# PwP: Panoramas without Panoramas Camera-Controlled 360° Video from Any View via Perspective-Only Adaptation

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

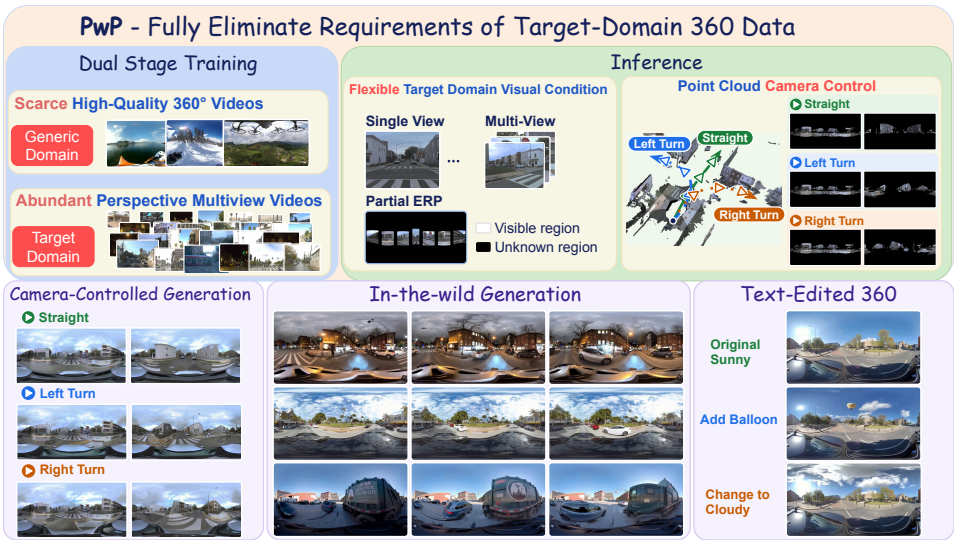


Figure 1: PwP demonstrates high-quality 360° video generation without requiring any 360° videos on the target domain during training and inference.

## Abstract

Adapting panoramic video generators to new domains is challenging because high-quality 360° videos are scarce, whereas calibrated perspective multiview videos are abundant. We introduce **PwP**, a two-stage framework for camera-controlled 360° video generation that adapts to a target domain without complete target-domain panoramas. Stage I learns a panoramic prior from curated 360° videos and trains the model to synthesize complete ERP videos from partial initial observations. Stage II adapts this prior with target-domain perspective multiview videos: we project camera observations into partial ERP videos with validity masks marking reliable RGB regions, and apply rectified-flow supervision only to valid latent regions, while rendered point-cloud control guides generation along a prescribed camera trajectory. At inference, PwP accepts any initial observation, from a single perspective view to multiple views or a complete ERP, and generates a complete 360° video. On held-out target-domain driving sequences and cross-dataset nuScenes evaluation, Stage-II adaptation improves generation quality and outperforms perspective-conditioned baselines even from a single view, while the Stage-I panoramic prior is well preserved on held-out 360-1M videos. These results show that abundant perspective observations can effectively supervise panoramic video generation without complete panoramic recordings in the target domain. Video results are available on our anonymous project website.

## 1 Introduction

To enable simulation for spatial intelligence in autonomous agents, 360° panoramic video generation has gained higher research interests (Luo et al., 2025; Wang et al., 2024a; Tan et al., 2025; Ji et al., 2025; Chen et al., 2026; Wu et al., 2026; Zhang et al., 2025; Xia et al., 2025), where the



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.

equirectangular projection (ERP) representation directly encodes full $360^\circ$ field of view (FoV) information in one view (Guo et al., 2025; Lee et al., 2024; Shin et al., 2025) to achieve omnidirectional world modeling with applications to training and evaluation of agents. The representation naturally ensures the generation consistency, compared with widely used multiview generation (Wu et al., 2025; Jiang et al., 2026a; Gao et al., 2025; Zhao et al., 2025b; Cao et al., 2025), which needs to additionally ensure cross-view generation consistency. Further, when changing the camera setup-calibration, intrinsics, or camera layout-perspective multiview videos need to re-generate, but using $360^\circ$ panorama enables us to directly crop out views from it without re-generation. This is practically useful for autonomous driving since it requires complex camera systems for safety and becomes expensive to change the sensor settings.

However, mainstream simulation frameworks for autonomous driving still generate perspective multiview videos rather than complete ERP videos, such as Nvidia Cosmos-2.5 (Ali et al., 2025), OmniDreams (Basant et al., 2026), and other driving video generators (Yang et al., 2026; Wang et al., 2024d; Zhou et al., 2026; Li et al., 2024; Zhu et al., 2025; Russell et al., 2025). Applying $360^\circ$ ERP video generation to this domain is challenging because suitable target-domain panoramic data are scarce. Although large-scale panoramic datasets such as 360-1M (Wallingford et al., 2024) contain abundant web videos, after quality filtering and restricting the data to driving scenes, only approximately 150 suitable panoramic videos remain. In contrast, autonomous-driving datasets are commonly captured using multiple limited-FoV cameras, such as Argoverse (Wilson et al., 2023), Waymo Open Dataset (Sun et al., 2020), and Waymo E2E (Xu et al., 2026). Under our temporal clip construction, these datasets provide nearly 100,000 perspective videos. Such perspective observations are abundant but are generally not used as target-domain supervision for full-ERP panoramic video generation in prior works (Luo et al., 2025; Wang et al., 2024a). This large data gap motivates learning a panoramic prior from general $360^\circ$ videos and adapting it to the driving domain using perspective multiview observations.

Alternatively, reprojecting multiple views into the canonical ERP space still leave unobserved regions, while differences in camera viewpoints can introduce projection artifacts. Though few generate panoramic videos not on the full ERP space (Wen et al., 2024; Teng et al., 2026), they sacrifice contents outside a specific pitch angle where traffic signs, large objects like trucks, or obstacles right under the ego car are cropped. Further, their incomplete ERP does not support rendering arbitrary camera views as it may fall outside the visible areas.

We introduce Panoramas without Panoramas (**PwP**; Fig. 1) to **eliminate the requirements of full $360^\circ$ ERP data on the target domain during both training and inference**. For training, we first train a camera-controlled $360^\circ$ video diffusion model (360-VDM) using general-domain panoramic videos. The 360-VDM takes an initial frame, text prompt, and geometric camera control signal as the condition. The geometric camera control signal is 3D point cloud built from panorama by a geometric foundation model (Wang et al., 2026), and we render the point cloud along the camera trajectory. We propose random ERP masking to mask out the initial frame on the full ERP while supervising with the ERP of full coverage, which gears the 360-VDM with ability to generate full ERP from incomplete initial ERP. To this end, we provide a data curation pipeline to filter out noisy $360^\circ$ raw data and process view-aligned geometric camera control signals. We also provide a foundation 360-VDM with arbitrary-shape initial frames, ready to post-train on target domains.

Then to post-train on a target domain without any full ERP data, PwP leverages available multiview observations to construct partially visible ERP as the training data. The proposed training formulation uses the partially visible areas as the only evidence to construct the initial frame and groundtruth supervision. In this way, 360-VDM retains its ability to imagine outside the visible areas and meanwhile becomes adaptable to the target domain. Compared with prior work, we are the first to adapt the full $360^\circ$ ERP video generation to the target domain without full ERP as the groundtruth.

Our main contributions are summarized as follows:

- We introduce *PwP*, a framework for camera-controlled $360^\circ$ video generation. The method is a perspective-only adaptation, which is the first work that removes all the requirements of full $360^\circ$ ERP data on the target domain during both the training and inference from practical aspects, along with ability to take flexible visual input as condition.
- We build a delicate data curation pipeline to filter out noisy $360^\circ$ raw videos and process view-aligned geometric camera control signals. We deliver a foundation $360^\circ$ video generation model that is ready to post-train on target domains.

- We provide a useful post-train recipe and leverage larger perspective multiview datasets for training on the target domain to train 360-VDM and provide a high-performing model. Extensive evaluations validate our design.

## 2 Related Work

**Panoramic Video Generation.** Recent video diffusion models (Blattmann et al., 2023; Yang et al., 2025b; Kong et al., 2024; Wan Team et al., 2025; Agarwal et al., 2025; Ali et al., 2025) have substantially advanced perspective video synthesis, motivating extensions to panoramic generation. Work on $360^\circ$ image generation explores multi-view diffusion, coupled perspective–panoramic denoising, cubemap representations, and hybrid perspective/panoramic training (Tang et al., 2023; Zhang et al., 2024; Kalischek et al., 2025; Feng et al., 2026). Recent panoramic video methods similarly adapt pretrained video generators to the panoramic domain (Wang et al., 2024a; Xia et al., 2025; Fang et al., 2025), while another line generates panoramas from narrow-FoV observations using perspective-conditioned architectures or multi-view attention (Tan et al., 2025; Xie et al., 2025; Li et al., 2026b; Luo et al., 2025; Wu et al., 2026; Jiang et al., 2026b).

Existing methods, however, generally either require panoramic video data for training or adaptation, or assume a predefined conditioning format such as text (Wang et al., 2024a; Xia et al., 2025; Zhang et al., 2025), a single perspective image/video (Tan et al., 2025; Luo et al., 2025; Xie et al., 2025; Li et al., 2026b; Wu et al., 2026; Jiang et al., 2026b), or a complete panorama (Li et al., 2025a; Lu et al., 2025; Zhang et al., 2026; Wang et al., 2025a; Chen et al., 2026). This does not match driving and robotic systems, where multiple limited-FoV cameras provide arbitrary and often incomplete spherical coverage. In contrast, **PwP** adapts a panorama-capable generator without requiring complete panoramic ground truth from the target domain and supports an arbitrary observed subset of the sphere, ranging from a single perspective view to multiple views or a complete panorama.

**Panoramic Training Data and Perspective Supervision.** Training data remains a major bottleneck for panoramic generative models. Large-scale web collections such as 360-1M (Wallingford et al., 2024) contain abundant panoramic videos, but require substantial curation to remove mislabeled perspective or $180^\circ$ content, stereoscopic and fisheye recordings, static or low-quality videos, and geometric artifacts (Luo et al., 2025; Zhang et al., 2026; Chen et al., 2026; Xia et al., 2025; Feng et al., 2026). They also lack camera annotations, requiring pose estimation when trajectories are needed (Zhang et al., 2026; Wu et al., 2026). Consequently, many methods rely on relatively small curated datasets (Wang et al., 2024a; Tan et al., 2025; Li et al., 2026b) or synthetic environments that provide controlled content and accurate geometry (Lu et al., 2025; Yin et al., 2026; Yang et al., 2025a; Wang et al., 2025a; Liu et al., 2026; Li et al., 2026a). However, obtaining complete panoramic videos that also match a specific target domain remains costly.

In contrast, domains such as autonomous driving provide large-scale calibrated multi-camera video datasets (Sun et al., 2020; Xu et al., 2026; Wilson et al., 2023; Caesar et al., 2020), which have supported extensive work on driving video generation (Wang et al., 2024c; Gao et al., 2024a; Wen et al., 2024; Gao et al., 2024b; Wang et al., 2024d; Li et al., 2024; Zhou et al., 2026; Ali et al., 2025; Basant et al., 2026). These cameras, however, cover only part of the viewing sphere, so their projection into a panoramic representation yields incomplete observations rather than panoramic ground truth. Prior work has explored combining perspective and panoramic supervision (Feng et al., 2026), generating panoramas from stitched multi-camera images (Teng et al., 2026), or learning from partial or corrupted observations (Ma et al., 2024; Yu et al., 2025a; Daras et al., 2023). **PwP** instead adapts an already panorama-capable video generator directly from perspective-derived, partially observed target-domain videos, without requiring complete panoramic ground truth.

**Camera- and Geometry-Controlled Video Generation.** Controllable video generation typically augments pretrained diffusion models with camera or geometric conditions (Zhang et al., 2023; NVIDIA et al., 2025). Camera control can be provided directly through pose sequences or ray-based embeddings (Wang et al., 2024e; He et al., 2025; Van Hoorick et al., 2024; Bai et al., 2025), while geometry-aware methods render reconstructed scene structure along target trajectories to guide synthesis (Ren et al., 2025; Yu et al., 2025b;a; Huang et al., 2025). Related approaches improve long-horizon consistency with explicit memory (Li et al., 2025b; Xiao et al., 2025), and feed-forward geometry models further make such geometric guidance practical by recovering scene structure and camera poses from sparse or unposed observations (Wang et al., 2024b; 2025b; 2026). Similar

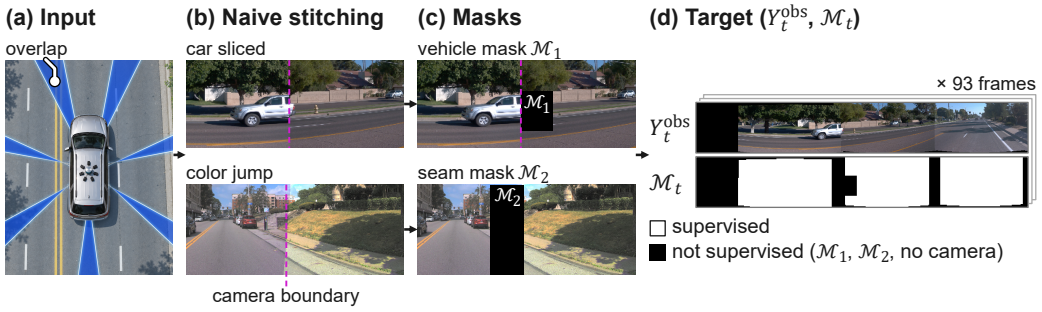


Figure 2: **Construction of partial panoramic targets.** (a) Driving agents widely adopt camera rings with adjacent views overlap. (b) On one ERP canvas, adjacent cameras may disagree at boundaries: a vehicle is sliced (Waymo Perception) and colors jump (Argoverse 2). (c) A vehicle mask $\mathcal{M}_1$ removes regions where the cut vehicle should exist; a seam mask $\mathcal{M}_2$ removes the boundary mismatch. (d) The target of the upper example: ERP frame $\mathbf{Y}_t^{\text{obs}}$ and validity mask $\mathbf{M}_t$, 93 frames per clip. Black is excluded from the loss, including directions no camera sees; no pixel is repaired.

geometry-conditioned generation has been explored for driving scenes (Yan et al., 2025; Wang et al., 2025c; Zhao et al., 2025a; Yang et al., 2026), but primarily in the perspective domain and therefore for specific camera viewpoints.

Panoramic methods extend these ideas through spherical pose or ray conditioning (Ji et al., 2025; Ye et al., 2024; Yin et al., 2026; Li et al., 2026a; Liu et al., 2026), action-conditioned exploration (Lu et al., 2025), or optical-flow-based motion control (Zhang et al., 2026), with several works explicitly rendering reconstructed geometry to guide trajectory-controlled generation (Yang et al., 2025a; Wang et al., 2025a; Chen et al., 2026). These methods generally assume a complete input panorama and panoramic video supervision. **PwP** instead constructs geometric guidance from any available subset of views, ranging from a single perspective image to multiple views or a complete panorama, and uses a panorama-capable prior to synthesize unobserved regions, while adapting to the target domain without complete panoramic ground truth.

## 3 Training Data Construction and Scaling

We use two complementary sources of training data. Stage I uses curated general-domain panoramic videos from 360-1M (Wallingford et al., 2024) to establish the panoramic generation prior. Stage II adapts this prior to the driving domain using calibrated perspective multiview videos from Argoverse 2 (Wilson et al., 2023), Waymo Open Dataset Perception (Sun et al., 2020), and Waymo Open Dataset for End-to-End Driving (Xu et al., 2026). We additionally process nuScenes (Caesar et al., 2020) using the same construction pipeline but reserve it for evaluation.

**Panoramic video curation.** We construct the Stage-I panoramic dataset from 360-1M (Wallingford et al., 2024). We apply automatic quality filtering to remove corrupted or discontinuous videos, insufficient motion, incorrect panoramic formats, repeated frames, and persistent static regions or overlays. The resulting curated general-domain panoramic videos are used for Stage I. Detailed filtering criteria are provided in Appendix A.1.

**Reliability masks.** After projecting the calibrated multi-camera observations onto a common ERP canvas, adjacent camera views may disagree near their boundaries because of viewpoint, timing, or appearance differences. These disagreements can produce color seams and truncated nearby objects (Fig. 2b). Details of the ERP projection are provided in Appendix A.2.

**Seam mask.** We construct a seam mask $\mathcal{M}_2$ that removes a band around each camera boundary (Fig. 2c, bottom). This prevents inconsistent colors or duplicated content in overlapping views from being used as direct RGB supervision.

**Vehicle mask.** Seam masking alone is insufficient for nearby vehicles, which may extend beyond the masked boundary because adjacent cameras observe them from different optical centers. We use YOLOv8 (Jocher et al., 2023) instance segmentation to identify vehicle regions and con-

struct a frame-dependent vehicle mask $\mathcal{M}_{1,t}$ over regions where a truncated vehicle should continue (Fig. 2c, top). This prevents background content from neighboring cameras from being treated as valid supervision for the missing vehicle region.

**Partial ERP observations and validity masks.** After applying the seam and vehicle masks, each frame provides a partial ERP observation $\mathbf{Y}_t^{\text{obs}}$ together with a binary validity mask $\mathbf{M}_t$. The mask identifies pixels supported by reliable RGB observations:

$$\mathbf{M}_t(u,v) = \begin{cases} 1, & \text{if } (u,v) \text{ is observed by a camera and } (u,v) \notin \mathcal{M}_{1,t} \cup \mathcal{M}_2, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

For a $T$-frame clip, we denote the resulting sequences as

$$\mathbf{Y}^{\text{obs}} = \{\mathbf{Y}_t^{\text{obs}}\}_{t=0}^{T-1}, \qquad \mathbf{M} = \{\mathbf{M}_t\}_{t=0}^{T-1}. \tag{2}$$

Because the projection is constructed from the calibration of each camera rig, the same procedure applies to all four datasets despite their different camera layouts (Fig. 6).

# 4 METHOD

## 4.1 OVERVIEW

PwP follows a two-stage training framework. Stage I trains a camera-controlled $360^\circ$ video diffusion model $G_{\theta_0}$ using curated general-domain panoramic videos. Stage II starts from $G_{\theta_0}$ and adapts the model to the driving domain using the perspective multiview training data constructed in Sec. 3, without requiring complete ERP videos from the driving domain. After adaptation, the model supports flexible initial visual observations together with a prescribed camera trajectory and a text prompt for complete ERP video generation.

## 4.2 STAGE I: PANORAMIC PRIOR LEARNING

Stage I establishes a panoramic generation prior $G_{\theta_0}$ using the curated general-domain $360^\circ$ videos from 360-1M (Wallingford et al., 2024) described in Sec. 3. Given a panoramic training video $\mathbf{Y} = \{\mathbf{Y}_t\}_{t=0}^{T-1}$, we train a diffusion transformer (DiT)-based model $G_{\theta_0}$ to generate a complete ERP video conditioned on an initial visual observation, a text prompt, and a geometric camera control sequence. At this stage, the complete ERP video $\mathbf{Y}$ is available as the training target. The resulting $G_{\theta_0}$ is used to initialize Stage II.

**Geometric camera control.** We use Pi3 (Wang et al., 2026) to reconstruct scene geometry and estimate camera poses from perspective views sampled from the panoramic video. The reconstructed point cloud is rendered along the camera trajectory to produce the ERP geometric control sequence. Details are provided in Appendix B.1.

**Masked conditioning.** To expose the model to incomplete visual observations, we sample a spatial mask from a predefined mask pool for each training sample. The sampled mask is fixed throughout the entire sequence, while different samples may use different spatial patterns. Let $\mathbf{M}^{\text{prior}}$ denote the sampled mask. The initial visual condition is

$$\mathbf{I}_0^{\text{prior}} = \mathbf{M}^{\text{prior}} \odot \mathbf{Y}_0. \tag{3}$$

The model is conditioned on this partial initial observation, the geometric control sequence $\mathbf{C}$, and the text prompt, while the complete panoramic video $\mathbf{Y}$ remains the training target. Thus, masking changes the conditioning but does not remove any region from Stage-I RGB supervision. Additional mask configurations are described in Appendix B.2.

**Stage I training objective.** Because complete panoramic supervision is available in Stage I, we apply the rectified-flow objective over the full video latent. The complete ERP video $\mathbf{Y}$ is encoded by the pretrained video VAE $\mathcal{E}$ as $\mathbf{x} = \mathcal{E}(\mathbf{Y})$. Following the rectified-flow formulation (Liu et al., 2023), we sample Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and an interpolation coefficient $s \in [0,1]$, and construct $\mathbf{x}_s = s\boldsymbol{\epsilon} + (1-s)\mathbf{x}$ and $\mathbf{v}^* = \boldsymbol{\epsilon} - \mathbf{x}$. Given Stage-I conditioning $\mathbf{c}_{\text{I}} = \left(\mathbf{I}_0^{\text{prior}}, \mathbf{C}, \mathbf{p}\right)$, the

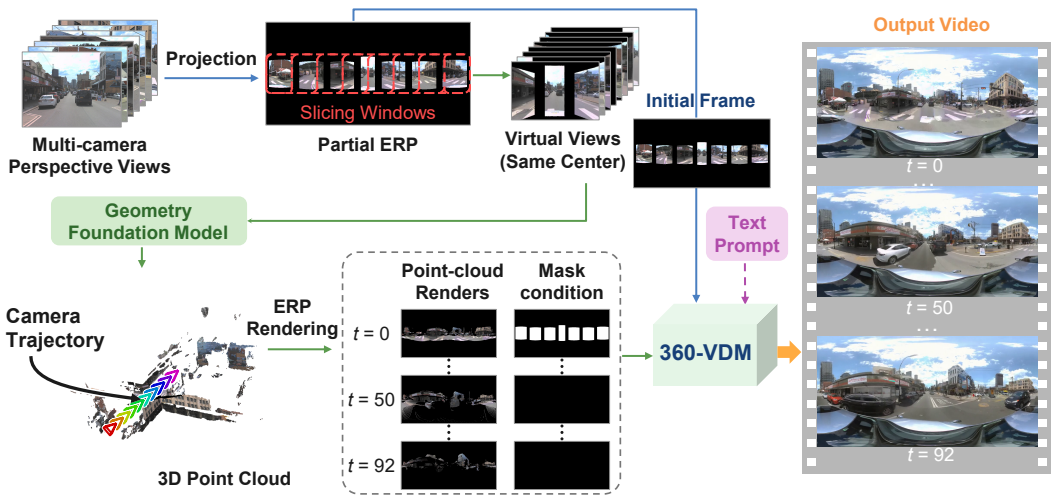


Figure 3: **Overview of Stage II**. PwP learns camera-controlled 360° video generation from any available views, using only partial ERP observations from perspective cameras.

model predicts $\mathbf{v}_\theta(\mathbf{x}_s, s; \mathbf{c}_\mathrm{I})$ and is trained with

$$\mathcal{L}_\mathrm{I} = \mathbb{E}\left[ w(s) \frac{1}{C_z|\Omega|} \sum_{c,\tau,i,j} \left( v_{\theta,c,\tau,i,j} - v^*_{c,\tau,i,j} \right)^2 \right], \tag{4}$$

where $\Omega$ denotes the full spatiotemporal latent grid, $C_z$ is the number of latent channels, and $w(s)$ is the rectified-flow time weighting inherited from the underlying video generator. Since Stage I uses complete ERP targets, all latent locations contribute to the loss.

### 4.3 Stage II: Perspective-Only Target-Domain Adaptation

Starting from the panoramic prior $G_{\theta_0}$ learned in Stage I, Stage II adapts the model to the driving domain using the perspective multiview data constructed in Sec. 3. Since complete ERP videos are unavailable as target-domain supervision, each training clip provides a partial ERP observation $\mathbf{Y}^\mathrm{obs}$ and its validity mask $\mathbf{M}$ (Fig. 3). Stage II retains the model architecture and geometric control from Stage I, while restricting RGB supervision to reliably observed regions.

**Observed-region supervision.** As defined in Sec. 3, regions where $\mathbf{M} = 0$ do not contain reliable RGB ground truth. We therefore construct

$$\widetilde{\mathbf{Y}} = \mathbf{M} \odot \mathbf{Y}^\mathrm{obs}, \qquad \mathbf{I}_0^\mathrm{obs} = \mathbf{M}_0 \odot \mathbf{Y}_0^\mathrm{obs}, \tag{5}$$

where $\odot$ denotes element-wise multiplication and $\mathbf{M}$ is broadcast across RGB channels. We encode the masked video with the pretrained video VAE as $\mathbf{x} = \mathcal{E}(\widetilde{\mathbf{Y}})$, while retaining $\mathbf{M}$ separately to determine valid latent supervision.

The geometric control sequence $\mathbf{C}$ remains available over the full video and is not masked by $\mathbf{M}$.

**Latent validity mask.** The RGB validity mask $\mathbf{M}$ is defined in pixel space, while the rectified-flow objective operates in the compressed video latent. Because the video VAE aggregates information spatially and temporally, directly resizing $\mathbf{M}$ may incorrectly supervise latent locations near invalid regions. We therefore expand the invalid regions before mapping the mask to latent resolution:

$$\mathbf{G} = 1 - \mathcal{A}_\mathrm{max}\left( \mathcal{D}_{r_s,r_t}(1 - \mathbf{M}) \right), \tag{6}$$

where $\mathcal{D}_{r_s,r_t}$ expands invalid regions with spatial and temporal margins $r_s$ and $r_t$, and $\mathcal{A}_\mathrm{max}$ maps the result to the spatiotemporal latent resolution. The resulting mask $\mathbf{G}$ identifies latent locations used for direct RGB supervision.

**Rectified-flow objective with valid supervision.** Stage II follows the same rectified-flow formulation as Stage I (Liu et al., 2023). Using $\mathbf{x}_s$ and $\mathbf{v}^*$ defined in Stage I, the conditioning is

$\mathbf{c}_{\mathrm{II}} = (\mathbf{I}_0^{\mathrm{obs}}, \mathbf{C}, \mathbf{p})$. Unlike Stage I, the velocity-prediction loss is restricted to latent locations marked valid by $\mathbf{G}$:

$$\mathcal{L}_{\mathrm{II}} = \mathbb{E}\left[ w(s) \frac{\sum_{c,\tau,i,j} G_{\tau,i,j} \left( v_{\theta,c,\tau,i,j} - v^*_{c,\tau,i,j} \right)^2}{C_z \sum_{\tau,i,j} G_{\tau,i,j}} \right]. \tag{7}$$

Here, $C_z$ and $w(s)$ follow the definitions in Stage I.

### 4.4 Flexible Inference

PwP supports flexible visual conditioning at inference time. The initial observation may be a single perspective view, multiple perspective views, a partial ERP, or a complete ERP, and is mapped to the common ERP representation to form $\mathbf{I}_0^{\mathrm{obs}}$. The camera trajectory is specified independently and converted into the geometric control sequence $\mathbf{C}$ using the Stage-I procedure. The adapted model then generates the complete ERP video as $\hat{\mathbf{V}}_{360} = G_{\theta^*}(\mathbf{I}_0^{\mathrm{obs}}, \mathbf{C}, \mathbf{p})$, where $\mathbf{p}$ is the text prompt. Thus, the same model supports different amounts of initial visual coverage and different prescribed camera trajectories.




Figure 4: Qualitative comparison with perspective-conditioned video-to-video baselines. The baselines operate from single-view inputs, while PwP uses the available multi-view observations together with geometry-derived point-cloud control.

## 5 Experiments

**Evaluation protocols.** We evaluate PwP under three settings, each using 50 93-frame sequences. First, a held-out target-domain test set (**Test**) with sequences from Waymo E2E (Xu et al., 2026), Waymo Perception (Sun et al., 2020), and Argoverse 2 (Wilson et al., 2023) evaluates adaptation to the driving domain. Second, nuScenes (Caesar et al., 2020) is excluded from Stage-II training and evaluates cross-dataset generalization. Third, held-out panoramic driving sequences from 360-1M (Wallingford et al., 2024) provide complete ERP references for full-panorama evaluation and measure whether Stage II preserves the Stage-I panoramic prior.

**Implementation.** PwP is based on Cosmos-Transfer-2.5-2B (Ali et al., 2025) architecture, and we train stage I and II models. All reported PwP outputs are generated in the ERP representation at a resolution of $1024 \times 512$ and contain 93 frames as one chunk generation. For evaluation, we use the camera poses associated with each held-out sequence as the prescribed trajectory so that the generated frames are temporally aligned with the corresponding reference views.

**Metrics.** We report PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), FID (Heusel et al., 2017), FVD (Unterthiner et al., 2019), and Q-Align (Wu et al., 2024). For the held-out target-domain and cross-dataset evaluations, complete panoramic ground truth is unavailable. We

Table 1: Quantitative comparison with external baselines.

| Evaluation | Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ | FVD ↓ | FID ↓ | Q-Align ↑ |
|---|---|---|---|---|---|---|---|
| Test | Argus (Luo et al., 2025) | 11.61 | 0.268 | 0.573 | 585.30 | 73.57 | 0.273 |
| | ViewPoint (Fang et al., 2025) | 9.40 | 0.175 | 0.701 | 1209.48 | 124.08 | 0.191 |
| | Imagine360 (Tan et al., 2025) | 10.41 | 0.206 | 0.609 | 1002.28 | 92.23 | 0.308 |
| | PwP (multiview) | **19.67** | **0.549** | **0.294** | **256.02** | **58.96** | **0.387** |
| nuScenes | Argus (Luo et al., 2025) | 11.51 | 0.288 | 0.575 | 985.82 | 104.68 | 0.252 |
| | ViewPoint (Fang et al., 2025) | 9.85 | 0.222 | 0.692 | 1530.29 | 147.28 | 0.180 |
| | Imagine360 (Tan et al., 2025) | 11.58 | 0.217 | 0.587 | 1062.69 | 92.66 | 0.296 |
| | PwP (multiview) | **18.49** | **0.494** | **0.315** | **303.83** | **55.84** | **0.385** |
| 360-1M | GenEx (Lu et al., 2025) | 14.82 | 0.554 | 0.438 | 2042.18 | 74.99 | 0.306 |
| | EvoWorld (Wang et al., 2025a) | 14.16 | 0.545 | 0.477 | 1955.93 | 90.83 | 0.288 |
| | OmniRoam (Liu et al., 2026) | 14.21 | 0.534 | 0.441 | 1952.53 | 62.51 | 0.261 |
| | PwP (Fullview) | **17.22** | **0.590** | **0.268** | **521.60** | **32.54** | **0.435** |

therefore project each generated ERP frame into the corresponding ground-truth camera views using the dataset calibration. PSNR, SSIM, and LPIPS are computed against temporally aligned real images over valid overlapping regions, while FID and FVD are computed in the perspective-view representation. For 360-1M, these reference-based metrics are computed directly on the complete ERP videos. Details of the Q-Align protocol are provided in Appendix C.1.

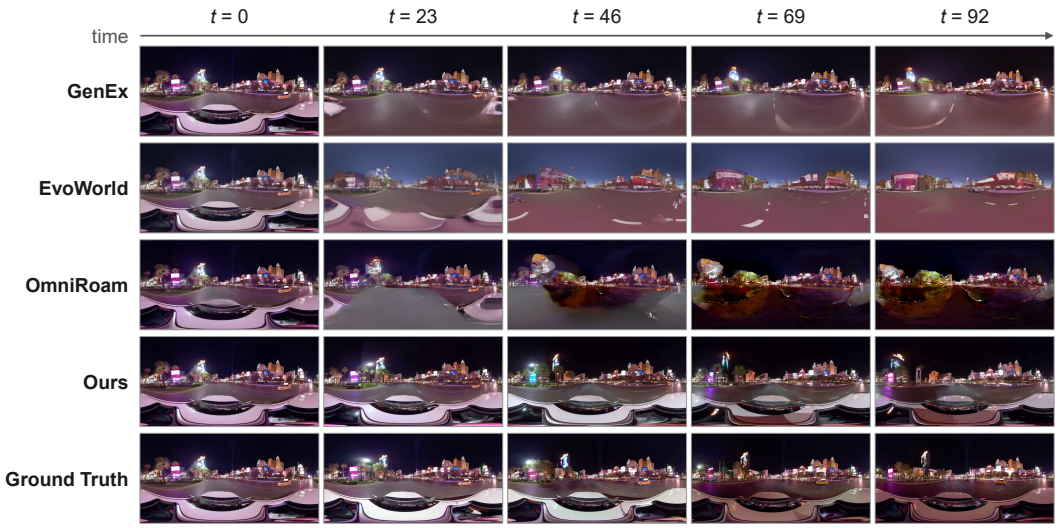


Figure 5: Qualitative comparison with full-panorama-conditioned image-to-video baselines across multiple time steps.

### 5.1 Main Results

**Comparison with external baselines.** Table 1 summarizes the quantitative comparison with all external baselines. For the held-out target-domain and cross-dataset nuScenes evaluations, we compare PwP with perspective-conditioned video-to-video methods. We evaluate the same Stage-II model under both single-view and multiview RGB conditioning. Under the single-view setting, PwP uses only the center perspective view at the first frame as its RGB observation, matching the input format of the perspective-conditioned baselines. PwP outperforms Argus, ViewPoint, and Imagine360 across the reported metrics on both evaluation settings, while the standard multiview setting further benefits from additional visual coverage.

For the full-panorama evaluation on 360-1M, we compare PwP with methods conditioned on a complete panoramic initial observation. PwP achieves the strongest reported results among GenEx, EvoWorld, and OmniRoam across the evaluated metrics.

Table 2: Ablation across the three evaluation settings.

| Evaluation | Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ | FVD ↓ | FID ↓ | Q-Align ↑ |
|---|---|---|---|---|---|---|---|
| Test | Cosmos (Ali et al., 2025) | 13.42 | 0.327 | 0.454 | 729.27 | 116.07 | 0.372 |
| | Stage I | 16.54 | 0.450 | 0.349 | 272.51 | 60.45 | 0.386 |
| | Direct Stage II | 15.80 | 0.429 | 0.366 | 345.37 | 70.35 | 0.359 |
| | PwP (Stage II) | **19.67** | **0.549** | **0.294** | **256.02** | **58.96** | **0.387** |
| nuScenes | Cosmos (Ali et al., 2025) | 14.40 | 0.400 | 0.487 | 609.74 | 94.07 | 0.348 |
| | Stage I | 16.67 | 0.406 | 0.392 | 370.15 | 61.10 | 0.381 |
| | Direct Stage II | 15.71 | 0.397 | 0.412 | 343.58 | 65.20 | 0.366 |
| | PwP (Stage II) | **18.49** | **0.494** | **0.315** | **303.83** | **55.84** | **0.385** |
| 360-1M | Cosmos (Ali et al., 2025) | 14.31 | 0.497 | 0.443 | 1090.28 | 90.96 | 0.381 |
| | Stage I | **17.31** | 0.587 | 0.271 | **497.00** | **31.41** | **0.436** |
| | Direct Stage II | 17.10 | **0.603** | 0.289 | 926.41 | 44.13 | 0.408 |
| | PwP (Stage II) | 17.22 | 0.590 | **0.268** | 521.60 | 32.54 | 0.435 |

Figure 4 provides qualitative comparisons under perspective-conditioned generation. PwP preserves observed appearance, object structure, and scene layout more consistently as the viewpoint changes. The additional multiview observations provide visual coverage beyond the center view, while the rendered point-cloud control supplies geometric guidance along the prescribed camera trajectory.

Figure 5 shows the corresponding full-panorama-conditioned comparison. PwP maintains scene appearance and spatial structure under camera motion while preserving visible temporal changes in dynamic scene elements.

### 5.2 Ablation Studies

**Effect of panoramic initialization and adaptation.** Table 2 compares the pretrained Cosmos model, Stage I, Direct Stage II, and the full PwP Stage-II model across all three evaluation settings. Direct Stage II uses the same target-domain training data and objective as PwP but initializes directly from the pretrained Cosmos checkpoint, without Stage I.

PwP achieves the strongest overall performance on both the held-out target-domain sequences and the unseen nuScenes dataset. Direct Stage II improves over the pretrained Cosmos model on several metrics, showing that partial target-domain supervision itself is effective, while initialization from the Stage-I panoramic prior provides a stronger starting point for adaptation. On 360-1M, PwP remains close to the Stage-I model, indicating that target-domain adaptation largely preserves the panoramic generation prior. A qualitative comparison of the pretrained Cosmos model, Stage I, Direct Stage 2, and the full PwP model on a representative nuScenes sequence is in Appendix Fig. 7.

Additional video results are available on our anonymous project website.

## 6 Conclusion

We presented PwP, a camera-controlled 360° video generator that works from any view, from a single perspective image to multiple views or a complete panorama, without complete target-domain panoramas for training or inference. PwP scales panoramic generation with abundant perspective data by turning multi-camera rigs into partial panoramic targets and supervising only reliably observed regions. We further contribute a curation pipeline for noisy web panoramas, a reusable 360° foundation model, and a driving-domain expert adapted from it. Experiments show consistent gains over the Stage-I model and perspective-conditioned baselines on held-out and unseen driving datasets, while the 360° prior is well preserved on 360-1M, Stage I's training distribution.

**Limitations.** Multi-camera rigs mainly cover the horizontal band around the vehicle, so regions near the poles, such as the ego vehicle's hood and roof, receive little supervision in Stage II. Unless these regions are visible in the initial observation, they may be generated with simplified textures.

### AI USE STATEMENT

In this work, we used generative AI tools to assist with developing supporting scripts for data processing and experimentation. We did not use generative AI tools to develop the research methodology, design experiments, formulate hypotheses, develop theoretical models or conceptual frameworks, or interpret experimental results. Additionally, we used generative AI tools for information retrieval and summarization, manuscript drafting and language polishing, and suggestions on LaTeX formatting and figure layout. All AI-assisted work was reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

This work uses existing research datasets and does not involve new collection of human-subject data. PwP is intended for research on panoramic video generation and simulation. As with other generative video models, the generated content may contain artifacts or inherit biases from the training data and pretrained models, and should not be treated as a substitute for real-world observations or safety validation. In particular, outputs from the model should not be used directly for safety-critical decision making without appropriate validation and human oversight. We encourage responsible use of the model and compliance with the terms and policies of the underlying datasets.

## REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical AI, 2025. URL https://arxiv.org/abs/2501.03575.

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. ReCamMaster: Camera-controlled generative rendering from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14834–14844, 2025.

Aarti Basant, Amlan Kar, Despoina Paschalidou, Fangyin Wei, Francesco Ferroni, Guillermo Garcia Cobo, Haithem Turki, Huan Ling, Jaewoo Seo, James Lucas, et al. Nvidia omnidreams: Real-time generative world model for closed-loop autonomous vehicle simulation. *arXiv preprint arXiv:2606.03159*, 2026.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. URL https://arxiv.org/abs/2311.15127.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11621–11631, 2020.

Chenjie Cao, Chaohui Yu, Shang Liu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Mvgenmaster: Scaling multi-view generation from any image via 3d priors enhanced diffusion model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6045–6056. IEEE, 2025.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://arxiv.org/abs/2605.25449.

Giannis Daras, Kulin Shah, Yuval Dagan, Aravind Gollakota, Alexandros G. Dimakis, and Adam Klivans. Ambient diffusion: Learning clean distributions from corrupted data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Zixun Fang, Kai Zhu, Zhiheng Liu, Yu Liu, Wei Zhai, Yang Cao, and Zheng-Jun Zha. ViewPoint: Panoramic video generation with pretrained diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2506.23513.

Haoran Feng, Dizhe Zhang, Xiangtai Li, Bo Du, and Lu Qi. DiT360: High-fidelity panoramic image generation via hybrid training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23367–23377, 2026. URL https://arxiv.org/abs/2510.11712.

Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. MagicDrive: Street view generation with diverse 3D geometry control. In *International Conference on Learning Representations (ICLR)*, 2024a.

Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28135–28144. IEEE, 2025.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Yuliang Guo, Sparsh Garg, S Mahdi H Miangoleh, Xinyu Huang, and Liu Ren. Depth any camera: Zero-shot metric depth estimation from any camera. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26996–27006. IEEE, 2025.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for video diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2404.02101.

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*, volume 30, 2017.

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

Chenhao Ji, Chaohui Yu, Junyao Gao, Fan Wang, and Cairong Zhao. CamPVG: Camera-controlled panoramic video generation with epipolar-aware diffusion, 2025. URL https://arxiv.org/abs/2509.19979.

Junpeng Jiang, Gangyi Hong, Miao Zhang, Hengtong Hu, Kun Zhan, Rui Shao, and Liqiang Nie. Dive: Efficient multi-view driving scenes generation based on video diffusion transformer. *IEEE Transactions on Multimedia*, 2026a.

Le Jiang, Xiangyu Bai, Bishoy Galoaa, Shayda Moezzi, Caleb James Lee, Tooba Imtiaz, Edmund Yeh, Jennifer Dy, Yanzhi Wang, and Sarah Ostadabbas. PanoWorld: Geometry-consistent panoramic video world modeling, 2026b. URL https://arxiv.org/abs/2605.15391.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8. https://github.com/ultralytics/ultralytics, 2023.

Nikolai Kalischek, Michael Oechsle, Fabian Manhardt, Philipp Henzler, Konrad Schindler, and Federico Tombari. CubeDiff: Repurposing diffusion-based image models for panorama generation. In *International Conference on Learning Representations (ICLR)*, 2025.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. HunyuanVideo: A systematic framework for large video generative models, 2024. URL https://arxiv.org/abs/2412.03603.

Suyoung Lee, Jaeyoung Chung, Jaeyoo Huh, and Kyoung Mu Lee. Odgs: 3d scene reconstruction from omnidirectional images with 3d gaussian splattings. *Advances in Neural Information Processing Systems*, 37:57050–57075, 2024.

Haoyuan Li, Dizhe Zhang, Yuemei Zhou, Xiangkai Zhang, Haoran Feng, Xiaofan Lin, Wenjie Jiang, Bo Du, Ming-Hsuan Yang, and Lu Qi. PanoWorld: Real-world panoramic generation, 2026a. URL https://arxiv.org/abs/2607.09661.

Lingen Li, Guangzhi Wang, Xiaoyu Li, Zhaoyang Zhang, Qi Dou, Jinwei Gu, Tianfan Xue, and Ying Shan. CubeComposer: Spatio-temporal autoregressive 4k 360-degree video generation from perspective video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32625–32635, 2026b. URL https://arxiv.org/abs/2603.04291.

Renjie Li, Panwang Pan, Bangbang Yang, Dejia Xu, Shijie Zhou, Xuanyang Zhang, Zeming Li, Achuta Kadambi, Zhangyang Wang, Zhengzhong Tu, and Zhiwen Fan. 4K4DGen: Panoramic 4D generation at 4K resolution. In *International Conference on Learning Representations (ICLR)*, 2025a. URL https://arxiv.org/abs/2406.13527.

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

Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. Drivingdiffusion: Layout-guided multi-view driving scenarios video generation with latent diffusion model. In *European Conference on Computer Vision*, pp. 469–485. Springer, 2024.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2209.03003.

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 *SIGGRAPH 2026 Conference Papers*, 2026. doi: 10.1145/3799902.3811180.

Taiming Lu, Tianmin Shu, Alan Yuille, Daniel Khashabi, and Jieneng Chen. GenEx: Generating an explorable world. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2411.11844.

Rundong Luo, Matthew Wallingford, Ali Farhadi, Noah Snavely, and Wei-Chiu Ma. Beyond the Frame: Generating 360-degree panoramic videos from perspective videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14336–14345, 2025. URL https://arxiv.org/abs/2504.07940.

Jingwei Ma, Erika Lu, Roni Paiss, Shiran Zada, Aleksander Holynski, Tali Dekel, Brian Curless, Michael Rubinstein, and Forrester Cole. VidPanos: Generative panoramic videos from casual panning videos. In *SIGGRAPH Asia 2024 Conference Papers*, 2024.

NVIDIA, Hassan Abu Alhaija, Jose Alvarez, Maciej Bala, Tiffany Cai, Tianshi Cao, Liz Cha, Joshua Chen, Mike Chen, Francesco Ferroni, et al. Cosmos-Transfer1: Conditional world generation with adaptive multimodal control, 2025. URL https://arxiv.org/abs/2503.14492.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132, 2025.

Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025.

Changha Shin, Woong Oh Cho, and Seon Joo Kim. Seam360gs: Seamless 360 gaussian splatting from real-world omnidirectional images. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–10. IEEE, 2025.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Jing Tan, Shuai Yang, Tong Wu, Jingwen He, Yuwei Guo, Ziwei Liu, and Dahua Lin. Imagine360: Immersive 360 video generation from perspective anchor. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. URL https://arxiv.org/abs/2412.03552.

Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. MVDiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Fei Teng, Kai Luo, Sheng Wu, Siyu Li, Pujun Guo, Jiale Wei, Jiaming Zhang, Kunyu Peng, and Kailun Yang. Hallucinating 360°: Panoramic street-view generation via local scenes diffusion and probabilistic prompting. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026. URL https://arxiv.org/abs/2507.06971.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation. In *ICLR 2019 Workshops: DeepGenStruct*, 2019.

Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. Generative camera dolly: Extreme monocular dynamic novel view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 313–331, 2024.

Matthew Wallingford, Anand Bhattad, Aditya Kusupati, Vivek Ramanujan, Matt Deitke, Sham Kakade, Aniruddha Kembhavi, Roozbeh Mottaghi, Wei-Chiu Ma, and Ali Farhadi. From an image to a scene: Learning to imagine the world from a million 360-degree videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. URL https://arxiv.org/abs/2412.07770.

Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Jiahao Wang, Luoxin Ye, TaiMing Lu, Junfei Xiao, Jiahan Zhang, Yuxiang Guo, Xijun Liu, Rama Chellappa, Cheng Peng, Alan Yuille, and Jieneng Chen. EvoWorld: Evolving panoramic world generation with explicit 3D memory, 2025a. URL https://arxiv.org/abs/2510.01183.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

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 (CVPR)*, pp. 6913–6923, 2024a.

Qitai Wang, Lue Fan, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. FreeVS: Generative view synthesis on free driving trajectory. In *International Conference on Learning Representations (ICLR)*, 2025c.

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

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. DriveDreamer: Towards real-world-driven world models for autonomous driving. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024c. doi: 10.1007/978-3-031-73195-2_4.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. $\pi^3$: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, volume 2026, pp. 10481–10497, 2026.

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

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

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*, pp. 1–11, 2024e. doi: 10.1145/3641519.3657518.

Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6902–6912, 2024.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023.

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 *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54015–54029, 2024.

Wei Wu, Xi Guo, Weixuan Tang, Tingxuan Huang, Chiyu Wang, and Chenjing Ding. Drivescape: High-resolution driving video generation by multi-view feature fusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17187–17196. IEEE, 2025.

Ziyi Wu, Daniel Watson, Andrea Tagliasacchi, David J. Fleet, Marcus A. Brubaker, and Saurabh Saxena. 360Anything: Geometry-free lifting of images and videos to 360°. In *ECCV*, 2026.

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)*, 2025. URL https://arxiv.org/abs/2505.22016.

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 (NeurIPS)*, volume 38, 2025.

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, 2025. URL https://arxiv.org/abs/2504.11389.

Runsheng Xu, Hubert Lin, Wonseok Jeon, Hao Feng, Yuliang Zou, Liting Sun, John Gorman, Kate Tolstaya, Sarah Tang, Brandyn White, Ben Sapp, Mingxing Tan, Jyh-Jing Hwang, and Dragomir Anguelov. Wod-e2e: Waymo open dataset for end-to-end driving in challenging long-tail scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3709–3718, 2026.

Yunzhi Yan, Zhen Xu, Haotong Lin, Haian Jin, Haoyu Guo, Yida Wang, Kun Zhan, Xianpeng Lang, Hujun Bao, Xiaowei Zhou, and Sida Peng. StreetCrafter: Street view synthesis with controllable video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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

Zhongqi Yang, Wenhang Ge, Yuqi Li, Jiaqi Chen, Haoyuan Li, Mengyin An, Fei Kang, Hua Xue, Baixin Xu, Yuyang Yin, Eric Li, Yang Liu, Yikai Wang, Hao-Xiang Guo, and Yahui Zhou. Matrix-3D: Omnidirectional explorable 3D world generation, 2025a. URL https://arxiv.org/abs/2508.08086.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations (ICLR)*, 2025b. URL https://arxiv.org/abs/2408.06072.

Weicai Ye, Chenhao Ji, Zheng Chen, Junyao Gao, Xiaoshui Huang, Song-Hai Zhang, Wanli Ouyang, Tong He, Cairong Zhao, and Guofeng Zhang. DiffPano: Scalable and consistent text to panorama generation with spherical epipolar-aware diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

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. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2509.24997.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. TrajectoryCrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 100–111, 2025a.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. ViewCrafter: Taming video diffusion models for high-fidelity novel view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2025b. doi: 10.1109/TPAMI.2025.3613256.

Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. Taming stable diffusion for text to 360° panorama image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6347–6357, 2024.

Cheng Zhang, Hanwen Liang, Donny Y. Chen, Qianyi Wu, Konstantinos N. Plataniotis, Camilo Cruz Gambardella, and Jianfei Cai. PanFlow: Decoupled motion control for panoramic video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, 2026. doi: 10.1609/aaai.v40i15.38231.

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

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

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. 1–12, 2025.

Guosheng Zhao, Chaojun Ni, Xiaofeng Wang, Zheng Zhu, Xueyang Zhang, Yida Wang, Guan Huang, Xinze Chen, Boyuan Wang, Youyi Zhang, Wenjun Mei, and Xingang Wang. Drive-Dreamer4D: World models are effective data machines for 4D driving scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12015–12026, 2025a.

Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10412–10420, 2025b.

Jiawei Zhou, Linye Lyu, Zhuotao Tian, Cheng Zhuo, and Yu Li. Smd: Multi-view safety-critical driving video generation in the real-world domain. In *Forty-third International Conference on Machine Learning*, 2026.

Zehao Zhu, Yuliang Zou, Chiyu Max Jiang, Bo Sun, Vincent Casser, Xiukun Huang, Jiahao Wang, Zhenpei Yang, Ruiqi Gao, Leonidas Guibas, et al. Scenecrafter: Controllable multi-view driving scene editing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6812–6822. IEEE, 2025.

# A Data Construction Details

## A.1 Panoramic Video Curation

Each video first undergoes basic decoding and integrity checks, followed by six automatic quality tests: scene cut detection, optical flow assessment, perspective format detection, dual-fisheye detection, repeated-frame detection, and static-region detection. A video is retained only if it passes all six tests. Together, these tests remove highly noisy, corrupted, or discontinuous videos, clips with insufficient motion, incorrect panoramic formats, repeated frames, and persistent static regions such as borders or virtual-object overlays (e.g., car dashboards).

## A.2 Multi-Camera ERP Projection

For each multi-camera clip, we project all camera views onto a common ERP canvas using the camera calibration and rotation only. For directions observed by multiple cameras, each ray takes its color from the camera whose viewing direction is closest to that ray. We do not apply blending, depth warping, or inpainting.

Because neighboring cameras have different optical centers and may also differ slightly in capture time and exposure, the projected views do not necessarily agree in their overlap regions. Colors may change abruptly across camera boundaries, while nearby or moving objects can appear truncated because adjacent cameras observe them at different positions. These artifacts motivate the reliability masks described in the main paper.

# B Stage-I Training Details

## B.1 Geometric Camera-Control Construction

For each panoramic training video, we extract eight overlapping perspective views from the initial ERP frame using a $90^\circ$ horizontal field of view and $45^\circ$ yaw intervals. This produces eight common-center views with $50\%$ overlap and full horizontal coverage of the panorama.

We combine these eight initial views with 92 temporally ordered views from the front direction, yielding 100 perspective images for a 93-frame video. Pi3 jointly reconstructs the scene geometry and estimates the camera poses from these images. The pose of the front-facing frame-0 view

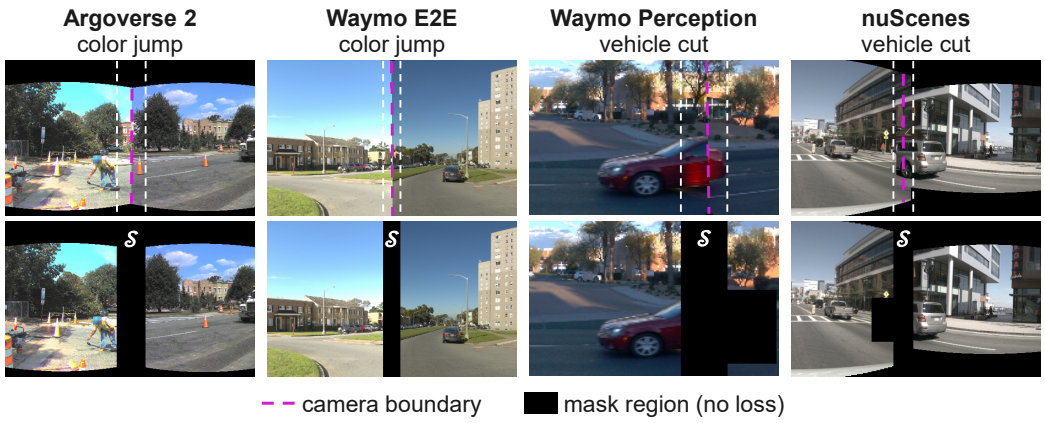


Figure 6: **Partial-ERP construction across different camera rigs.** Top: projected ERP observations with camera boundaries marked in magenta. Bottom: the corresponding masked targets, where black regions are excluded from direct RGB supervision. nuScenes is reserved for evaluation.

together with the poses of the 92 subsequent front views forms the 93-frame temporal camera trajectory used for rendering the ERP control sequence.

We denote the reconstructed point cloud as $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^{N}$ and the temporal camera trajectory as $\mathcal{T} = \{\mathbf{T}_t\}_{t=0}^{T-1}$, where $\mathbf{T}_t \in SE(3)$. The geometric control at time $t$ is

$$\mathbf{C}_t = \mathcal{R}_{\text{ERP}}\left(\mathcal{P}, \mathbf{T}_t\right), \qquad t = 0, \ldots, T-1, \tag{8}$$

where $\mathcal{R}_{\text{ERP}}$ denotes point-cloud rendering onto the ERP representation. The resulting sequence $\mathbf{C} = \{\mathbf{C}_t\}_{t=0}^{T-1}$ provides geometric guidance along the prescribed camera trajectory.

### B.2 Mask Conditioning

The Stage-I mask pool contains diverse spatial observation patterns. For each training sample, one mask is sampled at the sequence level and applied consistently throughout the entire video; the mask is not independently resampled across frames. Formally,

$$\mathbf{M}_t^{\text{prior}} = \mathbf{M}^{\text{prior}}, \qquad \forall t. \tag{9}$$

This sequence-level masking preserves temporal consistency of the conditioning while exposing the model to different spatial observation patterns across training samples.

For the Stage-I configuration that masks geometric control, the corresponding spatial regions of the point-cloud control sequence are masked consistently with the sampled visual mask.

## C Evaluation Details

### C.1 Q-Align Evaluation Protocol

For Q-Align (Wu et al., 2024), we do not evaluate the full ERP directly. Each generated ERP video is projected into eight equatorial perspective video clips with yaw angles spaced by $45°$ and a $90° \times 90°$ field of view. Each clip contains 25 temporally sampled frames and is evaluated using the Q-Align video scorer.

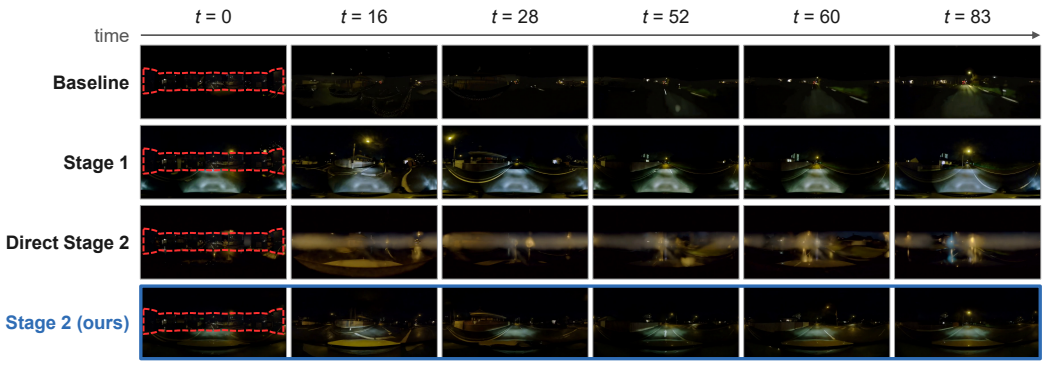


Figure 7: Qualitative ablation of panoramic initialization on nuScenes. Red boxes mark the visible regions in the initial frame.

# D Additional Qualitative Results

## D.1 Panoramic Initialization Ablation

Figure 7 complements the quantitative ablation in Table 2 with a representative sequence from the unseen nuScenes dataset.