SGA3R: SPHERICAL GEOMETRY-AWARE 3D RECONSTRUCTION FROM PANORAMAS

Anonymous authors
Paper under double-blind review

ABSTRACT

Panoramic cameras capture the surrounding scene in a single exposure and eliminate errors caused by external multi-view registration. Nevertheless, a critical limitation remains when adapting perspective-pretrained 3D reconstruction models to panoramic imagery: conventional planar projection and positional encodings do not adequately reflect spherical geometry. We present SGA3R, a framework that incorporates spherical geometry into view projection and attention for 3D reconstruction from single or multiple panoramas. To represent each panorama with a compact set of informative perspective views, we formulate view projection as information-guided spherical-cap optimization under explicit coverage and overlap constraints. To encode spherical relationships within the pretrained backbone, we introduce Dual-Path Harmonic Encoding (DPHE) that integrates Additive Harmonic Encoding and Rotary Harmonic Encoding, where the latter lifts RoPE from SO(2) to SO(3) relative rotation between patch directions via irreducible subspaces of multiple orders. We further extend SGA3R to multiple panoramas using alternating attention across view, frame, and global domains, together with frame-conditioned pose refinement, while preserving the single-frame architecture. Experiments on indoor datasets Matterport3D, Stanford 2D-3D-S, as well as the outdoor dataset Mapillary Metropolis show that SGA3R achieves the lowest absolute relative depth error (AbsRel) among the evaluated panoramic methods, and the full model reduces AbsRel on Matterport3D by 11.9% relative to the strongest evaluated baseline, while using only six views per panorama and no cross-view layering. Multi-frame experiments on Holo360D further demonstrate improved performance in point cloud reconstruction. Code will be released.

1 INTRODUCTION

Panoramic cameras stitch multiple fisheye fields of view into a single image that covers every direction, with no multi-view registration or extrinsic calibration. They are a natural input for 3D reconstruction and are widely used in robotics and autonomous driving, where the demand on reconstruction accuracy and robustness keeps rising. However, their imaging convention differs fundamentally from the pinhole model, so perspective reconstruction methods do not transfer directly.

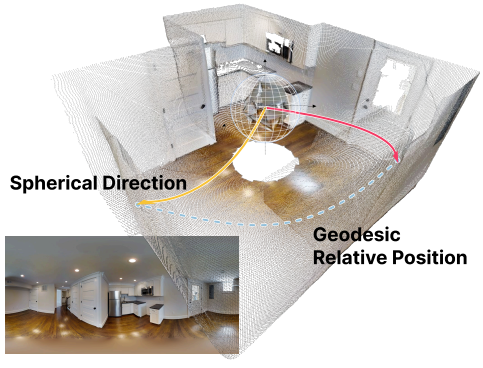

Figure 1: SGA3R is a panoramic 3D reconstruction framework that integrates spherical view optimization with spherical geometry embedding by associating each view and patch with spherical coordinates.

Recent feed-forward 3D geometry foundation models (Wang et al., 2025; 2026b) attain strong generalization and accuracy by training on vast data, but they learn from pinhole images whose planar-coordinate assumption is systematically mismatched to spherical geometry. Recent methods reuse such trunks along two routes. Route one reconstructs on the native ERP (Equirectangular Projection) representation. It keeps the uniform planar patch grid, or embeds spherical coordinates into the residual stream (Guo et al., 2026), or compensates polar distortion in loss space (Lin et al., 2026b). Such

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.

methods introduce spherical constraints only at the coordinate level and lack explicit geometric modelling; polar and equatorial patches also carry unequal information density, which limits local perception. Route two projects the panorama into sets of perspective views to suit the trunk. RPG360 (Jung et al., 2025) transfers through cubemap projection and scale alignment, and VGGT-360 (Yuan et al., 2026) selects views adaptively by per-view uncertainty. However, spherical geometry does not intrinsically enter attention to suit the panorama, which tends to produce depth artifacts due to low-texture views and cannot efficiently resolve cross-view layering.

Both mismatches share one source: a gap between the panoramic input and its encoding, where spherical geometry cannot run through both simultaneously. We therefore propose a unified principle: let spherical geometry guide the model at the two interfaces of view input and attention, and extend it to multiple frames, so that a trunk pretrained on perspective images transfers to panoramic reconstruction without relearning from scratch. Three gaps follow: how to express the whole sphere under a small yet semantically rich view budget; how to let pretrained features sense spherical geometry without destroying their representations; and how to extend single-frame reconstruction to multiple frames without breaking the single-frame prior.

SGA3R answers them in turn. At the view-projection interface, it models the panorama as a continuous smooth image field on the sphere and, through information-gradient-guided spherical-cap optimization under explicit constraints, covers the whole sphere with high information from few views. At the attention interface, it introduces Dual-Path Harmonic Encoding (DPHE), whose two paths act in synergy. The multiplicative Rotary Harmonic Encoding (RHE) lifts RoPE (Su et al., 2024) from $\mathrm{SO}(2)$ to $\mathrm{SO}(3)$ relative rotation between patch directions over irreducible subspaces of multiple orders, encoding the relative rotation geometry between views. The Additive Harmonic Encoding (AHE) injects spherical-harmonic features of each patch's absolute position into the residual stream, supplying global absolute geometry. For multiple frames, SGA3R builds alternating attention over view, frame, and global domains and, following V-DPM's (Sucar et al., 2026) frame-conditioned decoder with adaptive AdaLN modulation, refines frame-level poses.

SGA3R makes four contributions. First, we recast panoramic view projection as constrained coverage optimization on the sphere, solved under an adjustable lightweight view budget for spherical sampling with high information gain. Second, we propose Dual-Path Harmonic Encoding, whose multiplicative RHE and additive SHPE model relative spherical geometry and absolute spherical position respectively, adapting attention to spherical geometry. Third, we propose a plug-in cross-frame attention module that supports multi-frame reconstruction while preserving the single-frame architecture. Fourth, we build SGA3R, a panoramic reconstruction framework with spherical ray-consistent depth. It attains the lowest AbsRel on Matterport3D, Stanford 2D-3D-S and Mapillary Metropolis among the evaluated panoramic methods, and reduces the mean AbsRel on Matterport3D by 11.9% relative to the strongest baseline.

2 Related Work

Feed-forward 3D reconstruction foundations. Feed-forward models trained on large-scale perspective images show that a single network can carry a strong geometric representation. DUSt3R (Wang et al., 2024) recasts two-view reconstruction as point-map regression, MASt3R (Leroy et al., 2024) grounds matching in 3D, VGGT (Wang et al., 2025) predicts cameras, point maps and depth for hundreds of views in one pass, and π^3 (Wang et al., 2026b) removes VGGT's designated reference view. The Depth Anything line supplies monocular depth priors (Yang et al., 2024), and its iterative version generalizes the objective to arbitrary view settings (Lin et al., 2026a); reconstruction in dynamic scenes is likewise strong (Sucar et al., 2026; Zhang et al., 2026). All of them assume pinhole perspective input, which a panorama is not.

ERP-native panoramic reconstruction. A sphere cannot be faithfully mapped to the plane while mainstream vision architectures assume planar input, so the most direct choice is to feed the ERP image itself to the network. Traditional schemes build on sphere-native operators and adapt the sphere at the input or feature level (Cohen et al., 2018; Tateno et al., 2018; Facil et al., 2019; Zhang et al., 2022), yet they struggle to reuse a pretrained trunk. PanoVGGT (Guo et al., 2026) reuses a foundation model and encodes spherical coordinates in the residual stream, regressing depth, pose and point maps from a few panoramas. DAP (Lin et al., 2026b) and DA360 (Jiang et al., 2025) follow the same single-pass recipe for panoramic depth estimation; DA^2 (Li et al., 2026) proposes

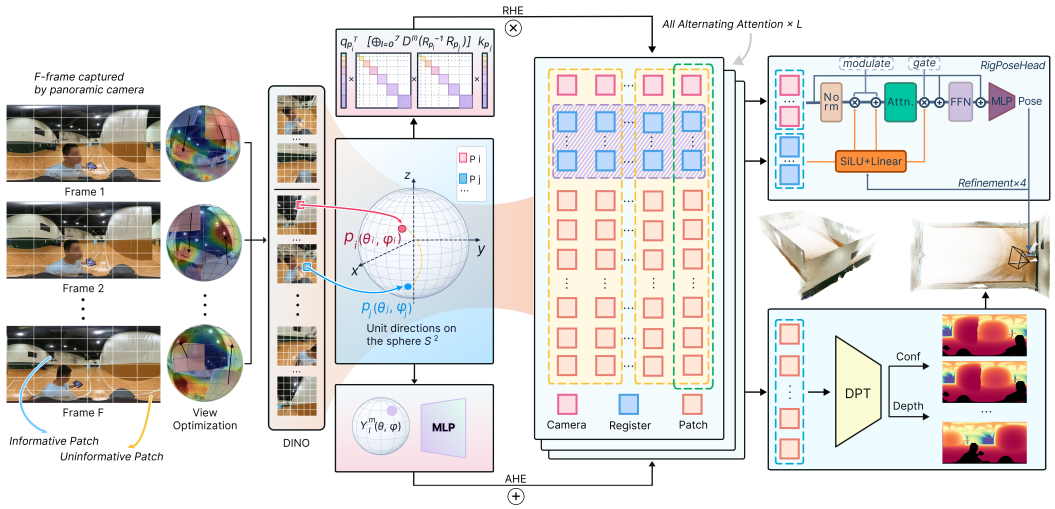

Figure 2: Overview of SGA3R. Each of panoramas is projected into V perspective views by adaptive spherical-cap view optimization (§3.2). Dual-Path Harmonic Encoding injects spherical geometry into the All Alternating Attention trunk. RigPoseHead decodes per-frame pose, and patch tokens are decoded by a DPT head into per-view depth and confidence, yielding a consistent point cloud.

SphereViT, letting image features query sine-cosine spherical embeddings through cross-attention. However, two limitations persist, a uniform planar patch grid that spends equal budget on unequal solid angles, and the scarcity of panoramic data.

Projection-based panoramic reconstruction. To avoid the planar distortion and geometric mismatch of ERP panoramas, another line projects the panorama into sets of perspective views. Some early methods focus on multi-projection feature fusion: BiFuse (Wang et al., 2020) fuses equirectangular and cubemap branches through a learnable mask, and Elite360D (Ai & Wang, 2024) takes ERP together with an undistorted icosahedral point set and lets pixel features cross-attend the global point set through dual attention. Recent methods reuse foundation-model trunks: RPG360 (Jung et al., 2025) reuses a perspective foundation model through cubemaps without training and aligns face scales by graph optimization, and UniK3D (Piccinelli et al., 2025) conditions on rays and spherical harmonics for arbitrary cameras. However, the six-view cubemap sampling of RPG360 and similar methods leaves views non-overlapping and without shared rays, and cannot optimize the sampling strategy by scene information. Methods such as Elite360D depend on many views, incurring high computational cost and low reconstruction efficiency, while shared-ray regions between views are prone to depth layering and inconsistent prediction.

3 Method

SGA3R is a spherical geometry-aware panoramic 3D reconstruction framework. It takes F panoramic frames, cuts each into V perspective views by adaptive spherical-cap view optimization, and models these views with an alternating-attention trunk as token sets on F spheres, producing per-frame depth and pose. This section follows three interfaces: the view-sampling interface (§3.2) optimizes the spherical-cap observation layout; the attention interface (§3.3.1) injects spherical geometry into the trunk; and the cross-frame interface (§3.3.2) extends single-frame reconstruction to multiple frames. Training objectives are given together in (§3.4). Figure 2 shows the pipeline.

3.1 Problem Formulation

A panoramic frame is an ERP image on the unit sphere $\mathbb{S}^2$. SGA3R takes F frames, each expressed by V perspective views, and learns a mapping

$$f : \{\mathcal{S}^t\}_{t=1}^{F} \longmapsto \{d^t, \xi^t\}_{t=1}^{F},$$

where $\mathcal{S}^t$ is the set of patch spherical coordinates of all views of frame t, d^t is the depth field, and $\xi^t \in \mathrm{SE}(3)$ is the pose. SGA3R project the sphere into V perspective images $\{\mathbf{X}^v\}$ as model input, defining the projection operator

$$\mathbf{X}^v = \mathrm{Proj}_{\mathbb{S}^2}(c_v, \phi_v), \quad c_v \in \mathbb{S}^2,\ \phi_v \in (0, \pi),$$

that is, a tangent-plane projection with centre direction c_v and field of view ϕ_v. Its intrinsics and extrinsics follow analytically from (c_v, ϕ_v) and need no calibration.

Modelling one sphere requires relating every patch back to it. Patch embedding turns each perspective view into an $\frac{H}{P} \times \frac{W}{P}$ token grid, and each patch centre maps to spherical coordinates (θ, φ) by back-projection and rotation, written $\Lambda_v : \mathbf{x} \mapsto (\theta, \varphi)$ (derivation in the appendix). Λ_v organizes the $F \times V$ perspective views onto F spheres and supplies a common geometric input to the adaptive spherical-cap view optimization of §3.2 and the Spherical Geometry-Aware Attention of §3.3.1.

3.2 Adaptive Spherical-Cap View Optimization

Covering a sphere with finitely many spherical caps is a classical problem: Gilbert (Gilbert, 1965) showed that for random caps of fixed angular radius, the uncovered area decays exponentially as the number of caps grows. Building on this, we treat coverage as a hard constraint and, under a given view budget, steer the view layout towards the high-information regions of the scene.

We represent the panorama as a continuous smooth image field on $\mathbb{S}^2$, from which a view with arbitrary intrinsics and extrinsics can be generated. Panoramic reconstruction needs both complete spherical coverage and avoidance of low-texture regions, where weak geometric supervision breeds depth artifacts. Given V, SGA3R seeks cap centres $c_v \in \mathbb{S}^2$ and fields of view ϕ_v that cover the sphere while concentrating on high-information regions. We state this as a constrained design objective:

$$\max_{\{c_v, \phi_v\}} \int_{\bigcup_v \mathrm{cap}_v} \rho(\lambda)\, d\Omega \quad \text{s.t.} \quad C \geq C_{\min},\ \eta \geq \eta_{\min},\ \phi_{\min} \leq \phi_v \leq \phi_{\max},$$

where C is the fraction of covered solid angle and η the minimum angular overlap between caps. $\eta_{\min}$ guarantees that neighbouring views share rays, the condition that supports cross-view geometric constraints. $\rho(\lambda)$ characterizes the density of spherical texture and structure information. Its operator implementation and ablation are in Appendix C.

SGA3R updates the layout by a heuristic ascent on the sphere: the update direction $\mathbf{d}_v$ fuses a normalized information attraction and a normalized coverage repulsion,

$$\mathbf{d}_v = \mathrm{norm}\big(\hat{\mathbf{g}}_v^{\mathrm{info}} + \hat{\mathbf{g}}_v^{\mathrm{cov}}\big), \qquad \mathbf{X}^v \leftarrow \mathrm{Proj}_{\mathbb{S}^2}\big(c_v + \alpha_c\, \mathbf{d}_v,\ \phi_v + \alpha_\phi \gamma_v\big),$$

where Proj retracts the cap centre onto $\mathbb{S}^2$, α_c sets the centre step, and $\alpha_\phi \gamma_v$ is the per-round increment of the field of view. The information attraction $\hat{\mathbf{g}}_v^{\mathrm{info}}$ pulls a view towards high information density within its own field of view, while the coverage repulsion $\hat{\mathbf{g}}_v^{\mathrm{cov}}$ suppresses excessive overlap between neighbouring caps. The two forces act together, concentrating views efficiently and covering the sphere evenly. Iteration stops at the first layout that satisfies all hard constraints, returning a feasible view-layout solution.

3.3 Spherical Geometry-Aware Attention

3.3.1 Dual-Path Harmonic Encoding (DPHE)

Rotary Harmonic Encoding (RHE). Axial RoPE encodes relative position only within the 2D grid of a single view. Existing spherical encodings mostly introduce geometric priors as attention bias, cross-attention, angular phase, or additive residual terms (Yun et al., 2023; Ren et al., 2025; Li et al., 2026; Guo et al., 2026). To adapt RoPE to the sphere, SGA3R lift it from a relative displacement to a relative group action. Planar RoPE uses the abelian group $\mathrm{SO}(2)$, applying $R(\theta_m)$ and $R(\theta_n)$ to Q/K so that attention depends on the relative action $R(\theta_n - \theta_m)$. Since $R^\top(\theta) = R(-\theta)$,

$$\langle R(\theta_m) q_m,\ R(\theta_n) k_n \rangle = q_m^\top R(\theta_n - \theta_m)\, k_n.$$

On the sphere, the planar coordinate is replaced by a direction $p \in \mathbb{S}^2$ and the acting group lifts to $\mathrm{SO}(3)$. Because the irreducible representations of $\mathrm{SO}(3)$ have dimension $2l+1$, one head decomposes exactly as $\bigoplus_{l=0}^{7} \mathcal{V}_l$ (with $\sum_{l=0}^{7}(2l+1) = 64$), encoding directional relations jointly across

several angular frequencies. Acting on tokens with the $\mathrm{SO}(3)$ orthogonal representation built from real Wigner-D matrices $D^{(l)}(R_p)$,

$$\langle D(R_p)q_p,\, D(R_{p'})k_{p'}\rangle = q_p^\top \Big(\bigoplus_{l=0}^{7} D^{(l)}\big(R_p^{-1}R_{p'}\big)\Big) k_{p'},$$

RHE is designed so that the attention score stays a function of relative rotation while the construction remains defined over the whole sphere. The score depends on the relative rotation $R_p^{-1}R_{p'}$ alone, which suits spherical continuity. The $l = 1$ block is the defining 3×3 rotation, whose $(3, 3)$ entry is the geodesic cosine $\cos\delta$, and cross-view attention is masked to pairs with $\cos\delta \geq \cos\tau$ ($\tau = 131°$), confining it to a geodesic neighbourhood. To fix the remaining gauge freedom, each R_p is anchored to the gravity axis e_z, so the relative rotation of a pair is uniquely determined and the score is invariant to rotations about e_z away from the two poles. The encoding thus inherits both the rotational structure of the ERP grid and its physical prior. To keep the construction defined at the poles, we fall back to a tangent-plane projection and truncate to $l \leq 1$, preserving the geodesic signal (see the appendix).

Additive Harmonic Encoding (AHE). RHE encodes pairwise geometry inside attention, whereas AHE supplies an absolute direction to the residual stream. We inject the real spherical harmonics of each patch direction through an MLP.

$$\mathbf{x}_i \leftarrow \mathbf{x}_i + \mathrm{MLP}\big(\big[Y_l^m(\Lambda_v(\mathbf{x}_i))\big]_{l\leq l_{\max}}\big).$$

At the poles only the $m = 0$ components remain, so the azimuthal singularity disappears automatically. RHE thus constrains relative geometry while AHE anchors absolute position.

3.3.2 All Alternating Attention (AAA)

Single-frame reconstruction builds on the VGGT-Ω trunk (Wang et al., 2026a), which encodes any number of perspective views with alternating intra-view/global, full/register attention. Since one panorama is expressed by V perspective views, F panoramas form $F \times V$ views that constitute topologically isomorphic view–frame–global domains, letting one self-attention exchange information across domains.

All Alternating Attention applies the same self-attention sequentially over three nested isomorphic domains (view $\subset$ frame $\subset$ global); the view domain is the patch grid of a single perspective view. The frame domain is the token set of the V views mapped onto one sphere by §3.1, sharing the spherical coordinate system and using the spherical positional encoding of §3.3.1; the global domain is the sequence of F frames, and global attention uses only AHE and sinusoidal encodings. The isomorphism of the three domains lets attention alternate over domains and token granularity (full/register), so information accumulates layer by layer along the nested structure. Figure 3 shows the three nested domains and the scope of attention at each level.

3.4 Spherical Decoding and Training

Spherical ray fusion. Multi-view predictions along a shared ray are naturally layered, and post-hoc scale alignment is costly without resolving them (Rey-Area et al., 2022; Jung et al., 2025). SGA3R moves confidence-weighted ray depth fusion (Merrell et al., 2007) into the forward pass. The depth and confidence of each view are scattered back onto the sphere, and observations falling on the same ray are averaged by confidence,

$$\hat{d}_{\mathrm{erp}}(c) = \frac{\sum_v \mathrm{conf}^v(c)\, d^v(c)}{\sum_v \mathrm{conf}^v(c)}, \qquad \mathrm{conf}^v(c) = 1 + \exp\big(\ell^v(c)\big),$$

where ℓ^v are the confidence logits of the DPT head, supervised together with the fusion. This removes layering along shared rays by construction and secures spherical depth consistency.

Pose decoding. To supply consistent poses for cross-frame inference, SGA3R follows the time-conditioned decoder of V-DPM (Sucar et al., 2026): the V register tokens of frame t are mapped by an MLP into a frame-condition vector, which is broadcast over the frame's tokens and injected into pose decoding through AdaLN; the vector replaces the scale and shift of LayerNorm and gates

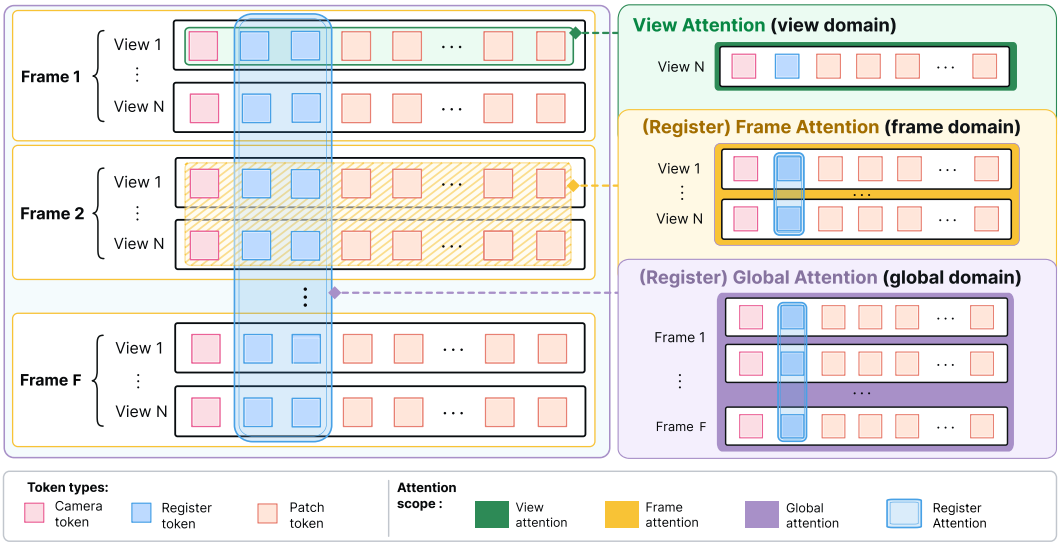

Figure 3: All Alternating Attention architecture: $F \times V$ perspective views form three nested isomorphic domains, view $\subset$ frame $\subset$ global.

the attention output in the DiT manner. Four refinement rounds produce the pose in the first frame's coordinate system.

Training objectives. The model is trained end-to-end with the total objective

$$\mathcal{L} = \mathcal{L}_{\text{erp}} + \lambda_{\text{depth}}\,\mathcal{L}_{\text{depth}} + \lambda_{\text{reproj}}\,\mathcal{L}_{\text{reproj}} + \mathbf{1}[F \geq 2]\,\big(\lambda_{\text{pose}}\,\mathcal{L}_{\text{pose}} + \lambda_{\text{cham}}\,\mathcal{L}_{\text{cham}}\big).$$

The core term $\mathcal{L}_{\text{erp}}$ computes regression, multi-scale gradient and surface-normal losses on the fused depth field, turning spherical consistency into supervision; the remaining terms assist with small weights: $\mathcal{L}_{\text{depth}}$ and $\mathcal{L}_{\text{reproj}}$ stabilize intra-view geometry and cross-view projection consistency, and for $F \geq 2$, $\mathcal{L}_{\text{pose}}$ and $\mathcal{L}_{\text{cham}}$ supervise the relative pose and align per-frame point clouds to a common frame.

Data augmentation. Motivated by the smoothness of the spherical projection and DPHE's need for directional diversity, we extend the VGGT-Ω augmentation by perturbing the view centres and fields of view of each panorama with Gaussian noise before re-projection, $\mathbf{X}^v = \text{Proj}_{\mathbb{S}^2}(\bar{c}_v + \epsilon_v,\ \phi_v + \beta_v)$ with $\epsilon_v \sim \mathcal{N}(0, \sigma_c^2)$ and $\beta_v \sim \mathcal{N}(0, \sigma_\phi^2)$; the perturbed centre is renormalized to $\mathbb{S}^2$ and the variants obey the coverage and overlap constraints of §3.2. This randomly augments the viewing directions under a fixed budget, so that DPHE sees richer view layouts.

4 Experiments

4.1 Setup

We evaluate the single-frame pipeline on Matterport3D (MP3D) (Chang et al., 2017), Stanford 2D-3D-S (2D3DS) (Armeni et al., 2017) and Mapillary Metropolis (Metropolis) (Mapillary, 2024). MP3D covers real indoor scenes under the official 61/11/18 building split. 2D3DS has six areas (Area 5 split into 5a and 5b), and we evaluate on a5a and a5b. Metropolis is an outdoor panoramic evaluation dataset. Holo360D (Ou et al., 2026) is an indoor-outdoor trajectory dataset used for multi-frame evaluation, from which we manually remove scenes with severe overexposure or motion blur. We only use its indoor sequences for training and evaluation. All methods are evaluated at their native resolution. Methods that take perspective views as input share six 512×512 views. π^3 and VGGT-Ω are evaluated view by view, and SGA3R is evaluated on the fused ERP grid after spherical ray fusion. ERP-native methods are evaluated at their own output resolution. Depths are placed on a common footing: metric depth at native scale; scale-invariant depth aligned by one global scale via panoramic IRLS (Cauchy kernel) (Guo et al., 2026); affine-invariant depth, also fitting a shift.

We adopt a two-stage training pipeline. In the single-frame stage, the model is initialized from pretrained VGGT-Ω (Wang et al., 2026a), with only the patch embedding and camera head frozen. We use AdamW with a weight decay of 0.05, four panoramas per GPU (24 views), gradient accumulation over two steps yielding an effective batch size of 8, bfloat16 precision, and gradient clipping at 1.0. In the multi-frame stage, the model is fine-tuned starting from the single-frame weights on indoor sequences on Holo360D. We sample six adaptive views per frame, where the sequence length $F \in \{2, 3\}$ and the sampling stride is uniformly sampled within the range $[0, 5]$. The total training time for both stages on four Nvidia RTX 5000 Pro GPUs is approximately 8 days.

4.2 View Optimization

To fix the view budget V, we use the pretrained VGGT-Ω weights, take $C_{\min} = 0.99$ and $\eta_{\min} = 5°$ as in §3.2, use patch histogram entropy (Moreau et al., 2026) as the information-density operator, and grow the field of view by $+2°$ per round. On 300 panoramas drawn at random from the validation split on MP3D, we compare fixed grids with adaptive projection at V=6/5/4 (Table 1).

Table 1: Effect of the view budget on reconstruction quality. fps is the inference speed for a single panorama under the given configuration, and init FOV is the starting field of view of the optimization. All other metrics are reported as means. Shading marks ■ best and ■ second best.

Configuration	init FOV	Coverage	AbsRel↓	RMSE↓	δ_1 ↑	fps
Fixed 6 views	115	0.9967	16.17	7.51	82.63	4.94
Fixed 5 views	130	1.0000	18.57	7.05	83.11	5.99
Fixed 4 views	140	0.9986	18.98	6.94	74.97	7.48
Adaptive 6 views	85	0.9921	14.03	8.32	87.27	1.73
Adaptive 5 views	105	0.9976	14.29	6.56	87.02	1.93
Adaptive 4 views	135	0.9974	18.33	6.85	75.43	3.63

Adaptive projection satisfies the constraints at all three budgets (lowest 0.9921) and leads fixed grid projection in reconstruction accuracy. V=6 is best across AbsRel and δ_1. Dropping to V=5 raises AbsRel by 1.9%, and V=4 degrades further. Therefore, we take V=6 to balance reconstruction quality and efficiency. The choice of information-density operator does not affect this conclusion; the comparison is in Appendix C.

4.3 Reconstruction

4.3.1 Single-Panorama

As shown in table 2, SGA3R has the lowest absolute relative depth error (AbsRel) on all three datasets and the highest mean δ_1 throughout. On Matterport3D, AbsRel is 11.9% below the runner-up, and δ_1 is 0.96 points higher. Notably, RPG360 does well on Metropolis but depends on graph optimization, taking 16 s to infer a single panorama, at high time complexity.

Table 2: Single-frame panoramic depth reconstruction, evaluated with affine-invariant depth (top), metric-depth (middle) and scale-invariant depth (bottom). Shading runs from dark to light for ■ best and ■ second best in each column within the scale-invariant depth protocol only. * marks six views from a fixed grid projection; † marks the same method re-run under scale alignment.

	Matterport3D						Stanford2D3D						Mapillary Metropolis					
	AbsRel↓		δ_1 ↑		RMSE↓		AbsRel↓		δ_1 ↑		RMSE↓		AbsRel↓		δ_1 ↑		RMSE↓	
Method	Mean	Med	Mean	Med	Mean	Med	Mean	Med	Mean	Med	Mean	Med	Mean	Med	Mean	Med	Mean	Med
PanDA	12.15	9.88	86.90	92.08	0.52	0.41	12.25	11.00	85.88	90.69	0.52	0.40	23.75	22.99	65.21	67.11	9.08	8.81
UniK3D	25.73	23.69	63.49	66.08	0.82	0.52	23.92	24.44	55.46	57.22	0.62	0.57	35.88	32.69	73.25	78.02	37.28	34.38
DAP	12.96	11.01	85.94	90.89	0.61	0.41	13.03	12.49	80.05	80.98	0.51	0.43	25.55	24.36	73.35	75.34	12.86	12.35
UniFuse	9.04	7.68	91.29	95.25	1.79	1.22	13.49	12.69	80.99	81.61	0.54	0.47	39.51	38.33	34.44	33.44	11.34	10.77
BiFuse	12.35	10.30	84.93	90.54	2.20	1.53	20.49	20.42	70.38	68.75	0.74	0.64	42.58	42.00	37.00	33.67	11.65	11.54
DA360	12.31	6.60	91.10	95.79	1.58	0.34	8.18	6.95	89.23	94.66	0.45	0.30	52.82	46.16	67.62	77.38	57.90	55.42
RPG360	13.86	12.14	83.18	87.92	2.17	1.83	11.64	9.26	85.80	88.65	0.52	0.46	16.89	16.52	77.18	80.06	7.06	6.69
PanoVGGT	4.70	4.16	96.97	97.83	1.18	1.01	4.19	3.58	97.26	98.39	0.33	0.25	17.52	15.98	75.47	80.57	8.23	7.59
DA^2	7.35	8.36	95.16	94.67	1.58	1.24	6.67	6.28	96.42	97.98	0.33	0.27	49.54	50.05	48.35	44.97	45.04	38.96
π^{3*}	9.12	7.90	88.58	90.25	1.19	0.93	9.30	7.84	86.08	86.50	0.30	0.26	32.92	30.37	60.17	60.19	20.72	20.09
VGGT-360	7.02	6.16	95.22	97.02	1.47	1.18	7.63	6.23	93.03	95.78	0.37	0.31	12.68	11.22	86.45	90.98	5.23	4.82
VGGT-Ω*	9.95	9.27	93.79	94.88	1.85	1.31	11.06	9.71	85.83	87.48	0.44	0.37	22.11	21.37	70.95	71.75	16.50	16.49
DAP†	9.63	7.89	90.62	93.42	2.18	1.39	10.77	10.44	84.79	88.89	0.46	0.37	24.77	24.29	69.55	75.00	11.75	11.65
PanDA†	30.42	29.15	48.53	49.24	3.62	3.31	44.82	44.68	30.55	29.50	12.50	12.04	24.55	24.03	63.35	66.09	8.88	8.73
UniK3D†	13.48	11.62	87.26	89.66	2.66	1.42	10.41	9.99	88.29	91.37	0.42	0.32	35.88	32.69	73.25	78.02	37.28	34.38
Ours	4.14	3.84	97.93	98.49	0.57	0.51	3.92	3.37	97.48	98.11	0.20	0.17	12.65	10.58	91.56	93.23	8.84	8.14

Qualitative results are in Figure 4. The two examples for PanoVGGT and DA^2 expose a shared weakness of ERP-native methods. In Fig. 1 the lamp beside the mirror is not reconstructed correctly, and in Fig. 2 the bed point cloud is broken. The first reflects insufficient local structural resolving capability under the uniform patch grid of the ERP-native paradigm, while the second shows that the paradigm cannot explicitly encode spherical continuity. The point cloud of VGGT-Ω shows layering across views, because each view predicts depth independently without cross-view reconciliation. SGA3R aggregates the per-view observations into one depth field by spherical ray fusion, preserving local structure in full while keeping depth consistent over the whole sphere.



Figure 4: Point cloud predictions on Matterport3D and Metropolis. VGGT-Ω back-projects the depth of six fixed-grid views with the prior extrinsics, whereas SGA3R reaches global consistency through spherical ray fusion. Red boxes mark baseline artifacts and green boxes the corresponding regions of our method.

4.3.2 Multi-Frame

Table 3 and the qualitative analysis in Figure 5 demonstrate that SGA3R yields improved and consistent point cloud and pose predictions. Pose is the inter-frame relative pose error in degrees, and AUC@30 follows VGGT's protocol: the area under the accuracy-threshold curve of the minimum of R and T at a 30° threshold. Figure 5 shows the difference qualitatively. The single-frame pipeline, run per frame with ICP alignment, leaves geometric structure drifting between frames, whereas cross-frame attention keeps geometry consistent.

Table 3: Multi-frame reconstruction on Holo360D. ■ marks the best value in each column over the test scenes. Overall denotes the average of Acc and Comp.

		Acc↓		Comp↓		Overall↓		AUC@30↑		T(°)↓		R(°)↓	
	Method	Mean	Med	Mean	Med	Mean	Med	Mean	Med	Mean	Med	Mean	Med
F=2	PanoVGGT	0.1747	0.1418	0.1525	0.1043	0.1636	0.1234	0.4990	0.5667	71.8003	74.5094	17.1238	13.9274
	Ours	0.1468	0.1027	0.0758	0.0480	0.1113	0.0747	0.6195	0.7000	56.0680	59.6594	13.0692	9.8829
F=3	PanoVGGT	0.1719	0.1394	0.1618	0.1077	0.1668	0.1207	0.4206	0.4444	72.1774	74.1281	21.8408	17.1852
	Ours	0.1638	0.1148	0.0728	0.0418	0.1183	0.0795	0.5802	0.5944	54.3613	55.1449	16.2073	12.3620

4.4 Ablations

Table 4 adds the single-frame trunk modules in turn, and every configuration is initialized from VGGT-Ω and trained for 15000 steps. Adding view optimization and spherical ray fusion switches supervision and evaluation from per-view depth to a sphere-consistent ERP depth field, which requires coordination along shared rays. Despite the larger freedom, AbsRel falls by 9.5% and δ_1 rises by 3.3 points, showing that the spherical-cap layout supports sphere-consistent supervision.

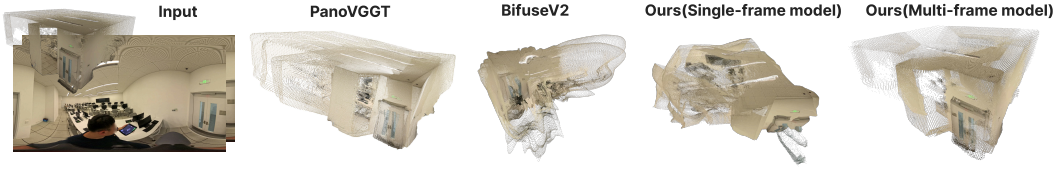

Figure 5: Point cloud results of multi-frame reconstruction (Holo360D). BiFuseV2, PanoVGGT and SGA3R(multi-frame model) obtains point clouds via depth-pose back-projection. SGA3R(single-frame model) reconstructs via ICP

Table 4: Design-path ablation on Metropolis. ✓ marks an enabled module; SAPE is the spherical positional embedding of PanoVGGT (Guo et al., 2026), and $l_{\max}$ applies to both AHE and RHE. Shading marks ■ best and ■ second best in each column within a task.

	Modules					AbsRel↓		δ_1 ↑		RMSE↓		fps
Configuration	VO+RF	SAPE	AHE	RHE	$l_{\max}$	Mean	Med	Mean	Med	Mean	Med	Mean
baseline					–	14.28	12.81	88.21	89.03	9.46	9.09	2.48
+VO+RF	✓				–	12.92	11.21	91.50	93.12	9.31	8.46	2.29
+SAPE	✓	✓			–	13.60	12.88	89.38	91.80	9.22	8.50	2.26
+AHE	✓		✓		7	18.28	17.47	72.37	74.84	6.71	6.55	2.15
+RHE	✓			✓	7	12.97	11.27	91.28	92.68	9.01	8.21	1.86
+AHE+RHE	✓		✓	✓	3	13.20	11.85	90.37	92.06	8.78	8.05	1.92
Full model	✓		✓	✓	7	12.75	11.01	91.39	92.78	8.85	8.17	1.72

For positional encoding, PanoVGGT's Spherical-aware Position Embedding (SAPE) raises AbsRel from 12.92 to 13.60 and AHE alone to 18.28, while RHE alone leaves it at 12.97. Their pairing gives the table-best 12.75/11.01, beating both, and within the paired rows raising the harmonic order from 3 to 7 accounts for the remaining gap. The ablation thus shows that the two paths are complementary and that angular frequency matters.

5 Conclusion

We present SGA3R, a panoramic 3D reconstruction framework that combines spherical view optimization with spherical geometry embedding. It equips a perspective-pretrained trunk with spherical geometry, so that the trunk reasons about a panoramic scene in spherical space and outputs a globally ray-consistent depth field. On Matterport3D, Stanford 2D-3D-S, Mapillary Metropolis, these designs attain the lowest AbsRel among the evaluated monocular panoramic methods while keeping point clouds and poses consistent across frames on Holo360D. Three limitations remain: the multi-frame pipeline is trained only on a few indoor sequences; geodesic sparse attention is realized only at the score level, without tensor-level memory optimization; and view optimization uses a conservative but adjustable budget, leaving more efficient layouts and faster solving open. We expect SGA3R to serve as an effective baseline for extending 3D foundation models to panoramas, to help make spherical geometric constraints a general prior for panoramic reconstruction, and to offer a new route towards unified 3D scene reconstruction.

AI Use Statement

We used generative artificial intelligence tools for tasks falling under mandatory disclosure in this work, including implementation of method code, assistance in theoretical derivations, and methodological feedback, particularly for examining conditions and boundary cases. Generative AI was employed to polish manuscript text and perform grammatical and phrasing checks. Generative AI tools were not used for the design of methods and experiments, formulation of paper claims and core proofs, or the drafting of the main text; these tasks were completed solely by the authors. Additionally, we used generative AI tools for tasks in the recommended disclosure category and preliminary

manuscript review, including editing the paper to improve readability, generating figures and LaTeX tables, and retrieving relevant literature. We have verified all AI-assisted outputs. All references and theoretical derivations were reviewed by the authors prior to inclusion. The authors take full responsibility for the final content of this work, including all text, claims, and artifacts produced with generative AI assistance.

Ethics Statement

This work uses four publicly available real-world architectural datasets: Matterport3D, Stanford 2D-3D-S, Mapillary Metropolis, and Holo360D. We do not collect any new imagery, and this study involves no human subjects. Metropolis consists of street-level imagery, while Holo360D contains captured indoor and outdoor scenes; both datasets may include passers-by, faces, and license plates captured incidentally. We use each dataset only in the form released by its respective distributors. We perform no re-identification or de-anonymization and do not redistribute any imagery. We do not release any data acquisition tools.

Reproducibility Statement

To ensure the full reproducibility of our work, we provide all necessary information in the main paper and supplementary materials. The source code of our proposed SGA3R, including model implementations, training configurations, and evaluation scripts, will be released publicly. Our experiments are conducted on the publicly available Matterport3D, Stanford 2D-3D-S, Mapillary Metropolis, and Holo360D datasets, and all data processing steps are detailed in the main text. The hyperparameters for Adaptive Spherical-Cap View Optimization and the information-density operators are provided in the appendix. Hyperparameters for Rotary Harmonic Encoding are also included in the appendix. These measures enable the research community to fully verify this work and build upon it for future research.

References

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

Hao Ai and Lin Wang. Elite360D: Towards efficient 360 depth estimation via semantic- and distance-aware bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9926–9935, 2024.

Iro Armeni, Sasha Sax, Amir R. Zamir, and Silvio Savarese. Joint 2D-3D-Semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

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. In *International Conference on 3D Vision (3DV)*, pp. 667–676, 2017. doi: 10.1109/3DV.2017.00081.

Taco S. Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical CNNs. In *International Conference on Learning Representations (ICLR)*, 2018.

Jose M. Facil, Benjamin Ummenhofer, Huizhong Zhou, Luis Montesano, Thomas Brox, and Javier Civera. CAM-Convs: Camera-aware multi-scale convolutions for single-view depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11818–11827, 2019. doi: 10.1109/CVPR.2019.01210.

E. N. Gilbert. The probability of covering a sphere with N circular caps. *Biometrika*, 52(3-4): 323–330, 1965. doi: 10.1093/biomet/52.3-4.323.

F. C. A. Groen, I. T. Young, and G. Ligthart. A comparison of different focus functions for use in autofocus algorithms. *Cytometry*, 6(2):81–91, 1985. doi: 10.1002/cyto.990060202.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36444–36453, 2026.

Hualie Jiang, Ziyang Song, Zhiqiang Lou, Rui Xu, and Minglang Tan. Depth anything in 360°: Towards scale invariance in the wild. *arXiv preprint arXiv:2512.22819*, 2025.

Dongki Jung, Jaehoon Choi, Yonghan Lee, and Dinesh Manocha. RPG360: Robust 360 depth estimation with perspective foundation models and graph optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, Main Conference, pp. 172213–172233, 2025. doi: 10.52202/085713-5730.

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3D with MASt3R. In *European Conference on Computer Vision (ECCV)*, pp. 71–91, 2024. doi: 10.1007/978-3-031-73220-1_5.

Haodong Li, Wangguandong Zheng, Jing He, Yuhao Liu, Xin Lin, Xin Yang, Ying-Cong Chen, and Chunchao Guo. DA^2: Depth anything in any direction. In *International Conference on Learning Representations (ICLR)*, 2026.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations*, volume 2026, pp. 141261–141285, 2026a.

Xin Lin, Meixi Song, Dizhe Zhang, Wenxuan Lu, Haodong Li, Bo Du, Ming-Hsuan Yang, Truong Nguyen, and Lu Qi. Depth any panoramas: A foundation model for panoramic depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26835–26844, 2026b.

Mapillary. Metropolis dataset, 2024. `https://www.mapillary.com/dataset/metropolis`, accessed 2025-11-01.

Paul Merrell, Amir Akbarzadeh, Liang Wang, Philippos Mordohai, Jan-Michael Frahm, Ruigang Yang, David Nister, and Marc Pollefeys. Real-time visibility-based fusion of depth maps. In *IEEE International Conference on Computer Vision (ICCV)*, 2007.

Arthur Moreau, Richard Shaw, Michal Nazarczuk, Jisu Shin, Thomas Tanay, Zhensong Zhang, Songcen Xu, and Eduardo Pérez-Pellitero. Off the grid: Detection of primitives for feed-forward 3D Gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11756–11766, 2026.

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.

José Luis Pech-Pacheco, Gabriel Cristóbal, Jesús Chamorro-Martínez, and Joaquín Fernández-Valdivia. Diatom autofocusing in brightfield microscopy: a comparative study. In *Proc. 15th International Conference on Pattern Recognition (ICPR)*, volume 3, pp. 314–317, 2000. doi: 10.1109/ICPR.2000.903548.

Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniK3D: Universal camera monocular 3D estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1028–1039, 2025. doi: 10.1109/CVPR52734.2025.00104.

Jiahui Ren, Mochu Xiang, Jiajun Zhu, and Yuchao Dai. PanoSplatt3R: Leveraging perspective pretraining for generalized unposed wide-baseline panorama reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28959–28969, 2025.

Manuel Rey-Area, Mingze Yuan, and Christian Richardt. 360MonoDepth: High-resolution 360° monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3752–3762, 2022. doi: 10.1109/CVPR52688.2022.00374.

Jianlin Su, Ahmed Murtadha, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Edgar Sucar, Eldar Insafutdinov, Zihang Lai, and Andrea Vedaldi. V-DPM: 4D video reconstruction with dynamic point maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14502–14511, 2026.

Keisuke Tateno, Nassir Navab, and Federico Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *European Conference on Computer Vision (ECCV)*, pp. 732–750, 2018. doi: 10.1007/978-3-030-01270-0_43.

Jay M. Tenenbaum. *Accommodation in Computer Vision*. PhD thesis, Stanford University, 1970. Stanford Computer Science Report STAN-CS-70-182; Stanford AI Project Memo AIM-134; DTIC AD748565.

Ch. Thum. Measurement of the entropy of an image with application to image focusing. *Optica Acta: International Journal of Optics*, 31(2):203–211, 1984. doi: 10.1080/713821475.

Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. BiFuse: Monocular 360 depth estimation via bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 459–468, 2020. doi: 10.1109/CVPR42600.2020.00054.

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)*, pp. 5294–5306, 2025. doi: 10.1109/CVPR52734.2025.00499.

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

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)*, pp. 20697–20709, 2024. doi: 10.1109/CVPR52733.2024.01956.

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

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jiayi Yuan, Haobo Jiang, De Wen Soh, and Na Zhao. Vggt-360: Geometry-consistent zero-shot panoramic depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Ilwi Yun, Chanyong Shin, Hyunku Lee, Hyuk-Jae Lee, and Chae Eun Rhee. EGformer: Equirectangular geometry-biased transformer for 360 depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6101–6112, 2023.

Chuhan Zhang, Guillaume Le Moing, Skanda Koppula, Ignacio Rocco, Liliane Momeni, Junyu Xie, Shuyang Sun, Rahul Sukthankar, Joëlle K. Barral, Raia Hadsell, Zoubin Ghahramani, Andrew Zisserman, Junlin Zhang, and Mehdi S. M. Sajjadi. Efficiently reconstructing dynamic scenes one D4RT at a time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7382–7392, 2026.

Jiaming Zhang, Kailun Yang, Chaoxiang Ma, Simon Reiss, Kunyu Peng, and Rainer Stiefelhagen. Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16896–16906, 2022. doi: 10.1109/CVPR52688.2022.01641.

A Patch Spherical Coordinate Derivation

Take a patch centre pixel $\mathbf{x} = (x_1, x_2)$ in view v. Back-projecting with the view intrinsics gives the viewing direction in the view camera frame,

$$\mathbf{n}_v(\mathbf{x}) = K_v^{-1} \begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix} \in \mathbb{R}^3,$$

and rotating it into the panorama frame by the view extrinsics R_v (the rotation determined by the view centre c_v; see §3.1) and normalizing yields the unit direction

$$\mathbf{r}_v(\mathbf{x}) = R_v \frac{\mathbf{n}_v(\mathbf{x})}{\|\mathbf{n}_v(\mathbf{x})\|} \in \mathbb{S}^2.$$

Its spherical coordinates are ($r_{v,i}$ is the i-th component of $\mathbf{r}_v(\mathbf{x})$)

$$\theta = \operatorname{atan2}\big(r_{v,2},\, r_{v,1}\big), \qquad \varphi = \arcsin(r_{v,3}),$$

which defines $\Lambda_v : \mathbf{x} \mapsto (\theta, \varphi)$.

B Rotary Harmonic Encoding Derivation

RoPE applies the $\mathrm{SO}(2)$ rotations $R(m\theta)$ and $R(n\theta)$ to the query at position m and the key at position n; by orthogonality $R(a)^\top R(b) = R(b-a)$, the inner product depends only on the relative displacement $n - m$. Lifting the same idea to the sphere replaces the position by a direction $p \in \mathbb{S}^2$ and the group $\mathrm{SO}(2)$ by $\mathrm{SO}(3)$. The irreducible representations of $\mathrm{SO}(3)$ are labelled by the angular-momentum quantum number l and have dimension $2l + 1$; the action of a group element R on the space of degree-l spherical harmonics is given by the Wigner-D matrix $D^{(l)}(R)$, and the space of spherical functions decomposes as the direct sum $\bigoplus_l (2l+1)$. Splitting the attention head dimension (64 here) along these subspaces and taking $l = 0, \ldots, 7$,

$$\sum_{l=0}^{7} (2l + 1) = 64,$$

fills the head dimension exactly, with no waste. For each patch, the spherical direction p together with the gravity axis e_z builds a rotation $R_p \in \mathrm{SO}(3)$: the third column is p, the first is $\|e_z \times p\|^{-1}(e_z \times p)$, and the second completes the right-handed frame. Roll is thereby locked to e_z, and $R_p e_z = p$. Applying it block-diagonally to query and key,

$$q'_p = \Big(\bigoplus_{l=0}^{7} D^{(l)}(R_p) \Big)\, q_p, \qquad k'_{p'} = \Big(\bigoplus_{l=0}^{7} D^{(l)}(R_{p'}) \Big)\, k_{p'}.$$

As an orthogonal representation of $\mathrm{SO}(3)$, the Wigner-D matrices guarantee that the attention score

$$\langle q'_p,\, k'_{p'} \rangle = q_p^\top \Big(\bigoplus_{l=0}^{7} D^{(l)}\big(R_p^{-1} R_{p'}\big) \Big)\, k_{p'}$$

depends only on the relative rotation $R_p^{-1} R_{p'}$ between the two directions. In particular, at $l = 1$ the Wigner-D matrix is the defining 3×3 rotation matrix, whose $(3,3)$ entry is $e_z^\top R_p^\top R_{p'} e_z = (R_p e_z) \cdot (R_{p'} e_z) = p \cdot p' = \cos\delta$, where δ is the geodesic distance between the two directions.

Symmetry boundary. Because R_p is built from e_z as well as p, a global rotation S that fixes e_z gives $R_{Sp} = S R_p$, hence $R_{Sp}^{-1} R_{Sp'} = R_p^{-1} R_{p'}$ and the attention score is exactly invariant away from the two poles.

Pole guard and high-latitude truncation. The construction of R_p takes $\hat{u} = \widehat{e_z \times p}$ as its first column, and at the poles $p = \pm e_z$ the cross product $e_z \times p$ is the zero vector, leaving $\hat{u}$ undefined. This degeneracy has the same origin as the polar singularity of the ERP representation itself: meridians converge at the poles and azimuth is undefined there, so the degenerate boundary of the encoding coincides with that of the representation. In practice a tangent-plane fallback replaces the first column when $\|e_z \times p\| < \varepsilon = 10^{-6}$: we take $\hat{u} = \mathrm{norm}\big(e_x - (e_x^\top p)\, p\big)$ and fall back to the same projection of e_y if $p \parallel e_x$, so that R_p is everywhere a proper rotation. The construction runs in double precision and returns single precision. The second is high-latitude truncation. Azimuth is undefined in the ERP representation near the poles anyway. We use $\varphi_{\mathrm{pole}} = 85^\circ$ throughout. For tokens near the poles we therefore set the higher-order blocks to zero, keeping the $l \le 1$ subrepresentation. Because $\bigoplus_{l \le 1} \mathcal{V}_l$ is a subrepresentation, the projection commutes with D, so the score remains a function of the relative rotation $R_p^{-1} R_{p'}$ alone. The cost is a loss of azimuthal resolution near the poles; we do not claim this is resolved.

C View Optimization Details

Ascent direction. The information attraction is the density-weighted centroid of the patch directions inside the field of view,

$$\hat{\mathbf{g}}_v^{\mathrm{info}} = \mathrm{norm}\Big(\sum_{p \in \text{view } v} \rho(p)\, p \Big),$$

pulling the view towards the denser information within its own field of view. The coverage repulsion is an inverse-square repulsion between neighbouring caps,

$$\hat{\mathbf{g}}_v^{\mathrm{cov}} = \mathrm{norm}\Big(\sum_{v' \neq v} \frac{c_v - c_{v'}}{\|c_v - c_{v'}\|^3}\, \mathbf{1}\big[\|c_v - c_{v'}\| < r_0\big] \Big),$$

where $\|c_v - c_{v'}\|$ is the spherical chord distance and r_0 a cutoff radius, taken as the current field of view $r_0 = \phi_v$: distant views exert no force on each other and only crowded caps push apart. The two forces are combined with equal weights, $w_c{:}w_i = 10{:}10$.

Retraction and guards. After the centre update the centre is renormalized onto the unit sphere,

$$c_v \leftarrow \big(c_v + \alpha_c\, \mathbf{d}_v\big) / \|c_v + \alpha_c\, \mathbf{d}_v\|,$$

a projection-style retraction on the spherical manifold (Absil et al., 2008); the centre elevation is additionally clamped to $\pm 85^\circ$ so that no view centre sits at a pole. The field of view grows by γ each round from its initial value up to a budget-dependent upper limit; iteration stops at the first point where both $C \ge C_{\min}$ and $\eta \ge \eta_{\min}$ hold. Coverage C is computed as the solid angle of the union of caps: the fraction of Fibonacci quasi-uniform sample points falling inside any cap ($p^\top c_v \ge \cos(\phi_v/2)$), on the same footing as the coverage in Table 5. The angular overlap η is the thinnest *barrel* overlap between cap pairs: writing $\theta_{ij} = \arccos(c_i^\top c_j)$ for the great-circle distance between two cap centres, the overlap of caps i and j is $\big[\frac{1}{2}(\phi_i + \phi_j) - \theta_{ij}\big]^+$, and η is the minimum of this quantity over pairs with positive overlap. $\eta_{\min}$ therefore bounds the thinnest shared-ray region between neighbouring views: too thin an overlap leaves too few shared rays for the cross-view matching that the geometric constraints rely on.

Procedure. The V cap centres are initialised on a Fibonacci lattice, with initial fields of view that depend on the view budget, and each round applies the update above with centre step $\alpha_c = 0.12$, increasing every field of view by a constant increment up to a budget-dependent upper limit. Coverage and overlap are evaluated on $S = 2 \times 10^5$ Fibonacci sample points, and iteration stops at the first round in which both constraints hold. In practice the number of rounds is capped at $N = 5$ or 10 according to the time budget. The optimization runs per panorama at inference time rather than being learned once: because the layout depends on the scene's information density, each panorama receives its own layout, which is why adaptive projection costs more wall-clock time than a fixed grid (Table 1).

Information-density operator. Under the earlier protocol, with the field of view growing by $+5^\circ$ per step and $C_{\min} = 0.99$, we compared three operators at three budgets on MP3D single-building panoramas (Table 5). FOV is the mean / max field of view needed to satisfy both constraints, time is

the wall clock to first feasibility, and information gain is the density improvement over a geometric initialization at the same field of view.

The three operators are drawn from the gradient family (Tenengrad (Tenenbaum, 1970)), the entropy family (patch histogram entropy (Thum, 1984; Moreau et al., 2026)) and the Laplacian family (patch-level neighbourhood dissimilarity Groen et al., 1985; Pech-Pacheco et al., 2000).

Table 5: Effect of view budget $\times$ operator on constraint satisfaction (MP3D).

V	Operator	FOV (°)	Time (s)	Information gain (%)	Coverage
6	laplacian	105.0 / 105.0	0.13	$+3.73$	0.992 / 0.991
	shannon_hist	105.0 / 105.0	0.15	$+1.80$	0.994 / 0.993
	tenengrad	105.1 / 110.0	0.14	$+4.70$	0.992 / 0.990
5	laplacian	130.7 / 135.0	0.26	-1.34	0.998 / 0.990
	shannon_hist	129.9 / 135.0	0.29	$+0.25$	0.999 / 0.992
	tenengrad	130.5 / 135.0	0.27	$+1.16$	0.997 / 0.990
4	laplacian	135.0 / 135.0	0.24	-0.78	0.992 / 0.990
	shannon_hist	135.0 / 135.0	0.25	-0.12	0.993 / 0.991
	tenengrad	135.3 / 140.0	0.25	-0.94	0.992 / 0.990

Fewer views require a wider field of view to satisfy the constraints ($V{=}6/5/4$ need about $105°/130°/135°$) and yield lower information gain, which turns negative across the board at $V{=}4$; optimization takes under 0.3 s in all cases, and $V{=}6$ is fastest (0.13–0.15 s). At $V{=}6$ the three operators satisfy the constraints at the same field of view and differ only in gain, so the operator affects the strength of the information guidance, not the feasibility of the layout.