revisit metrics over panoramic modeling alone (w/ Pano) in both tasks (see Table 3). For I2V, rotation error drops from 9.000 to 6.604 and background PSNR rises from 12.69 to 13.63 dB, while trajectory error decreases from 0.4645 to 0.3960 for I2V and from 0.5201 to 0.4956 for video continuation. However, R-ID remains inconsistent, dropping from 0.3256 to 0.2794 for I2V and changing marginally for continuation. These results confirm the value of GeoAnch, yet show that synchronized correspondence alone is insufficient for identity recovery after out-of-view intervals.

Table 3. **Component ablation across both tasks.** The evaluations are conducted using CameraCtrl [11] on $s = 4$.

Configuration	Overall Quality			Camera Controllability				BG Revisit Fidelity			Dynamic Subject Persistence			
	FID↓	FVD↓	FVD$_c$↓	Rot↓	Trans↓	CamMC↓	ATE↓	PSNR↑	SSIM↑	LPIPS↓	ID↑	R-ID↑	Vis.↑	Traj.↓
Image-to-video ($K = 1$)														
w/o Pano	23.60	288.5	295.1	13.086	0.531	0.677	0.594	12.29	0.4415	0.618	0.5834	0.2325	0.6519	0.4753
w/ Pano	26.16	288.1	281.0	9.000	0.467	0.562	0.519	12.69	0.4473	0.612	0.5975	0.3256	0.6700	0.4645
w/ Pano & GeoAnch	23.07	275.2	265.9	6.604	0.438	0.508	0.493	13.63	0.4817	0.589	0.6163	0.2794	0.7245	0.3960
w/ Pano & GeoAnch & TemComp	22.18	262.3	252.6	5.556	0.399	0.458	0.446	13.58	0.4754	0.573	0.6395	0.3355	0.7290	0.3830
Video continuation ($K = 21$)														
w/o Pano	22.61	227.5	223.5	11.704	0.451	0.597	0.489	13.43	0.4701	0.579	0.4946	0.2494	0.6053	0.5161
w/ Pano	27.20	246.0	232.7	8.453	0.462	0.560	0.497	13.64	0.4682	0.581	0.4890	0.3167	0.6104	0.5201
w/ Pano & GeoAnch	23.89	228.5	227.0	5.905	0.424	0.493	0.456	14.82	0.5162	0.547	0.5007	0.3173	0.6465	0.4956
w/ Pano & GeoAnch & TemComp	21.59	194.9	201.9	4.345	0.401	0.450	0.433	15.22	0.5238	0.515	0.5534	0.3771	0.6678	0.4289

Table 4. Per-step denoising cost. Each entry reports FLOPs (T) / time (ms) / peak allocated memory (GB). The first and second rows correspond to CameraCtrl [11] and UCPE [60], respectively.

Baseline	+OmniPersist			
	$s = 10$	$s = 4$	$s = 2$	$s = 1$
102 / 420 / 11.2	128 / 547 / 12.2	155 / 670 / 12.3	203 / 838 / 12.5	313 / 1185 / 13.1
103 / 442 / 10.2	131 / 571 / 10.9	159 / 689 / 11.0	209 / 865 / 11.2	326 / 1249 / 11.6

Incorporating TemComp further improves all visual quality, camera-control, and dynamic-persistence metrics over the GeoAnch-only variant across both tasks. Specifically, R-ID rises from 0.2794 to 0.3355 for I2V and from 0.3173 to 0.3771 for video continuation, with trajectory error dropping from 0.3960 to 0.3830 and from 0.4956 to 0.4289. Although the GeoAnch-only variant achieves slightly higher I2V background PSNR and SSIM, the full model attains the best overall quality, camera control, and dynamic persistence in both tasks, confirming the complementary roles of GeoAnch and TemComp.

Effect of Panoramic Sampling Density. We evaluate each mixed-interval checkpoint across $s \in \{1, 2, 4, 10\}$ without retraining, varying the panoramic length over 81, 41, 21, and 9 frames (see Tables 1 and 2). For CameraCtrl, $s = 4$ achieves the lowest FID and rotation error in both tasks, whereas denser sampling benefits dynamic-persistence metrics. Conversely, UCPE benefits more from denser panoramic prediction, with $s = 1$ yielding the lowest FID and rotation error. Meanwhile, $s = 2$ maximizes R-ID in both tasks, increasing from 0.3694 at $s = 4$ to 0.3839 for I2V and from 0.3849 to 0.4667 for continuation. Although trends are non-monotonic across metrics, even $s = 10$ (9 frames) improves all 14 metrics over the baselines for both interfaces and tasks.

Analysis of Computational Cost. Table 4 reports per-step denoising FLOPs, runtime, and peak memory for an $81 \times 480 \times 832$ perspective video and an $(80/s+1) \times 512 \times 1024$ panoramic video, including camera-control and cross-view modules. Sparse panoramic evolution achieves substantial gains with modest overhead. At $s = 10$, the CameraCtrl and UCPE variants increase FLOPs by only 25.5% and 27.2%, runtime by 30.2% and 29.2%, and peak memory by only 1.0 GB and 0.7 GB over their baselines. At $s = 4$, they reduce FLOPs by 50.5% and 51.2% and runtime by 43.5% and 44.8% compared with $s = 1$, while adding only 1.1 GB and 0.8 GB in memory over the baselines. Furthermore, GeoAnch and TemComp each add $\sim$1 TFLOP (less than 0.7% of the total cost) in $s = 4$, maintaining efficiency. Thus, OmniPersist offers a practical trade-off among visual fidelity, re-entry consistency, and computation.

5. Conclusion

In this paper, we address the critical challenge of dynamic object persistence in camera-controlled world models, where out-of-view moving entities frequently mutate or vanish upon viewpoint revisits. To resolve this limitation, we present OmniPersist, a single-stage framework that couples perspective video generation with 360° panoramic evolution within a unified diffusion model, requiring no initial panoramic images or auxiliary models. By operating the panoramic stream at a lower frame rate, OmniPersist efficiently supports plausible evolution of out-of-view content. We seamlessly bridge the two streams across space and time via Geometric Anchoring (GeoAnch) for synchronous spatial correspondence and Temporal Compensation (TemComp) for cross-frame temporal aggregation. Extensive experiments show that OmniPersist achieves consistent performance gains across both latent-level (CameraCtrl) and attention-level (UCPE) camera conditioning in visual fidelity, camera controllability, background revisit consistency, and dynamic persistence.

Limitations and Future Work. While OmniPersist sustains dynamic persistence during inference without requiring panoramic inputs, its training currently relies on panoramic video data. Nevertheless, with the increasing prevalence of omnidirectional capture hardware and realistic 3D simulation platforms, high-quality panoramic video assets are expected to expand rapidly. In future work, we plan to extend this framework toward real-time streaming and long-horizon video generation, and investigate its applications in robotic navigation.

OmniPersist: Persistent Dynamic World Modeling via Coupled Panoramic Evolution

Supplementary Material

The content of the appendix involves:

- Sec. A describes the dataset construction pipeline.
- Sec. B specifies evaluation metric implementations.
- Sec. C details the implementation of our framework.
- Sec. D provides additional ablation studies, quantitative evaluations, and qualitative results.

A. Dataset Construction

A.1. Overview of the Pipeline

Our curation pipeline converts high-resolution raw panoramic footage into synchronized pairs of equirectangular projection (ERP) and perspective video clips. The panoramic stream is standardized by downsampling the footage to a resolution of 1024×512, with its camera-to-world poses calibrated using ViPE [15].

To simulate realistic camera viewing trajectories with intentional leave-and-revisit patterns, we ground camera motion on moving foreground subjects. Specifically, we employ SAM 3.1 [4] to track dynamic entities directly across the ERP frames. The initial camera orientation is anchored to a reliably tracked subject exhibiting sufficient pixel area near the panoramic center, establishing a stable visual prefix. If the subject is initially located off-center, we apply a horizontal rotation to the panorama to bring it to the central meridian. Guided by the motion paths of tracked subjects, we synthesize four categories of camera trajectories: object tracking, single revisit, double revisit, and full $\sim$360° camera rotations.

Perspective video clips are rendered from the panoramic footage along the generated viewing trajectories and scaled to a resolution of 832×480. To preserve visual fidelity and respect source-resolution constraints, the horizontal field of view (FoV) is capped at a maximum of 120°, while the minimum horizontal FoV is restricted to ensure that the natively sampled resolution is no less than half of 832×480 (limiting upscaling to a factor of two). The vertical FoV is determined by the $832 : 480$ aspect ratio.

The absolute camera pose of each perspective frame is computed by compounding the calibrated panoramic pose with the simulated relative viewport transformation. Specifically, at discrete frame index t, the camera-to-world poses satisfy the geometric relationship:

$$\mathbf{C}_t^{\mathrm{x}} = \mathbf{C}_t^{\mathrm{p}} \mathbf{V}_t, \qquad \mathbf{C}_t^{\mathrm{x}}, \mathbf{C}_t^{\mathrm{p}}, \mathbf{V}_t \in \mathbb{R}^{4\times4}, \tag{A}$$

Table A. **Camera trajectory programs.** The ten parameterized programs are categorized into the four semantic trajectory types described in the main paper.

Trajectory Type	Notation	Trajectory Pattern
Object tracking	$\mathcal{T}_{\mathrm{trk}}$	Continuous subject tracking
Single revisit	$\mathcal{T}_{\mathrm{sr},1}$	Horizontal pan and return
	$\mathcal{T}_{\mathrm{sr},2}$	Asymmetric turn with pitch tilt
	$\mathcal{T}_{\mathrm{sr},3}$	Vertical detour
	$\mathcal{T}_{\mathrm{sr},4}$	Return overshoot
Double revisit	$\mathcal{T}_{\mathrm{dr},1}$	Double horizontal pan and return
	$\mathcal{T}_{\mathrm{dr},2}$	Double asymmetric turn with tilt
	$\mathcal{T}_{\mathrm{dr},3}$	Double vertical detour
	$\mathcal{T}_{\mathrm{dr},4}$	Double return overshoot
Full $\sim$360° rotation	$\mathcal{T}_{\mathrm{rot}}$	$\sim$360° rotation

where $\mathbf{C}_t^{\mathrm{x}}$ denotes the absolute camera-to-world pose of the perspective frame, $\mathbf{C}_t^{\mathrm{p}}$ represents the calibrated pose of the panoramic frame, and $\mathbf{V}_t$ is the homogeneous pure-rotation matrix governing the relative viewport orientation. Both pose sequences are defined in a shared world coordinate system anchored to the coordinate frame of the initial perspective camera, such that $\mathbf{V}_0 = \mathbf{I}$.

Because $\mathbf{V}_t$ contains solely rotational components, it steers the perspective viewing direction while strictly preserving the optical center trajectory of the source panorama. Consequently, physical translations of the panoramic capture device are faithfully preserved in the perspective clips, whereas the rendering process introduces no artificial translational offsets or synthesized motion parallax, guaranteeing physically grounded perspective observations.

A.2. Camera Trajectory Synthesis

We synthesize perspective camera trajectories across four semantic categories, defined by ten parameterized programs $\mathcal{T}$ as summarized in Table A. These categories include object tracking ($\mathcal{T}_{\mathrm{trk}}$) as a control baseline with continuous tracking, single revisit ($\mathcal{T}_{\mathrm{sr},1\ldots4}$) where the camera turns away from the subject once and returns, double revisit ($\mathcal{T}_{\mathrm{dr},1\ldots4}$) with two leave-and-return cycles, and full rotation ($\mathcal{T}_{\mathrm{rot}}$) performing an approximately 360° panoramic turn.

To generate smooth and realistic camera motion, each trajectory interpolates yaw, pitch, and roll keyframes using piecewise quintic Hermite splines, ensuring continuous angular velocity and acceleration. Crucially, the reference camera orientation dynamically follows the moving subject

rather than remaining static in world coordinates. Camera offsets are added on top of this moving reference orientation. Consequently, when the camera returns, it points to the current position of the subject instead of resetting to the initial orientation at the start of the video.

Single-revisit trajectories share an out-and-back structure with four motion patterns. Horizontal pan-and-return trajectories ($\mathcal{T}_{\text{sr},1}$) place the maximum yaw turn at the midpoint, producing symmetric departure and return phases. Asymmetric turns ($\mathcal{T}_{\text{sr},2}$) shift the peak turn away from the temporal midpoint and add a small pitch tilt, producing unequal departure and return durations alongside vertical angle changes. Vertical detours ($\mathcal{T}_{\text{sr},3}$) combine the horizontal sweep with alternating positive and negative pitch waypoints, guiding the camera through an out-of-plane detour rather than a flat path. Return overshoots ($\mathcal{T}_{\text{sr},4}$) swing the camera past the reference heading to the opposite side before correcting back to the target direction.

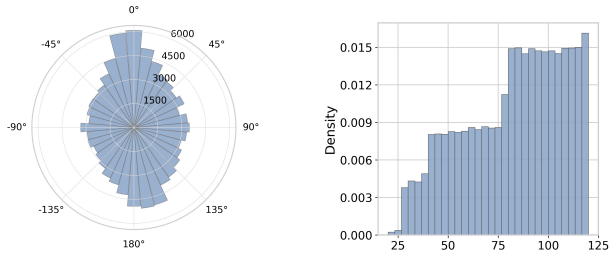
Double revisits ($\mathcal{T}_{\text{dr},1\ldots4}$) repeat this process twice, separating the two turns with a brief pause. To avoid visual distortion during fast turns, we limit the maximum angular speed to 90°/s for single revisits ($\mathcal{T}_{\text{sr}}$), 105°/s for double revisits ($\mathcal{T}_{\text{dr}}, \mathcal{T}_{\text{trk}}$), and 120°/s for full rotation ($\mathcal{T}_{\text{rot}}$). If a planned trajectory exceeds these speed limits, pauses and transition times are adjusted. If the speed is still too high, the rotation angle is scaled down accordingly.

A.3. Dataset Statistics

Dataset Scale. The curated dataset comprises 125K training clips and 300 evaluation clips, with strictly disjoint source videos between the two splits. Multiple perspective clips are rendered from individual panoramic source videos under different camera trajectories. Across the dataset, the ten parameterized trajectory programs in Table A are sampled uniformly, ensuring balanced representation.

Camera Motion Direction Distribution. We analyze the distribution of camera motion directions across the training set to assess the diversity of generated camera paths. For each clip, the camera motion direction is computed from the net horizontal rotation between the first and last frames. As illustrated in Figure A(a), the motion directions span the complete 360° range with broad coverage. Because revisit trajectories where the camera returns toward the subject constitute a major proportion of the dataset, the distribution naturally exhibits higher density around 0° and 180°, while maintaining substantial representation across all lateral angles.

Field of View Distribution. Figure A(b) illustrates the horizontal field of view (FoV) distribution across the training set. Because source panoramic videos vary in resolution and perspective sampling strictly enforces an upscaling limit, the resulting FoV distribution exhibits a tiered density rather than a strictly uniform spread. Nevertheless, the horizontal FoV spans broadly from approximately 25° to the preset upper limit of 120°, providing rich variations in focal lengths and visual perspective. The distribution maintains substantial coverage across both standard viewing angles and wide-angle configurations. This broad FoV coverage prevents models from overfitting to a fixed focal length and encourages generalization across diverse perspective geometries.



(a) Motion direction (b) Horizontal FoV

Figure A. **Camera statistics of the training set.** (a) Polar distribution of camera motion directions, where radial bar lengths denote sample counts across 360° heading angles. (b) Probability density of the horizontal perspective FoV in degrees. The dataset exhibits broad and balanced coverage across both motion directions and viewing angles.

B. Evaluation Metrics

During evaluation, the known prefix $\mathbf{X}_{0:K-1}$ provides conditioning context but is excluded from quantitative metric computation, where evaluated frame indices span 1–80 for image-to-video (I2V) and 21–80 for video continuation.

B.1. Overall Quality

We evaluate overall visual quality using FID [12], FVD [37], and center-cropped FVD (FVD$_c$). Specifically, FID is computed on 2048-dimensional feature representations from Inception-v3 [31], while FVD is calculated on 400-dimensional classification logits from a pretrained I3D network [5]. Standard FVD resizes video frames directly to 224×224, whereas FVD$_c$ preserves the aspect ratio by scaling the shorter spatial dimension to 224 pixels and applying a central crop. Both variants compute the Fréchet distance between Gaussian distributions fitted to feature representations of generated and ground-truth videos.

B.2. Camera Controllability

We evaluate camera controllability using rotation error (Rot), translation error (Trans) [11], CamMC [46], and absolute trajectory error (ATE). ViPE [15] independently reconstructs the camera extrinsics of both generated and ground-truth (GT) videos using known camera intrinsics. All pose metrics operate on relative trajectories anchored at the initial frame. Let $\hat{\mathbf{R}}_t, \mathbf{R}_t \in \mathbb{R}^{3\times3}$ and $\hat{\mathbf{p}}_t, \mathbf{p}_t \in \mathbb{R}^3$

denote the anchored rotations and translations for the generated and ground-truth sequences, respectively.

Following UCPE [60], translations are independently normalized by their maximum trajectory radius prior to score computation:

$$\hat{\mathbf{p}}_t^\star = \frac{\hat{\mathbf{p}}_t}{\max_{0\le r<T} \|\hat{\mathbf{p}}_r\|_2 + 10^{-9}}, \qquad \mathbf{p}_t^\star = \frac{\mathbf{p}_t}{\max_{0\le r<T} \|\mathbf{p}_r\|_2 + 10^{-9}}. \tag{B}$$

For the N evaluated suffix frames, the quantitative errors are given by

$$E_{\mathrm{Rot}} = \frac{180}{\pi N} \sum_t \arccos\left(\frac{\mathrm{tr}(\hat{\mathbf{R}}_t^\top \mathbf{R}_t) - 1}{2}\right), \tag{C}$$

$$E_{\mathrm{Trans}} = \frac{1}{N} \sum_t \|\hat{\mathbf{p}}_t^\star - \mathbf{p}_t^\star\|_2, \tag{D}$$

$$E_{\mathrm{ATE}} = \sqrt{\frac{1}{N} \sum_t \|\hat{\mathbf{p}}_t^\star - \mathbf{p}_t^\star\|_2^2}, \tag{E}$$

$$E_{\mathrm{CamMC}} = \frac{1}{N} \sum_t \|\,\mathrm{vec}(\hat{\mathbf{C}}_t^\star) - \mathrm{vec}(\mathbf{C}_t^\star)\|_2, \tag{F}$$

where the argument of the arccosine is numerically clamped to $[-1, 1]$. Here, $\mathbf{C}_t^\star = [\mathbf{R}_t, \mathbf{p}_t^\star] \in \mathbb{R}^{3\times 4}$ concatenates rotation and normalized translation (and analogously for $\hat{\mathbf{C}}_t^\star$), which $\mathrm{vec}(\cdot)$ flattens into a 12-dimensional vector.

B.3. Background Revisit Fidelity

Revisit timestamps are identified deterministically from the ground-truth camera trajectory, independent of generated video content. For each candidate frame $t \ge 32$, we identify its closest conditioning view $j_t^\star = \arg\min_{j<K} d(t, j)$, with $K = 1$ for I2V and $K = 21$ for video continuation, using ground-truth camera poses and the normalized pose distance

$$d(t, j) = \frac{\angle(\mathbf{R}_t, \mathbf{R}_j)}{5^\circ} + \frac{\|\mathbf{p}_t - \mathbf{p}_j\|_2}{0.05D}, \tag{G}$$

where $\angle(\cdot, \cdot)$ denotes the geodesic rotation angle, $\mathbf{p}_t$ represents the camera center, and $D = \max_{a,b} \|\mathbf{p}_a - \mathbf{p}_b\|_2$ is the maximum pairwise distance along the ground-truth trajectory. We rank candidate frames by $d(t, j_t^\star)$ and greedily select up to five frames, enforcing a minimum temporal separation of four frames between chosen timestamps. Frame selections are precomputed separately for the two conditioning lengths and shared across all evaluated methods.

At each selected timestamp, we evaluate PSNR, SSIM [45], and VGG-based [29] LPIPS [62] between generated frame and the corresponding ground-truth frame, restricting comparison to the intersection of their static-background masks. Frame pairs with less than 10% valid background coverage are excluded from evaluation.

B.4. Dynamic Subject Persistence

We quantify dynamic persistence using overall subject identity (ID), re-entry identity (R-ID), visibility F1-score (Vis.), and normalized subject trajectory error (Traj.).

Instance Tracking and Event Annotation. We first track foreground objects across ground-truth (GT) videos using SAM 3.1 [4]. For each video, we select up to five salient instances present in the initial frame that successfully reappear following camera revisits. To determine visibility transitions, we track the same instances across the ground-truth panoramic video and reproject each panoramic instance mask onto the perspective view using the relative camera transformation at that frame. When panoramic tracking is reliable but the reprojected mask area on the perspective image plane is zero, the timestamp is labeled as an out-of-view event. Conversely, a non-zero reprojected area signifies an in-view state, establishing continuous visibility event annotations throughout the sequence.

Subject and Re-Entry Identity. In the corresponding generated video, independent tracking is initialized using identical first-frame instance prompts. Overall subject identity (ID) measures the cosine similarity between foreground feature representations extracted via DINOv2 [25] from the generated and the corresponding ground-truth instances. If the ground-truth mask is invalid at a given timestamp, we search for the nearest valid ground-truth frame within a temporal window of ± 10 frames to maintain reliable feature comparisons. Re-entry identity (R-ID) specifically isolates the critical phase of subject recovery after an out-of-view period. The evaluation interval commences at the ground-truth re-entry timestamp and extends until the subject exits the field of view again. Within this interval, we collect up to the earliest eight valid ground-truth frames to evaluate re-entry identity, directly assessing whether the subject preserves its appearance immediately upon returning.

Visibility F1-Score. In the generated video, an instance is considered visible when object tracking produces a non-empty mask with valid tracking quality. The instance is confirmed to be out of frame when the mask touches the image boundary before disappearing. The visibility F1-score is defined as $2\mathrm{TP}/(2\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$, evaluated between generated and GT visibility states across frame indices with deterministic GT states.

Subject Trajectory Error. Because generated videos are primarily intended for viewing in 2D space, we employ the bounding-box center discrepancy between generated and GT objects to measure the accuracy of subject motion prediction in the main text. The metric compares bounding-box centers after rescaling predicted coordinates to the GT

image dimensions, defined as

$$e_t^{\mathrm{2D}} = \begin{cases} \|\hat{\mathbf{c}}_t - \mathbf{c}_t\|_2 / \sqrt{W^2 + H^2}, & \text{reliable prediction,} \\ 1, & \text{otherwise,} \end{cases} \tag{H}$$

where $\hat{\mathbf{c}}_t$ and $\mathbf{c}_t$ denote the predicted and GT bounding-box centers at frame t, respectively, and W and H represent the width and height of the GT frame. Missing generated subjects and failed track matches receive a penalty of 1, whereas frames with unavailable GT positions are excluded from evaluation.

B.5. Subject 3D Displacement Error

The subject trajectory error reported in the main paper may be affected by the accuracy of camera pose predictions. Therefore, we additionally report the relative 3D displacement error of each object to evaluate dynamic subject persistence. Quantitative results are presented in Sec. D.2, while this section details the metric formulation.

We report **3D Displacement Error (3D-DE↓)** and **3D Displacement Coverage (3D-DC↑)** as two complementary metrics based on independently reconstructed geometry for ground-truth and generated videos. We use UniDepth V2 [27] to estimate optical-axis depth while reusing existing camera poses from ViPE. For each instance, we back-project the pixels with valid positive depth within its mask into the initial-camera coordinate system of the video. Their coordinate-wise median defines a robust surface-position proxy $\mathbf{m}_t$. Let

$$\boldsymbol{\delta}_t = \mathbf{m}_t - \mathbf{m}_0, \quad \hat{\boldsymbol{\delta}}_t = \hat{\mathbf{m}}_t - \hat{\mathbf{m}}_0, \quad L = (\mathbf{m}_0)_z > 0, \tag{I}$$

where unaccented and hatted quantities denote ground truth and prediction, respectively, and L represents the depth of the GT object at frame zero.

Let $\mathcal{T}$ contain the evaluated suffix frames with confirmed ground-truth subject presence and reliable ground-truth tracking and geometry at frames zero and t. For $t \in \mathcal{T}$, let $v_t = 1$ if the prediction also has reliable tracking and valid 3D positions at both timestamps, and $v_t = 0$ otherwise. We define

$$e_t^{\mathrm{3D}} = \begin{cases} \min\left(\dfrac{\|\hat{\boldsymbol{\delta}}_t - \boldsymbol{\delta}_t\|_2}{L}, 1 \right), & v_t = 1, \\ 1, & v_t = 0, \end{cases} \tag{J}$$

and the per-subject metrics are

$$\text{3D-DE} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} e_t^{\mathrm{3D}}, \qquad \text{3D-DC} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} v_t. \tag{K}$$

Normalizing by the initial depth L of the ground-truth subject rather than its motion amplitude avoids instability for nearly stationary subjects. Coverage captures the availability of valid 3D displacement estimates. To some extent, higher values reflect more consistent subject presence and more reliable 3D geometry in the generated video. These metrics therefore provide evidence complementary to the 2D trajectory and visibility scores.

Table B. **Panoramic temporal RoPE for an 81-frame perspective video.** The perspective stream consistently has 21 latent frames. Panoramic RoPE positions are assigned to align with the perspective latent timeline.

s	Pano RGB J	Pano latent F_{p}	Pano RoPE positions
1	81	21	$0, 1, \ldots, 20$
2	41	11	$0, 2, \ldots, 20$
4	21	6	$0, 4, 8, 12, 16, 20$
10	9	3	$0, 10, 20$

C. Implementation Details

This section details temporal indexing and latitude weighting in loss calculation.

C.1. Temporal Indexing

Under the $4\times$ temporal compression of the shared VAE [16], a T-frame perspective video and its J-frame panoramic counterpart, sampled at interval s with $J = (T-1)/s + 1$, yield latent lengths

$$F_{\mathrm{x}} = \frac{T-1}{4} + 1, \qquad F_{\mathrm{p}} = \frac{J-1}{4} + 1 = \frac{T-1}{4s} + 1. \tag{L}$$

Perspective latent index i and panoramic latent index j correspond to original-video reference times $4i$ and $4sj$, respectively. These serve as reference coordinates for temporally encoded features.

Temporal RoPE Alignment. Temporal RoPE [30] coordinates i and sj for perspective and panoramic latent frames are constructed separately before concatenating the two token sequences. Meanwhile, spatial RoPE retains the two-dimensional token grid of each view. Table B summarizes the resulting panoramic RoPE positions across sampling intervals for $T = 81$.

Camera Alignment. In Eq. (1) of the main paper, the camera intrinsics $\mathbf{A}_i$ and rotations $\mathbf{R}_i^{\mathrm{x}}$ attached to perspective latent frame i are sampled from perspective RGB frame $4i$, whereas the panoramic rotation $\mathbf{R}_j^{\mathrm{p}}$ attached to panoramic latent frame j is sampled from panoramic RGB frame $4j$. At synchronized indices $i = sj$, both views refer to original time $4sj$ and share a camera center.

C.2. Latitude Weighting

Let H_{p} be the panoramic VAE latent height, and let $y \in \{0, \ldots, H_{\mathrm{p}} - 1\}$ denote the row index. Following Argus [24], we construct the latitude-dependent row weights

Table C. **Comparison with a single static panoramic state.** The experiments are conducted based on CameraCtrl [11].

Configuration	Overall Quality			Camera Controllability				BG Revisit Fidelity			Dynamic Subject Persistence			
	FID↓	FVD↓	FVD$_c$↓	Rot↓	Trans↓	CamMC↓	ATE↓	PSNR↑	SSIM↑	LPIPS↓	ID↑	R-ID↑	Vis.↑	Traj.↓
Image-to-video ($K = 1$)														
w/o Pano	23.60	288.5	295.1	13.086	0.531	0.677	0.594	12.29	0.4415	0.618	0.5834	0.2325	0.6519	0.4753
Single panoramic state	31.19	295.5	309.0	12.628	0.508	0.649	0.564	12.68	0.4451	0.628	0.5816	0.2575	0.6526	0.4870
OmniPersist ($s = 10$)	22.91	266.6	276.4	5.695	0.418	0.478	0.465	13.37	0.4731	0.580	0.6429	0.2840	0.7333	0.3783
OmniPersist ($s = 4$)	22.18	262.3	252.6	5.556	0.399	0.458	0.446	13.58	0.4754	0.573	0.6395	0.3355	0.7290	0.3830
Video continuation ($K = 21$)														
w/o Pano	22.61	227.5	223.5	11.704	0.451	0.597	0.489	13.43	0.4701	0.579	0.4946	0.2494	0.6053	0.5161
Single panoramic state	34.58	274.7	280.5	10.291	0.468	0.590	0.507	13.58	0.4635	0.600	0.4773	0.3293	0.5971	0.5411
OmniPersist ($s = 10$)	21.69	201.0	200.6	4.661	0.401	0.455	0.435	15.20	0.5221	0.518	0.5559	0.3492	0.6835	0.4287
OmniPersist ($s = 4$)	21.59	194.9	201.9	4.345	0.401	0.450	0.433	15.22	0.5238	0.515	0.5534	0.3771	0.6678	0.4289

as

$$q_y = \frac{|y - (H_\mathrm{p}/2 - 1/2)|}{H_\mathrm{p}/2}, \tag{M}$$

$$a_y = \sqrt{1 - q_y^2} + \alpha, \qquad \alpha = 0.25, \tag{N}$$

$$w_y = \frac{a_y}{\sum_{r=0}^{H_\mathrm{p}-1} a_r / H_\mathrm{p}}, \tag{O}$$

where q_y represents the normalized vertical distance from the equator, a_y is the unnormalized row weight with a constant offset $\alpha = 0.25$ to prevent zero weights near the poles, and w_y denotes the normalized row weight. These weights w_y emphasize equatorial rows while retaining positive weights near the poles.

In Eq. (4) of the main paper, the latitude-dependent weight $\mathbf{W}_\mathrm{p}$ inside the squared norm is defined row-wise as $(\mathbf{W}_\mathrm{p})_y = \sqrt{w_y}$. Thus, the panoramic loss $\mathcal{L}_\mathrm{p}$ can also be expressed as

$$\mathcal{L}_\mathrm{p} = \mathrm{mean}\left[w_y(\hat{\mathbf{u}}^\mathrm{p} - \mathbf{u}^\mathrm{p})^2\right], \tag{P}$$

where $\hat{\mathbf{u}}^\mathrm{p}$ and $\mathbf{u}^\mathrm{p}$ denote the predicted and target panoramic velocity fields, respectively.

D. Additional Results

D.1. Comparison with Static Panoramic State

We examine whether a single initial panoramic state can replace panoramic evolution. Table C compares four CameraCtrl-based configurations on both tasks, namely the perspective-only baseline, joint prediction with a single panoramic state at original time $t = 0$, and OmniPersist at $s = 10$ and $s = 4$. The single-state variant predicts only one panoramic frame throughout denoising. Although this configuration improves re-entry identity over the baseline in both tasks, it worsens FID, FVD, and subject trajectory error. Initial panoramic support alone therefore provides inconsistent benefits across generation quality and dynamic persistence. In contrast, evolving panoramic states at both $s = 10$ and $s = 4$ outperform the single-state configuration across all 14 metrics in both tasks.

D.2. 3D Displacement Evaluation

Tables D–F extend main-paper Tables 1–3 with 3D displacement evaluation. We report 3D displacement error (3D-DE, lower is better) and 3D tracking coverage (3D-DC, higher is better), as defined in Sec. B.5.

At the primary setting $s = 4$, OmniPersist lowers 3D-DE and raises 3D-DC relative to both camera-control baselines in both tasks, aligning with the visibility and 2D trajectory gains in the main paper. Specifically, 3D-DE in I2V drops from 0.6380 to 0.5447 for CameraCtrl and from 0.6094 to 0.5304 for UCPE, while 3D-DC in video continuation under UCPE rises from 0.4741 to 0.5889. Both $s = 4$ variants also surpass all released baselines on both metrics in video continuation and achieve lower error than LingBot-World 2.0 in I2V, though with slightly lower coverage. These 3D results support the persistence gains in the main comparison. Gains over the perspective-only baselines persist across all four sampling intervals in both tasks, with even $s = 10$ achieving lower 3D-DE and higher 3D-DC, confirming that sparse panoramic evolution remains effective.

The component ablation in Table F aligns with the main-paper ablation. Adding a panoramic stream alone yields limited I2V gains and slightly degrades both 3D metrics in video continuation. Incorporating GeoAnch and TemComp progressively improves both measures across tasks, with TemComp reducing 3D-DE in video continuation from 0.6462 to 0.6031 and raising 3D-DC from 0.5404 to 0.5853, confirming the value of both modules.

D.3. Additional Qualitative Results

Perspective Camera Revisits. Due to space limitations, we include only Figure 3 in the main paper for qualitative comparison and present additional results here. Because LiveWorld [8], LingBot-World 2.0 [10] (for I2V), and SANA-

Table D. **3D displacement evaluation for image-to-video (I2V).** Rows correspond to main-paper Table 1. 3D-DE measures 3D displacement error, and 3D-DC reports 3D tracking coverage.

Configuration	3D-DE↓	3D-DC↑
HyDRA (1.3B) [6]	0.6150	0.6153
Matrix-World 3.0 (5B) [47]	0.6929	0.4682
DreamX-World (5B) [34]	0.6702	0.5254
SANA-WM (2.6B+17B) [64]	0.6380	0.5562
LiveWorld (14B) [8]	0.6755	0.5425
LingBot-World 2.0 (28B) [10]	0.5490	0.6913
CameraCtrl† [11]	0.6380	0.5792
CameraCtrl+OmniPersist ($s = 1$)	0.5313	0.6740
CameraCtrl+OmniPersist ($s = 2$)	0.5359	0.6785
CameraCtrl+OmniPersist ($s = 4$)	0.5447	0.6735
CameraCtrl+OmniPersist ($s = 10$)	0.5358	0.6740
UCPE† [60]	0.6094	0.5748
UCPE+OmniPersist ($s = 1$)	0.5326	0.6672
UCPE+OmniPersist ($s = 2$)	0.5433	0.6795
UCPE+OmniPersist ($s = 4$)	0.5304	0.6863
UCPE+OmniPersist ($s = 10$)	0.5565	0.6572

Table E. **3D displacement evaluation for video continuation.** Rows correspond to main-paper Table 2. 3D-DE measures 3D displacement error, and 3D-DC reports 3D tracking coverage.

Configuration	3D-DE↓	3D-DC↑
HyDRA (1.3B) [6]	0.8949	0.2622
Matrix-World 3.0 (5B) [47]	0.8412	0.2570
SANA-WM (2.6B+17B) [64]	0.7979	0.3800
LiveWorld (14B) [8]	0.7560	0.3806
CameraCtrl† [11]	0.6847	0.4965
CameraCtrl+OmniPersist ($s = 1$)	0.5775	0.6077
CameraCtrl+OmniPersist ($s = 2$)	0.5984	0.5964
CameraCtrl+OmniPersist ($s = 4$)	0.6031	0.5853
CameraCtrl+OmniPersist ($s = 10$)	0.6039	0.5847
UCPE† [60]	0.6795	0.4741
UCPE+OmniPersist ($s = 1$)	0.5752	0.5916
UCPE+OmniPersist ($s = 2$)	0.5677	0.6006
UCPE+OmniPersist ($s = 4$)	0.5853	0.5889
UCPE+OmniPersist ($s = 10$)	0.6080	0.5565

Table F. **3D displacement evaluation for component ablation.** Rows correspond to main-paper Table 3. 3D-DE measures 3D displacement error, and 3D-DC reports 3D tracking coverage.

Configuration	3D-DE↓	3D-DC↑
Image-to-video ($K = 1$)		
w/o Pano	0.6380	0.5792
w/ Pano	0.6101	0.5952
w/ Pano & GeoAnch	0.5548	0.6644
w/ Pano & GeoAnch & TemComp	0.5447	0.6735
Video continuation ($K = 21$)		
w/o Pano	0.6847	0.4965
w/ Pano	0.6876	0.4858
w/ Pano & GeoAnch	0.6462	0.5404
w/ Pano & GeoAnch & TemComp	0.6031	0.5853

WM [64] (for video continuation) perform relatively well among the released models, we select these models for comparison with the CameraCtrl [11] and UCPE [60] baselines, as well as their corresponding CameraCtrl+OmniPersist ($s = 4$) and UCPE+OmniPersist ($s = 4$) variants. Figures B and C compare a pre-departure observation with early and later post-return observations at identical frame indices across these methods. The examples illustrate a consistent background, subject appearance, and continued motion after re-entry. Specifically, Figure B presents I2V comparisons for two skaters and two vehicles, while Figure C provides video continuation comparisons for people walking on a bridge and a boy on a swing.

Panoramic Evolution. Figure D shows the generation results of CameraCtrl+OmniPersist at $s = 4$ on the I2V task. We observe that the panoramic stream continues to evolve even when moving subjects leave the perspective field of view. Across the six examples, the panoramic states retain these subjects and depict changes in their positions during out-of-view intervals, providing evolving scene context for subsequent perspective revisits.

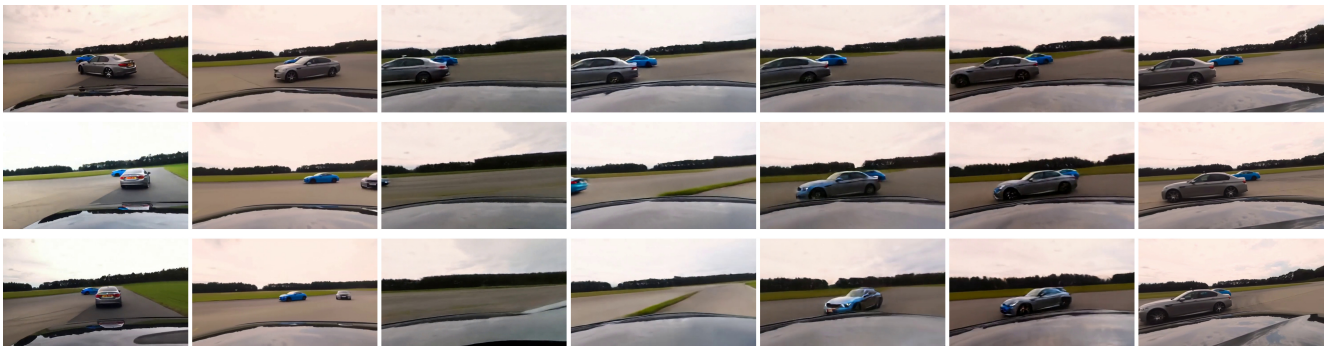
LLM Usage Statement

During the preparation of this manuscript, Gemini 3.8 Flash was used for language polishing, and GPT-6 was employed for code debugging. The authors have reviewed all content and retain full responsibility for all methodological claims, implementation descriptions, experimental results, references, and the final text.



LiveWorld LingBot-World 2.0 CameraCtrl UCPE CameraCtrl+OmniPersist UCPE+OmniPersist Ground Truth

(a) Prompt overview: *A video shows the skater in the white helmet and the skater in the grey helmet are positioned side-by-side, moving forward along the road.*



LiveWorld LingBot-World 2.0 CameraCtrl UCPE CameraCtrl+OmniPersist UCPE+OmniPersist Ground Truth

(b) Prompt overview: *A video shows a silver sedan and a blue coupe driving on a paved track. The vehicles maintain a consistent distance from one another as they travel along the straight section of the track.*

Figure B. **Generation results on I2V across camera revisits.** From top to bottom in each example, we show a frame before the camera turns away and two revisit frames after it returns to illustrate identity preservation and continued motion. Frames in each row share the same frame index across all methods and ground truth.

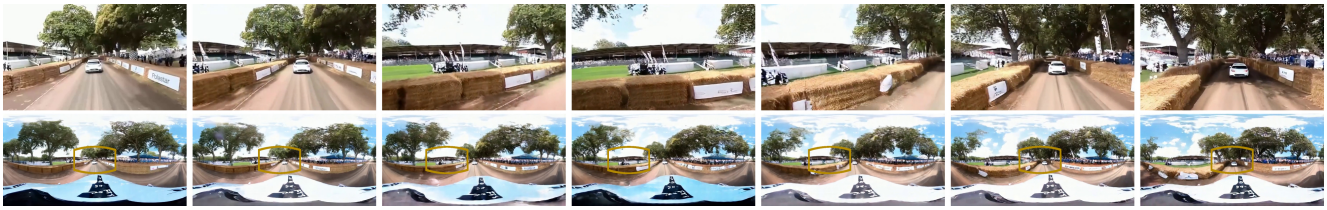


(a) Prompt overview: *A video shows a group of people walking along a white bridge with intricate carvings under a clear blue sky.*



(b) Prompt overview: *A video shows a young boy in a blue shirt sitting on a swing, moving back and forth while holding the blue frame.*

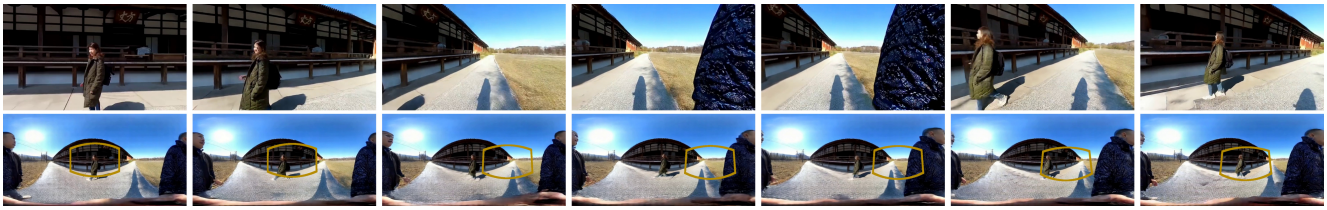
Figure C. **Generation results on video continuation across camera revisits.** From top to bottom in each example, we show a frame before the camera turns away and two revisit frames after it returns to illustrate identity preservation and continued motion. Frames in each row share the same frame index across all methods and ground truth.



(a) Prompt overview: *A video shows a white car driving forward on a dirt track lined with hay bales.*



(b) Prompt overview: *A video shows a silver hatchback car driving away from the camera on a paved road.*



(c) Prompt overview: *A video shows a woman with shoulder-length brown hair walking along a paved path next to a long, traditional wooden building with a dark roof.*



(d) Prompt overview: *A video shows a young boy in a blue shirt sitting on a swing, moving back and forth while holding the blue frame.*



(e) Prompt overview: *A video shows two men walking away from the viewer along a wide concrete pier.*



(f) Prompt overview: *A video shows a woman in a reddish-orange dress standing in a living room decorated for Christmas. The woman finally sits on the sofa.*

Figure D. **Joint perspective generation and panoramic evolution under CameraCtrl+OmniPersist ($s = 4$) on the I2V task.** Each subfigure shows generated perspective frames at $t = 0, 16, 28, 40, 52, 64, 80$ in the top row and the corresponding panoramic frames in the bottom row. With panoramic sampling interval $s = 4$, the panoramic-video frame indices are $r = 0, 4, 7, 10, 13, 16, 20$, respectively, satisfying $t = sr$. Each perspective frame corresponds to the gold-boxed region in its synchronized panorama.

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. In *ICLR*, pages 66712–66737, 2025. 2

[2] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *ICCV*, pages 14834–14844. IEEE, 2025. 2

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, pages 22563–22575. IEEE, 2023. 1

[4] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *ICLR*, pages 138846–138923, 2026. 9, 11

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 4724–4733. IEEE, 2017. 10

[6] Kaijin Chen, Dingkang Liang, Xin Zhou, Yikang Ding, Xiaoqiang Liu, Pengfei Wan, and Xiang Bai. Out of sight but not out of mind: Hybrid memory for dynamic video world models. *arXiv preprint arXiv:2603.25716*, 2026. 1, 2, 5, 6, 7, 14

[7] Ting-Hsuan Chen, Ying-Huan Chen, Tao Tu, Jie-Ying Lee, Cho-Ying Wu, Fangzhou Lin, Hengyuan Zhang, David Paz, Xinyu Huang, Yuliang Guo, et al. Pantheon360: Taming digital twin generation via 3d-aware 360° video diffusion. *arXiv preprint arXiv:2605.25449*, 2026. 3

[8] Zicheng Duan, Jiatong Xia, Zeyu Zhang, Wenbo Zhang, Gengze Zhou, Chenhui Gou, Yefei He, Feng Chen, Xinyu Zhang, and Lingqiao Liu. LiveWorld: Simulating out-of-sight dynamics in generative video world models. *arXiv preprint arXiv:2603.07145*, 2026. 1, 3, 6, 7, 13, 14

[9] Zixun Fang, Kai Zhu, Zhiheng Liu, Yu Liu, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Panoramic video generation with pretrained diffusion models. *NeurIPS*, 38:12486–12508, 2026. 3

[10] Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, et al. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026. 1, 2, 5, 6, 7, 13, 14

[11] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for video diffusion models. In *ICLR*, pages 100433–100464, 2025. 2, 5, 6, 7, 8, 10, 13, 14

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 6, 10

[13] Or Hirschorn, Aaron Olender, Eli Alshan, Ianir Ideses, Lior Fritz, and Sagie Benaim. Spherope: Zero-shot optimization-free 360 panorama generation with spherical rope. *arXiv preprint arXiv:2606.32033*, 2026. 4

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 5

[15] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025. 6, 9, 10

[16] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 3, 5, 12

[17] Chunyang Li, Yuanbo Yang, Jiahao Shao, Hongyu Zhou, Katja Schwarz, and Yiyi Liao. Rerope: Repurposing rope for relative camera control. *arXiv preprint arXiv:2602.08068*, 2026. 2

[18] Haoyuan Li, Dizhe Zhang, Yuemei Zhou, Xiangkai Zhang, Haoran Feng, Xiaofan Lin, Wenjie Jiang, Bo Du, Ming-Hsuan Yang, and Lu Qi. Panoworld: Real-world panoramic generation. *arXiv preprint arXiv:2607.09661*, 2026. 3

[19] Lingen Li, Guangzhi Wang, Xiaoyu Li, Zhaoyang Zhang, Qi Dou, Jinwei Gu, Tianfan Xue, and Ying Shan. Cubecomposer: Spatio-temporal autoregressive 4k 360° video generation from perspective video. *arXiv preprint arXiv:2603.04291*, 2026. 3

[20] Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026. 1

[21] Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, Chuanyun Deng, Yepan Xiong, Min Chen, et al. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. In *ICCV*, pages 28785–28796. IEEE, 2025. 2

[22] Yingyan Li, Shuyao Shang, Weisong Liu, Bing Zhan, Haochen Wang, Yuqi Wang, Yuntao Chen, Xiaoman Wang, Yasong An, Chufeng Tang, et al. Drivevla-w0: World models amplify data scaling law in autonomous driving. In *ICLR*, pages 7890–7911, 2026. 1

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[24] Rundong Luo, Matthew Wallingford, Ali Fahardi, Noah Snavely, and Wei-Chiu Ma. Beyond the frame: Generating 360 panoramic videos from perspective videos. In *ICCV*, pages 14336–14345. IEEE, 2025. 3, 5, 12

[25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024. 6, 11

[26] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4172–4182. IEEE, 2023. 1

[27] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool.

Unidepthv2: Universal monocular metric depth estimation made simpler. *IEEE TPAMI*, 2025. 12

[28] 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 *CVPR*, pages 6121–6132. IEEE, 2025. 2

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 11

[30] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 3, 4, 12

[31] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 10

[32] Jiaming Tan, Zhen Li, Shuwei Shi, Minggui Teng, Siqi Yang, Yuwei Wu, Bo Zheng, Chuanhao Li, and Kaipeng Zhang. Mugen: Interactive panoramic world exploration via camera control. *arXiv preprint arXiv:2609.38077*, 2026. 3

[33] Jiaming Tan, Mingliang Zhai, Zhen Li, Yuwei Wu, Chuanhao Li, and Kaipeng Zhang. Alayavista: Streaming world modeling from panoramic states to perspective video. *arXiv preprint arXiv:2609.14462*, 2026. 2, 3

[34] DreamX Team, Yancheng Bai, Rui Chen, Xiangxiang Chu, Rujing Dang, Hao Dou, Bingjie Gao, Qiwen Gu, Siyu Hong, Jiachen Lei, et al. Dreamx-world 1.0: A general-purpose interactive world model. *arXiv preprint arXiv:2606.16993*, 2026. 1, 2, 5, 6, 14

[35] InSpatio Team, Donghui Shen, Guofeng Zhang, Haomin Liu, Haoyu Ji, Hujun Bao, Hongjia Zhai, Jialin Liu, Jing Guo, Nan Wang, et al. Inspatio-world: A real-time 4d world simulator via spatiotemporal autoregressive modeling. *arXiv preprint arXiv:2604.07209*, 2026. 2

[36] Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026. 1, 2

[37] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6, 10

[38] Matthew Wallingford, Anand Bhattad, Aditya Kusupati, Vivek Ramanujan, Matt Deitke, Sham Kakade, Aniruddha Kembhavi, Roozbeh Mottaghi, Wei-Chiu Ma, and Ali Farhadi. From an image to a scene: Learning to imagine the world from a million 360 videos. *NeurIPS*, 37:17743–17760, 2024. 5

[39] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 3, 5

[40] Hanlin Wang, Hao Ouyang, Qiuyu Wang, Wen Wang, Qingyan Bai, Ka Leong Cheng, Yue Yu, Yixuan Li, Yihao Meng, Zichen Liu, et al. Worlddirector: Building controllable world simulators with persistent dynamic memory. *arXiv preprint arXiv:2607.02517*, 2026. 1, 3

[41] Jiahao Wang, Luoxin Ye, TaiMing Lu, Junfei Xiao, Jiahan Zhang, Yuxiang Guo, Xijun Liu, Rama Chellappa, Cheng Peng, Alan Yuille, et al. Evoworld: Evolving panoramic world generation with explicit 3d memory. *arXiv preprint arXiv:2510.01183*, 2025. 3

[42] Qian Wang, Weiqi Li, Chong Mou, Xinhua Cheng, and Jian Zhang. 360dvd: Controllable panorama video generation with 360-degree video diffusion model. In *CVPR*, pages 6913–6923. IEEE, 2024. 3

[43] Weijie Wang, Haoyu Zhao, Yifan Yang, Feng Chen, Zeyu Zhang, Yefei He, Zicheng Duan, Donny Y Chen, Yuqing Yang, and Bohan Zhuang. Latent spatial memory for video world models. *arXiv preprint arXiv:2606.09828*, 2026. 2

[44] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *ECCV*, pages 55–72. Springer, 2024. 1

[45] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 6, 11

[46] 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*, pages 1–11, 2024. 2, 6, 10

[47] Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026. 1, 2, 5, 6, 7, 14

[48] Ziyi Wu, Daniel Watson, Andrea Tagliasacchi, David J Fleet, Marcus A Brubaker, and Saurabh Saxena. 360anything: Geometry-free lifting of images and videos to 360°. In *ECCV*, pages 135–155. Springer, 2026. 3, 5

[49] Yifei Xia, Shuchen Weng, Siqi Yang, Jingqi Liu, Chengxuan Zhu, Minggui Teng, Zijian Jia, Han Jiang, and Boxin Shi. Panowan: Lifting diffusion video generation models to 360° with latitude/longitude-aware mechanisms. In *NeurIPS*, pages 148027–148046, 2026. 3

[50] Kevin Xie, Amirmojtaba Sabour, Jiahui Huang, Despoina Paschalidou, Greg Klar, Umar Iqbal, Sanja Fidler, and Xiaohui Zeng. Videopanda: Video panoramic diffusion with multi-view attention. *arXiv preprint arXiv:2504.11389*, 2025. 3

[51] Tianshuo Xu, Yichen Xie, Depu Meng, Chensheng Peng, Quentin Herau, Bo Jiang, Yihan Hu, and Wei Zhan. Teaching video generators to remember: Eliciting dynamic memory for out-of-sight state evolution. *arXiv preprint arXiv:2605.25333*, 2026. 3

[52] Yu Yang, Zhilu Zhang, Xiang Zhang, Yihan Zeng, Hui Li, and Wangmeng Zuo. Physworld: From real videos to world

models of deformable objects via physics-aware demonstration synthesis. *arXiv preprint arXiv:2510.21447*, 2025. 1

[53] Zhongqi Yang, Wenhang Ge, Yuqi Li, Jiaqi Chen, Haoyuan Li, Mengyin An, Fei Kang, Hua Xue, Baixin Xu, Yuyang Yin, et al. Matrix-3d: Omnidirectional explorable 3d world generation. *arXiv preprint arXiv:2508.08086*, 2025. 3

[54] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, pages 83048–83077, 2025. 1

[55] Minghao Yin, Jiahao Lu, Wenbo Hu, Wang Zhao, Ying Shan, and Kai Han. Scope: Sightline-coordinate positional encoding for video diffusion transformers. *arXiv preprint arXiv:2606.27345*, 2026. 2

[56] Jiwen Yu, Jianxiong Gao, Jianhong Bai, Yiran Qin, Kaiyi Huang, Quande Liu, Xintao Wang, Pengfei Wan, Kun Gai, and Xihui Liu. Memlearner: Learning to query context memory for video world models. In *ECCV*, pages 96–114. Springer, 2026. 1, 2

[57] Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *ICCV*, pages 100–111. IEEE, 2025. 2

[58] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *IEEE TPAMI*, 2025. 2

[59] Wei Yu, Runjia Qian, Yumeng Li, Liquan Wang, Songheng Yin, Dennis Anthony, Yang Ye, Yidi Li, Weiwei Wan, Animesh Garg, et al. Mosaicmem: Hybrid spatial memory for controllable video world models. *arXiv preprint arXiv:2603.17117*, 2026. 2

[60] 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 *CVPR*, pages 38027–38037, 2026. 1, 2, 5, 6, 7, 8, 11, 14

[61] Muyang Zhang, Yuzhi Chen, Rongtao Xu, Changwei Wang, JinMing Yang, Weiliang Meng, Jianwei Guo, Huihuang Zhao, and Xiaopeng Zhang. Panodit: Panoramic videos generation with diffusion transformer. In *AAAI*, pages 10040–10048, 2025. 3

[62] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595. IEEE, 2018. 6, 11

[63] Sixiao Zheng, Minghao Yin, Wenbo Hu, Xiaoyu Li, Ying Shan, and Yanwei Fu. Versecrafter: Dynamic realistic video world model with 4d geometric control. In *CVPR*, pages 40277–40290, 2026. 2

[64] 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. 1, 2, 5, 6, 7, 14