# PANOWORLD: REAL-WORLD PANORAMIC GENERATION

**Anonymous authors**
Paper under double-blind review

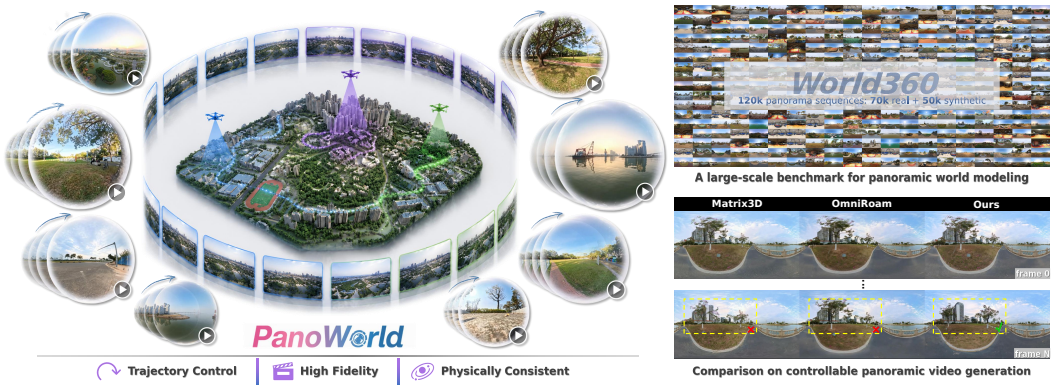


Figure 1: **PanoWorld** is a novel framework for high-fidelity and controllable panoramic video generation. Our approach achieves precise trajectory control across complex movements while maintaining high-fidelity visual synthesis with physical consistency in diverse real-world environments.

## ABSTRACT

In this work, we aim to address the challenge of controllable panoramic world modeling by leveraging the rotation-equivariant properties of omnidirectional representations, where camera rotation can be naturally modeled as an implicit geometric transformation. Building on this insight, we propose PanoWorld, which simplifies camera trajectories into translations via fixed headings for current-action modeling and long-horizon consistency through Dense Panoramic Ray-Conditioning (DPRC) and Geometry-aware Memory Augmentation (GMA). Then, a three-stage training pipeline is introduced to progressively optimize each component. To better evaluate physical consistency under large-scale spatial variations and diverse real-world environments with varying lighting scenarios, we construct **World360**, a large-scale dataset consisting of both real-world video clips collected via panoramic unmanned aerial vehicles and high-quality simulated clips generated by AirSim360. Extensive experiments on World360 demonstrate the effectiveness of PanoWorld, achieving superior generation quality while maintaining strong controllability. Our models, training code, and dataset will be publicly available.

## 1 INTRODUCTION

Recently, world models have attracted significant attention for modeling dynamic environments and enabling controllable generation (Tang et al., 2025; Gao et al., 2026; Fang et al., 2026). They have shown strong potential in various robotic applications, including autonomous driving and unmanned aerial vehicles (Huang et al., 2025b; Sun et al., 2025a; Wu et al., 2026a).

Among various world models, panoramic representations, particularly equirectangular projections (ERP), have emerged as a mainstream paradigm by capturing the full 360° field of view (FoV) for each frame in a specific trajectory (Lin et al., 2025). Despite recent advancements, achieving physical consistency, including geometric coherence and stable visual appearance, across space and time



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

remains challenging for panoramic world models, where such issues are amplified by the full-view nature of panoramic representations (Lu et al., 2025; Zhang et al., 2025a).

Most existing work, including 3DGS-based and video-generation paradigms, adopts memory mechanisms (e.g., 3D points or KV caches) to retrieve past information for spatiotemporal consistency (Yang et al., 2025; Chou et al., 2025; Schwarz et al., 2025). However, these solutions rely on temporal memory to preserve information beyond the current field of view, whereas panoramic representations provide persistent 360° scene visibility, reducing the reliance on long-term memory. Thus, one question is raised: how can memory mechanisms better adapt panoramic representations?

To address this issue, we begin by analyzing the properties of the equirectangular projection (ERP), which is rotation-equivariant, where rotations mainly alter the distortion pattern while preserving the underlying scene content. This observation inspires us to simplify camera motion by treating rotation as a geometric transformation and modeling only translation explicitly. Building on this insight, we propose PanoWorld, a diffusion-based framework that introduces Dense Panoramic Ray-Conditioning (DPRC) for explicit camera-motion modeling and Geometry-aware Memory Augmentation (GMA) for retrieving consistent panoramic observations, while decoupling translation and rotation via fixed headings. To stably optimize each component, we employ a progressive, three-stage training strategy.

Existing datasets are predominantly indoor or simulation-based (Wang et al., 2024a; Yang et al., 2025; Xia et al., 2025), where physical conditions are relatively stable and environmental diversity is limited, making them less suitable for modeling and evaluating real-world physical consistency. To bridge this gap, we construct **World360**, a large-scale dataset comprising 70K real-world video clips and 50K high-fidelity simulated clips generated with the AirSim360 platform (Ge et al., 2025). Extensive experiments on World360 demonstrate that our approach consistently outperforms existing methods by a substantial margin, validating its effectiveness for high-quality and controllable panoramic video generation.

Our contributions are summarized as follows:

- We propose PanoWorld, a diffusion-based panoramic world model that improves long-horizon trajectory consistency by exploiting the inherent geometric structure of omnidirectional representations.
- PanoWorld comprises two core modules, DPRC and GMA, together with a progressive three-stage training pipeline. DPRC provides dense geometric conditioning for panoramic motion modeling, while GMA enhances memory retrieval and preserves spatiotemporal consistency over long horizons.
- We construct **World360**, a large-scale benchmark for panoramic world modeling under diverse real-world physical variations. Extensive experiments demonstrate the effectiveness of PanoWorld in both panoramic video generation quality and camera controllability.

## 2 Related Work

**Panoramic Video Generation.** Panoramic vision provides a holistic geometric perspective for globally coherent scene synthesis (Xie et al., 2025). Recent works such as WorldPrompter (Zhang et al., 2025b) demonstrate that panoramic generation can serve as an effective intermediate representation for improving structural consistency. Early foundational frameworks, including 360DVD (Wang et al., 2024a), pioneered the text-to-panoramic video generation pipeline and introduced the WEB360 dataset for benchmarking. To improve data scale and generation quality, PanoWan (Xia et al., 2025) builds upon large-scale immersive datasets, such as 360-1M (Wallingford et al., 2024). Meanwhile, recent methods, including 4K4DGen (Li et al., 2025a) and DynamicScaler (Liu et al., 2025), focus on enhancing spatiotemporal and geometric consistency in panoramic video generation. However, despite these advances, existing panoramic video generators still struggle in complex outdoor environments, where maintaining structural coherence and visual consistency over long camera trajectories remains challenging.

**Camera-controlled Video Diffusion Models.** Camera-motion control has become a key direction for decoupling scene dynamics from camera motion (He et al., 2024; Yu et al., 2024; Wang et al., 2024b; Bai et al., 2025; Zhang et al., 2026). For panoramic video generation, existing works

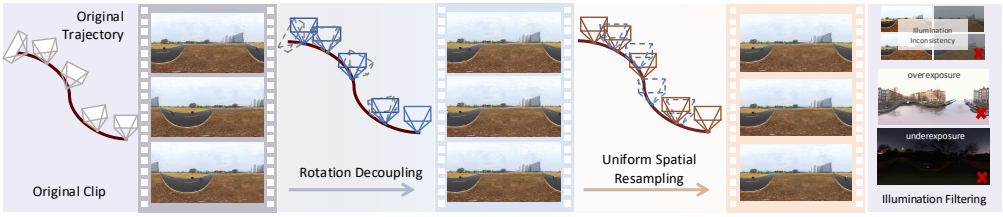


Figure 2: **The Data Curation Pipeline.** The pipeline converts raw panoramic clips into high-quality sequences by: (1) *Rotation Decoupling* to isolate pure translation; (2) *Uniform Spatial Resampling* to standardize motion scales; and (3) *Exposure Quality Filtering* to ensure exposure consistency.

Table 1: Comparison of panorama datasets for world generation.

| Dataset | Source | Samples | Motion Space | Pose | Depth | Text | Res. | Video |
|---|---|---|---|---|---|---|---|---|
| Imagine360 (Tan et al., 2025) | Real | 10k | Planar | no | no | ✓ | Mixed | ✓ |
| Web360 (Wang et al., 2024a) | Real | 2k | Planar | no | no | ✓ | 720p | ✓ |
| Argus (Luo et al., 2025) | Real | 283k | Planar | no | no | no | Mixed | ✓ |
| 360-1M (Wallingford et al., 2024) | Real | 1,076k | Planar | ✓ | no | no | Mixed | ✓ |
| PanoWan (Xia et al., 2025) | Real | 13k | Planar | no | no | ✓ | Mixed | ✓ |
| Matrix-Pano (Yang et al., 2025) | Syn. | 116k | 3D Space‡ | ✓ | ✓ | ✓ | 2K | ✓ |
| **World360 (Ours)** | **Real+Syn.** | **120k** | **Multi-Alt.** | ✓ | **w/ Syn.** | ✓ | **2K** | ✓ |

‡ Existing large-scale real datasets primarily consist of **planar street-level motion**, where altitude information is ambiguous or fixed. In contrast, **World360** provides diverse **aerial trajectories** across multiple altitudes.

explore spherical motion conditioning (Wang et al., 2024a), perspective-video conditioned panoramic synthesis (Ji et al., 2025), geometry-aware view synthesis (Yang et al., 2025; Zhou et al., 2026; Chen et al., 2026), real–generated frame fusion (Liu et al., 2026), and trajectory-aware diffusion transformers for explorable (Yin et al., 2025) and whole-house (Jia et al., 2026) generation. Ray-based camera representations have also been explored in perspective settings (Zhang et al., 2026; Yin et al., 2026; Wu et al., 2026b), while methods retaining explicit scene representations rely on geometric reconstruction or persistent world memory (Yang et al., 2025; Wang et al., 2025b). However, adapting camera-motion representations to panoramic domains remains challenging due to non-planar ray geometry and long-horizon motion. Our framework addresses this challenge through ray-conditioned motion modeling with ray-space memory without explicit reconstruction.

**Memory-Augmented Video Synthesis.** Long-horizon video generation is challenged by scene drift arising from limited persistent memory (Song et al., 2025). Existing approaches enhance temporal consistency by leveraging extended context (Yu et al., 2025; Chou et al., 2025; Xiao et al., 2025) or implicit memory representations (Li et al., 2025c); however, preserving consistent scene appearance and geometry during extended camera movements remains challenging. For panoramic video generation, methods based on explicit scene representations leverage geometric reconstruction to maintain consistency (Yang et al., 2025; Wang et al., 2025b). In contrast, our framework leverages the inherent rotation-equivariant geometry of panoramic representations to improve long-horizon consistency and camera controllability without explicit global reconstruction.

## 3 Data Pipeline and Benchmark

In this section, we introduce our data engine for large-scale controllable panoramic video synthesis. We first describe the curation pipeline for temporal and geometric consistency, followed by World360, a benchmark covering diverse real-world environments and lighting scenarios.

**Data Curation** The raw panoramic data is horizon-leveled by the acquisition system, with yaw as the primary rotational degree of freedom. Based on this property, our data curation pipeline adopts a three-fold alignment strategy (Fig. 2). **Yaw Rotation Decoupling** factors out relative yaw rotations and re-project each sequence to a unified heading, isolating translational motion for consistent learning of depth and motion parallax in the panoramic ray-conditioned space. **Uniform Spatial Resampling** then replaces conventional fixed-time sampling with a constant spatial increment ($\Delta d$), ensuring a uniform geometric scale across trajectories despite variations in velocity. Finally,

**Exposure Quality Filtering** removes sequences with severe exposure artifacts, such as flickering or extreme underexposure, while preserving natural illumination diversity. Together, these processes reduce acquisition artifacts and geometric inconsistencies, providing a clean foundation for physically consistent world modeling.

**World360** Based on the aforementioned pipeline, we construct World360, a comprehensive benchmark for panoramic world modeling under real-world physical variations. As summarized in Table 1, World360 comprises 120,000 high-quality sequences, unifying 70,000 curated real-world clips with 50,000 high-fidelity simulations from AirSim360 (Ge et al., 2025). Unlike datasets focusing primarily on planar camera motions, World360 introduces diverse multi-altitude aerial trajectories with precise camera poses and depth information. This benchmark bridges the gap between constrained modeling and unconstrained panoramic synthesis, enabling the evaluation of structural integrity across complex 3D paths.

# 4 Our Method

Given an initial panoramic observation and a target camera trajectory, our framework first removes yaw rotational ambiguity to simplify the motion space, then injects geometry-aware motion control and historical memory into a unified diffusion backbone, and finally decodes the fused representation into a controllable panoramic video. This section is organized into two parts: Sec. 4.1 presents the core modeling design underlying motion control and memory modeling, and Sec. 4.2 describes the three-stage optimization strategy used to train the full framework.

## 4.1 Model Design

This section presents three core modeling designs that promote physically consistent generation, as illustrated in Fig. 3. We first simplify the learning problem through motion decoupling (4.1.1), then model view-dependent translation with DPRC (4.1.2), and finally enforce long-horizon scene persistence through geometry-aware memory modeling (4.1.3).

### 4.1.1 Motion Decoupling Principle

Conventional models often learn translation and rotation jointly, which leads to structural warping in panoramic video generation. To maintain geometric consistency, we observe that camera rotation in a panoramic sequence is independent of scene depth, and therefore isolate it as a precise geometric transformation, allowing the diffusion model to focus exclusively on translation-induced parallax. To bridge the gap between pre-trained perspective priors and panoramic geometry, we further fine-tune the backbone on panoramic datasets with LoRA, enabling it to capture intrinsic equirectangular properties such as polar distortion and horizontal continuity while preserving the generative quality of the base model. Combined with explicit motion decoupling, this adaptation removes artifacts that violate $360^\circ$ projection rules and preserves structural integrity across camera movements.

### 4.1.2 Action Modeling: DPRC

To achieve unconstrained motion control while enforcing joint consistency across temporal and geometric domains, we propose the DPRC mechanism. Unlike traditional perspective models that rely on pixel-level optical flow, DPRC leverages equirectangular geometry to model the generative process as a dynamic light-field evolution, aligning kinematic priors with the physical origin of panoramic rays.

**Spherical Ray Unprojection.** We first map each pixel $(i, j)$ of the latent grid to a unit ray direction $\mathbf{r}^{\text{cam}} \in \mathbb{S}^2$. Given the latent grid height $H$ and width $W$, the longitude $\theta$ and latitude $\phi$ are:

$$\theta = \frac{j + 0.5}{W} 2\pi - \pi, \qquad \phi = \frac{\pi}{2} - \frac{i + 0.5}{H} \pi. \tag{1}$$

Following our dataset convention, the corresponding camera-space ray is defined as:

$$\mathbf{r}^{\text{cam}} = [-\cos\phi\cos\theta,\ -\cos\phi\sin\theta,\ -\sin\phi]^\top. \tag{2}$$

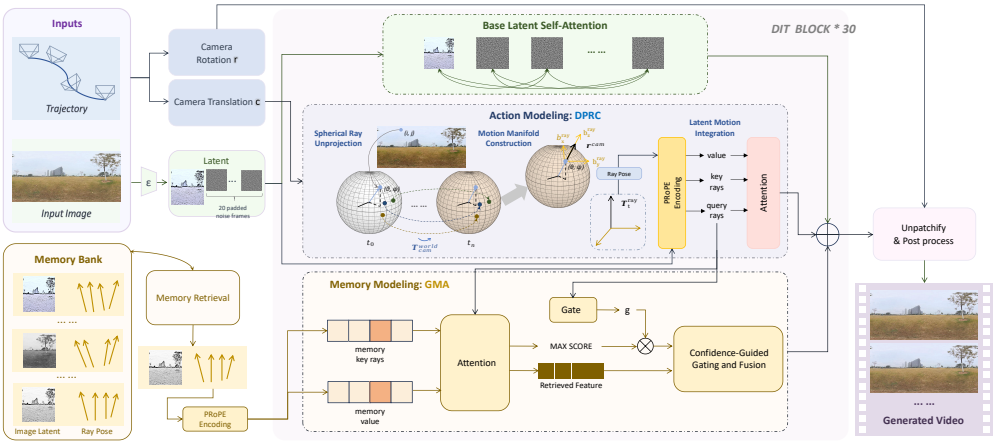


Figure 3: **PanoWorld Network Architecture.** Given an initial panoramic image, camera motion, and a memory bank, PanoWorld combines three components within each DiT block: visual self-attention models the latent video content; DPRC encodes dense ray-conditioned motion features for precise action control; and GMA retrieves geometrically aligned memory features and injects them through confidence-guided gating for long-horizon consistency. The fused representation is then decoded into a trajectory-controlled panoramic video.

This spherical parameterization preserves the inherent $360^\circ$ topology and avoids the non-uniform distortion introduced by planar projections.

**Motion Manifold Construction.** To maintain geometric consistency during camera translation, the model needs to capture how each panoramic ray evolves under spatial displacement. Since our representation decouples rotation from translation, we set the camera rotation to $\mathbf{R}_t = \mathbf{I}$ and parameterize the camera motion solely by the translation $\mathbf{c}_t$. We construct a ray-conditioned motion representation through two successive transformations.

**Step 1: Ray-aligned Orthonormal Basis.** For each panoramic ray, we construct a right-handed ray-aligned orthonormal basis $\mathbf{B}^{\text{ray}} = [\mathbf{b}_x^{\text{ray}}, \mathbf{b}_y^{\text{ray}}, \mathbf{b}_z^{\text{ray}}]$, where $\mathbf{b}_z^{\text{ray}} = \mathbf{r}^{\text{cam}}$. The remaining axes are constructed from the world-up direction $\mathbf{u}^{\text{world}} = [0, 0, -1]^\top$ by setting $\mathbf{b}_y^{\text{ray}} = \text{normalize}(\mathbf{b}_z^{\text{ray}} \times \mathbf{u}^{\text{world}})$ and $\mathbf{b}_x^{\text{ray}} = \mathbf{b}_y^{\text{ray}} \times \mathbf{b}_z^{\text{ray}}$.

**Step 2: Ray-conditioned Transformation.** We transform the global camera translation into each ray-aligned coordinate frame:

$$\mathbf{T}_t^{\text{ray}} = \begin{bmatrix} (\mathbf{B}^{\text{ray}})^\top & -(\mathbf{B}^{\text{ray}})^\top \mathbf{c}_t \\ \mathbf{0}^\top & 1 \end{bmatrix} \in SE(3). \tag{3}$$

The translation term $-(\mathbf{B}^{\text{ray}})^\top \mathbf{c}_t$ expresses the camera displacement in the corresponding ray-aligned coordinate system. Since each ERP ray is associated with a distinct transformation, $\mathbf{T}_t^{\text{ray}}$ naturally provides spatially varying geometric conditions across the panoramic latent grid.

**Latent Motion Integration.** We directly use the ray-conditioned transformations $\mathbf{T}_t^{\text{ray}}$ as the transformations in PRoPE (Li et al., 2025b). For each panoramic latent token, its corresponding ERP ray determines a transformation that is directly applied to the $q/k/v/o$ features in the camera-conditioned self-attention pathway. This formulation enables ray-wise modeling of relative geometry under camera motion across the full $360^\circ$ field of view.

### 4.1.3 Memory Modeling: GMA

To ensure robust spatio-temporal consistency, we propose the GMA mechanism. Leveraging the rotation-equivariant panoramic ray space, GMA builds a reconstruction-free implicit geometric memory by associating previously observed regions through ray correspondence. The retrieved ray-aligned memory provides geometry-aware priors during generation, improving consistency under continuous camera motion.

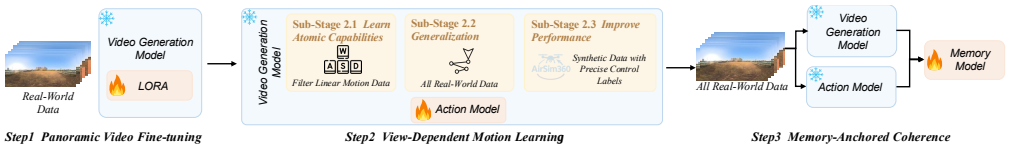


Figure 4: Overview of the progressive three-stage pipeline. *Stage 1* focuses on geometric adaptation for panoramic video generation; *Stage 2* targets view-dependent motion learning ; and *Stage 3* enables memory-anchored coherence to enforce long-term temporal consistency.

**Unified Projection for Query and Memory.** GMA unifies motion conditioning and memory retrieval within a shared geometric coordinate system by encoding query latents and memory bank features $\{\mathbf{x}_m\}$ into a common PRoPE space. Memory keys and values are formulated as :

$$[\mathbf{K}_{mem}, \mathbf{V}_{mem}] = \text{PRoPE}(\text{Linear}(\mathbf{x}_m), \hat{\mathbf{r}}_m), \tag{4}$$

where $\hat{\mathbf{r}}$ is the same ray-based representation used for motion control. This alignment enables the attention mechanism to retrieve historical features based on 3D ray correspondence rather than explicit image-space warping. Consequently, when the camera revisits previously observed locations, the retrieved features provide geometric and appearance consistency with the original observations.

**Confidence-Guided Gating and Fusion.** Since PRoPE injects ray information into memory tokens, the attention weights provide a ray-conditioned correspondence signal. We define the retrieval confidence for query token $i$ as

$$\alpha_{ij} = \text{softmax}_j \left( (\mathbf{Q}_i \mathbf{K}_{\text{mem},j}^\top)/\sqrt{d} \right), \tag{5}$$

The retrieval confidence is then defined as the maximum attention weight:

$$c_i = \max_j \alpha_{ij}. \tag{6}$$

The final feature is obtained through confidence-weighted fusion:

$$\mathbf{F}_{\text{final}}(i) = \mathbf{F}_{\text{base}}(i) + (g_i c_i) \odot \mathbf{F}_{\text{mem}}(i), \tag{7}$$

where $g_i$ is the learned adaptive gate. By anchoring memory injection to high-confidence historical ray correspondences, the proposed mechanism helps reduce visual flickering and structural drift during long-horizon generation, thereby improving the consistency of scene appearance throughout the trajectory.

### 4.2 Training Strategy

This section presents how the above components are optimized progressively rather than all at once, as illustrated in Fig. 4. We first adapt the backbone to panoramic geometry, then learn motion control, and finally activate memory modeling, which stabilizes optimization and helps each component acquire its intended role before full integration.

**Stage 1: Panoramic Video Fine-tuning.** The goal of this stage is to adapt the pre-trained video model to the panoramic domain. We use LoRA to fine-tune the backbone on panoramic datasets, enabling it to learn equirectangular geometric properties such as polar distortions and horizontal wrap-around consistency. To account for the varying pixel area in equirectangular projections, we employ a latitude-aware reconstruction loss:

$$\mathcal{L}_{\text{stage1}} = \mathbb{E}\left[ \cos(\phi) \cdot \|\mathbf{v}_\theta(\mathbf{x}_t, t) - \mathbf{v}^*\|_2^2 \right], \tag{8}$$

where $\phi$ denotes the latitude coordinate and $\mathbf{v}^*$ is the target velocity in the flow-matching formulation. The resulting backbone is adapted to the panoramic domain, providing a foundation for geometrically consistent synthesis and reducing structural tearing and boundary artifacts.

**Stage 2: View-Dependent Motion Learning.** In this stage, we freeze the LoRA weights and train the Action Stream via DPRC to master 3D translational control. By removing yaw rotation from training trajectories, we force the model to isolate depth-dependent parallax learning. DPRC models ego-motion as a dense re-indexing of panoramic rays, allowing the generator to learn intensity evolution relative to 3D spatial coordinates. This ray-geometry alignment enforces geometric consistency, ensuring stable scene appearance and structural integrity during complex maneuvers.

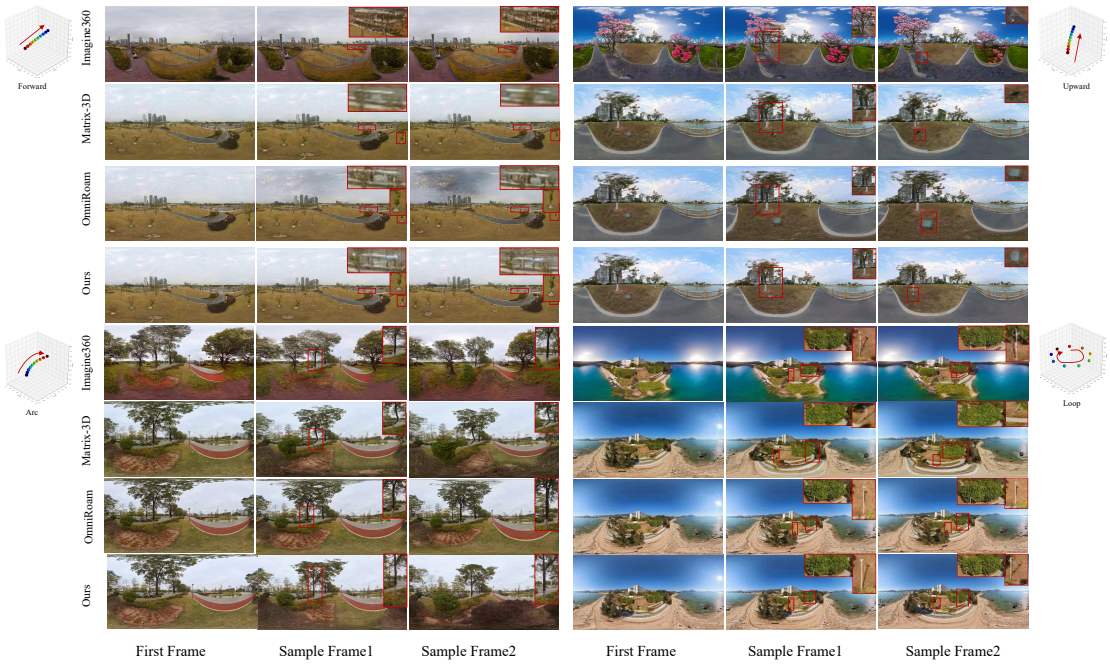


Figure 5: Qualitative comparison on real-world outdoor sequences.

**Stage 3: Memory-Anchored Coherence.** The final stage activates the GMA module, transitioning from autoregressive generation to memory-anchored world synthesis. Through confidence-guided gating, the model selectively incorporates historical observations as geometric anchors. By maintaining consistency with previously observed scene regions during camera revisits, GMA improves long-horizon scene coherence and reduces temporal artifacts. This memory-based formulation enables stable panoramic world generation under extended camera trajectories.

# 5 Experiments

## 5.1 Experiment Settings

**Implementation Detail** We implement PanoWorld by fine-tuning Wan2.2-5B (Wang et al., 2025a) with LoRA while preserving its pretrained priors. Following Sec. 4.2, action-conditioning and memory modules are optimized separately with the remaining parameters frozen. We provide models at both 480p and 720p resolutions under the same training protocol, and further enable real-time generation via Causal Forcing (Zhu et al., 2026). Detailed settings are provided in the Appendix.

**Evaluation Metrics** We evaluate our framework in terms of distributional fidelity, visual quality, and trajectory-conditioned generation fidelity. Specifically, we report FID (Heusel et al., 2017) and its regional variants ($\text{FID}_{\text{pole}}$ and $\text{FID}_{\text{equ}}$) for panoramic distributional fidelity (Sun et al., 2025b), FAED (Oh et al., 2022) for frame-level distributional similarity, and NIQE (Mittal et al., 2013) together with Q-Align ($\text{QA}_{\text{qual.}}$) (Wu et al., 2024) for no-reference visual quality. To evaluate trajectory controllability, we condition all methods on identical ground-truth (GT) trajectories following Ji et al. (2025); Liu et al. (2026) and compute temporally windowed PSNR between generated and GT sequences, where higher values indicate closer correspondence to the specified trajectory.

**Baselines** We compare PanoWorld with three panoramic video generation methods: Imagine360 (Tan et al., 2025), Matrix-3D (Yang et al., 2025), and OmniRoam (Liu et al., 2026). Matrix-3D and OmniRoam are evaluated following their original inference protocols. For OmniRoam, we further fine-tune the model on World360 for a fair comparison. We provide corresponding 480p and 720p

Table 2: Quantitative results on panoramic video generation. Best and second-best results are highlighted in red and yellow, respectively.

| Method | Resolution | Distribution Fidelity ↓ | | | Alignment ↓ | Perceptual Quality | |
|---|---|---|---|---|---|---|---|
| | | FID ↓ | $FID_{pole}$ ↓ | $FID_{equ}$ ↓ | FAED ↓ | NIQE ↓ | $QA_{qual.}$ ↑ |
| Imagine360 | 480p | 81.18 | 113.95 | 60.99 | 4.50 | 3.15 | 4.09 |
| Matrix-3D | 480p | 34.63 | 67.88 | 68.64 | 3.39 | 4.74 | 3.00 |
| OmniRoam | 480p | 62.50 | 87.10 | 46.20 | 3.48 | 3.52 | 3.68 |
| OmniRoam† | 480p | 60.77 | 85.25 | 45.92 | 3.36 | 3.39 | 3.75 |
| **Ours** | 480p | 27.64 | 47.21 | 26.00 | 2.63 | 3.85 | 4.02 |
| Matrix-3D | 720p | 35.35 | 70.27 | 44.06 | 2.86 | 3.63 | 3.75 |
| OmniRoam | 720p | 46.17 | 91.50 | 84.10 | 3.18 | 3.88 | 3.99 |
| OmniRoam† | 720p | 44.89 | 89.95 | 83.24 | 3.06 | 3.74 | 4.07 |
| **Ours** | 720p | 16.93 | 35.46 | 18.86 | 1.18 | 3.24 | 4.14 |

† denotes the model fine-tuned on the **World360** dataset.

Table 3: Motion control evaluation. "–" indicates metrics unavailable in the original implementation of the corresponding baseline.

| Method | Resolution | $PSNR_{F20-F25}$ | $PSNR_{F50-F55}$ | $PSNR_{F70-F75}$ | $PSNR_{F75-F80}$ |
|---|---|---|---|---|---|
| Imagine360 | 480p | – | – | – | – |
| Matrix-3D | 480p | 20.47 | 18.80 | 18.13 | 18.02 |
| OmniRoam | 480p | 18.22 | 17.31 | 17.06 | 16.80 |
| OmniRoam† | 480p | 18.51 | 17.59 | 17.34 | 17.02 |
| **Ours** | 480p | 22.83 | 21.65 | 21.04 | 20.92 |
| Matrix-3D | 720p | 20.50 | 18.86 | 18.26 | 18.16 |
| OmniRoam | 720p | 18.24 | 17.40 | 17.15 | 16.95 |
| OmniRoam† | 720p | 18.63 | 17.78 | 17.37 | 17.30 |
| **Ours** | 720p | 22.94 | 21.71 | 21.10 | 20.99 |

$PSNR_{F20-F25}$ denotes the mean PSNR computed over frames 20–25; the other temporal windows are defined analogously.

PanoWorld models under the same framework. Imagine360 outputs are resized and temporally interpolated to match the 81-frame evaluation setting.

### 5.2 Results and Comparisons

**Qualitative Results.** Figure 5 evaluates visual fidelity and motion consistency across diverse real-world trajectories. PanoWorld maintains precise trajectory adherence and high structural stability without distortion. In contrast, Matrix-3D suffers from severe motion blur and disocclusion artifacts during altitude shifts, while OmniRoam is constrained by its strong fixed-height training prior, failing to yield motion parallax during translation. Overall, PanoWorld better mitigates structural drift and preserves environmental persistence during complex maneuvers.

**Quantitative Analysis of Visual Quality.** As shown in Table 2, PanoWorld achieves the best performance on the distribution-fidelity metrics, obtaining the lowest FID (27.64), $FID_{pole}$ (47.21), and $FID_{equ}$ (26.00). These results demonstrate strong 360° structural fidelity and improved consistency across polar and equatorial regions. For visual quality, PanoWorld achieves a competitive $QA_{qual.}$ score of 4.02, close to Imagine360's 4.09.

**Quantitative Analysis of Trajectory Control.** As Table 3 shows, PanoWorld consistently outperforms Matrix-3D and OmniRoam across all temporal windows with superior PSNR and path adherence. Notably, OmniRoam suffers substantial structural drift during elevation shifts. Camera poses reconstructed via ViPE (Huang et al., 2025a) further confirm PanoWorld's tightest alignment with ground-truth trajectories (see Appendix for full ViPE pose-estimation metrics), verifying its robustness in executing precise maneuvers in large-scale environments.

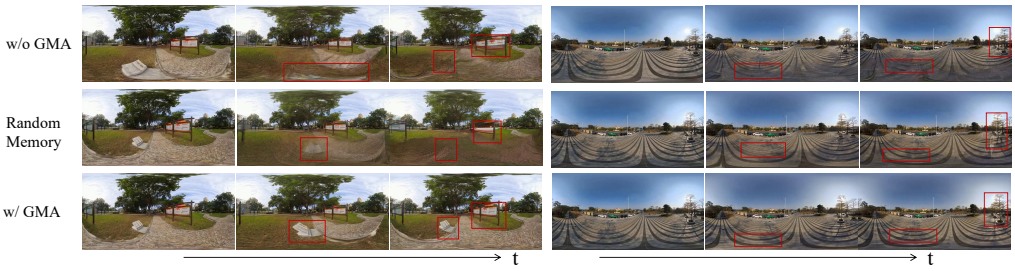


Figure 6: Qualitative ablation results of the GMA module. Comparison of synthesized video frames under control: (a) w/o GMA baseline exhibits significant artifacts; (b) Random Memory variant results in broken geometry; (c) Full setting with GMA maintains consistent generation.

Table 4: Ablation study of camera-motion injection and memory modules. All values are PSNR (↑).

| Condition Representation | Injection | 20–25 | 50–55 | 70–75 | 75–80 |
|---|---|---|---|---|---|
| *Camera-Motion Injection (F0–F80)* | | | | | |
| Frame Pose | Concat | 16.49 | 15.35 | 14.88 | 14.61 |
| ERP Coord. | Concat | 17.01 | 16.16 | 15.67 | 15.26 |
| Ray-feature | XAttn. | 18.42 | 17.33 | 16.31 | 16.19 |
| Ray-wise SE(3) | AdaLN | 18.67 | 17.42 | 16.82 | 16.71 |
| Ray-wise SE(3) | Concat | 20.45 | 19.32 | 18.68 | 18.46 |
| Ray-wise SE(3) | **DPRC** | **22.83** | **21.65** | **21.04** | **20.92** |
| *Memory Module (F81–F161)* | | | | | |
| w/o GMA | | 20.41 | 19.12 | 18.53 | 18.36 |
| Random Memory | | 14.06 | 12.28 | 11.97 | 12.11 |
| **w/ GMA** | | **21.35** | **20.19** | **19.61** | **19.44** |

### 5.3 Ablative Studies

We conduct ablation experiments on camera-motion injection and memory modeling in Table 4. For camera-motion injection, DPRC consistently outperforms alternative mechanisms, demonstrating the effectiveness of ray-wise SE(3) motion conditioning. It also outperforms frame-level pose and ERP-coordinate representations, highlighting the benefit of fine-grained ray-wise geometric conditioning.

For memory modeling, GMA improves scene consistency under extended camera trajectories, achieving a $+1.08$ dB PSNR gain over *w/o GMA*. The model first generates an 81-frame segment to initialize the memory before generating subsequent frames autoregressively. For *Random Memory*, the same injection process and gate are used, with only the memory content randomly replaced.

Qualitative comparisons are shown in Fig. 6. Without GMA, videos suffer from accumulated drift, while *Random Memory* degrades due to irrelevant memory content under the same injection process. In contrast, GMA preserves more stable scene structures through geometry-aware memory. Further analyses, including revisit consistency evaluation and additional quantitative studies, are provided in the Appendix.

## 6 Conclusion

In this work, we presented PanoWorld, a framework for controllable panoramic video generation. By leveraging the rotation-equivariance of omnidirectional representations, we decouple rotations to simplify motion learning. Based on the unique properties of panoramic rays, our proposed DPRC and GMA modules ensure robust geometric and temporal consistency for controllable panoramic world generation. To support this research, we introduced World360, a large-scale dataset featuring both real-world UAV captures and high-quality simulations. Extensive experiments on this benchmark demonstrate that PanoWorld significantly outperforms state-of-the-art methods in scene persistence and structural integrity, offering a robust solution for unconstrained panoramic world synthesis.

### AI USE STATEMENT

Large Language Models (LLMs) were used to assist manuscript preparation, limited to language refinement, grammar checking, and readability improvements. Specifically, ChatGPT-5.5 and Qwen3.8-Max were used for grammar checking and language polishing. All AI-assisted modifications were reviewed and verified by the authors, who take full responsibility for the final content of this work, including all text, claims, and artifacts.

LLMs were not involved in the development of research ideas, algorithm design, experimental setup, data analysis, or interpretation of results. In addition, a pretrained vision-language model, Qwen3-VL, was used only as a fixed annotation tool with frozen weights and predefined prompts to generate environment-focused captions for the World360 dataset (Appendix A.3, Step 1). These generated captions serve as dataset annotations and do not constitute scientific claims or contributions of this work.

No LLM is listed as an author, and all hypotheses, experiments, and conclusions were developed and verified by the authors.

## ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. Our **World360** dataset contains real-world panoramic videos captured from an aerial imaging platform, which may include individuals, vehicles, buildings, and other location-specific visual information. We recognize that such data may introduce privacy and geoprivacy risks, particularly when panoramic imagery is associated with spatial information.

To mitigate these risks, all data were collected in locations where the required flight permissions were obtained and in accordance with applicable regulations. Before inclusion in the dataset, sensitive personal information, including faces and license plates, was anonymized. Precise geographic information was also removed or anonymized to prevent direct association between released imagery and specific real-world locations. The released data therefore retain only the relative spatial information necessary for the research tasks, without exposing precise personal location information.

The proposed framework is intended for research on panoramic video generation, scene understanding, and controllable camera motion. We acknowledge that realistic panoramic world modeling is potentially dual-use and could be applied beyond these research objectives. We therefore discourage the use of the released data and models for privacy-invasive applications, unauthorized surveillance, or other uses that violate applicable laws, safety requirements, or ethical standards. We also provide a mechanism for reporting and requesting the removal of content that may present unforeseen privacy or safety concerns.

All dataset collection, processing, and release procedures are conducted with attention to privacy, data-use requirements, and applicable regulations.

## REPRODUCIBILITY STATEMENT

We have made every effort to ensure that the results presented in this paper are reproducible. The experimental setup, including training configurations, model architectures, optimization details, and evaluation protocols, is described in the paper and supplementary material. We provide detailed descriptions of PanoWorld, including the proposed DPRC and GMA modules and the progressive three-stage training pipeline, to facilitate reproduction of our approach.

For data reproducibility, we provide a comprehensive description of the construction process of **World360**, including data collection, preprocessing, filtering criteria, and benchmark settings. The public datasets and resources used in this work are properly documented to ensure consistent evaluation. The code, training configurations, and necessary resources will be released in an anonymous repository to facilitate replication and verification. We believe these efforts will enable future researchers to reproduce our results and further advance panoramic world modeling.

## REFERENCES


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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 2

Ting-Hsuan Chen, Ying-Huan Chen, Tao Tu, Jie-Ying Lee, Cho-Ying Wu, Fangzhou Lin, Hengyuan Zhang, David Paz, Xinyu Huang, Yuliang Guo, Yu-Lun Liu, Yue Wang, and Liu Ren. Pantheon360: Taming digital twin generation via 3D-aware 360° video diffusion. *arXiv preprint arXiv:2605.25449*, 2026. 3

Yu-Cheng Chou, Xingrui Wang, Yitong Li, Jiahao Wang, Hanting Liu, Cihang Xie, Alan Yuille, and Junfei Xiao. Captain safari: A world engine. *arXiv preprint arXiv:2511.22815*, 2025. 2, 3

Chuan Fang, Lingteng Qiu, Yixun Liang, Rui Chen, Kunming Luo, Zhaohua Zheng, Tongyuan Bai, Feipeng Tian, Zilong Dong, Zihan Zhou, et al. Spatialcrafter: Single image world modeling with generative 3d proxies. *arXiv preprint arXiv:2608.27073*, 2026. 1

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. 1

Xian Ge, Yuling Pan, Yuhang Zhang, Xiang Li, Weijun Zhang, Dizhe Zhang, Zhaoliang Wan, Xin Lin, Xiangkai Zhang, Juntao Liang, et al. Airsim360: A panoramic simulation platform within drone view. *arXiv preprint arXiv:2512.02009*, 2025. 2, 4

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 2

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 *Advances in neural information processing systems*, 2017. 7

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*, 2025a. 8

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *ACM Transactions on Graphics (TOG)*, 44(6): 245:1–245:15, 2025b. 1

Chenhao Ji, Chaohui Yu, Junyao Gao, Fan Wang, and Cairong Zhao. Campvg: Camera-controlled panoramic video generation with epipolar-aware diffusion. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 23:1–23:12, 2025. 3, 7

Jinrang Jia, Zhenjia Li, Yijiang Hu, and Yifeng Shi. Panoworld: A generative spatial world model for consistent whole-house panorama synthesis. *arXiv preprint arXiv:2605.17916*, 2026. 3

Renjie Li, Panwang Pan, Bangbang Yang, Dejia Xu, Shijie Zhou, Zeming Li, Achuta Kadambi, Zhangyang Wang, Zhengzhong Tu, Zhiwen Fan, et al. 4k4dgen: Panoramic 4d generation at 4k resolution. In *International Conference on Learning Representations*, 2025a. 2

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. In *Advances in Neural Information Processing Systems*, 2025b. 5

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25690–25699. IEEE, 2025c. 3

Xin Lin, Xian Ge, Dizhe Zhang, Zhaoliang Wan, Xianshun Wang, Xiangtai Li, Wenjie Jiang, Bo Du, Dacheng Tao, Ming-Hsuan Yang, et al. One flight over the gap: A survey from perspective to panoramic vision. *arXiv preprint arXiv:2509.04444*, 2025. 1

Jinxiu Liu, Shaoheng Lin, Yinxiao Li, and Ming-Hsuan Yang. Dynamicscaler: Seamless and scalable video generation for panoramic scenes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6144–6153, 2025. 2

Yuheng Liu, Xin Lin, Xinke Li, Baihan Yang, Chen Wang, Kalyan Sunkavalli, Yannick Hold-Geoffroy, Hao Tan, Kai Zhang, Xiaohui Xie, Zifan Shi, and Yiwei Hu. OmniRoam: World wandering via long-horizon panoramic video generation. In *ACM SIGGRAPH Conference Papers*, pp. 34:1–34:12, 2026. doi: 10.1145/3799902.3811180. 3, 7

TaiMing Lu, Tianmin Shu, Alan Yuille, Daniel Khashabi, and Jieneng Chen. Generative world explorer. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8NlUL0Cv1L. 2

Rundong Luo, Matthew Wallingford, Ali Fahardi, Noah Snavely, and Wei-Chiu Ma. Beyond the frame: Generating 360 panoramic videos from perspective videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14336–14345. IEEE, 2025. 3

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2013. 7

Changgyoon Oh, Wonjune Cho, Yujeong Chae, Daehee Park, Lin Wang, and Kuk-Jin Yoon. Bips: Bi-modal indoor panorama synthesis via residual depth-aided adversarial learning. In *European Conference on Computer Vision*, volume 13676, pp. 352–371, 2022. 7

Katja Schwarz, Norman Mueller, and Peter Kontschieder. Generative gaussian splatting: Generating 3d scenes with video diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27510–27520, 2025. 2

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 52810–52835, 2025. 3

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025a. 1

Xiancheng Sun, Mai Xu, Shengxi Li, Senmao Ma, Xin Deng, Lai Jiang, and Gang Shen. Spherical manifold guided diffusion model for panoramic image generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5824–5834. IEEE, 2025b. 7

Jing Tan, Shuai Yang, Tong Wu, Jingwen He, Yuwei Guo, Ziwei Liu, and Dahua Lin. Imagine360: Immersive 360 video generation from perspective anchor. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 3, 7

Junshu Tang, Jiacheng Liu, Jiaqi Li, Longhuang Wu, Haoyu Yang, Penghao Zhao, Siruis Gong, Xiang Yuan, Shuai Shao, Linfeng Zhang, et al. Hunyuan-gamecraft-2: Instruction-following interactive game world model. *arXiv preprint arXiv:2511.23429*, 2025. 1

Matthew Wallingford, Anand Bhattad, Aditya Kusupati, Vivek Ramanujan, Matt Deitke, 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. In *Advances in Neural Information Processing Systems*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/1fa3d6ccbcd15f7285fee666b2bd57be-Abstract-Conference.html. 2, 3

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, Xiaofeng Meng, 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*, 2025a. 7

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*, 2025b. 3

Qian Wang, Weiqi Li, Chong Mou, Xinhua Cheng, and Jian Zhang. 360dvd: Controllable panorama video generation with 360-degree video diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6913–6923, 2024a. 2, 3

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, 2024b. 2

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54015–54029, 2024. 7

Wei Wu, Fan Lu, Yunnan Wang, Shuai Yang, Shi Liu, Fangjing Wang, Qian Zhu, He Sun, Yong Wang, Shuailei Ma, et al. A pragmatic vla foundation model. *arXiv preprint arXiv:2601.18692*, 2026a. 1

Yu Wu, Minsik Jeon, Jen-Hao Rick Chang, Oncel Tuzel, and Shubham Tulsiani. Rayrope: Projective ray positional encoding for multi-view attention. *arXiv preprint arXiv:2601.15275*, 2026b. 3

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 *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. 2, 3

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*, 2025. 3

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. 2

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. 2, 3, 7

Minghao Yin, Jiahao Lu, Wenbo Hu, Wang Zhao, Shan Ying, and Kai Han. Raype: Ray-space positional encoding for 3d-aware video generation. *arXiv preprint arXiv:2606.27345*, 2026. 3

Yuyang Yin, HaoXiang Guo, Fangfu Liu, Mengyu Wang, Hanwen Liang, Eric Li, Yikai Wang, Xiaojie Jin, Yao Zhao, and Yunchao Wei. Panoworld-x: Generating explorable panoramic worlds via sphere-aware video diffusion. *arXiv preprint arXiv:2509.24997*, 2025. 3

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*, pp. 19:1–19:11, 2025. 3

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. *arXiv preprint arXiv:2409.02048*, 2024. 2

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*, pp. 38027–38037, 2026. 2, 3

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 10040–10048, 2025a. 2

Zhaoyang Zhang, Yannick Hold-Geoffroy, Miloš Hašan, Ziwen Chen, Fujun Luan, Julie Dorsey, and Yiwei Hu. Generating 360° video is what you need for a 3d scene. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 111:1–111:12, 2025b. 2

Yang Zhou, Ziheng Wang, Yuqin Lu, Haofeng Liu, Jun Liang, Shengfeng He, and Jing Li. Moverse: Real-time video world modeling with panoramic gaussian scaffold. *arXiv preprint arXiv:2606.13376*, 2026. 3

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026. 7

# Supplementary Material for PanoWorld: Real-World Panoramic Generation

## A Datasets Pre-processing

We construct two complementary datasets for model training, collectively referred to as *World360*, as illustrated in Fig. 7. The first is *Scene-Reality Data* (real-world data), comprising large-scale panoramic videos captured across diverse physical environments using a 360° aerial imaging platform. This dataset provides natural aerial camera motions and diverse visual appearances, improving the generalization of panoramic world modeling in real scenarios. The second is *Trajectory-Precise Data* (synthetic data), generated with AirSim360. By leveraging high-fidelity simulation and accurate trajectory annotations, this dataset provides precise motion supervision and geometric consistency for controllable panoramic video generation.

The overall statistics of World360, including the number of environments, flight sequences, raw frames, and temporally sampled training clips, are summarized in Table 5. To further characterize the motion diversity within individual generation windows, we report clip-level altitude variation statistics in Table 6, where altitude variations are defined as the difference between the maximum and minimum altitude within each clip, while the non-planar ratio denotes the proportion of clips whose altitude variation exceeds 0.5 m.

### A.1 Scene-Reality Real Data

The Scene-Reality Learning Dataset is a large-scale synchronized video-inertial dataset captured by panoramic Unmanned Aerial Vehicles (UAVs) in real-world environments. It contains 381 flight sequences from 300 outdoor scenes, with over 6.8 million panoramic frames and corresponding camera trajectories estimated by a high-precision GPS-aided Inertial Navigation System (INS). The trajectories are horizontally rectified to remove roll and pitch variations, retaining translation and yaw rotation for consistent panoramic motion modeling. The dataset covers diverse environments, including parks, roadways, artificial lakes, and architectural complexes, while containing diverse real-world aerial trajectories with varying spatial layouts and altitude variations.

**Temporal Alignment.** To ensure strict temporal consistency between the high-rate kinematic data and the panoramic visual stream, we perform a multi-stage synchronization process. This is necessitated by the asynchronous logging of the panoramic camera (59.94 FPS) and the GPS-fused system (200 Hz). The primary objective is to establish a unified timeline by accurately anchoring the video sequence within the continuous pose stream.

In the first step, we perform initial timestamp alignment by matching the camera's exposure start signal with the corresponding entry in the pose sequence to determine the absolute reference time $t_0$. Once the global starting point is anchored, we reconstruct the discrete timeline for the subsequent video frames using the constant frame rate:

$$t_i = t_0 + i \cdot \frac{1}{\text{FPS}}, \tag{9}$$

where $t_i$ represents the target timestamp for the $i$-th frame. Given that the 200 Hz pose data is recorded on an independent clock, we synchronize the two modalities by matching each calculated $t_i$ to the raw pose sequence via nearest-neighbor search. To mitigate sampling offsets and ensure a smooth trajectory, the discrete pose sequence is resampled to these exact video timestamps using linear interpolation:

$$\mathbf{p}(t_i) = \text{Interp}(t_i, \mathcal{P}_{\text{raw}}), \tag{10}$$

where $\mathcal{P}_{\text{raw}}$ denotes the raw trajectory. Finally, any leading video frames that precede the first available pose record are discarded to guarantee that every frame is paired with a valid, synchronized 6-DoF state.

**Rotational Decoupling and Coordinate Alignment** To eliminate excessive camera self-rotation and establish a geometrically consistent mapping between visual content and camera motion, we

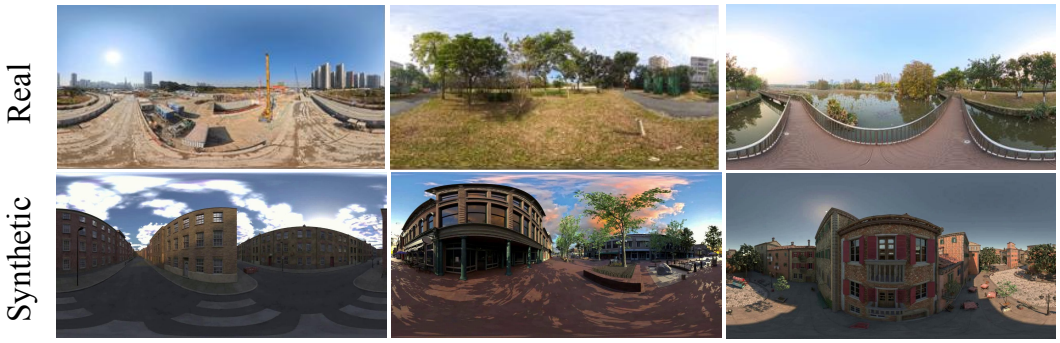


Figure 7: Visualization of the **World360** dataset. The first row shows representative real-world sequences collected by a 360° aerial imaging platform, while the second row presents simulated sequences generated with AirSim360.

Table 5: Statistics of the World360 dataset. Flight sequences denote raw UAV recordings or simulated trajectories, while clips are generated through spatial sliding-window sampling for different training configurations.

| Source | Environments | Flight sequences | Raw frames | 81-frame clips | 161-frame clips |
|---|---|---|---|---|---|
| Real UAV | 300 | 381 | 6.85M | 70.0K | 50.1k |
| AirSim360 | 20 | 4318 | 8.00M | 50.0K | 30.2k |

perform rotational decoupling and coordinate alignment. Due to the unconstrained nature of aerial flight, raw panoramic videos may contain substantial ego-rotation, including rapid yaw changes and irregular rotational variations, which complicate the correspondence between camera motion and scene evolution. Meanwhile, the raw 6-DoF poses are logged in the Earth-fixed East-North-Up (ENU) coordinate system, which is not aligned with the camera's initial viewing direction. We therefore transform the trajectories into a unified canonical camera coordinate system and decouple rotational motion from translation.

Specifically, we first establish a canonical coordinate frame by aligning the ENU coordinate system with the camera coordinate system of the initial frame. Let $\mathbf{R}_t^{\mathrm{ENU}} \in SO(3)$ denote the camera orientation at frame $t$ obtained from the INS metadata. The camera orientation in the canonical frame is given by

$$\mathbf{R}_t = \left(\mathbf{R}_0^{\mathrm{ENU}}\right)^\top \mathbf{R}_t^{\mathrm{ENU}}, \tag{11}$$

where $\mathbf{R}_0^{\mathrm{ENU}}$ specifies the initial alignment between the ENU and canonical camera frames. The corresponding relative translation is defined as

$$\mathbf{t}_t = \left(\mathbf{R}_0^{\mathrm{ENU}}\right)^\mathsf{T} \left(\mathbf{t}_t^{\mathrm{ENU}} - \mathbf{t}_0^{\mathrm{ENU}}\right). \tag{12}$$

This canonicalization removes the sequence-dependent global orientation and origin while preserving the relative camera motion.

To mitigate high-frequency sensor noise, we further refine the camera motion through a hybrid visual-inertial optimization. Sparse feature correspondences $(u, v)$ between consecutive ERP frames are lifted onto the unit sphere $\mathbb{S}^2$ according to the spherical coordinate convention adopted throughout the dataset. For an ERP frame of width $W$ and height $H$, the longitude $\theta$ and latitude $\phi$ corresponding to pixel $(u, v)$ are defined as

$$\theta = \left(\frac{u+0.5}{W} - \frac{1}{2}\right) 2\pi, \qquad \phi = \left(\frac{1}{2} - \frac{v+0.5}{H}\right) \pi. \tag{13}$$

The associated unit ray in the canonical camera frame is

$$\mathbf{p}^{\mathrm{cam}} = \begin{bmatrix} -\cos\phi\cos\theta \\ -\cos\phi\sin\theta \\ -\sin\phi \end{bmatrix}. \tag{14}$$

Table 6: Clip-level altitude variation statistics of World360. $|\Delta z|$ denotes the height variation range computed as the difference between the maximum and minimum altitude of each clip.

| Source | Clip Length | $\|\Delta z\|$ Median (m) | $\|\Delta z\|$ 95% (m) | Non-Planar Ratio |
|---|---|---|---|---|
| Real UAV | 81 frames | 0.76 | 1.69 | 50.5% |
| | 161 frames | 1.25 | 3.19 | 59.9% |
| AirSim360 | 81 frames | 0.98 | 2.00 | 59.7% |
| | 161 frames | 2.05 | 4.00 | 60.5% |

For a matched correspondence, $\mathbf{p}_i^{\mathrm{cam}}$ and $\mathbf{q}_i^{\mathrm{cam}}$ denote the corresponding unit ray directions in two consecutive ERP frames. The relative camera motion is then refined by minimizing the spherical reprojection error jointly with the inertial measurements, thereby suppressing high-frequency orientation noise while preserving consistency with the INS-derived motion.

All spherical rays, camera poses, and subsequent geometric operations follow this canonical camera-frame convention. In particular, the third coordinate axis defines the vertical direction, and the above ray parameterization is used consistently for spherical unprojection, pose transformation, and geometry-aware conditioning.

The residual yaw offset is estimated using a RANSAC-based alignment initialized by the INS yaw measurement:

$$\Delta\psi^\star = \arg\min_{\Delta\psi} \sum_i \|R_z(\Delta\psi)\mathbf{p}_i - \mathbf{q}_i\|_2^2 , \tag{15}$$

where $R_z(\Delta\psi)$ denotes the rotation around the gravity axis. In feature-deficient environments (e.g., open water surface or sky regions), the system automatically falls back to an INS-prior yaw-only estimate combined with latitude-weighted phase correlation along the longitudinal direction. The calculated rotation sequence is smoothed via a one-dimensional constant-velocity Kalman filter. Finally, the rotationally decoupled and stabilized visual stream is generated by applying the inverse rotation matrix to the spherical sampling grid and reprojecting it back to the canonical ERP format via grid remapping, successfully removing camera self-rotation while ensuring that the processed 6-DoF poses are strictly aligned with the canonical video frames.

**Pose Noise Suppression** Due to aerodynamic turbulence, motor vibrations, and occasional GPS multi-path effects during outdoor UAV flights, the raw synchronized 6-DoF poses inevitably contain high-frequency fluctuations and sensor-induced jitter. To obtain physically plausible trajectories while preserving meaningful aerial motion patterns, we apply a two-stage pose refinement strategy.

First, a hysteresis filter is applied to suppress isolated low-amplitude fluctuations and stabilize the trajectory baseline. Second, a one-dimensional Gaussian smoothing kernel is applied along the temporal dimension to ensure continuous and differentiable motion. Considering that horizontal UAV movements usually contain stronger high-frequency oscillations due to flight dynamics and maneuvering, we adopt a larger smoothing scale for horizontal components $(x, y)$ with $\sigma = 1.0$, while using a smaller scale for the vertical component $(z)$ with $\sigma = 0.3$ to preserve genuine altitude variations.

**Spatial-Uniform Trajectory Resampling and Visual Alignment** In real-world data collection, UAVs inevitably operate at varying velocities due to path-following adjustments, manual control, and aerodynamic effects. Consequently, uniform temporal sampling leads to non-uniform spatial distributions, with high-speed segments becoming spatially sparse and low-speed or hovering segments densely sampled. Such stationary frames introduce temporal redundancy and provide limited supervision for motion-conditioned learning. To obtain a consistent spatial motion scale, we re-index and resample synchronized video–trajectory pairs in the spatial arc-length domain rather than the temporal domain, removing prolonged stationary intervals and ensuring that consecutive frames correspond to a fixed physical displacement. The complete engineering pipeline for spatial normalization and sequence generation is implemented through the following structured steps:

**Step 1: Trajectory Filtering and 3D Cumulative Arc-Length Calculation.** To successfully map the dataset into the spatial domain, we must first establish a continuous, monotonically increasing

spatial metric that reflects the true total distance traversed by the drone. Since our platform is an omnidirectional panoramic UAV, it undergoes complex 3D aerial dynamics that include extensive vertical climbs and descents along the gravitational axis, meaning that a simplified 2D planar distance would fail to capture the vertical motion and distort the alignment. Therefore, to accurately quantify the physical movement, we explicitly integrate the $z$-component into the metric computation and calculate the cumulative three-dimensional Euclidean arc length $s_i$ from the initial frame to the $i$-th frame as follows:

$$s_i = \sum_{k=1}^{i} \Delta s_k = \sum_{k=1}^{i} \sqrt{(x_k - x_{k-1})^2 + (y_k - y_{k-1})^2 + (z_k - z_{k-1})^2}, \tag{16}$$

where $\mathbf{P}_i = (x_i, y_i, z_i)^T$ represents the discrete 3D position states, and the boundary condition is initialized as $s_0 = 0$.

**Step 2: Uniform Spatial Re-indexing and Pose Interpolation.** Before spatial re-indexing, we first remove invalid trajectory segments caused by stationary periods and temporal inconsistencies. Specifically, intervals with negligible spatial displacement are filtered out to avoid redundant supervision, and segments with inconsistent timestamps between visual frames and pose measurements are discarded. With the continuous metric prepared, our next objective is to generate a normalized trajectory benchmark where the kinematic states are uniformly distributed in space. By filtering out temporal steps where $\Delta s_k \approx 0$, this uniform spatial grid inherently compresses the redundant static segments into a single spatial point, thereby balancing the distribution of dynamic motion increments across the entire dataset. To implement this, we define a target spatial grid with a fixed physical interval of $\Delta s = 0.05\,\text{m}$, formulating the discrete sampling points $s'_j$ as:

$$s'_j = j \cdot \Delta s, \quad j \in \left\{0, 1, \ldots, \left\lfloor \frac{s_{\max}}{\Delta s} \right\rfloor \right\}. \tag{17}$$

We then map all continuous kinematic state variables from the original temporal reference frame onto this newly defined spatial grid via continuous piecewise linear interpolation:

$$f(s'_j) = \text{Interp}(s'_j, \mathcal{S}_{\text{raw}}, f(\mathcal{S}_{\text{raw}})), \quad f \in \{x, y, z, \psi\}, \tag{18}$$

where $\mathcal{S}_{\text{raw}} = \{s_0, s_1, \ldots, s_N\}$ represents the raw non-uniform spatial coordinate set, and $\psi$ denotes the corresponding yaw orientation angle. Note that since the trajectory has already been rotationally decoupled and stabilized in the upstream stage, the Yaw angles can be directly interpolated without phase-wrapping artifacts.

**Step 3: Image Frame Differentiation and Pose-Guided Blending.** Once the spatially uniform re-sampling positions are determined from the pose stream, the corresponding visual frames are aligned to these target spatial locations. Since the target positions generally do not coincide with the discrete exposure times of the original video, directly selecting the nearest raw frame may introduce spatial quantization errors and temporal judder. We therefore identify the two adjacent raw frames that bracket each target position and perform linear blending between them. Specifically, for a target position $s$ lying between two consecutive raw frames at positions $s_k$ and $s_{k+1}$, the interpolated frame is computed as

$$I(s) = (1 - \alpha) I_k + \alpha I_{k+1}, \qquad \alpha = \frac{s - s_k}{s_{k+1} - s_k}, \quad \alpha \in (0.1, 0.9), \tag{19}$$

where $I_k$ and $I_{k+1}$ denote the corresponding raw video frames, and the positions $s_k$, $s_{k+1}$, and $s$ are determined from the camera pose stream. Thus, the interpolation weight is derived from the relative spatial displacement rather than the temporal frame index. This procedure aligns the visual stream with the uniformly sampled spatial trajectory, yielding a target inter-frame displacement of $0.05$,m. The resulting sequences are subsequently encoded using automatic codec selection to support cross-platform deployment.

**Step 4: Fixed-Length Trajectory Slicing.** To support different training objectives of our framework, the continuous spatially-uniformized video–trajectory sequences are partitioned into fixed-length clips using a sliding-window strategy. Specifically, we extract clips with length $L_1 = 81$ frames for motion-conditioned video generation, while longer clips with length $L_2 = 161$ frames are extracted for optimizing the geometry-aware memory module.

For a target sequence length $L \in \{81, 161\}$, a clip starting from the $i$-th resampled state is defined as:

$$\mathcal{S}_i = \{\mathbf{P}_j \mid j = i, \ldots, i + L - 1\}, \tag{20}$$

where $\mathbf{P}_j$ denotes the resampled trajectory state at the $j$-th spatial position. The sliding stride is configured according to the corresponding training requirements, allowing partially overlapping clips to improve temporal coverage and sample diversity.

After slicing, trajectory segments shorter than the required clip length are discarded to ensure consistent temporal supervision. For the remaining tail regions where sufficient frames are available, a backward-aligned window is additionally retained to maximize the utilization of valid trajectory portions:

$$\mathcal{S}_{\text{tail}} = \{\mathbf{P}_j \mid j = M - L, \ldots, M - 1\}. \tag{21}$$

### A.2 Trajectory-Precise Synthetic Data

To complement our real-world dataset and provide pristine, noise-free geometric supervision, we curate a trajectory-precise synthetic dataset sourced from the AirSim360 simulator across 20 diverse outdoor environments. Unlike real-world capture pipelines that suffer from sensor noise, clock drift, and interpolation residues, the simulation platform inherently guarantees perfect, hardware-level spatiotemporal synchronization between the panoramic visual frames and the ground-truth 6-DoF poses. Leveraging this synchronized framework, we develop an efficient, high-throughput automated data collection engine designed to achieve strictly spatial-uniform data acquisition.

To implement efficient 3D trajectory planning while preventing the camera agent from getting trapped in structurally fragmented spaces or wasting rendering resources on uninformative regions, we devise a hierarchical layer-wise filtering and voxelization strategy. First, the continuous 3D free space of each outdoor scene is discretized into a uniform 3D macro-grid with a resolution of $1\,\text{m} \times 1\,\text{m} \times 1\,\text{m}$. Since complex aerial trajectories naturally involve extensive elevation changes, the efficiency of traditional unconstrained 3D exploration heavily degrades in altitude layers that are overly cluttered or lack traversable continuity. To resolve this and maximize the throughput of the collection engine, our pipeline pre-calculates the exact number of traversable grid cells for each discrete altitude layer. If the count of valid candidate grids within a specific altitude layer falls below a predefined density threshold, indicating insufficient space for meaningful aerial maneuvers, the engine **automatically bypasses the entire altitude layer**. This proactive layer-wise filtering forces the system to allocate computation exclusively to open, geometrically rich elevations.

For the selected valid altitude layers, candidate anchor points are populated on a dense sub-grid with a refined resolution of $0.05\,\text{m}$ to ensure precise spatial-uniform sampling. To realistically mimic real-world aerodynamic flight statistics and maximize trajectory diversity, the engine is programmed to operate under three complementary trajectory generation modes:

- **Mode 1: Multi-Directional Homogeneous Maneuver.** To establish highly structured multi-view illumination and motion consistency from a unified initial state, the agent departs from a fixed initial waypoint $\mathbf{P}_{\text{start}}$ and executes flight paths along three distinct directional headings, exploring a diverse repertoire of motion profiles (including straight lines, horizontal/vertical curves, and continuous 3D arcs). Each directional flight continues until an environmental obstacle within the $1\,\text{m}$ grid framework is detected, at which point the agent transitions into a stationary hovering state. Crucially, to filter out trivial fragments and ensure sufficient spatial scale for learning long-horizon consistency, a multi-directional sequence is considered valid and logged into the corpus if and only if the trajectories generated along all three directional branches concurrently exceed a predefined threshold length $\mathcal{L}_{\text{min}}$.
- **Mode 2: High-Coverage Autonomous Exploration.** To encapsulate intricate 3D maneuvers and maximize environmental familiarity across complex topography, the agent executes active scene coverage from a randomized spatial initialization $\mathbf{P}_{\text{rand}}$. Utilizing a 3D A* algorithm optimized over the $1\,\text{m}$ macro-grid map, the agent dynamically calculates collision-free paths to explore the environment. The trajectory generation process automatically terminates once the cumulative visibility volume covers at least 70% of the total traversable free space within the current scene, ensuring massive spatial diversity.
- **Mode 3: Randomized Multi-Goal Navigation.** To simulate unconstrained point-to-point transit and capture generic 6-DoF flight statistics, the engine randomly samples both the

initial waypoint $\mathbf{P}_{\text{rand_start}}$ and the target destination $\mathbf{P}_{\text{rand_end}}$ across the unoccluded 3D grid space. A globally optimal, collision-free trajectory connecting these two randomized terminal states is then generated using the 3D A* pathfinder.

To better approximate the smooth, continuous pose transitions observed in real aerial flights and eliminate sharp velocity discontinuities, the piecewise linear edges generated between consecutive anchor points in Mode 2 and Mode 3 are modified using a geometric smoothing procedure. Near each anchor point, adjacent linear segments are replaced with continuous circular arcs constructed to be tangent to the edge vectors while rigorously validating against 3D collision constraints. Since the vicinity spheres of adjacent anchor points may overlap, the tangent offset for each circular arc is limited to 0.3 times the corresponding edge length, ensuring continuous velocity profiles and realistic 6-DoF angular transitions. Finally, continuous panoramic video sequences are rendered along these normalized trajectories at a precise, constant spatial increment of $\Delta s = 0.05\,\text{m}$ directly from the underlying 3D scene representations, yielding high-fidelity, strictly spatial-uniform benchmarks with complex elevation changes and rich geometric constraints.

### A.3 Prompt Construction and Filtering Pipeline

To enable robust text-conditioned control and multi-modal alignment within the generative world model, the processed video-trajectory sequences are paired with high-quality, semantically rich textual descriptions. Rather than relying on motion-centric captions that introduce redundancy with the explicit 6-DoF trajectory inputs, we construct an automated textual supervision pipeline focusing on environmental characteristics, spatial layouts, and scene appearance. This pipeline bridges visual scene content with textual conditions through the following structured operations:

**Step 1: Environment-Focused Scene Captioning.** To synthesize a comprehensive textual representation without introducing computational redundancy from dense frame processing, we uniformly extract a sparse sequence of 21 frames from each video clip. These frames are sequentially fed into a vision-language model (Qwen3-VL) with a tailored system prompt that commands the model to describe the background environment while strictly omitting any description of camera ego-motion or dynamic object movements. By observing the multi-frame sequence simultaneously, the model successfully distills stationary scene semantics, structural attributes, weathering conditions, and material textures into a single, cohesive environmental narrative.

**Step 2: Dual-Level Environmental Granularity.** To support different scales of textual conditioning and facilitate both fast semantic anchoring and intricate scene learning, the environment-focused descriptions are adaptively split into a dual-level framework based on word count. Descriptions with a length strictly below 35 words are categorized as *Concise Prompts*, which capture essential high-level scene semantics such as global terrain types (e.g., dense pine forests, rocky canyons) and macro-climatic profiles. Conversely, descriptions exceeding 35 words are preserved as *Detailed Prompts*, providing dense details regarding spatial geometry, architectural styles, fine-grained surface textures, and complex ambient light interactions (e.g., volumetric sun shafts, diffuse overcast shadowing). This multi-granularity design allows the downstream world model to optimize for both coarse-grained environment matching and fine-grained visual-text correspondences simultaneously.

**Step 3: Exposure Quality Assessment.** To remove acquisition artifacts while preserving natural illumination diversity, the visual frames undergo an automated exposure quality screening process based on global pixel intensity statistics:

- *Under-Exposure Detection:* To identify severe shadow noise and completely uninformative black regions, we compute the frame-level average pixel intensity $\mu_I$ across the RGB channels. Frames exhibiting a mean brightness value below a strict lower bound, $\mu_I < 30$ (dead dark), are flagged.
- *Low-Contrast Tracking:* To capture washed-out scenes with compressed dynamic range, we evaluate the image contrast $\sigma_I$ via pixel intensity standard deviation. Sequences displaying low contrast ($\sigma_I < 20$) are flagged, which typically indicates poorly lit environments when coupled with low mean intensity.
- *Over-Exposure Ratio Estimation:* To filter out clipped highlights where geometric textures are entirely lost, we calculate the over-exposure pixel ratio $R_{\text{over}}$. An individual pixel is

considered saturated if its intensity exceeds 250, and a video clip is marked as over-exposed if the saturated region occupies a massive portion of the frame, formulated as $R_{\text{over}} > 40\%$.

The identifiers of all video sequences exhibiting severe exposure artifacts are systematically logged into an exposure anomaly catalog.

**Step 4: Perceptual Quality Scoring.** Beyond traditional pixel-level brightness statistics, we further consider rare perceptual failures such as severe blur, compression artifacts, and visual distortions. We employ Q-Align, a learned blind image quality assessment (BIQA) model, solely for identifying extreme low-quality cases. Q-Align produces a native quality score on a 1–5 scale, which we normalize to $\mathcal{Q} \in [0, 1]$ only for dataset filtering.

**Step 5: Strategic Dual-Stage Filtering.** To balance processing efficiency with dataset purity across different training scales, our pipeline provides two alternative filtering schemes to generate the clean subset:

- *Exposure-Only Filtering (Fast Mode):* This strategy targets rapid data throughput by exclusively pruning sequences documented in the exposure anomaly catalog (`exposure_or_dark.txt`). It is optimized for scenarios requiring quick dataset scaling while removing severe exposure artifacts.
- *Dual-Attribute Filtering (Complete Mode):* This strategy jointly applies exposure and perceptual quality constraints. After removing exposure-compromised sequences, we further filter low-fidelity clips with a perceptual quality threshold $\mathcal{Q} < 0.1$, excluding 1.1% of the remaining corpus. The final subset contains video sequences with enhanced visual consistency after quality filtering. Notice that Q-Align is used only for removing extremely low-quality samples during dataset curation, rather than optimizing or ranking the dataset according to Q-Align scores.

**Step 6: Dataset Partitioning.** With the purified dataset established, our final objective is to partition the corpus into distinct training and testing benchmarks (`train.csv` and `test.csv`). We adopt a dynamic dataset splitting strategy where the testing set is assigned a dynamic ratio ranging from 10% to 20% of the clean dataset, with the remaining majority allocated to the training set. Crucially, the selection priority for the testing set is strictly bound to video clips that possess comprehensive, high-accuracy ground-truth pose annotations, safeguarding the evaluation phase against trajectory tracking errors and establishing a solid evaluation foundation.

**Step 7: Dataset Split Details.** We construct dataset splits at the scene level. All flights and derived clips from the same scene are assigned to the same partition, ensuring that no scene overlap exists among the training, validation, and test sets. Since all clips inherit the split assignment of their source flights, clips from the same flight or scene never appear across different splits. Therefore, the evaluation set is both scene-disjoint and flight-disjoint, preventing spatial and temporal leakage during evaluation.

# B Experiments Settings

## B.1 Video Model Training Pipeline

We adopt WAN2.2-5B as our foundational visual generator and integrate Low-Rank Adaptation (LoRA) with a rank configuration of 64 for parameter-efficient fine-tuning. The models are trained using the DiffSynth-Studio framework on the 70K real-world video-trajectory samples. Synthetic data are additionally incorporated only during the motion-control enhancement stage to improve camera motion controllability. We provide models at both 480p and 720p resolutions, where visual sequences are resized to their corresponding target resolutions. The 720p model is directly fine-tuned at the native high resolution rather than obtained through post-processing upscaling. Following the official implementation of WAN2.2-5B, the initial frame of the sequence is concatenated directly with the latent noise along the frame dimension to serve as a strong spatial conditioning anchor, which is subsequently excluded during loss computation to focus the optimization gradient purely on synthesized content. To enhance the model's geometric robustness across multi-modal inputs, we apply random online yaw-angle rotations to the video streams and randomly toggle between concise and detailed environmental prompts. The network is optimized for exactly 2 epochs, with

the latitude-aware reconstruction loss enabled to strictly enforce spherical geometric consistency. To accommodate hardware memory constraints while maintaining a stable optimization trajectory, the batch size is configured to 1, coupled with a gradient accumulation step of 4, and the complete distributed training pipeline is deployed on 8 NVIDIA H20 GPUs.

### B.2 Action Model Training Pipeline

To preserve the pretrained visual generation capability and avoid gradient interference during kinematic alignment, we completely freeze the pretrained video generation backbone and optimize only the newly introduced camera-conditioned attention modules. The target camera trajectory is first converted into camera poses, from which ray-conditioned transformations are constructed for the ERP representation. Each panoramic latent token is associated with an ERP ray and its corresponding transformation $\mathbf{T}^{\mathrm{ray}}$. These transformations are incorporated into PRoPE and directly applied to the $q/k/v/o$ features of the camera-conditioned self-attention pathway. The resulting camera-conditioned attention output is introduced as a parallel branch to the original self-attention pathway, thereby providing explicit conditioning on the target camera motion while keeping the pretrained visual backbone unchanged.

The training protocol is organized into three progressive stages, with each stage executed for exactly 2 epochs. In the first stage, we initialize the camera-conditioning modules with basic directional motion priors using the filtered planar-line dataset. To increase the spatial coverage of the training samples and reduce overfitting to specific flight headings, we apply structured data augmentation by jointly rotating the visual frames and their corresponding trajectory yaw angles by $90^\circ$, $-90^\circ$ and $180^\circ$. The learning rate is set to $1 \times 10^{-4}$.

In the second stage, we train on the full 70K real-world samples to adapt the camera-conditioning modules to diverse and unconstrained physical camera motions. The learning rate remains $1 \times 10^{-4}$. During this stage, synchronized random yaw perturbations are applied online to both the video sequences and their corresponding 6-DoF trajectories, ensuring that the visual observations and camera-motion conditions remain geometrically consistent after augmentation.

In the third stage, we further fine-tune the camera-conditioning modules on 50K trajectory-precise synthetic samples. This stage further refines the correspondence between the conditioned camera motion and the target 6-DoF trajectories, including motion patterns that are less frequently observed in the real-world data. The learning rate is reduced to $5 \times 10^{-5}$ to limit deviations from the motion priors established in the preceding stages. The same synchronized random-yaw augmentation strategy is retained.

For all three stages, the batch size is set to 1 with a gradient accumulation step of 4, resulting in an effective batch size of 4. Throughout the entire training process, gradients are propagated only through the newly introduced camera-conditioned attention modules, while the pretrained video generation backbone remains frozen.

### B.3 Memory Model Training Pipeline

To preserve the pretrained visual generation and action control capabilities, the video generation backbone and action model are frozen during memory training. Only the geometry-aware memory module is optimized. The training procedure consists of two progressive stages.

**Stage 1: Geometric Warm-up.** We first perform a geometric warm-up stage using 20k frame pairs sampled from the collected panoramic videos. Each pair contains a context frame and a target frame with a randomly masked region. Given their relative camera displacement, the memory module learns to reconstruct the masked visual content by retrieving geometric information from the context observation. The displacement between paired frames is constrained within $[0.2, 2.0]$ m to encourage learning geometric correspondence rather than local appearance interpolation.

**Stage 2: Memory-guided Video Training.** We then optimize GMA using full panoramic video sequences. For each 161-frame clip, an 81-frame generation window is randomly sampled. Memory frames are retrieved from preceding observations according to the spatial sampling strategy described in Eq. (22). If valid memory frames are unavailable, the memory branch is disabled for the current

sample. The selected memory features and their corresponding ray representations are precomputed offline and injected through PRoPE-based cross-attention with confidence-guided gating.

$$\mathcal{C}(s) = \{t \mid t < s,\ d_{\min} \leq \|\mathbf{p}_t - \mathbf{p}_s\|_2 \leq d_{\max}\}. \tag{22}$$

During training, the memory features and conditioning frame are perturbed with Gaussian noise to improve robustness. The memory module is optimized with a learning rate of $1 \times 10^{-4}$, batch size 1, and gradient accumulation step 4.

### B.4 Baseline Methods

Existing efforts on controllable panoramic video generation remain limited. We select publicly reproducible representatives across different paradigms. Matrix-3D and OmniRoam serve as the primary baselines for video quality and camera controllability, while Imagine360 is additionally included to evaluate panoramic generation quality. Methods relying on explicit per-scene 3D geometry are represented by Matrix-3D, which is further used to analyze the reconstruction trade-off in revisit consistency experiments. CamPVG adopts a different conditioning setting based on perspective-video inputs, and PanoWorld-X does not release inference weights; therefore, we cite them for related discussion but do not include them in quantitative comparisons. Matrix-3D releases inference code only and is evaluated following its original protocol. OmniRoam releases training and inference code but no training data and is primarily trained on indoor scenes; therefore, we fine-tune it on World360 for a fair comparison (OmniRoam$^{\dagger}$).

**OmniRoam** OmniRoam is a method for long-horizon panoramic video generation conditioned on camera trajectories, adopting a ReCamMaster-style pipeline that combines real and generated frames for consistent view synthesis. Since OmniRoam is primarily trained on indoor data, we adapt the released mo$f$del to World360 for a fair comparison. OmniRoam is built upon Wan2.1-1.3B, whereas our method is based on Wan2.2-5B. Due to their intrinsically different image-to-video architectures, directly replacing the OmniRoam backbone with Wan2.2-5B is not straightforward and would require substantial architectural modifications. To ensure a fair comparison, we train OmniRoam using the same number of training steps and a comparable scale of computational resources as our method. OmniRoam supports output resolutions of both 480p and 720p and additionally provides a real-time inference pipeline based on Self-Forcing. We evaluate all OmniRoam variants using the same set of visual-quality and camera-controllability metrics as our method whenever applicable.

**Matrix-3D** Matrix-3D is a method for 3D world generation that supports panoramic video generation from a single panoramic image conditioned on a camera trajectory. Since the complete training recipe and training pipeline are not publicly available, we do not introduce an additional World360-specific fine-tuning procedure, avoiding an implementation-dependent retraining setting that may differ from the original method. We use the officially released inference code to generate videos, which consists of three stages: mesh generation using MoGe, 3D-to-2D rendering, and final video generation. Matrix-3D supports both 480p and 720p output resolutions, using a Wan2.1 backbone for 480p generation and a Wan2.2-5B backbone for 720p generation. We evaluate the generated videos using the full set of metrics, including both video quality and camera controllability.

**Imagine360** Imagine360 is a method for panoramic video generation conditioned on a perspective video sequence. Since its training code is not publicly available and the model is trained on outdoor data, we use the officially released inference code and model for evaluation without additional fine-tuning. For each panoramic test video, we extract a perspective video clip with a $90^{\circ}$ field of view and use it as the input to Imagine360. The model generates 32 frames at a resolution of $512 \times 1024$. To align with our evaluation protocol, we temporally interpolate the generated videos to 81 frames and spatially downsample them to $480 \times 960$. Since Imagine360 does not take camera trajectories or action controls as input, we evaluate it only in terms of panoramic video quality.

### B.5 Evaluation Protocols

We split the dataset by scene, ensuring that all test sequences come from scenes unseen during training. During the inference and benchmarking phase, we sample 200 representative video-trajectory sequences from the World360 dataset to serve as a comprehensive evaluation baseline,

ensuring no overlap with the training data. This evaluation corpus covers a diverse taxonomy of 3D maneuvers, including forward, backward, lateral, vertical (up/down) translations, continuous arcs, and complex loop trajectories. To rigorously quantify model performance, we establish a multi-dimensional evaluation suite spanning distribution fidelity, omnidirectional visual quality, and translational trajectory controllability:

**Visual and Geometric Fidelity Metrics:** We employ the Fréchet Inception Distance (FID), together with panoramic regional variants ($\text{FID}_{\text{pole}}$ and $\text{FID}_{\text{equ}}$), to assess visual fidelity, structural integrity, and projection consistency across different regions of the equirectangular representation. To account for the inherent geometric distortion of equirectangular projections, we further evaluate perceptual quality and aesthetic consistency using the Natural Image Quality Evaluator (NIQE) and Q-Align-based quality metrics ($\text{QA}_{\text{qual.}}$), with latitude-aware pixel weighting to account for the non-uniform area distortion across latitudes. Furthermore, temporal consistency and motion artifacts are evaluated using Fréchet Auto-Encoder Distance (FAED)-based metrics, providing a comprehensive evaluation framework tailored to panoramic video synthesis.

For all frame-level and video-level metrics, we exclude the first frame, which is directly provided as the conditioning input, and compute the metrics using all subsequent generated frames. All metrics are computed independently at the native evaluation resolution of each setting; specifically, 480p and 720p results are evaluated separately without resizing the generated frames to a common resolution.

**Trajectory Conditioning and Visual Alignment:** To ensure a fair and deterministic comparison, all competing methods are conditioned on the same ground-truth (GT) camera trajectories. This unified conditioning avoids biases caused by differences in motion scales or trajectory estimation noise from conventional Structure-from-Motion (SfM) pipelines. We compute the average Peak Signal-to-Noise Ratio (PSNR) between the generated videos and the GT sequences over multiple progressive temporal windows to evaluate visual alignment and structural consistency under explicit camera motion.

# C More results

## C.1 Camera-Motion Conditioning Injection Ablation

To further analyze how camera-motion information should be injected into the video generation backbone, we compare six alternative camera-conditioning mechanisms: *PRoPE*, *Concat*, *AdaLN*, *Pose*, *ERP*, and *XAttn*. The comparison is implemented within a unified camera-conditioned architecture, such that the injection mechanism is the primary variable across the variants.

**Common Architecture.** All variants introduce a camera-conditioned branch in parallel with the original self-attention pathway of each DiT block. Given the input feature $\mathbf{x}$ and camera condition $\mathbf{c}$, the resulting residual is formulated as

$$\mathbf{h}_{\text{out}} = \mathbf{x} + \mathcal{A} * \text{self}(\mathbf{x}) + \mathcal{A} * \text{cam}(\mathbf{x}, \mathbf{c}), \tag{23}$$

where $\mathcal{A} * \text{self}$ denotes the original self-attention pathway and $\mathcal{A} * \text{cam}$ denotes the newly introduced camera-conditioned pathway. All variants use the same camera-conditioned self-attention module and the same parallel connection to the original DiT block. The output projection of the camera-conditioned branch is initialized to zero, such that the modified network is equivalent to the pretrained video backbone at initialization.

The camera condition is derived from the target camera trajectory. For each sampled camera pose and each ERP location, we construct a ray-conditioned transformation $\mathbf{T}_i^{\text{ray}} \in SE(3)$. The resulting transformation is associated with the corresponding panoramic latent token. Following the ERP motion representation described in Sec. 4.1.1, the current action condition models camera translation through ray-conditioned transformations, while camera rotation is decoupled from the learned motion condition through the ERP coordinate transformation. In addition, the spatial re-sampling procedure uses a fixed inter-frame distance of $0.05$,m and removes stationary segments; therefore, the learned condition does not explicitly model instantaneous velocity, acceleration, or hovering.

For all six variants, the same training data, optimization schedule, batch size, and inference configuration are used. The same 2D positional encoding for the ERP latent patches is retained across all

variants. For the non-PRoPE variants, the SE(3) component of the PRoPE transformation is replaced by the identity transformation, while the 2D patch positional encoding is retained. This prevents the comparison from being confounded by the presence or absence of an additional positional encoding mechanism.

**DPRC(PRoPE.)** Our proposed camera-conditioning mechanism directly incorporates the ray-conditioned transformation into the attention features. For the $i$-th panoramic token, the corresponding transformation $\mathbf{T}_i^{\mathrm{ray}}$ is applied to the $q/k/v/o$ features as

$$q_i \leftarrow (\mathbf{T}_i^{\mathrm{ray}})^\top q_i, \qquad k_i, v_i \leftarrow (\mathbf{T}_i^{\mathrm{ray}})^{-1}(k_i, v_i), \qquad o_i \leftarrow \mathbf{T}_i^{\mathrm{ray}} o_i. \tag{24}$$

Consequently, the attention interaction depends on the relative transformation between the corresponding rays. This provides ray-wise geometric conditioning that varies spatially across the full ERP representation.

**Concat.** For the *Concat* baseline, the ray-conditioned transformation is represented explicitly as a feature vector. We flatten the rotational and translational components of $\mathbf{T}_i^{\mathrm{ray}}$ into a 12-dimensional representation and encode it using a lightweight ray encoder:

$$\mathbf{c}_i = \mathrm{LN}\left(\mathrm{MLP}\left(\mathbf{f}_i^{\mathrm{ray}}\right)\right), \qquad \mathbf{f}_i^{\mathrm{ray}} \in \mathbb{R}^{12}, \tag{25}$$

where the MLP has dimensions $12 \to 256 \to 256$. The encoded camera feature is concatenated with the visual feature before the $q/k/v$ projections:

$$q_i, k_i, v_i = \mathrm{Proj}\left([\mathbf{x}_i; \mathbf{c}_i]\right). \tag{26}$$

Thus, unlike PRoPE, the camera geometry is introduced as an additive feature representation rather than through a multiplicative transformation of the attention features.

**AdaLN.** The *AdaLN* baseline uses the same ray encoder as *Concat* but injects the resulting camera feature through feature-wise modulation. Specifically, a modulation head maps the 256-dimensional camera feature to scaling and shifting parameters:

$$(\boldsymbol{\gamma}_i, \boldsymbol{\beta}_i) = \mathrm{ModHead}(\mathbf{c}_i), \tag{27}$$

where the modulation head maps $256 \to 2d$ and $d$ denotes the hidden dimension of the DiT block. The visual feature is then modulated as

$$\mathbf{x}_i' = \mathbf{x}_i \odot (1 + \boldsymbol{\gamma}_i) + \boldsymbol{\beta}_i. \tag{28}$$

The modulation head is zero-initialized so that the camera-conditioning branch does not perturb the pretrained network at initialization.

**Pose.** The *Pose* baseline investigates whether frame-level camera poses are sufficient without explicitly modeling the ray-wise geometry of the ERP representation. Instead of assigning an individual ray transformation to each spatial token, we encode the camera pose of each frame using a pose encoder with the same architecture as the ray encoder. The resulting frame-level feature is then broadcast to all spatial tokens belonging to that frame and concatenated with the corresponding visual features before the $q/k/v$ projections.

Importantly, the pose encoder is implemented as a separate module from the ray encoder. This prevents checkpoints trained with different injection variants from being loaded interchangeably and ensures that the comparison isolates the difference between ray-wise and frame-wise camera conditioning.

**ERP.** The *ERP* baseline replaces the ray-local SE(3) representation with explicit ERP coordinates together with the frame-level camera pose. For each ERP token, its spherical location is represented by

$$\mathbf{e}_i = [\sin\theta_i, , \cos\theta_i, , \sin\phi_i, , \cos\phi_i], \tag{29}$$

where $\theta_i$ and $\phi_i$ denote the longitude and latitude of the corresponding ERP location. This 4-dimensional spherical coordinate is concatenated with the 12-dimensional frame-level camera pose, resulting in a 16-dimensional input to the ERP encoder:

$$\mathbf{c}_i = \mathrm{LN}\left(\mathrm{MLP}\left([\mathbf{e}_i; \mathbf{p}_t]\right)\right), \qquad [\mathbf{e}_i; \mathbf{p}_t] \in \mathbb{R}^{16}. \tag{30}$$

The resulting 256-dimensional feature is concatenated with the visual feature before the $q/k/v$ projections. This baseline therefore evaluates whether explicit ERP coordinates and frame-level poses can provide camera-motion conditioning without constructing ray-local SE(3) transformations.

Table 7: ViPE-based camera trajectory evaluation on generated panoramic videos. ATE and $\text{RPE}_{trans}$ are reported in meters after recovering trajectories with ViPE.

| Method | ATE Mean ↓ (m) | ATE Median ↓ (m) | $\text{RPE}_{trans}$ ↓ (m) |
|---|---|---|---|
| Matrix-3D | 0.1153 | 0.0985 | 0.0516 |
| OmniRoam† | 0.1612 | 0.1058 | 0.0619 |
| **Ours** | 0.0995 | 0.0836 | 0.0518 |

**XAttn.** The *XAttn* baseline introduces the camera condition through cross-attention. The visual features are used as queries, while encoded camera features provide keys and values:

$$\mathbf{h}_{\text{cam}} = \text{Attn}\left(Q(\mathbf{x}), K(\mathbf{c}), V(\mathbf{c})\right). \tag{31}$$

The camera condition is therefore injected through an explicit cross-attention operation rather than modifying the visual feature before self-attention or transforming the attention features through PRoPE.

### C.2 Additional Qualitative Results

We provide additional qualitative results on real-world outdoor sequences from the test set. As shown in Fig. 8, Fig. 9, and Fig. 10, PanoWorld consistently produces photorealistic panoramic video clips with cleaner geometric structures and faithful camera motion. Notably, our method accurately follows complex camera trajectories beyond planar movements, successfully handling challenging 3D motions involving continuous elevation and altitude changes.

### C.3 Additional Quantitative Analysis of Trajectory Control

To further evaluate trajectory-following fidelity, we use ViPE as an external pose estimator to recover camera trajectories from generated panoramic videos and compare them with the ground-truth (GT) trajectories. As shown in Fig. 11, PanoWorld achieves closer global trajectory alignment with the GT under diverse camera motions, while baseline methods exhibit larger accumulated deviations. Quantitative results in Table 7 show that PanoWorld achieves the lowest global trajectory deviation measured by ATE, while maintaining comparable local motion consistency measured by RPE. The comparable RPE values indicate that PanoWorld preserves local motion consistency while achieving lower global trajectory deviation under long-range camera movements.

### C.4 Additional Revisit Consistency Analysis

To evaluate revisit consistency of panoramic world generation, we further conduct experiments under loop camera motions. All methods are evaluated using identical revisit trajectories and camera motion patterns. Since the first frame is provided as the conditioning frame, revisit PSNR measures the preservation of the initial scene appearance after camera movements under the given generation conditions.

For OmniRoam, we follow its original two-stage generation pipeline. Specifically, OmniRoam first predicts an 81-frame preview sequence under the target trajectory, which is subsequently refined into a temporally denser output using its default refinement process. Temporal refinement only increases frame density without modifying the underlying camera trajectory. Therefore, all methods are compared using the same spatial travel distance rather than the number of generated frames.

As shown in Table 8, PanoWorld achieves higher revisit PSNR across all motion patterns. The comparison between Ours and Ours w/o GMA indicates that GMA provides additional benefits for revisit consistency during extended camera movements by leveraging previously observed content. Note that reconstruction-based methods such as Matrix-3D benefit from explicit geometric representations during revisits, while this evaluation mainly focuses on consistency achieved by generative panoramic world models under identical trajectory conditions.

### C.5 Ablation Study on Key Components

We further analyze the contribution of individual components in Table 9. Removing dense DPRC degrades trajectory controllability across all temporal ranges, highlighting the importance of dense

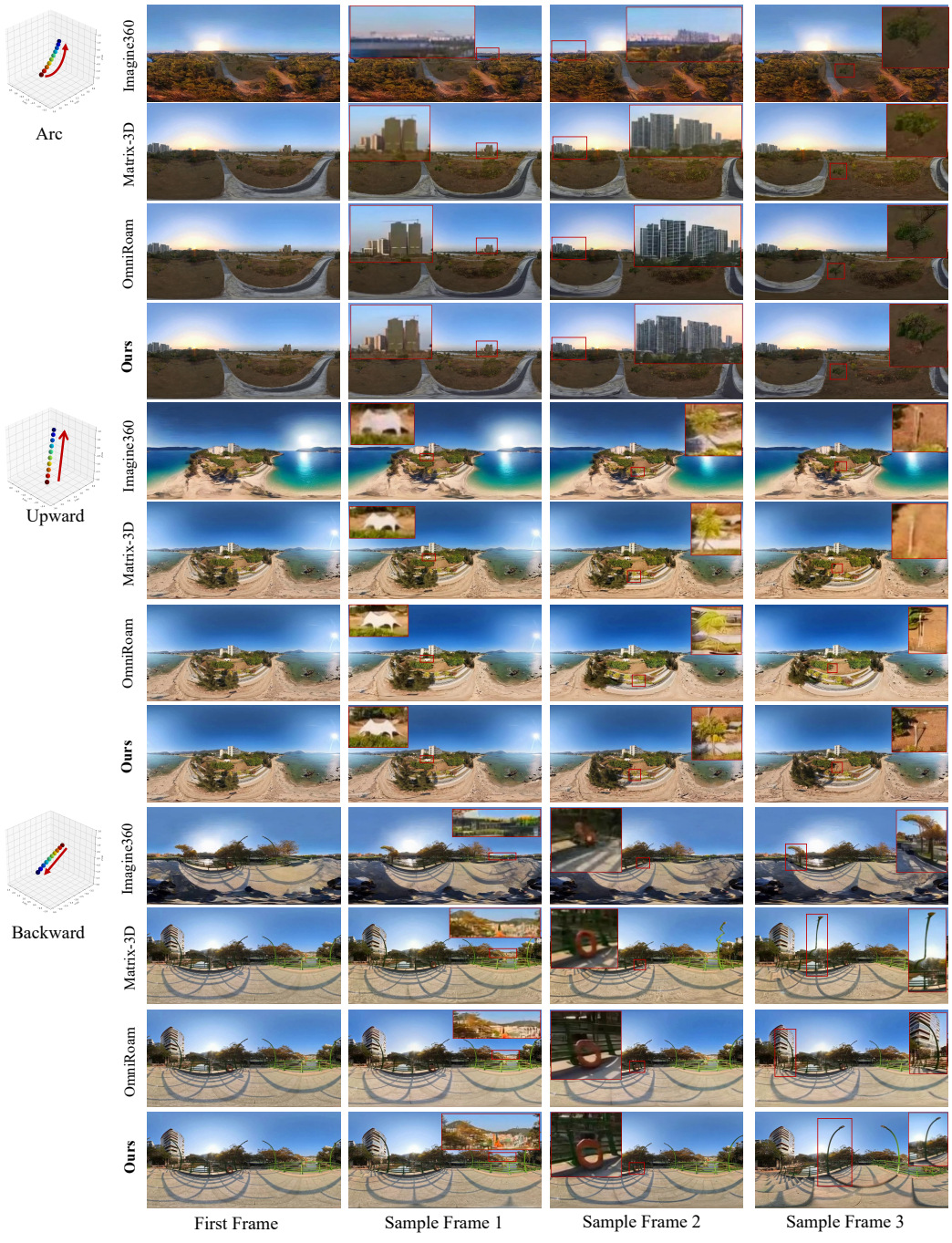


Figure 8: Additional qualitative results on real-world outdoor sequences: compound 3D maneuvers in structured urban and coastal scenes.

ray-level motion conditioning. Removing the latitude-aware loss mainly affects omnidirectional visual fidelity, demonstrating its role in preserving panoramic structural quality. Disabling the three-stage training strategy leads to degraded trajectory controllability and reduced overall performance, indicating the importance of progressive optimization for jointly learning motion and memory components. These results validate the effectiveness of the proposed design choices and training strategy.

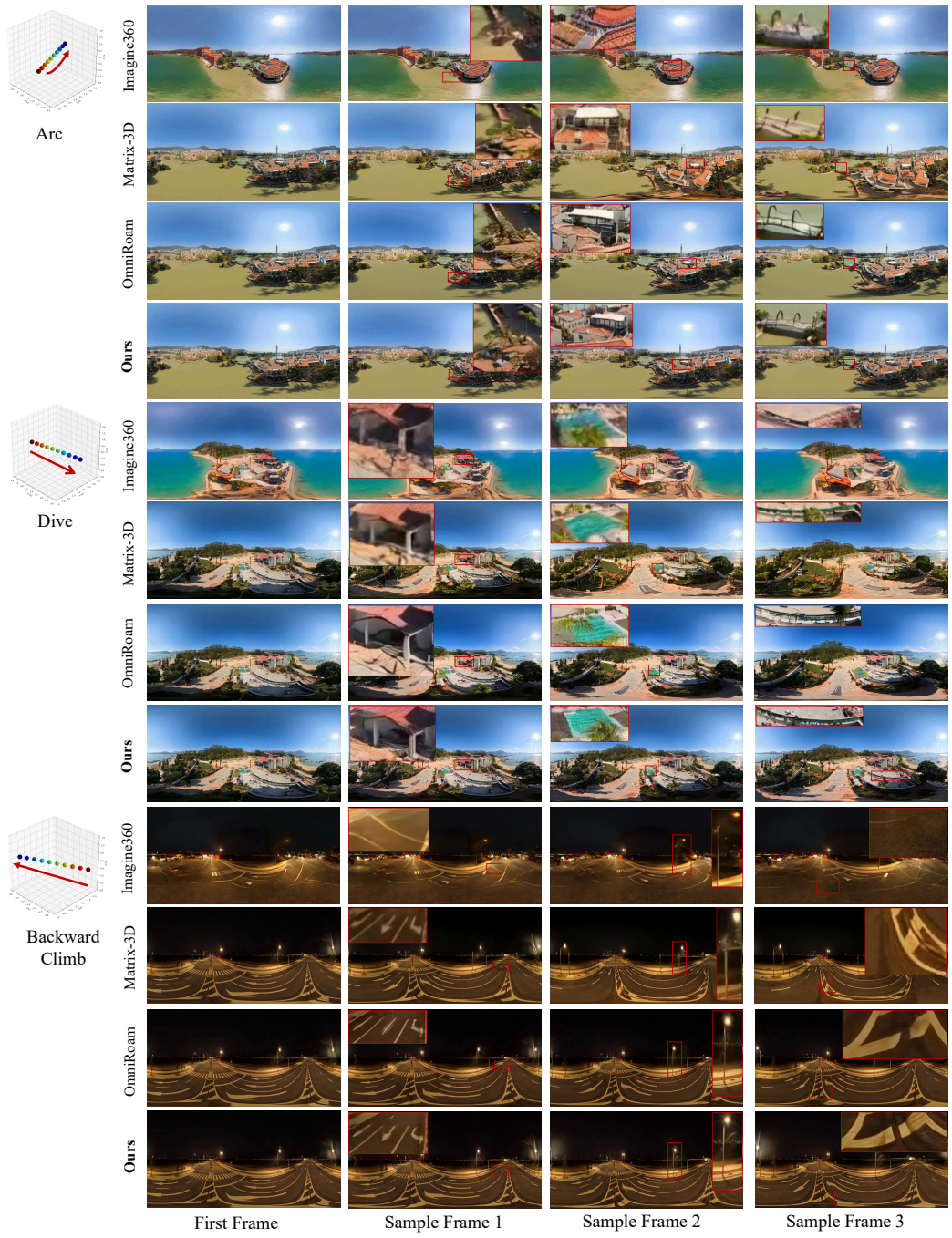


Figure 9: Additional qualitative results on real-world outdoor sequences: steep elevation changes and low-light conditions.

### C.6 Yaw-Translation Conditioning Ablation

We further investigate the design choice of camera-motion representation under the yaw-only setting. Specifically, the panoramic videos are first horizon-rectified, where pitch and roll variations are removed, and yaw remains as the primary rotational component. We compare three motion-conditioning strategies: (1) joint yaw-translation modeling, where rotational and translational signals are jointly encoded; (2) translation-only conditioning without rotational rectification; and (3) our decoupled yaw-translation representation, which explicitly separates heading changes from translational motion.

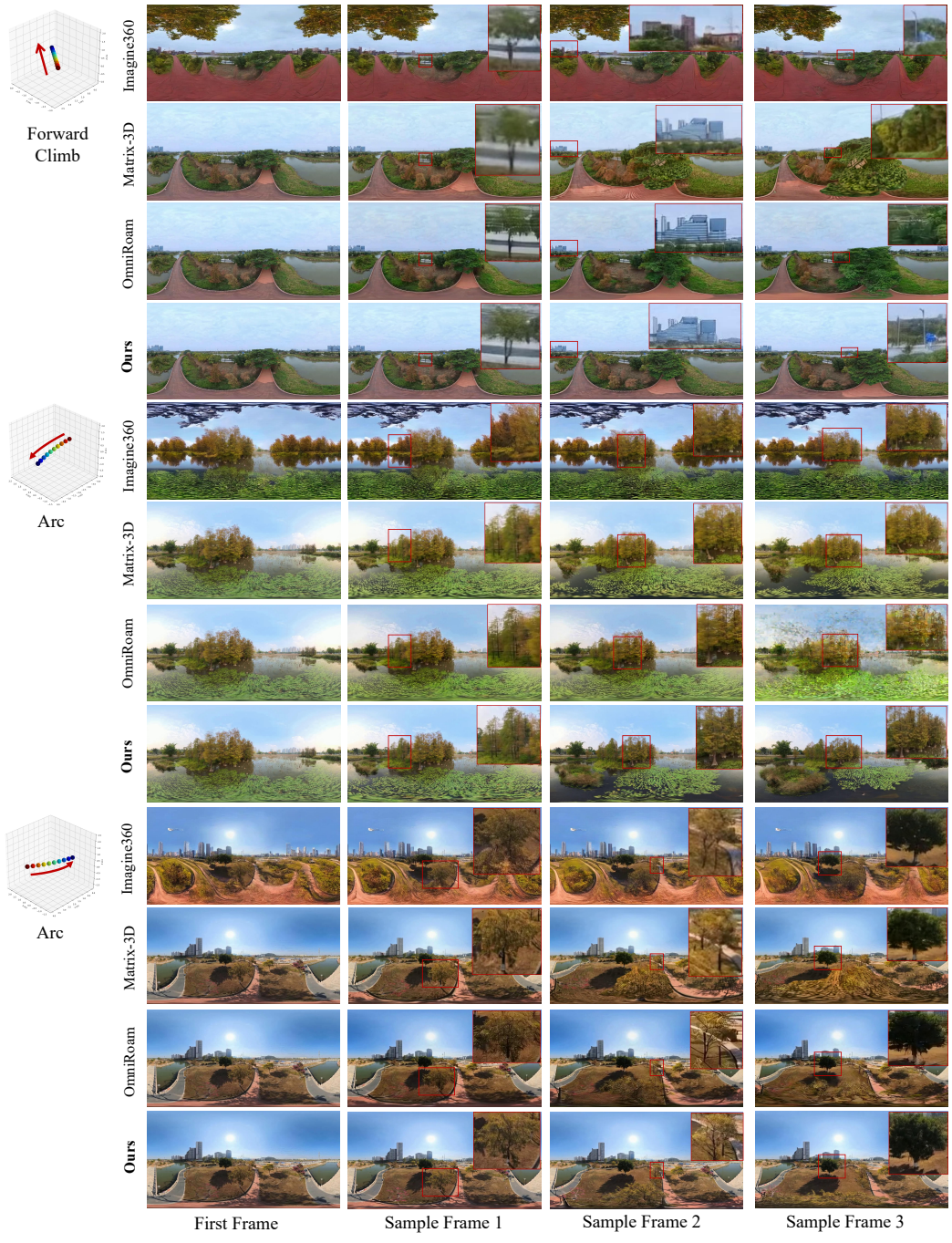


Figure 10: Additional qualitative results on real-world outdoor sequences: reflective and texture-repetitive surfaces.

As shown in Table 10, both joint modeling and translation-only conditioning degrade visual quality and trajectory controllability. In contrast, the proposed decoupled representation achieves consistently better performance across all metrics, indicating that separating yaw rotation from translation reduces the learning difficulty of controllable panoramic video generation.

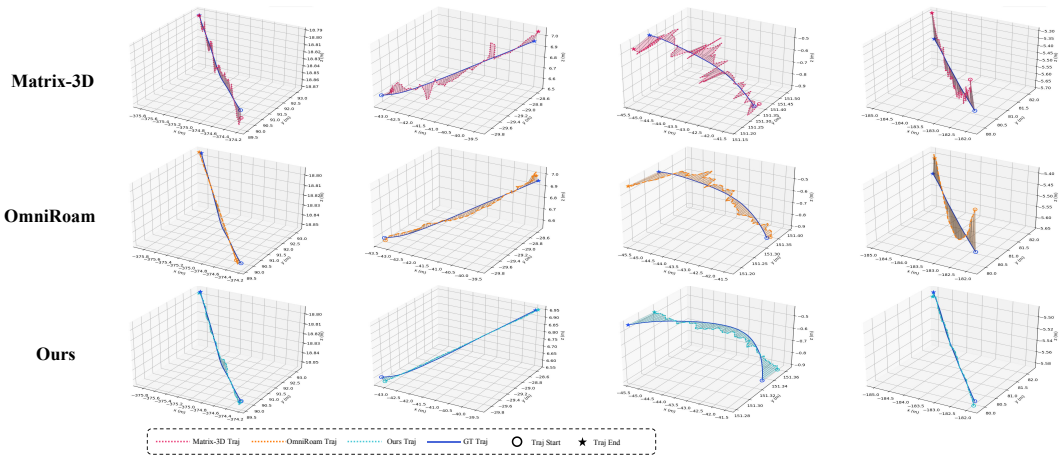


Figure 11: Qualitative trajectory comparison using ViPE. ViPE is used to recover camera trajectories from generated panoramic videos, which are compared with the corresponding GT trajectories.

Table 8: Revisit consistency comparison under identical loop trajectories. All methods are evaluated using trajectories with approximately 8.0 m accumulated camera travel distance, and each motion pattern contains 40 sequences. PSNR is measured between the initial conditioning frame and the final frame after returning to the starting location.

| Method | Straight (PSNR↑) | Lateral (PSNR↑) | Circle (PSNR↑) | Average (PSNR↑) |
|---|---|---|---|---|
| Matrix-3D | 16.48 | 16.81 | 16.55 | 16.61 |
| OmniRoam† | 15.85 | 15.79 | 15.70 | 15.78 |
| Ours w/o GMA | 16.49 | 16.74 | 16.62 | 16.62 |
| **Ours** | **18.41** | **18.84** | **18.55** | **18.60** |

### C.7 Backbone Scale Comparison

To further investigate the impact of backbone scale, we train an additional 1.3B version of PanoWorld using the Wan2.1-1.3B backbone, matching the backbone size of OmniRoam. As shown in Table 11, PanoWorld* consistently outperforms both OmniRoam variants across visual quality and trajectory controllability metrics under the same 1.3B backbone. Increasing the backbone scale from 1.3B to 5B further improves both video quality and camera controllability, indicating that our proposed motion and memory modeling remains effective across different backbone scales.

## D Application

### D.1 Real-Time Application via Causal-Forcing

**Qualitative Results.** We present additional qualitative results for real-time video generation utilizing our developed panoramic I2V world model. In contrast to previous methods that suffer from severe structural blurring or motion inconsistencies due to architectural gaps, our framework—built upon Causal Forcing and advanced to the Wan2.2-TI2V-5B backbone—demonstrates superior capabilities. Given a single input panoramic image and a sequence of camera trajectories, our model successfully generates temporally coherent, high-fidelity panoramic scene videos. Thanks to the chunk-wise teacher forcing and causal ODE distillation, the generated previews exhibit exceptionally stable scene geometry, precise perspective transitions, and strict adherence to the specified camera motion over extended temporal horizons.

**Technical Implementation** To build a high-fidelity, controllable interactive environment simulator, we adapt the Wan2.2-TI2V-5B diffusion model, transforming its original dual-stream architecture into a causal autoregressive framework via Causal Forcing.

Table 9: Ablation study on key components of PanoWorld

| Method | Visual Quality | | | | | Traj. Controllability | | |
|---|---|---|---|---|---|---|---|---|
| | FID↓ | $FID_p$↓ | $FID_e$↓ | FAED↓ | NIQE↓ | $P_{F20-F25}$↑ | $P_{F50-F55}$↑ | $P_{F75-F80}$↑ |
| **PanoWorld** | 27.64 | 47.21 | 26.00 | 2.63 | 3.85 | 22.83 | 21.65 | 20.92 |
| w/o dense DPRC | 29.51 | 48.39 | 27.23 | 2.56 | 3.80 | 18.99 | 17.65 | 16.90 |
| w/o Latitude-aware loss | 30.59 | 50.65 | 28.95 | 2.61 | 3.91 | 21.68 | 20.44 | 19.65 |
| w/o three-stage training | 35.08 | 58.70 | 30.57 | 3.00 | 4.01 | 16.78 | 14.34 | 13.52 |

Table 10: Ablation study on yaw-translation motion conditioning.

| Method | Visual Quality | | | | | Traj. Controllability | | |
|---|---|---|---|---|---|---|---|---|
| | FID↓ | $FID_p$↓ | $FID_e$↓ | FAED↓ | NIQE↓ | $P_{F20-F25}$↑ | $P_{F50-F55}$↑ | $P_{F75-F80}$↑ |
| Joint Yaw+T Modeling | 36.25 | 63.32 | 32.80 | 3.05 | 4.10 | 19.34 | 18.07 | 17.23 |
| T-only w/o Rectification | 37.02 | 56.88 | 32.55 | 4.21 | 4.29 | 17.46 | 16.24 | 15.72 |
| **Decoupled Yaw+T (Ours)** | 27.64 | 47.21 | 26.00 | 2.63 | 3.85 | 22.83 | 21.65 | 20.92 |

We implement a chunk-wise autoregressive strategy that partitions the 21 latent frames into a $[1] + [4 \times 5]$ structure. The initial frame serves as the clean I2V conditioning anchor and is excluded from loss computation. The remaining 20 frames are trained with teacher forcing in 4-frame chunks. To preserve the dual-stream nature of the backbone while enforcing causality, the full GT latent sequence is concatenated with the noisy sequence to form the input ($clean_x$), and an I2V-specific teacher forcing attention mask is constructed via FlexAttention. This mask guarantees that: (i) the noisy tokens of the conditioning frame only attend to their own clean counterparts, and (ii) the noisy tokens within a specific generation chunk attend solely to the clean GT tokens from preceding chunks, preventing future information leakage within the same chunk. Camera poses are duplicated along the token dimension to precisely align with this clean/noisy dual-stream sequence structure.

We follow the standard three-stage pipeline of Causal Forcing to compress the autoregressive Wan2.2 model into an efficient real-time generator: (1) autoregressive diffusion training with teacher forcing, (2) causal ODE initialization to bridge the architectural gap, and (3) asymmetric distribution matching distillation (DMD). At 480p resolution, Stage-1 is trained for 3,000 steps, and Stage-3 DMD is optimized for 10,500 steps, yielding an ultra-fast 4-step generator with discretized timesteps at $\{1000, 750, 500, 250\}$.

**Long-video distillation & Rolling Forcing Inference.** To transcend the 21-latent-frame (81 pixel frames) training limit, we extend chunkwise DMD to 161-pixel videos (41 latent frames) via a Rolling Forcing mechanism. During training, rollouts autoregressively generate up to 41 causal latent frames using a sliding-window KV cache, while DMD supervision is applied to the last 21 frames for compatibility with frozen bidirectional score networks. We distill the model on 1,998 pre-encoded panoramic clips. At inference, Rolling Forcing generates all 41 latent frames in a single pass conditioned on the first frame, text, and camera trajectories (synthetic or dataset poses), yielding 161-frame videos with a 4-step distilled scheduler. To eliminate contamination of the clean cache within the moving denoising windows, a prefix snapshot-and-restore mechanism for the KV caches is utilized. Furthermore, for memory efficiency during long-sequence generation, chunked VAE decoding with a 1-latent-frame (corresponding to 4 pixel frames) overlap blending mechanism is implemented.

**Quantitative Results.** We report the quantitative evaluation on panoramic video generation and motion controllability in Table 12. As demonstrated, our full model (**Ours**) significantly outperforms existing state-of-the-art methods, achieving the best scores across almost all visual quality (e.g., FID, FAED), perceptual quality (QA), and trajectory controllability (PSNR) metrics.

When introducing the causal-forcing variant, it exhibits only a marginal and acceptable drop in both visual fidelity and trajectory control compared to our full model. However, this negligible quality gap is accompanied by a dramatic, orders-of-magnitude improvement in inference efficiency. As illustrated in Figure 13, on a single NVIDIA H20 GPU, our full model requires 4 minutes and 48 seconds to complete the generation, whereas competitive methods like Matrix-3D and OmniRoam suffer from prohibitive computational costs, taking up to nearly 16.5 and 31 minutes, respectively. In

Table 11: Comparison of different backbone scales.

| Method | Backbone | Visual Quality | | | | | Traj. Controllability | | |
|---|---|---|---|---|---|---|---|---|---|
| | | FID↓ | $FID_p$↓ | $FID_e$↓ | FAED↓ | NIQE↓ | $P_{F20-F25}$↑ | $P_{F50-F55}$↑ | $P_{F75-F80}$↑ |
| OmniRoam | W2.1-1.3B | 62.50 | 87.10 | 46.20 | 3.48 | 3.52 | 18.22 | 17.31 | 16.80 |
| OmniRoam† | W2.1-1.3B | 60.77 | 85.25 | 45.92 | 3.36 | 3.39 | 18.51 | 17.59 | 17.02 |
| PanoWorld* | W2.1-1.3B | 31.00 | 51.20 | 28.74 | 2.83 | 3.86 | 21.69 | 20.35 | 19.98 |
| PanoWorld | W2.2-5B | 27.64 | 47.21 | 26.00 | 2.63 | 3.85 | 22.83 | 21.65 | 20.92 |

† denotes the model fine-tuned on the **World360** dataset.

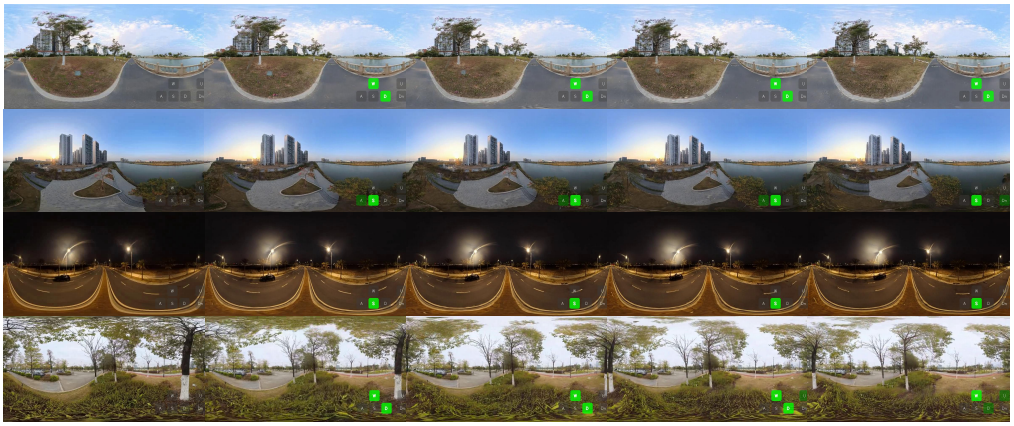

Figure 12: Real-time generation. The initial input frame is shown in the left column, followed by consecutive frames generated under specified keyboard controls.

stark contrast, our causal-forcing model successfully slashes the generation time to a mere 8 seconds. This speedup and the efficiency trade-off shown in Fig. 13 and Fig. 12 demonstrate the practicality of causal-forcing for real-time generation.

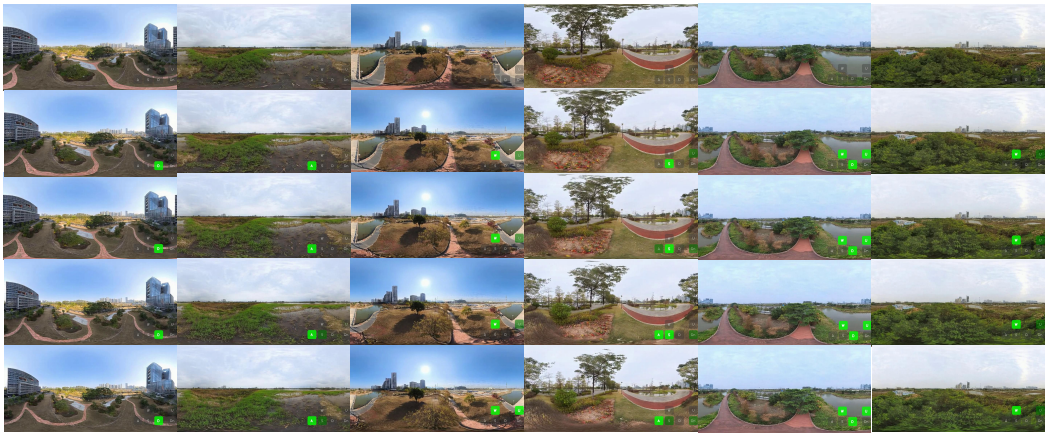

Figure 13: Additional qualitative results of real-time generation.

# E Discussion

## E.1 Limitation

Despite its effectiveness in controllable panoramic video generation, our current framework has several limitations. First, the initial frame is directly preserved as the conditioning frame to provide

Table 12: Quantitative comparison of panoramic video generation quality, motion controllability, and real-time generation efficiency.

| Method | Gen. Time | Visual Quality | | | | Perceptual Quality | | Traj. Controllability | |
|---|---|---|---|---|---|---|---|---|---|
| | | FID↓ | $FID_p$↓ | $FID_e$↓ | FAED↓ | NIQE↓ | $QA_q$↑ | $P_{F20-F25}$↑ | $P_{F50-F55}$↑ |
| Matrix-3D | 16m36s | 34.63 | 67.88 | 68.64 | 3.39 | 4.74 | 3.00 | 20.47 | 18.80 |
| OmniRoam† | 30m46s | 60.77 | 85.25 | 45.92 | 3.36 | 3.39 | 3.75 | 18.51 | 17.59 |
| + Self-Forcing | 16s | 67.32 | 90.35 | 47.98 | 3.79 | 3.73 | 3.57 | 17.32 | 16.62 |
| **Ours** | 4m48s | 27.64 | 47.21 | 26.00 | 2.63 | 3.85 | 4.02 | 22.83 | 21.65 |
| + Causal-Forcing | 8s | 28.07 | 56.66 | 27.27 | 2.81 | 4.06 | 3.96 | 21.93 | 20.32 |

a strong spatial anchor and avoid unnecessary reconstruction of the observed scene. However, this design introduces an inherent asymmetry between the observed conditioning frame and the subsequently generated frames, as the model is required to maintain consistency while extending the scene beyond the initial observation.

Second, our current memory mechanism mainly focuses on preserving previously observed scene information during camera movements. Handling more extensive environmental changes, such as large-scale scene transitions or previously unseen regions with limited visual references, remains challenging and requires further exploration of more adaptive world memory mechanisms.

### E.2 Future Work

To address the aforementioned limitations, we envision several promising directions for future exploration. First, to bridge the domain gap caused by directly injecting the input image, we plan to investigate lightweight boundary-refinement modules or latents-space alignment techniques. Instead of hard replication, optimizing the transition from the real input to the generated latent sequence could significantly preserve visual fidelity in subsequent frames. Second, to mitigate the heavy dependency on the initial frame and improve long-term temporal stability, we aim to develop a dynamic memory management mechanism. By continuously updating the memory pool with newly generated, high-confidence frames and introducing active forgetting strategies, the model can adaptively refresh its scene representation. This would reduce error accumulation and better support navigation through highly dynamic, large-scale environments over extended horizons.