R0: FEED-FORWARD 3D RECONSTRUCTION FOR ARBITRARY CENTRAL CAMERAS

Anonymous authors
Paper under double-blind review

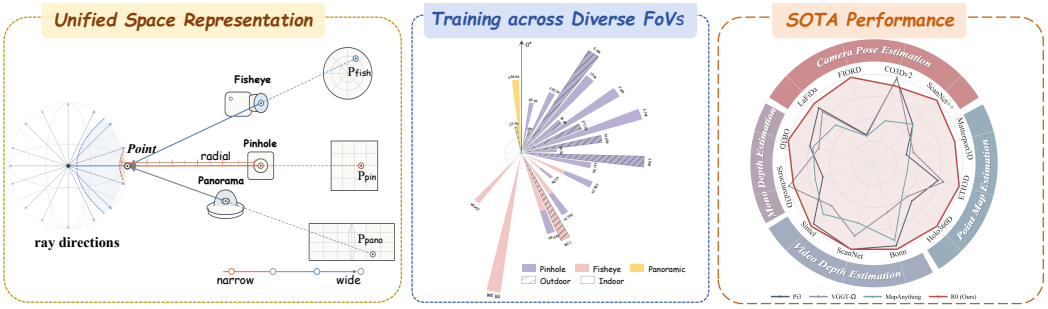

Figure 1: **One ray space unifies diverse camera projections.** R0 represents pinhole, fisheye, and panoramic cameras using shared ray directions and radial distances, learns across continuous fields of view, and achieves strong pose, reconstruction, and depth performance across projections.

ABSTRACT

Autonomous driving, robotics, and immersive VR/AR systems increasingly rely on fisheye and panoramic cameras for perception. However, existing feed-forward 3D reconstruction models are built on pinhole projection and degrade significantly on non-pinhole cameras. The root cause is that different cameras are treated as discrete categories, whereas all central cameras actually share the same ray space and differ only in how they unfold rays onto the image plane. We present R0, to our knowledge the first camera-agnostic feed-forward reconstruction model trained at scale across pinhole, fisheye, and panoramic cameras. By predicting a continuous ray field, constructing spherical neighborhood sampling, and conditioning cross-view attention on relative geometry, R0 unifies sampling, interaction, and output before projection. It natively handles different fields of view and unfolding functions, without requiring camera intrinsics, projection type, or undistortion at inference. Experiments on pinhole, fisheye, panoramic, and heterogeneous cameras show that R0 achieves state-of-the-art results on non-pinhole data while remaining competitive on pinhole data. Code, models, and large-scale data processing details will be released.

1 INTRODUCTION

Feed-forward 3D reconstruction foundation models are reshaping how machines recover geometry from images. Models such as VGGT Wang et al. (2025) and its successors Keetha et al. (2026); Wang et al. (2026c); Xie et al. (2026); Wang et al. (2026a) can infer dense geometry and camera poses from a collection of images in a single forward pass, turning what once required costly per-scene optimization into an efficient and generalizable prediction problem. Yet this remarkable progress has largely unfolded within the world of pinhole cameras. Real-world perception systems often see through much wider fields of view. Fisheye cameras are routinely deployed on autonomous vehicles and mobile robots for surround perception, while panoramic cameras provide compact observations for indoor and outdoor scene reconstruction Ji et al. (2026), embodied navigation, and immersive spatial capture. These sensors are not rare exceptions to the pinhole setting. In many applications, they are precisely what enables comprehensive and robust perception. Unfortunately, current foundation models do not readily inherit this advantage. Their geometric representations and

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.

learned priors are closely tied to perspective projection, where image coordinates follow a specific mapping to camera rays. Fisheye and panoramic cameras parameterize these rays in fundamentally different ways, breaking the correspondence between image patterns, ray directions, and 3D structure learned from pinhole imagery. As a result, a model that appears geometrically general across scenes may remain surprisingly specific to the camera through which those scenes are observed.

If camera generalization were merely a matter of data diversity, mixing pinhole, fisheye, and panoramic images during training should provide a naturally straightforward solution. We test this hypothesis by fine-tuning Pi3 Wang et al. (2026c) on a mixture of these camera types. Surprisingly, mixed training degrades the model's original performance on pinhole images, yet still falls short of camera-specific models separately fine-tuned on fisheye or panoramic data. This failure reveals that the obstacle lies deeper than the training distribution. Current models reason primarily in the image plane, where image coordinates and visual patterns are coupled with projection-specific ray geometry. The same image-plane evidence can therefore carry incompatible geometric meanings under different camera models. Learning a new pixel-to-ray relationship interferes with the perspective prior already encoded by the network, creating a trade-off between retaining existing capabilities and acquiring new ones. Overcoming this trade-off requires more than exposing the model to additional camera types. It requires reconsidering the space in which geometric information is represented and exchanged. This leads to the central question of our work: *Is there a projection-agnostic space shared by all central cameras, where heterogeneous images can interact without geometric conflict?*

We find this common representation in ray space. For any central camera, each pixel corresponds to a ray emanating from the optical center, and a 3D point is specified by its ray direction and radial distance. This underlying geometry remains the same across camera types. Pinhole, fisheye, and panoramic cameras differ only in which directions they observe and how those directions are mapped onto the image domain. Projection therefore changes the parameterization of rays, but not the rays themselves. From this perspective, a camera can be characterized continuously by its directional coverage and projection function, rather than by a discrete camera label. Heterogeneous cameras are no longer separate geometric categories, but different parameterizations of the same ray space.

This unified view of cameras directly shapes the design of our model. Since camera variation can be expressed continuously through directional coverage and projection, a general model should neither rely on discrete camera labels nor be tied to a fixed projection. It should instead infer the ray associated with each pixel and carry out geometric reasoning directly in ray space. Following this principle, we introduce **R0** (Reconstruction-Zero), a feed-forward reconstruction model built upon the permutation-equivariant multi-view architecture of Pi3 Wang et al. (2026c). R0 replaces the three-channel point-map head of Pi3 with a four-channel ray–radial head that disentangles 3D geometry into viewing direction and radial distance. At its core, a *Ray Field Predictor* estimates a continuous ray field that places pixels from heterogeneous cameras in a shared geometric coordinate system. This field then guides geometric reasoning throughout the network. At the local level, *Ray Canonical Sampling* constructs neighborhoods on the viewing sphere, making local aggregation independent of how rays are unfolded onto the image plane. At the global level, *Geometry Attention* conditions cross-view matching on ray directions, allowing observations under different projections to interact through their underlying geometry. Given one or multiple images, R0 jointly predicts a viewing-ray direction and radial distance for every pixel. Their combination directly recovers the corresponding 3D point. For multi-view inputs, R0 additionally estimates the relative camera poses.

As summarized in Figure 1, experiments validate the effectiveness of the ray-space representation across camera projections and tasks. On FIORD, R0 improves pose AUC@30 from 44.401 to 81.638, while reducing RRE and RTE from the best competing values of 18.256 and 16.056 to 3.689 and 5.392, respectively. On ScanNet++, it raises AUC@30 from 51.500 to 67.200. For point-map reconstruction on OB3D, R0 improves Acc., Comp., and NC from the best competing value for each metric, $2.166/3.330/0.655$, to $0.670/0.471/0.800$. Meanwhile, it retains strong performance on pinhole inputs, achieving Acc./Comp. errors of $0.160/0.110$ on ETH3D. These consistent gains across camera geometries and tasks indicate that ray space provides a shared geometric basis for heterogeneous cameras and are consistent with complementary supervision across projections.

Our main contributions are as follows: **(i) Unified ray-space representation**: we propose to define 3D reconstruction in ray space before projection, unifying arbitrary central cameras into a continuous direction–distance representation, without discrete camera category labels. **(ii) Continuous ray field for projection-agnostic interaction.** We represent camera geometry as a continuous ray

field parameterized by spherical harmonics and refined through ray residual correction. This shared geometric coordinate further guides canonical neighborhood sampling and cross-view attention, enabling consistent geometric reasoning across camera projections without sacrificing performance on pinhole images. **(iii) Strong cross-camera generalization**: We train and evaluate R0 on multiple datasets spanning pinhole, fisheye, panoramic, and heterogeneous multi-view cameras. R0 substantially outperforms general pinhole feed-forward models on non-pinhole data and even surpasses the specific methods directly fine-tuned on the corresponding camera types, while remaining competitive with same-scale baselines across multiple pinhole benchmarks. These results suggest that unified modeling in ray space allows knowledge learned from large-scale pretraining to transfer effectively across camera geometries.

2 Related Work

2.1 Feed-Forward Reconstruction

Feed-forward 3D reconstruction has progressed rapidly since DUSt3R Wang et al. (2024) formulated two-view reconstruction as direct point map regression. VGGT Wang et al. (2025) jointly predicts cameras, depth, point maps, and tracks from arbitrary images in a single forward pass, while Pi3 Wang et al. (2026c) removes the fixed reference frame and achieves permutation-equivariant reconstruction with local point maps and relative poses. MapAnything Keetha et al. (2026) and Depth Anything 3 Wang et al. (2026c) further explore decomposed representations. Recent work scales these models to longer sequences and larger corpora: Scal3R Xie et al. (2026) introduces test-time global context memory, LingBot-Map Chen et al. uses a geometric context transformer for streaming reconstruction, and InfiniteVGGT Yuan et al. (2026) extends VGGT to endless streams. VGGT-Ω Wang et al. (2026a) reduces training memory with a compact dense head and register attention, while FastVGGT Shen et al. (2026) improves inference efficiency.

2.2 Reconstruction under Diverse Cameras

Non-pinhole depth estimation handles ERP distortion or fisheye projection through multi-projection fusion, tangent-plane decomposition, or self-supervised reprojection Wang et al. (2026b); Jiang et al. (2021); Wang et al. (2026c); Shen et al. (2022); Ai & Wang (2024); Zhao et al. (2025); Deng et al. (2025), and more recently generalizes across camera models with pitch-aware transformations, spherical representations, or relative depth and scale Guo et al. (2026); Piccinelli et al. (2025); Ganesan et al. (2026). These methods are largely monocular and often assume a known camera model or calibration.

Multi-view reconstruction under diverse cameras has also been explored. Wid3R Jung et al. (2026) and Fisheye3R Duan et al. (2026) condition on camera-model or calibration tokens, CAM3R Guruprasad et al. (2026) predicts rays and radial distances from uncalibrated image pairs, and PanoVGGT Guo et al. (2026), Argus Li et al. (2026b), and FastPano3D Li et al. (2026a) design panorama-specific representations. These methods rely on discrete conditions, calibration, or projection-specific representations. No unified shared representation across camera models exists yet, and no large-scale pretrained feed-forward model natively handles diverse central cameras within a single framework.

3 Method

R0 performs all geometric reasoning in ray space. A shared encoder extracts features, a Ray Field Predictor estimates a continuous ray field, Ray Canonical Sampling constructs spherical neighborhoods, and Geometry Attention performs cross-view interaction. Output heads predict ray directions, radial distances, point maps, confidence, and relative poses. Section 3.1 defines the unified representation, Section 3.2 describes the Ray Field Predictor, Section 3.3 introduces Ray Canonical Sampling and Geometry Attention, and Section 3.4 presents training and inference.

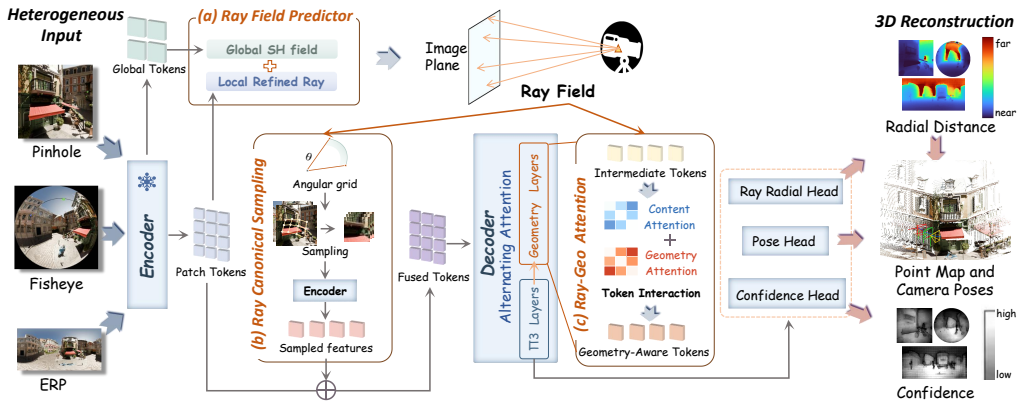

Figure 2: **Overview of R0.** The model takes RGB images from arbitrary central cameras, predicts a continuous ray field, and performs canonical sampling and geometry attention in ray space before producing ray directions, radial distances, point maps, confidence, and relative poses.

3.1 Unified Representation

When a camera captures a 3D scene, a point on an object is mapped along a ray from the optical center onto the image plane, becoming a pixel. Conversely, each pixel corresponds to a ray from the optical center. A 3D point is uniquely determined by the unit direction $\hat{\mathbf{r}} \in \mathbb{S}^2$ of this ray and the radial distance $\rho \in \mathbb{R}^+$ along it:

$$\mathbf{X} = \rho\hat{\mathbf{r}}. \tag{1}$$

We call the space formed by all such direction-distance pairs the ray space. It is camera-agnostic: different cameras only differ in how pixels are mapped to rays.

Given an unordered set of RGB images $\mathcal{I} = \{I_v\}_{v=1}^V$ captured by arbitrary central cameras, for a pixel $\mathbf{p} = (u, v)$ in view v, the model predicts a unit viewing direction $\hat{\mathbf{r}}_v(\mathbf{p})$ and a radial distance $\hat{\rho}_v(\mathbf{p})$. The 3D point is recovered as $\hat{\mathbf{X}}_v(\mathbf{p}) = \hat{\rho}_v(\mathbf{p})\hat{\mathbf{r}}_v(\mathbf{p})$. The full model F_Θ outputs, for each view, a ray field $\hat{\mathbf{R}}_v$, a radial-distance map $\hat{\rho}_v$, a point map $\hat{\mathbf{X}}_v$, a confidence map $\hat{\mathbf{C}}_v$, and a relative camera pose $\hat{\mathbf{T}}_v$:

$$F_\Theta(\mathcal{I}) = \left\{\hat{\mathbf{R}}_v,\ \hat{\rho}_v,\ \hat{\mathbf{X}}_v,\ \hat{\mathbf{C}}_v,\ \hat{\mathbf{T}}_v\right\}_{v=1}^{V}. \tag{2}$$

We implement this representation on Pi3 by replacing its point map head with a four-channel ray–radial head that outputs rays and distances.

3.2 Ray Field Predictor

Ray-space geometry requires the viewing direction associated with every pixel. Our Ray Field Predictor estimates a dense, continuous ray field from the global and local features of a frozen ViT encoder, providing the geometric conditioning used by subsequent modules.

We first predict expansion parameters $\mathbf{c}_v$ from the class token, including the field of view, spherical center, and continuous parameters controlling the radial curve $r(\theta)$. These parameters define a differentiable mapping from pixel coordinates $\mathbf{u}$ to spherical directions $\boldsymbol{\omega}_v(\mathbf{u})$. On this sphere, the ray field is represented by a third-order real spherical-harmonic expansion:

$$\mathbf{R}_{0,v}(\mathbf{u}) = \mathrm{norm}\left(\sum_{l=1}^{3}\sum_{m=-l}^{l} \mathbf{a}_{v,lm}\, Y_l^m(\boldsymbol{\omega}_v(\mathbf{u}))\right), \tag{3}$$

where Y_l^m denotes a real spherical-harmonic basis and $\mathbf{a}_{v,lm} \in \mathbb{R}^3$ is the corresponding vector-valued coefficient. We refer to this global field as the SH Ray Field $\mathbf{R}_{0,v}$.

To model higher-order lens distortion and local deviations not captured by the harmonic expansion, we predict a per-pixel 3D residual Δ_v from the patch tokens of the frozen encoder. It is projected onto the tangent plane of $\mathbf{R}_{0,v}(\mathbf{u})$ to obtain a tangent displacement $\mathbf{t}_v(\mathbf{u})$, whose angular magnitude is smoothly bounded by a maximum correction angle τ along spherical geodesics. The bounded displacement is then mapped back to the unit sphere with the exponential map. Adding this residual yields the Refined Ray Field $\hat{\mathbf{R}}_v$, which is used by all subsequent modules.

3.3 Ray-Space Interaction

By associating every pixel with a unit direction, the Refined Ray Field enables the model to construct consistent local observations within each view and to establish long-range geometric correspondences across views. Implementation details are provided in Appendix A.1.

Ray Canonical Sampling. A fixed image-plane neighborhood subtends different solid angles under different projections, preventing local features from aligning across cameras. We therefore define neighborhoods in ray space using angular rather than pixel radii, so that the same radius covers the same spherical region for every camera. Specifically, we select center directions within the valid ray domain, construct multi-scale directional grids on their tangent planes, and map these directions back to the image through differentiable inverse projection. The neighborhoods are defined in ray space while their values are sampled from the 2D image. A shared encoder processes each sampled region, and the resulting features are injected into the native patch tokens through a learned gate whose strength combines a geometric-reliability prior with a learned residual.

Geometry Attention. Ray Canonical sampling provides consistent spherical neighborhoods within each view, but cross-view correspondence still requires explicit geometric coupling. We introduce Geometry Attention, which exchanges information between tokens across positions and views according to their relative ray geometry. It is applied to the final 18 Transformer blocks of the decoder.

Content attention in existing pinhole feed-forward models relies only on appearance and 2D position. We augment it with geometric matching. The geometric condition $\mathcal{G}_i$ of token i contains its ray direction, 3D point, centered log radial distance, and validity indicator. Each token encodes only its own geometric state, and attention inner products induce relative geometric relations. The attention score combines content and geometric matching:

$$A_{ij}^{\text{ray}} = \mathbf{q}_i^{\text{content}\top}\mathbf{k}_j^{\text{content}} + \sum_g \gamma_g K_g(\mathcal{G}_i, \mathcal{G}_j). \tag{4}$$

where γ_g is a learned per-kernel weight and K_g is one of three relative kernels. These kernels measure the angular cosine between ray directions, the relative distance between 3D points, and the difference between centered log radial distances, respectively.

3.4 Training and Inference

Supervision and Losses. Ground-truth rays $\mathbf{r}^*$ are generated from the camera projection model. For pinhole data, the ground-truth radial distance is

$$\rho^*(\mathbf{p}) = \frac{z^*(\mathbf{p})}{r_z^*(\mathbf{p})}, \tag{5}$$

where r_z^* is the component of $\mathbf{r}^*$ along the pinhole optical axis. Radial distances for fisheye and panoramic data are obtained directly from their projection models. Over the valid region Ω, the ray and radial-distance losses are

$$\mathcal{L}_{\text{ray}} = \frac{1}{|\Omega|}\sum_{\mathbf{p}\in\Omega}(1 - \hat{\mathbf{r}}^\top\mathbf{r}^*). \tag{6a}$$

$$\mu = \text{median}_{\mathbf{p}\in\Omega}(\log\hat{\rho} - \log\rho^*),$$
$$\mathcal{L}_{\text{radial}} = \frac{1}{|\Omega|}\sum_{\mathbf{p}\in\Omega}|\log\hat{\rho} - \log\rho^* - \mu|. \tag{6b}$$

The total objective is

$$\mathcal{L} = \lambda_p\mathcal{L}_{\text{point}} + \lambda_r\mathcal{L}_{\text{ray}} + \lambda_\rho\mathcal{L}_{\text{radial}} + \lambda_{\text{pose}}\mathcal{L}_{\text{pose}} + \lambda_c\mathcal{L}_{\text{conf}}, \tag{7}$$

Table 1: **Point cloud reconstruction across camera projections.**

Method	Matterport3D			ETH3D			OB3D			Holo360D		
	Acc ↓	Comp ↓	NC ↑	Acc ↓	Comp ↓	NC ↑	Acc ↓	Comp ↓	NC ↑	Acc ↓	Comp ↓	NC ↑
VGGT Wang et al. (2025)	**0.271**	1.986	0.537	0.293	0.338	0.859	3.368	3.905	0.655	17.919	8.845	0.557
Pi3 Wang et al. (2026c)	0.801	1.076	0.554	0.197	0.213	0.882	4.287	3.554	0.627	7.265	2.066	0.574
MapAnything Keetha et al. (2026)	0.482	1.298	0.543	0.480	0.575	0.802	3.495	3.970	0.617	7.284	4.332	0.535
Depth Anything 3 Lin et al. (2025)	0.909	1.052	0.543	0.317	0.391	0.825	2.705	3.330	0.608	9.767	2.936	0.549
Pi3X Wang et al. (2026c)	0.513	1.361	0.546	0.184	0.201	**0.900**	8.200	5.508	0.622	73.883	3.430	0.569
VGGT-Ω Wang et al. (2026a)	0.335	1.384	0.550	0.173	0.191	0.896	2.166	3.385	0.651	8.956	2.473	0.576
R0 (Ours)	0.391	**0.210**	**0.802**	**0.160**	**0.110**	0.875	**0.670**	**0.471**	**0.800**	**2.702**	**1.299**	**0.724**

where $\mathcal{L}_{\text{point}}$, $\mathcal{L}_{\text{pose}}$, $\mathcal{L}_{\text{conf}}$, and the loss weights follow Pi3.

Training Strategy. The vision encoder is initialized from a pretrained DINOv2 Oquab et al. (2024) model and kept frozen throughout training. Our decoder is the 36-layer alternating-attention backbone. It is initialized from pretrained Pi3 weights and fully fine-tuned. The original Pi3 point map head is replaced by a four-channel ray–radial head, while the pose and confidence heads are retained. Training proceeds in two stages: Stage 1 trains the Ray Field Predictor jointly with the reconstruction model, and Stage 2 freezes the Predictor and gradually switches the geometric input from ground truth to its predictions. To improve generalization across scenes and camera models, we train on 31 data sources spanning indoor, outdoor, object-centric, driving, and dynamic scenes, with both real and synthetic data. The full list and split protocol are provided in Appendix A.2.

Inference. R0 takes only RGB images and predicts rays, radial distances, point maps, confidence, and relative poses, requiring no camera intrinsics, projection type, or undistortion.

4 Experiments

We evaluate dense point maps and multi-view reconstruction (Section 4.1), camera pose (Section 4.2), and video and monocular depth (Section 4.3) across pinhole, fisheye, panoramic, and heterogeneous cameras. R0 achieves state-of-the-art or competitive performance. Qualitative reconstruction results are shown in Figure 3. Evaluation and sampling protocols are detailed in Appendix A.3, which also specifies the metric units used in the tables.

4.1 Dense Point Map and Multi-View Reconstruction

We evaluate multi-view point-map reconstruction on Matterport3D, ETH3D, OB3D, and Holo360D, spanning pinhole and panoramic cameras as well as indoor, outdoor, object-centric, and scene-level settings (Table 1). For all reconstruction datasets, predicted point maps are aligned to the ground truth by a proper Sim(3) transformation followed by point-to-point ICP. We report accuracy (Acc.), completeness (Comp.), and normal consistency (NC). Dataset-specific sampling and result-averaging protocols are detailed in Appendix A.3.

Results and Analysis. Table 1 summarizes the 3D point cloud reconstruction performance across diverse benchmark datasets and projection geometries. Overall, R0 establishes new state-of-the-art results across all evaluated benchmarks, demonstrating superior geometric accuracy, surface completeness, and normal consistency. The advantage of our model is most pronounced on challenging non-pinhole benchmarks such as OB3D and Holo360D. On OB3D, R0 outperforms all feed-forward baselines by a large margin, reducing reconstruction error and incompleteness from the previous best results of 2.166 and 3.330 to 0.670 and 0.471, respectively, while boosting the normal consistency from 0.655 to 0.800. Similarly on Holo360D, R0 cuts the Chamfer-based errors by more than half compared to the strongest competitors, driving Acc down to 2.702 and Comp down to 1.299, alongside a marked improvement in NC to 0.724. On Matterport3D, while VGGT obtains a numerically lower Acc (0.271), this comes at the expense of severe geometric incompleteness, as evidenced by its poor Comp (1.986) and low NC (0.537). In contrast, R0 achieves an unprecedented completeness score of 0.210 and a normal consistency of 0.802, an improvement of over 0.25 in NC, and surface-aligned reconstructions.

Table 2: **Camera pose estimation on LaFiDa, FIORD, and OB3D.**

Method	LaFiDa			FIORD			OB3D		
	AUC@30° ↑	RRE (°)↓	RTE (°)↓	AUC@30° ↑	RRE (°)↓	RTE (°)↓	AUC@30° ↑	RRE (°)↓	RTE (°)↓
Pi3 Wang et al. (2026c)	13.795	22.224	31.650	44.401	18.273	16.056	21.657	26.474	29.838
Pi3X Wang et al. (2026c)	8.001	52.431	42.246	40.468	19.945	17.485	24.282	26.398	26.598
VGGT Wang et al. (2025)	3.133	80.048	46.803	40.580	20.207	18.445	22.520	25.673	27.400
VGGT-Ω Wang et al. (2026a)	15.169	31.128	**30.613**	42.912	18.727	16.842	17.136	28.119	34.512
MapAnything Keetha et al. (2026)	8.548	36.608	35.920	42.230	18.256	17.824	20.903	26.952	28.656
Depth Anything 3 Lin et al. (2025)	10.034	32.253	35.416	41.695	18.463	16.503	16.379	31.076	35.182
R0 (Ours)	**15.184**	**17.193**	35.655	**81.638**	**3.689**	**5.392**	**57.275**	**4.295**	**18.191**

Table 3: **Camera pose estimation on Sintel, CO3Dv2, and ScanNet++.**

Method	Sintel			CO3Dv2			ScanNet++ (DSLR)		
	ATE ↓	RPE-T ↓	RPE-R ↓	AUC@30° ↑	RRE (°)↓	RTE (°)↓	AUC@30° ↑	RRE (°)↓	RTE (°)↓
Pi3 Wang et al. (2026c)	**0.074**	0.039	0.279	87.704	3.048	4.286	50.700	8.697	15.693
Pi3X Wang et al. (2026c)	0.125	0.041	0.295	87.080	3.039	4.404	51.400	8.564	15.206
VGGT Wang et al. (2025)	0.172	0.062	0.471	**89.987**	**2.334**	**3.951**	38.500	12.390	20.487
VGGT-Ω Wang et al. (2026a)	0.100	**0.031**	**0.217**	87.867	3.028	4.122	46.100	12.188	15.225
MapAnything Keetha et al. (2026)	0.226	0.077	0.642	71.442	8.833	9.479	51.500	8.575	16.009
Depth Anything 3 Lin et al. (2025)	0.114	0.056	0.469	88.445	3.077	4.357	32.400	17.592	20.142
R0 (Ours)	0.175	0.087	0.462	85.803	3.344	4.814	**67.200**	**3.915**	**10.629**

Crucially, these substantial advancements on non-standard projections do not sacrifice canonical pinhole reconstruction quality. On the standard ETH3D benchmark, R0 secures the leading Acc (0.160) and Comp (0.110) while maintaining highly competitive normal consistency (0.875), verifying that our unified formulation simultaneously masters wide-angle distortion, panoramic fields of view, and standard perspective cameras.

4.2 Camera Pose Estimation

We evaluate camera poses predicted jointly with scene geometry without providing camera intrinsics or projection labels, as summarized in Table 2 and Table 3. Across challenging non-pinhole setups, R0 sets a new state of the art by substantial margins. Most notably, on FIORD and OB3D, R0 nearly doubles the pose accuracy over the best baselines, raising AUC@30° to 81.638 and 57.275 while reducing rotation errors (RRE) from over 18° and 25° to just 3.689° and 4.295°. On real-world ScanNet++ scenes, R0 likewise achieves the top AUC@30° of 67.200 and cuts the rotation error to 3.915°, more than halving baseline errors. Meanwhile, R0 remains competitive on canonical benchmarks such as Sintel and CO3Dv2, confirming that its robust cross-projection capabilities do not compromise standard camera pose tracking. Dataset protocols and setup details are documented in Appendix A.3.

4.3 Depth Estimation

We evaluate video depth using AbsRel and δ_1, and additionally report RMSE for monocular depth. These two settings test complementary properties: video depth requires a consistent scale over an entire sequence, whereas monocular evaluation measures the quality of each independently predicted depth map. For ERP images, all metrics are weighted by equal solid angle to avoid overemphasizing pixels near the poles.

Video depth estimation. We align each predicted sequence using a shared scale and shift to preserve cross-frame temporal consistency, evaluating on Bonn, ScanNet, and Sintel in Table 4. R0 consistently attains top-tier results across all setups. On Bonn, R0 reaches the lowest RMSE of 0.205 alongside the highest threshold accuracy ($\delta_1 = 0.976$). On ScanNet, it matches the best AbsRel (0.026) while securing the leading δ_1 of 0.992. Furthermore, on Sintel, R0 significantly reduces global trajectory drift, yielding the lowest RMSE of 3.611 and outperforming prior methods.

Table 4: **Video depth estimation on Bonn, ScanNet, and Sintel.**

Method	Bonn			ScanNet		Sintel		
	AbsRel ↓	RMSE ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	RMSE ↓	δ_1 ↑
VGGT Wang et al. (2025)	0.049	0.249	0.974	0.038	0.984	0.226	5.565	0.683
Pi3 Wang et al. (2026c)	0.044	0.221	0.975	**0.026**	0.990	0.210	3.978	0.728
Pi3X Wang et al. (2026c)	**0.038**	0.211	0.975	0.026	0.991	**0.188**	4.027	**0.737**
MapAnything Keetha et al. (2026)	0.050	0.239	0.975	0.053	0.972	0.290	4.585	0.688
Depth Anything 3 Lin et al. (2025)	0.049	0.223	0.973	0.053	0.966	0.266	5.490	0.654
VGGT-Ω Wang et al. (2026a)	0.123	0.357	0.930	0.036	0.980	0.357	5.843	0.632
R0 (Ours)	0.042	**0.205**	**0.976**	0.026	**0.992**	0.207	**3.611**	0.732

Table 5: **Monocular depth estimation on Structured3D and OB3D.**

Method	Structured3D			OB3D		
	AbsRel ↓	RMSE ↓	δ_1 ↑	AbsRel ↓	RMSE ↓	δ_1 ↑
VGGT Wang et al. (2025)	**0.155**	**0.278**	0.889	0.203	1.672	0.672
Pi3 Wang et al. (2026c)	0.480	1.017	0.895	0.183	1.425	0.697
MapAnything Keetha et al. (2026)	0.356	0.867	0.880	0.190	1.583	0.692
Depth Anything 3 Wang et al. (2026c)	0.177	0.307	0.882	0.215	1.707	0.641
VGGT-Ω Wang et al. (2026a)	0.170	0.304	0.883	0.212	1.803	0.650
Pi3X Wang et al. (2026c)	0.237	0.505	**0.901**	0.209	1.627	0.650
R0 (Ours)	0.193	0.636	0.894	**0.100**	**1.418**	**0.905**

Monocular depth estimation. We evaluate monocular predictions by aligning each depth map independently in Table 5. While maintaining competitive accuracy on Structured3D, R0 demonstrates overwhelming advantages on the OB3D benchmark. Specifically, it cuts the relative error in half down to an AbsRel of 0.100, while boosting δ_1 accuracy from below 0.70 across all baselines to 0.905, marking an absolute gain of over 20%. Detailed protocols are provided in Appendix A.3.

4.4 Ablation Study

We ablate ray-space sampling and interaction on Holo360D-ERP, Holo360D-Fisheye, OmniScene-Heterogeneous, and ScanNet-Pinhole, where Holo360D-Fisheye is derived by projecting a subset of Holo360D scenes. The four variants are: R0 w/o Samp. & Attn. replaces only the output head; R0 w/o Samp. removes Ray Canonical Sampling; R0 w/o Attn. removes Geometry Attention; and R0 (all) enables all components. Pi3-FT uses the same training data and optimization settings, isolating the effect of architecture from that of additional training. Table 6 reports reconstruction metrics, and additional comparisons of the Ray Field Predictor are provided in Appendix A.4.

Pi3-FT substantially degrades ScanNet-Pinhole accuracy from 0.053 to 0.215, confirming the geometric conflict caused by image-plane representations under mixed training. Replacing only the ray/radial output head recovers pinhole accuracy from 0.215 to 0.075; adding Ray Canonical Sampling and Geometry Attention further improves it to 0.063, while substantially improving all non-pinhole benchmarks. The consistent gains from adding sampling and interaction indicate that the three components make complementary contributions. Overall, the complete model achieves the best or near-best results across the evaluated camera types while retaining competitive pinhole reconstruction.

4.5 Qualitative Comparisons

Figure 3 illustrates the zdmulti-view 3D reconstruction results across fisheye, standard pinhole, and panoramic cameras. Existing feed-forward baselines such as Pi3, MapAnything, and VGGT struggle to generalize beyond canonical camera geometries, whereas our method consistently yields complete and structurally coherent geometry across all camera setups. For fisheye inputs with extreme radial distortion, Pi3 truncates large portions of the room, and both MapAnything and VGGT suffer

Table 6: **Ablation on point cloud reconstruction.**

Method	Holo360D-ERP				Holo360D-Fisheye				OmniScene-Heterogeneous				ScanNet-Pinhole			
	Acc ↓	Comp ↓	NC ↑	CD ↓	Acc ↓	Comp ↓	NC ↑	CD ↓	Acc ↓	Comp ↓	NC ↑	CD ↓	Acc ↓	Comp ↓	NC ↑	CD ↓
Pi3 Wang et al. (2026c)	2.435	1.535	0.547	3.970	0.426	1.115	0.613	1.541	0.541	2.588	0.526	3.129	**0.053**	**0.053**	**0.925**	**0.106**
Pi3-FT	0.348	0.476	0.642	0.824	0.252	0.144	0.692	0.396	0.357	**0.310**	0.814	0.667	0.215	0.152	0.800	0.368
R0 w/o Samp. & Attn.	0.193	0.174	0.853	0.367	0.160	0.118	0.874	0.278	0.280	0.451	0.825	0.731	0.093	0.075	0.848	0.168
R0 w/o Samp.	0.183	0.155	0.879	0.338	0.138	0.110	0.910	0.248	0.287	0.485	0.838	0.772	0.082	0.069	0.868	0.151
R0 w/o Attn.	0.158	0.139	0.901	0.297	0.114	0.098	0.926	0.212	**0.227**	0.375	0.872	0.602	0.068	0.061	0.890	0.129
R0 (all)	**0.145**	**0.133**	**0.907**	**0.278**	**0.107**	**0.094**	**0.931**	**0.201**	0.241	0.361	**0.873**	**0.602**	0.063	0.059	0.894	0.122

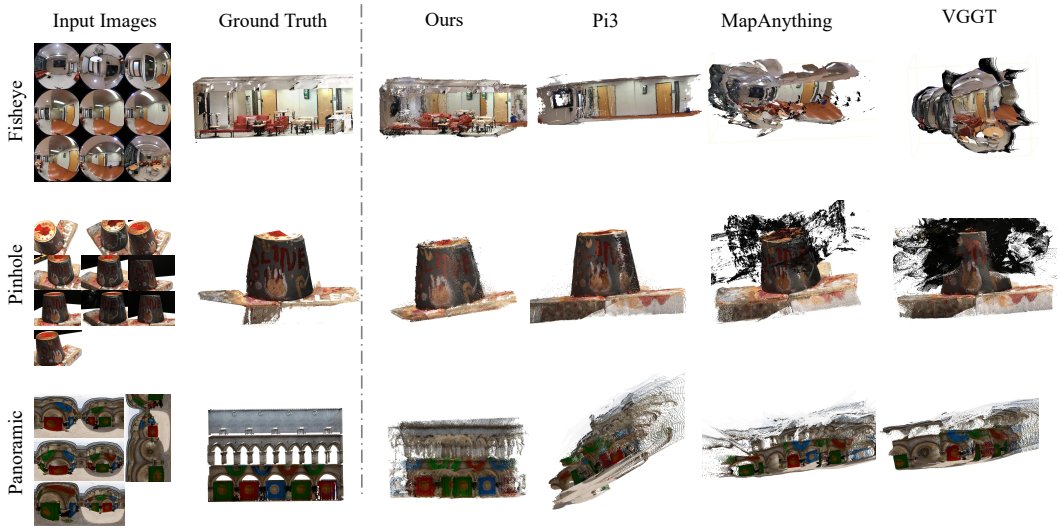

Figure 3: **Qualitative comparison of multi-view 3D reconstruction on zero-shot images with the same camera viewpoint.** On fisheye and panoramic inputs, existing feed-forward methods (Pi3, MapAnything, VGGT) fail to recover coherent geometry, whereas R0 produces complete reconstructions. R0 also remains clean on pinhole data.

from catastrophic structural collapse with heavy floaters. In contrast, our model reliably recovers the entire indoor layout with clear planar walls. Under canonical pinhole views, our approach effectively suppresses the dense background noise observed in MapAnything and VGGT while capturing crisper object boundaries than Pi3. Furthermore, on panoramic images spanning an ultra-wide field of view, competing methods either produce severely tilted scenes or fail entirely, while ours accurately preserves the vertical pillars and repetitive arches without geometric warping.

5 Conclusion

We presented R0, a unified 3D reconstruction model formulated in ray space. R0 unifies sampling, interaction, and output in a single pre-projection space, treating central cameras not as separate categories but as different ways of unfolding a common ray space. Experiments on pinhole, fisheye, panoramic, and heterogeneous multi-view data show that R0 substantially outperforms pinhole-pretrained baselines on non-pinhole imagery while remaining competitive on pinhole data. These results suggest that ray-space geometry offers a projection-agnostic representation that can serve as a shared foundation for cross-camera 3D reconstruction. Future work will extend this representation to dynamic scenes, modeling moving objects and scene flow in the same geometric space.

AI USAGE STATEMENT

We used large language models to assist with code generation. All generated content was reviewed, revised, and verified by the authors, who take full responsibility for the final code and the entire paper.

REPRODUCIBILITY STATEMENT

We provide the model architecture and module placement in Appendix A.1, the optimization schedule, supervision, data composition, and preprocessing in Appendix A.2, and the evaluation splits, deterministic sampling rules, alignment procedures, and aggregation protocols in Appendix A.3. Appendix A.4 further specifies the training budgets and controlled variants used in the ablation study. The experiments can be reproduced by applying these settings to the cited official dataset splits, constructing the fixed evaluation groups with the reported deterministic sampling rules, and evaluating them with the stated alignment and aggregation procedures.

REFERENCES

Hao Ai and Lin Wang. Elite360d: Towards efficient 360 depth estimation via semantic-and distance-aware bi-projection fusion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9926–9935. IEEE, 2024.

Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1534–1543, 2016.

Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan Frost, Luke Holland, Campbell Orme, et al. Scenescript: Reconstructing scenes with an autoregressive structured language model. In *European Conference on Computer Vision*, pp. 247–263. Springer, 2024.

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19640–19648. IEEE, 2023.

Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *European conference on computer vision*, pp. 611–625. Springer, 2012.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

LZ Chen, J Gao, Y Chen, KL Cheng, Y Sun, L Hu, N Xue, X Zhu, Y Shen, Y Yao, et al. Geometric context transformer for streaming 3d reconstruction. arxiv 2026. *arXiv preprint arXiv:2604.14141*.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2432–2443. IEEE, 2017.

Afshin Dehghan, Gilad Baruch, Zhuoyuan Chen, Yuri Feigin, Peter Fu, Thomas Gebauer, Daniel Kurz, Tal Dimry, Brandon Joffe, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *NeurIPS Datasets and Benchmarks*, 2(6):16, 2021.

Jiaxi Deng, Yushen Wang, Haitao Meng, Zuoxun Hou, Yi Chang, and Gang Chen. Omnistereo: Real-time omnidireactional depth estimation with multiview fisheye cameras. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1003–1012. IEEE, 2025.

Ruxiao Duan, Erin Hong, Dongxu Zhao, Eric Turner, Alex Wong, and Yunwen Zhou. Fisheye3r: Adapting unified 3d feed-forward foundation models to fisheye lenses. In *European Conference on Computer Vision*, pp. 1–19. Springer, 2026.

Michael Fonder and Marc Van Droogenbroeck. Mid-air: A multi-modal dataset for extremely low altitude drone flights. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 553–562. IEEE, 2019.

Girish Chandar Ganesan, Yuliang Guo, Liu Ren, and Xiaoming Liu. Unidac: universal metric depth estimation for any camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26953–26963, 2026.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361. IEEE, 2012.

Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, et al. A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*, 2020.

Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2485–2494, 2020.

Ulas Gunes, Matias Turkulainen, Xuqian Ren, Arno Solin, Juho Kannala, and Esa Rahtu. Fiord: A fisheye indoor-outdoor dataset with lidar ground truth for 3d scene reconstruction and benchmarking. In *Scandinavian Conference on Image Analysis*, pp. 3–17. Springer, 2025.

Yijing Guo, Mengjun Chao, Luo Wang, Tianyang Zhao, Haizhao Dai, Yingliang Zhang, Jingyi Yu, and Yujiao Shi. Panovggt: Feed-forward 3d reconstruction from panoramic imagery. *arXiv preprint arXiv:2603.17571*, 2026.

Namitha Guruprasad, Abhay Yadav, Cheng Peng, and Rama Chellappa. Cam3r: Camera-agnostic model for 3d reconstruction. *arXiv preprint arXiv:2603.22631*, 2026.

Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2821–2830. IEEE, 2018.

Shintaro Ito, Natsuki Takama, Toshiki Watanabe, Kent Selwyn The, Koichi Ito, Hwann-Tzong Chen, and Takafumi Aoki. Ob3d: A new dataset for benchmarking omnidirectional 3d reconstruction using blender. *IEEE Access*, 13:205646–205666, 2025.

Yuzhou Ji, Xiaotian Yang, and Zhipeng Zhang. Casamaestro: Multi-view panoramas for house-scale 3d reconstruction. *arXiv preprint arXiv:2606.31086*, 2026.

Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360 panorama depth estimation. *IEEE Robotics and Automation Letters*, 6(2):1519–1526, 2021.

Dongki Jung, Jaehoon Choi, Adil Qureshi, Somi Jeong, Dinesh Manocha, and Suyong Yeon. Wid3r: Wide field-of-view 3d reconstruction via camera model conditioning. *arXiv preprint arXiv:2602.05321*, 2026.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction; map-anything. github. io. In *2026 International Conference on 3D Vision (3DV)*, pp. 499–509. IEEE, 2026.

Jianqiang Li, Liumei Zhang, Wenjia Guo, Tianlong Feng, Yongzhi Liao, Di Lu, Hanchi Ren, and Jingjing Deng. Fastpano3d: Feed-forward indoor panoramic 3d reconstruction from a single image. *arXiv preprint arXiv:2606.30352*, 2026a.

Xi Li, Linyuan Li, Yan Wu, Tong Rao, Kai Zhang, Xinchen Hui, and Cihui Pan. Argus: Metric panoramic 3d reconstruction for indoor scenes. In *European Conference on Computer Vision*, pp. 628–647. Springer, 2026b.

Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3182–3192. IEEE, 2023.

Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2041–2050. IEEE, 2018.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169. IEEE, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=a68SUt6zFt`. Featured Certification.

Jing Ou, Zidong Cao, Yinrui Ren, Zhuoxiao Li, Jinjing Zhu, Tongyan Hua, Shuai Zhang, Hui Xiong, and Wufan Zhao. Holo360d: A large-scale real-world dataset with continuous trajectories for advancing panoramic 3d reconstruction and beyond. *arXiv preprint arXiv:2604.22482*, 2026.

Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguere, and Cyrill Stachniss. Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7855–7862. IEEE, 2019.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20076–20086. IEEE, 2023.

Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unik3d: Universal camera monocular 3d estimation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1028–1039. IEEE, 2025.

Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10881–10891. IEEE, 2021.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10892–10902. IEEE, 2021.

Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3260–3269, 2017.

You Shen, Zhipeng Zhang, Yansong Qu, Xiawu Zheng, Jiayi Ji, Shengchuan Zhang, and Liujuan Cao. Fastvggt: Fast visual geometry transformer. In *The Fourteenth International Conference on Learning Representations*, 2026.

Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: panorama transformer for indoor 360 depth estimation. In *European Conference on Computer Vision*, pp. 195–211. Springer, 2022.

Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2930–2937, 2013.

Meixi Song, Dizhe Zhang, Hao Ren, Ruiyang Zhang, Bo Du, Ming-Hsuan Yang, and Lu Qi. Unisharp: Universal sharp monocular view synthesis. *arXiv preprint arXiv:2606.07514*, 2026.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

Steffen Urban and Boris Jutzi. Lafida—a laserscanner multi-fisheye camera dataset. *Journal of Imaging*, 3(1):5, 2017.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. Vggt-■. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21486–21499, June 2026a.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *Advances in Neural Information Processing Systems*, 38:35928–35959, 2026b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709. IEEE, 2024.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4909–4916. IEEE, 2020.

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 C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 10481–10497, 2026c. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/11a09e0aaa74867c6b0719c639fc09f8-Paper-Conference.pdf.

Changhee Won, Jongbin Ryu, and Jongwoo Lim. Omnimvs: End-to-end learning for omnidirectional stereo matching. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8986–8995. IEEE, 2019.

Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22378–22389. IEEE, 2024.

Tao Xie, Peishan Yang, Yudong Jin, Yingfeng Cai, Wei Yin, Weiqiang Ren, Qian Zhang, Wei Hua, Sida Peng, Xiaoyang Guo, et al. Scal3r: Scalable test-time training for large-scale 3d reconstruction. *arXiv preprint arXiv:2604.08542*, 2026.

Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1787–1796. IEEE, 2020.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12–22. IEEE, 2023.

Shuai Yuan, Yantai Yang, Xiaotian Yang, Xupeng Zhang, Zhonghao Zhao, Lingming Zhang, and Zhipeng Zhang. Infinitevggt: Visual geometry grounded transformer for endless streams. *arXiv preprint arXiv:2601.02281*, 2026.

Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3712–3722, 2018.

Guoyang Zhao, Yuxuan Liu, Weiqing Qi, Fulong Ma, Ming Liu, and Jun Ma. Fisheyedepth: A real scale self-supervised depth estimation model for fisheye camera. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3780–3787. IEEE, 2025.

Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *European Conference on Computer Vision*, pp. 519–535. Springer, 2020.

Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19798–19808. IEEE, 2023.

Heng Zhou, Shuhong Liu, Yonghao He, Bohao Zhang, Fa Fu, Chenhui Hou, Xianbao Hou, Lijun Han, and Wei Sui. X-lens: Real-time metric depth estimation with heterogeneous cameras. *arXiv preprint arXiv:2607.12993*, 2026.

A Appendix

A.1 Architecture Details

The backbone follows Pi3 , except that we replace its three-channel local point-map head with a four-channel ray–radial head. The first three channels predict per-pixel ray directions, and the fourth predicts radial distance along each ray.

We add three modules. The Ray Field Predictor reads global and spatial patch tokens from layers $\{4, 11, 17, 23\}$ of the frozen encoder and contains three heads. Its spherical-harmonic head comprises a linear layer and two self-attention blocks and predicts the spherical domain together with 15 spherical-harmonic bases, each with a 3D coefficient. Two-layer MLP residual and support heads predict per-pixel tangent-plane residuals and the valid imaging region, respectively. Ray Canonical Sampling is placed between the frozen encoder and the main decoder and injects sampled features into the native image tokens. Geometry Attention is inserted into the final 18 decoder blocks in parallel with native content attention.

A.2 Training Details

We train on 31 data sources and derive fisheye data from a subset of them. Training has two stages and 25,000 optimizer updates in total. In Stage 1 (16,000 steps), the Ray Field Predictor and reconstruction model are trained jointly, using ground-truth rays for sampling and geometry attention. In Stage 2 (9,000 steps), the Predictor is frozen and the geometric input is gradually switched from ground truth to its predictions. The encoder remains frozen throughout. We use AdamW with cosine learning-rate schedules, peak learning rates of 5×10^{-5} for new modules and 5×10^{-6} for the decoder, 150–400 warmup updates, and gradient clipping at norm 1.0.

Predictor Supervision. The Predictor is supervised by two types of ground truth: per-pixel unit rays $\mathbf{r}^*$ and a projection support mask s^*. Its total loss is

$$\mathcal{L}_{\text{pred}} = \mathcal{L}_{\text{final ray}} + 0.25\,\mathcal{L}_{\text{anchor ray}} + 0.10\,\mathcal{L}_{\text{support}} + 0.001\,\mathcal{L}_{\text{residual}} + 0.05\,\mathcal{L}_{\text{no-regret}}.$$

The final-ray and anchor-ray losses use a periodic polar loss following UniK3D Piccinelli et al. (2025): polar-angle error with weight 3, azimuthal error wrapped to 2π, and both computed only inside the ground-truth support and the resize-valid region. The support loss is a per-pixel BCE against the projection support mask, with resize padding excluded. The residual term penalizes the squared tangent-plane angle, and the no-regret term penalizes the case where the final ray is worse than the SH anchor, using the geodesic angle $\arccos(\hat{\mathbf{r}}^\top \mathbf{r}^*)$.

The Predictor is trained in three sub-stages. In the anchor stage, only the SH anchor is trained with the anchor-ray loss. In the residual-auxiliary stage, the anchor is frozen and the residual and support heads are trained with the final-ray, support, residual, and no-regret losses. In the joint stage, the anchor, residual, and support are trained jointly with the full loss, with the anchor gradient scaled by 0.25. The 15 three-dimensional SH coefficients and the per-pixel residual have no direct labels; both are learned end-to-end through the anchor-ray and final-ray losses.

Images retain their original aspect ratios: the long side is resized to 518 pixels and the short side is adjusted to the nearest multiple of 14. We use a dataset-specific variable-view protocol, dynamically sampling the number of input views from a range defined for each dataset. The loss uses non-metric shared-gauge alignment: ground-truth point maps and pose translations share one scale, while predicted point maps and pose translations share another; radial distances are log-centered per sample. Training uses 64 GPUs and approximately 2,560 views per optimizer update. It runs at roughly $10\,\mathrm{s/update}$ and takes about five days.

Training datasets. We train on 31 data sources spanning indoor, outdoor, object-centric, driving, and dynamic scenes, with both real and synthetic data. We distinguish datasets for which we evaluate on held-out test data from those used entirely for training. For the former, we use only the official training splits of CO3Dv2 Reizenstein et al. (2021), ScanNet Dai et al. (2017), ARKitScenes Dehghan et al. (2021), DDAD Guizilini et al. (2020), Holo360D Ou et al. (2026), Matterport3D Chang et al. (2017), Stanford 2D-3D-S Armeni et al. (2016), Structured3D Zheng et al. (2020), ScanNet++ Yeshwanth et al. (2023), and OmniScene Zhou et al. (2026). The remaining sources are used

in full for training: BlendedMVS/BlendedMVG Yao et al. (2020), MegaDepth Li & Snavely (2018), HyperSim Roberts et al. (2021), MVS-Synth Huang et al. (2018), PointOdyssey Zheng et al. (2023), Virtual KITTI 2 Cabon et al. (2020), WildRGB-D Xia et al. (2024), Aria Digital Twin Pan et al. (2023), TartanAir Wang et al. (2020), MatrixCity Li et al. (2023), Taskonomy Zamir et al. (2018), Mid-Air Fonder & Van Droogenbroeck (2019), KITTI Raw Geiger et al. (2012), HM3D Ramakrishnan et al. (2021), DL3DV-10K Ling et al. (2024), Replica Straub et al. (2019), OmniRooms Song et al. (2026), OmniThings Won et al. (2019), OmniHouse Won et al. (2019), Aria Synthetic Environments Avetisyan et al. (2024), and A2D2 Geyer et al. (2020). HM3D Ramakrishnan et al. (2021) and Replica Straub et al. (2019) are additionally rendered with fisheye projection from their official scene assets.

A.3 Experimental Setup

We use *frame* for a single image, *group* for a set of views evaluated jointly in one multi-view trial, and *anchor frame* for the reference center frame within a group. For all reconstruction datasets, metrics are computed after a proper Sim(3) alignment followed by point-to-point ICP. The ICP correspondence threshold is 0.1 m for indoor scenes; for outdoor scenes, it is adjusted according to the physical scale of each scene. Pose and depth evaluations do not use ICP: pose uses its own trajectory alignment, while depth uses metric or median-scale alignment as specified by the corresponding metric. Dataset-specific aggregation below describes only how per-sequence or per-scene results are averaged.

Metric units. All values are rounded to three decimal places in the tables. Reconstruction distances are reported in meters for 7-Scenes, ScanNet, ETH3D, the non-pinhole benchmarks, and the ablation study. NC, AbsRel, δ_1, and PointRel are unitless. AUC is also unitless but is displayed as a percentage ($100 \times \mathrm{AUC}$) in the pose tables. RRE, RTE, and RPE-R are in degrees. Sintel ATE and RPE-T, as well as depth RMSE, are reported in meters. LaFiDa ATE also uses meters, whereas FIORD and Zip-NeRF ATE inherit their independently reconstructed COLMAP scene units and should not be interpreted as metric distances.

Reconstruction and depth. Following Pi3 Wang et al. (2026c), Sintel Butler et al. (2012) video depth is evaluated on 14 Final-pass training sequences comprising 643 frames. A single L1 scale and shift is fitted jointly over all valid pixels in each sequence, predictions are clipped at 70 m, and the dataset metrics are aggregated according to the number of valid pixels in each sequence. The same complete, stride-one sequences are used for camera pose evaluation with per-sequence Sim(3) trajectory alignment followed by an equal average over the 14 sequences. For both tasks, the scoring support is determined by $0 < d_{\mathrm{gt}} < 70$ and finite predictions, without confidence, predicted-validity, or edge masking. Bonn Palazzolo et al. (2019) video depth follows the Pi3 protocol on five sequences: `balloon2`, `crowd2`, `crowd3`, `person_tracking2`, and `synchronous`. We use frames 30–139 of each sequence, process all 110 frames jointly, and fit one L1 scale and shift over all valid pixels of the complete sequence. Depth PNGs are converted to meters by dividing by 5000; the valid range is $0 < d_{\mathrm{gt}} < 70$ m, and results are weighted by the number of valid pixels per sequence. On ScanNet Dai et al. (2017), we uniformly sample 20 frames along each scan in the official validation split and weight physical scenes equally. On Holo360D Ou et al. (2026), we sample three groups of eight native ERP frames from the beginning, middle, and end of every official test sequence; reconstruction and video-depth results are averaged over sequences and then over indoor and outdoor subsets. On Matterport3D Chang et al. (2017), monocular depth uses all valid panoramic frames, whereas multi-view evaluation selects three anchor frames per test building from the covisibility graph and eight frames around each anchor; buildings receive equal weight.

On Structured3D Zheng et al. (2020), we generate one 8-frame ERP group and one 10-frame pinhole group per official test scene, evaluate the two projections independently, and weight scenes equally; monocular depth is evaluated per frame with independent alignment. OmniScene Zhou et al. (2026) reconstruction uses complete synchronized 6-camera groups from the official test split; we average over groups, sequences, and scenes in this order. ETH3D Schops et al. (2017) follows the Pi3 high-resolution kf5 manifest over all 13 scenes. Predictions and ground truth use the same pixels satisfying depth $> 10^{-4}$, without confidence or predicted-validity masking; each scene is aligned by a proper correspondence-based Umeyama Sim(3), refined by point-to-point ICP with a 0.1 m threshold, and then weighted equally.

For OB3D Ito et al. (2025) depth, we uniformly sample 20 frames from each of the 12 non-egocentric test scenes, perform median-scale alignment per frame, and weight scenes equally. OB3D pose and reconstruction share three non-overlapping 8-frame groups per scene, giving 36 frozen groups. Pose errors are averaged within each scene, while reconstruction evaluates all points in each group; the 12 scenes are then weighted equally. On the 18 official 7-Scenes Shotton et al. (2013) test sequences, the sparse protocol takes every 200th frame (86 frames total), while the dense protocol takes every 40th frame (426 frames total). Both use RGB-registered depth, per-sequence alignment and scoring, and equal sequence weighting.

Camera pose. For CO3Dv2 Reizenstein et al. (2021), we uniformly sample 10 frames per sequence from the official test split, average first over sequences, and then weight categories equally. On the 50 public ScanNet++ Yeshwanth et al. (2023) validation scenes, we select three anchors per scene and 24 covisible frames per anchor for pose evaluation, with equal scene weighting; monocular depth is evaluated per frame with independent alignment. On Holo360D Ou et al. (2026), pose is evaluated on the same three groups of eight native ERP frames sampled from the beginning, middle, and end of each official test sequence; results are averaged over sequences and then over indoor and outdoor subsets. LaFiDa Urban & Jutzi (2017) uses six sequences with 32 groups each, for 192 groups and 1,152 raw fisheye images. Each group contains two timestamps and the complete three-camera rig (cam0/cam1/cam2), ordered first by time and then camera index; sequences receive equal weight. For FIORD Gunes et al. (2025), we use Kitchen_In, meetingroom, and parakennus, uniformly sample 10 continuous fisheye sequences per scene with a fixed deterministic pattern spanning 4–24 frames, and exhaustively pair all frames, yielding 30 sequences and 1,495 image pairs. Zip-NeRF Barron et al. (2023) pose evaluation uses 40 uniformly sampled continuous fisheye sequences following a fixed deterministic pattern spanning 29–44 frames. It exhaustively pairs all frames, yielding 26,009 image pairs. FIORD and Zip-NeRF pose results use a global average over all image-pair errors.

A.4 Ablation Details

Ablations are trained on a single server with eight RTX 4090 GPUs. The training sources are Holo360D, ScanNet, and OmniScene. We project 30% of the Holo360D training scenes into two-view fisheye images with a 220° FoV to form Holo360D-Fisheye; the four datasets are sampled at a 1:1:1:1 ratio. Pi3-FT uses a global batch size of 32 and is early-stopped at 17,000 steps. R0 w/o Samp. & Attn. replaces only the output head with the ray/radial representation, loads no Predictor, and is early-stopped at 9,000 steps. R0 w/o Samp. retains the Predictor and Geometry Attention; R0 w/o Attn. retains the Predictor and Ray Canonical Sampling; and R0 (all) enables all three. These three variants use a global batch size of 24 in both stages and converge by 15,000 steps.

A.5 Additional Results

Continuous-FoV ray prediction. Figure 4 compares three ray-field predictors over continuously sampled fields of view. All three predictors are independently trained with five random seeds using the same protocol. The training images are synthesized from the Holo360D training split with continuously sampled FoVs, covering pinhole, equidistant-fisheye, and ERP projections. All methods share the same frozen Pi3 image encoder, ray and support-mask supervision, and optimization budget. During both training and evaluation, the predictors receive only RGB images; projection family, FoV, camera intrinsics, dataset identity, and ground-truth rays are not provided as inputs.

Ours RFP is the Ray Field Predictor introduced in Section 3.2. *Fixed Classification* predicts one of the three projection families and assigns the corresponding frozen class-level ray-field prototype estimated exclusively from the training split. *Direct MLP* is a parameter-matched baseline that directly regresses per-pixel rays from frozen image features and pixel coordinates, without a spherical-harmonic anchor or tangent-plane residual.

Ours RFP keeps the angular error low across the entire FoV range, whereas the parameter-matched Direct MLP fails to model the structured spherical ray field. Fixed Classification performs well only near the FoVs represented by its class prototypes, with substantially larger errors elsewhere. These results suggest that the gain of Ours RFP is primarily due to its structured continuous ray-field representation, rather than model capacity or discrete projection recognition.

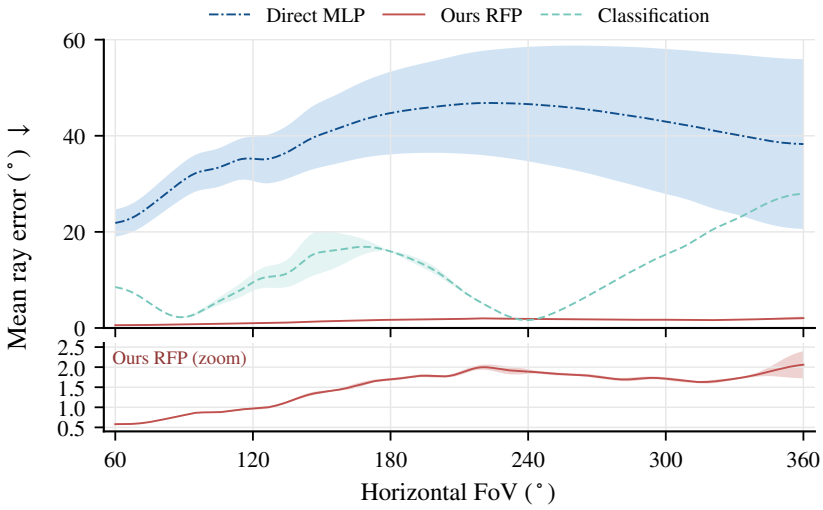

Figure 4: **Continuous-FoV ray-field prediction.** Curves and shaded regions show the mean and one standard deviation over five seeds. The lower panel magnifies the error range of Ours RFP; lower is better.

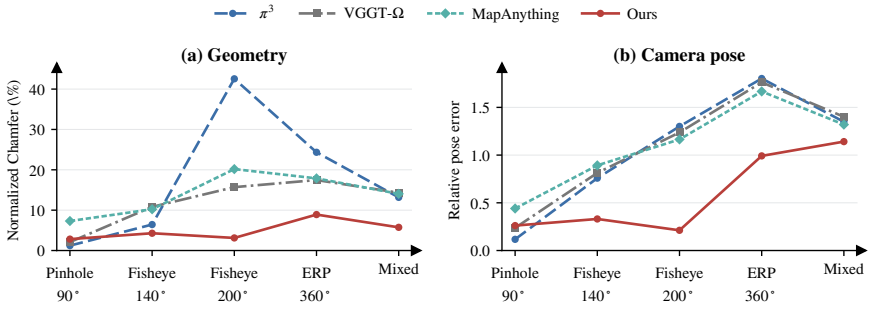

Figure 5: **Cross-projection consistency.** Geometry (a) and camera-pose (b) errors across projection types and FoVs. R0 remains substantially more stable as the input changes from pinhole to fisheye, panoramic, and mixed projections; lower is better.

Cross-projection consistency. Figure 5 evaluates geometry and camera-pose consistency as the input changes from standard pinhole projection to increasingly wide fisheye, panoramic, and mixed-projection settings.

Despite the FoV increasing from $140°$ to $200°$, R0 reduces its normalized Chamfer error from 4.28% to 3.12% and its pose error from 0.331 to 0.212. At $200°$, these errors are respectively 80.1% and 81.8% lower than the strongest baseline results of 15.70% and 1.164.

The advantage remains substantial for panoramic and heterogeneous inputs. On ERP-$360°$, R0 achieves geometry and pose errors of 8.91% and 0.992, improving over the strongest baseline results of 17.47% and 1.667 by 49.0% and 40.5%, respectively. On mixed-projection inputs, R0 obtains 5.75% and 1.141, corresponding to reductions of 56.4% and 13.5%.

Averaged across all five settings, R0 reduces the normalized Chamfer error from 12.08% for VGGT-Ω to 4.98%, a relative reduction of 58.8%. It similarly reduces the mean pose error from 1.066 for Pi3 to 0.587, an improvement of 44.9%. Although R0 is not the strongest method under the standard pinhole-$90°$ setting, its pronounced gains on fisheye, panoramic, and mixed-projection inputs demonstrate the benefit of unified modeling for wide-FoV distortion and heterogeneous projections.

A.6 LIMITATIONS

R0 assumes that every input is captured by a central camera, so all rays pass through a common optical center. This covers common pinhole, fisheye, and panoramic cameras, but cannot exactly represent non-central cameras whose optical center varies across pixels, such as some catadioptric or stitched multi-camera systems. R0 also targets static scenes and does not explicitly model dynamic objects or scene flow; substantial motion may therefore violate multi-view geometric consistency.