PANOREG: FROM PIXEL MATCHING TO VIEW REASONING FOR ROBUST PANORAMIC REGISTRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Conventional panoramic registration typically follows a *match-then-RANSAC* paradigm, establishing pixel-level association followed by RANSAC-style pose estimation. However, under challenging conditions such as low overlap and visual ambiguity, high-quality correspondences often become scarce, leading to limited robustness. This paper proposes **PanoREG**, a novel *viewpose-then-Bayesian* paradigm for robust, training-free panoramic registration. Instead of relying on pixel-level matching, PanoREG recovers the relative pose between two panoramas by associating and reasoning over pose estimates from their decomposed sub-views. Specifically, PanoREG comprises two core steps: *calibrated view-pose association* and *robust Bayesian pose estimation*. The calibrated view-pose association step first constructs multiple panorama-level pose hypotheses from sub-view pose predictions. We decompose each panorama into calibrated perspective views and employ a 3D foundation model to estimate per-view poses. We then associate cross-panorama view-pose pairs through the predefined view-to-panorama extrinsics, yielding multiple noisy hypotheses of the panorama-level relative pose. Given these hypotheses, the robust Bayesian pose estimation step defines an outlier-robust Cauchy likelihood over them and derives the posterior distribution of the panorama pose via Bayesian inference. This enables robust maximum a posteriori (MAP) pose estimation by suppressing highly inconsistent hypotheses. Additionally, to improve pose estimation in challenging cases, we introduce *confidence-controlled adaptive verification*, where the confidence of the MAP pose estimate is quantified using posterior uncertainty. Reliable MAP estimates are directly accepted, whereas low-confidence cases trigger additional point-level geometric refinement. Extensive experiments on challenging panoramic benchmarks demonstrate that PanoREG substantially outperforms existing state-of-the-art methods, particularly under low-overlap conditions.

1 INTRODUCTION

Panoramic images, with the omnidirectional field of view and rich scene context, have become an increasingly vital sensing modality in applications such as large-scale visual localization (Caruso et al., 2015; Kim et al., 2024), autonomous mapping (Kim et al., 2024), and panoramic 3D reconstruction (Guo et al., 2026b). A fundamental problem underpinning these downstream tasks is panoramic relative pose estimation, which seeks to recover the rigid camera transformation between two panoramic frames. Despite the comprehensive visual coverage offered by omnidirectional sensing, reliable registration remains challenging in practice, where large viewpoint changes, spherical distortion, visual ambiguities, and limited scene overlap severely hinder accurate pose estimation.

To tackle this challenge, existing panoramic registration approaches predominantly adhere to a *match-then-RANSAC* paradigm (Cruz-Mota et al., 2012; Zhao et al., 2015; Murrugarra-Llerena et al., 2022; Gava et al., 2023; Jung et al., 2025). These methods typically extract local keypoints or dense descriptors across panoramic images, project them onto tangent planes or spherical grids to alleviate distortion, and subsequently estimate the relative pose via RANSAC-style robust solver. While effective in texture-rich scenes with substantial overlap, this paradigm fundamentally relies on the availability of high-quality correspondences. Under challenging conditions, such as wide camera baselines, occlusions, or large rotations, the co-visible region shrinks substantially. In such low-overlap regimes, correspondence evidence becomes sparse, spatially clustered, or corrupted by vi-

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.

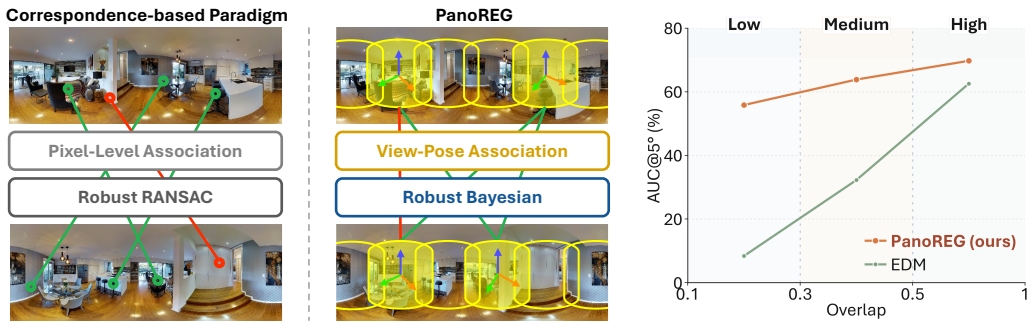


Figure 1: Conventional methods (*e.g.*, EDM (Jung et al., 2025)) follow a *match-then-RANSAC* paradigm based on pixel-level association, which collapses when visual overlap is limited (right plot). By contrast, **PanoREG** introduces a novel *viewpose-then-Bayesian* paradigm that couples calibrated view-pose association with robust Bayesian pose estimation, achieving remarkable accuracy and robustness even under extreme low-overlap conditions.

sual ambiguities, causing conventional geometric solvers to degenerate or fail catastrophically (as highlighted in Figure 1). More importantly, treating fragile pixel-level matches as the primary source of pose evidence overlooks the inherent geometric structure of panoramic imaging: *a panorama can be analytically decomposed into multiple calibrated perspective sub-views whose relative orientations and shared optical center are known exactly by construction.*

This geometric structure presents a natural opportunity: instead of relying on brittle pixel matches, can we formulate panoramic registration as reasoning over view-level pose estimates? This perspective is particularly compelling with the emergence of perspective 3D foundation models (Wang et al., 2025; 2026), which demonstrate remarkable zero-shot spatial reasoning on pinhole images. However, effectively transferring these perspective priors to equirectangular panoramas remains nontrivial. Existing approaches typically rely on task-specific retraining or architectural adaptation (Ren et al., 2025; Guo et al., 2026b), which incurs substantial computational overhead and may compromise out-of-domain generalizability. Conversely, simply decomposing panoramas into perspective crops and aggregating per-view predictions from a pretrained 3D foundation model introduces different challenges: local occlusions, textureless regions, and perspective cropping artifacts inevitably yield erroneous, conflicting view poses. As a result, naive aggregation schemes (*e.g.*, uniform averaging) are highly susceptible to these pose outliers, leading to severe estimation errors.

To address these challenges, this paper presents **PanoREG**, a novel training-free framework that shifts the panoramic registration paradigm from conventional *match-then-RANSAC* paradigm to the *viewpose-then-Bayesian* paradigm. Specifically, PanoREG operates through two tightly coupled steps: *calibrated view-pose association* and *robust Bayesian pose estimation*. In the first step, each panorama is decomposed into calibrated perspective crops at uniformly spaced azimuths. By jointly feeding the resulting multi-view set into a frozen 3D foundation model within a unified coordinate system, each cross-panorama view-pose pair analytically induces a candidate hypothesis for the global panoramic transformation through the known view-to-panorama extrinsics. Crucially, since perspective views within the same panorama share a common optical center and fixed angular offsets, we introduce an intra-panorama consistency metric that analytically quantifies the *geometric reliability* of each pose hypothesis in a fully training-free manner.

In the second step, we consolidate these multi-view hypotheses through a reliability-aware Bayesian framework. To suppress the potential view-pose outliers, we define the pose likelihood using heavy-tailed Cauchy distributions over $SO(3)$ geodesic and S^2 angular residuals. Unlike standard Gaussian objectives that penalize deviations quadratically, the Cauchy likelihood induces a logarithmic penalty, effectively mitigating the influence of large discrepancies. The optimal panoramic pose is then recovered via Maximum A Posteriori (MAP) estimation, weighted by the pose hypothesis reliability priors. Beyond pose estimation, we further apply a Laplace approximation around the MAP solution to estimate local posterior uncertainty and quantify pose confidence. This confidence score governs an *uncertainty-guided adaptive verification* mechanism: reliable estimates are accepted directly, while uncertain cases selectively trigger point-level geometric verification using 3D pointmaps predicted by the foundation model. To summarize, our main contributions are as follows:

- We propose **PanoREG**, a training-free framework that shifts panoramic relative pose estimation from the correspondence-dependent *match-then-RANSAC* paradigm to a novel *viewpose-then-Bayesian* reasoning scheme, without requiring task-specific retraining.
- We introduce a *calibrated view-pose association* strategy that shifts pose hypothesis construction from conventional pixel-level association to view-pose association, leveraging known view-to-panorama extrinsics and pretrained 3D foundation models. We further exploit the inherent intra-panorama geometry to analytically assess the reliability of these hypotheses.
- We formulate a reliability-aware *robust Bayesian pose estimator* together with *uncertainty-guided adaptive verification*, using heavy-tailed Cauchy likelihoods over rotation and translation-direction to suppress outliers, while leveraging posterior uncertainty to selectively trigger point-level refinement for low-confidence cases.
- Extensive evaluations on indoor/outdoor benchmarks and downstream visual odometry demonstrate PanoREG's clear superiority, particularly under extreme low overlap (0.1–0.3), where it improves AUC@5 on Stanford2D3D from 9.13% to 56.61% without any panoramic fine-tuning.

2 Related Work

Panoramic Relative Pose Estimation. Panoramic relative pose estimation has conventionally been cast as a correspondence-driven pipeline. To mitigate spherical distortions, early works devised sphere-aware features (Hansen et al., 2007; Cruz-Mota et al., 2012; Arican & Frossard, 2012; Zhao et al., 2015; Guan & Smith, 2017; Gava et al., 2013) or projected panoramas onto tangent planes for perspective matching (Eder et al., 2020; Jiang et al., 2023; Murrugarra-Llerena et al., 2022). Recent advances introduce learned spherical matching, either operating directly on the sphere (Gava et al., 2023) or tailoring dense correspondences for equirectangular projection (Jung et al., 2025). Nonetheless, these methods fundamentally bottleneck at RANSAC-based verification, which easily degrades under low visual overlap. Alternatively, learning-based methods bypass traditional matching by leveraging structural layout priors (Hutchcroft et al., 2022; Su et al., 2023) or directly regressing camera motion (Tu et al., 2024). However, their reliance on panorama-specific supervision severely compromises cross-domain generalization. In contrast, our framework is entirely *correspondence-free* and *training-free*. By reasoning over coupled pose hypotheses derived from intrinsic sub-view geometry, it enables robust zero-shot generalization to unseen scenes without panorama-specific supervision.

Foundation Models for Panoramic Vision. Recent 3D foundation models, such as DUSt3R (Wang et al., 2024), VGGT (Wang et al., 2025), and π^3 (Wang et al., 2026), exhibit remarkable visual geometry reasoning by directly inferring point maps and camera poses from uncalibrated perspective views. However, their perspective-centric pretraining struggles with equirectangular panoramas due to severe spherical distortion. To bridge this domain gap, existing works typically resort to panorama-specific adaptation or retraining. For instance, Pano3R (Song et al., 2025) combines correspondence-guided perspective pairing with DUSt3R-based global alignment, PanoSplatt3R (Ren et al., 2025) adds spherical adaptation for 3DGS, and PanoVGGT (Guo et al., 2026b) retrains VGGT omnidirectionally. While effective, these strategies incur heavy retraining overhead or risk degrading the models' generic open-world priors. In contrast, PanoREG keeps the 3D foundation model fully *frozen*, resolving the spherical gap geometrically at inference time via rigid sub-view constraints and robust probabilistic pose estimation.

3 Our Method

3.1 Calibrated View-Pose Association

Conventional panoramic registration establishes *pixel-level associations* between two panoramas and recovers their relative pose through geometric verification. However, under challenging conditions, reliable correspondences often become scarce, leading to limited robustness. We instead seek to establish associations at the *view level*: each perspective view provides a view-level pose estimate, and different cross-panorama view pairs induce multiple estimates of the same panorama-level relative pose. This shifts pose hypothesis generation from conventional pixel-level association to view-pose association for subsequent robust Bayesian-based pose estimation.

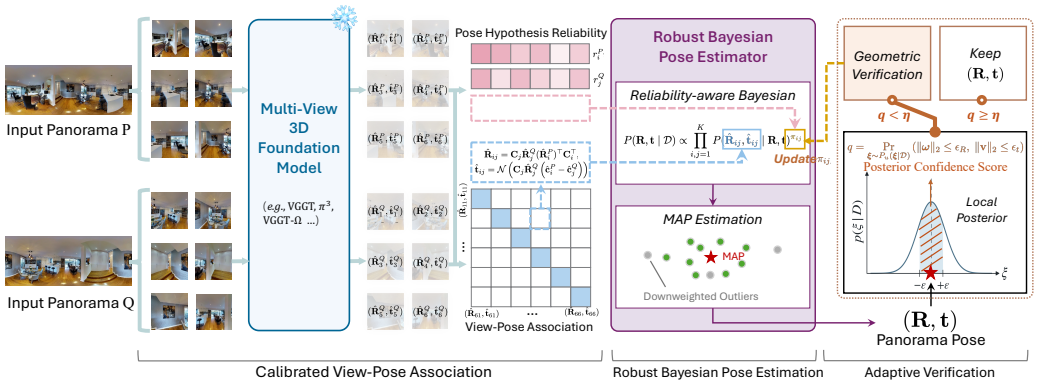


Figure 2: **Overview of PanoREG.** Given two panoramas, we first perform *calibrated view-pose association* to infer sub-view poses via a frozen 3D foundation model, construct panorama-level hypotheses, and assess their reliability via intra-panorama geometric constraints. Next, *robust Bayesian pose estimation* fuses these weighted hypotheses to recover the maximum a posteriori (MAP) pose estimate and posterior uncertainty, which enables *adaptive verification* to selectively trigger point-level refinement only for uncertain cases.

Calibrated View-Pose Inference. Given two equirectangular panoramas $\mathbf{P}$ and $\mathbf{Q}$, we uniformly project each into K horizontal perspective views, $V_P = \{\mathbf{p}_1, ..., \mathbf{p}_K\}$ and $V_Q = \{\mathbf{q}_1, ..., \mathbf{q}_K\}$, at evenly spaced azimuths $\phi_i = 2\pi(i-1)/K$. Since both panoramas follow the same projection rule, the view-to-panorama rotation depends only on the viewing direction and is exactly known as

$$\mathbf{C}_i = \begin{bmatrix} \cos\phi_i & 0 & \sin\phi_i \\ 0 & 1 & 0 \\ -\sin\phi_i & 0 & \cos\phi_i \end{bmatrix} \in \mathrm{SO}(3). \tag{1}$$

These calibrated perspective views allow us to directly leverage pretrained perspective-based 3D foundation reconstruction models for multi-view pose inference. To establish a common reference frame across the two panoramas, we jointly feed the combined view set $V_P \cup V_Q$ into a pretrained model (*e.g.*, VGGT (Wang et al., 2025) and π^3 (Wang et al., 2026)). Through joint multi-view feature interaction, the model aggregates geometric evidence across all $2K$ views and predicts their camera poses in a common coordinate frame: $\{(\hat{\mathbf{R}}_i^P, \hat{\mathbf{t}}_i^P)\}_{i=1}^K$ and $\{(\hat{\mathbf{R}}_j^Q, \hat{\mathbf{t}}_j^Q)\}_{j=1}^K$.

Cross-View Pose Association. Through the above calibrated view-pose inference, we obtain the camera poses of all perspective views from both panoramas in a common reference frame. However, these pose predictions remain *view-level*, while panoramic registration requires a single relative pose between the two complete panoramas. To bridge this gap, we exploit a fundamental property of panoramic decomposition (*i.e.*, *intra-panorama geometric constraints*): although the perspective views differ in viewing direction and local coordinate frame, they are not independent cameras; *all views from the same panorama share a common optical center and have exactly known relative orientations*. Consequently, every cross-panorama view pair $(\mathbf{p}_i, \mathbf{q}_j)$ induces a pose hypothesis $(\hat{\mathbf{R}}_{ij}, \hat{\mathbf{t}}_{ij})$ for the same underlying panorama-to-panorama transformation $(\mathbf{R}, \mathbf{t})$, using the known view-to-panorama rotations, as follows:

$$\hat{\mathbf{R}}_{ij} = \mathbf{C}_j \hat{\mathbf{R}}_j^Q (\hat{\mathbf{R}}_i^P)^\top \mathbf{C}_i^\top \in \mathrm{SO}(3), \quad \hat{\mathbf{t}}_{ij} = \mathcal{N}\left(\mathbf{C}_j \hat{\mathbf{R}}_j^Q \left(\hat{\mathbf{c}}_i^P - \hat{\mathbf{c}}_j^Q\right)\right) \in \mathbb{S}^2, \tag{2}$$

where $\hat{\mathbf{c}}_i^P = -(\hat{\mathbf{R}}_i^P)^\top \hat{\mathbf{t}}_i^P$ and $\hat{\mathbf{c}}_j^Q = -(\hat{\mathbf{R}}_j^Q)^\top \hat{\mathbf{t}}_j^Q$ represent the predicted camera centers, and $\mathcal{N}(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|_2$ denotes ℓ_2 normalization. Considering all $K \times K$ cross-panorama view-pose pairs yields K^2 candidate hypotheses for the same panorama-level relative pose.

Pose Hypothesis Reliability. Ideally, these K^2 pose hypotheses above should agree on a common panorama-level relative pose. In practice, however, errors in individual view-pose predictions can inevitably produce noisy hypotheses, motivating an explicit assessment of pose hypothesis reliability before subsequent pose fusion. We therefore reuse the *intra-panorama geometric constraints* introduced above to assess hypothesis reliability. Specifically, for a view $\mathbf{p}_i$ in panorama $\mathbf{P}$, we

measure its rotation and center inconsistencies with respect to the remaining views $\{\mathbf{p}_l \mid l \neq i\}$ as:

$$e_i^{R,P} = \frac{1}{K-1}\sum_{l\neq i} d_{\mathrm{SO}(3)}\left(\hat{\mathbf{R}}_l^P(\hat{\mathbf{R}}_i^P)^\top, \mathbf{C}_l^\top \mathbf{C}_i\right), \qquad e_i^{c,P} = \frac{1}{K-1}\sum_{l\neq i}\left\|\hat{\mathbf{c}}_i^P - \hat{\mathbf{c}}_l^P\right\|_2, \tag{3}$$

where $d_{\mathrm{SO}(3)}(\cdot,\cdot)$ denotes the geodesic distance on $\mathrm{SO}(3)$. The first term measures the deviation of the predicted relative rotation from that prescribed by the panoramic decomposition, while the second measures the deviation from the shared-camera-center constraint. Smaller values therefore indicate stronger geometric consistency and a more reliable view-pose prediction. Consequently, we normalize these inconsistencies into a view-level reliability score as:

$$r_i^P = \exp\left(-\frac{e_i^{R,P}}{\sigma_R} - \frac{e_i^{c,P}}{\sigma_c}\right), \tag{4}$$

where $i \in \{1, ..., K\}$, and σ_R and σ_c control the sensitivity to rotation and center inconsistencies, respectively. The reliability scores $\{r_1^Q, ..., r_K^Q\}$ are computed analogously for panorama $\mathbf{Q}$. Since each pose hypothesis $(\hat{\mathbf{R}}_{ij}, \hat{\mathbf{t}}_{ij})$ is jointly induced by views $\mathbf{p}_i$ and $\mathbf{q}_j$, we define its reliability as $\pi_{ij} = r_i^P r_j^Q$, which is subsequently used as the reliability weight for robust Bayesian pose fusion.

3.2 Robust Bayesian Pose Estimator

The view-level associations in Sec. 3.1 provide K^2 noisy pose hypotheses $\mathcal{D} = \{(\hat{\mathbf{R}}_{ij}, \hat{\mathbf{t}}_{ij})\}$, together with their reliability priors π_{ij} for the same panorama-level relative pose $(\mathbf{R}, \mathbf{t})$. We next seek a robust panorama-level estimator that can consolidate reliable hypotheses into a consistent global estimate while effectively suppressing severe pose outliers. To this end, we formulate panorama-level pose estimation as reliability-aware Bayesian fusion, jointly considering the geometric reliability of each hypothesis and its consistency with the global consensus for robust pose recovery.

Robust Pose Likelihood. Given a view-association pose hypothesis $(\hat{\mathbf{R}}_{ij}, \hat{\mathbf{t}}_{ij}) \in \mathcal{D}$, we treat it as a noisy observation of the panorama-level relative pose $(\mathbf{R}, \mathbf{t})$ and define its likelihood according to their geometric consistency. Specifically, we quantify this consistency by the rotational geodesic distance $d_{\mathrm{SO}(3)}(\mathbf{R}, \hat{\mathbf{R}}_{ij})$ and the translation-direction angular distance $d_{\mathbb{S}^2}(\mathbf{t}, \hat{\mathbf{t}}_{ij})$. In particular, we note that in challenging real-world cases, some hypotheses may exhibit large residuals and should therefore have limited influence on the global estimate. To robustly model such pose outliers, we adopt heavy-tailed Cauchy distributions over the two residuals:

$$P(\hat{\mathbf{R}}_{ij}, \hat{\mathbf{t}}_{ij} \mid \mathbf{R}, \mathbf{t}) \propto \left(1 + \frac{d^2_{\mathrm{SO}(3)}(\mathbf{R}, \hat{\mathbf{R}}_{ij})}{\tau_R^2}\right)^{-1}\left(1 + \frac{d^2_{\mathbb{S}^2}(\mathbf{t}, \hat{\mathbf{t}}_{ij})}{\tau_t^2}\right)^{-1}, \tag{5}$$

where τ_R and τ_t are the corresponding Cauchy scale parameters. Unlike a Gaussian likelihood, whose negative log-likelihood grows quadratically with the residual, the Cauchy likelihood induces a slowly growing logarithmic penalty for large residuals, thereby limiting the influence of severely inconsistent hypotheses.

Reliability-Aware Bayesian Pose Fusion. Given the robust pose likelihood above, we formulate panorama-level pose estimation as Bayesian inference over all pose hypotheses $\mathcal{D}$. Specifically, by Bayes' rule, the posterior of the panorama-level pose is given by

$$P(\mathbf{R}, \mathbf{t} \mid \mathcal{D}) \propto P(\mathcal{D} \mid \mathbf{R}, \mathbf{t}) P(\mathbf{R}, \mathbf{t}), \tag{6}$$

where $P(\mathbf{R}, \mathbf{t})$ denotes the pose prior. Assuming conditional independence among the pose hypotheses given $(\mathbf{R}, \mathbf{t})$, the joint likelihood factorizes over individual hypotheses. In addition to likelihood-based consistency with the global pose, each hypothesis is equipped with the reliability prior π_{ij}. We incorporate this prior by weighting the contribution of each hypothesis according to its constituent view poses, yielding the reliability-aware posterior as follows:

$$P(\mathbf{R}, \mathbf{t} \mid \mathcal{D}) \propto P(\mathbf{R}, \mathbf{t}) \prod_{i=1}^{K}\prod_{j=1}^{K} P(\hat{\mathbf{R}}_{ij}, \hat{\mathbf{t}}_{ij} \mid \mathbf{R}, \mathbf{t})^{\pi_{ij}}. \tag{7}$$

In the absence of additional prior knowledge, we adopt a uniform pose prior, which is constant and can therefore be omitted. As such, the posterior is determined entirely by the reliability-weighted

cross-view evidence. The final panorama-level pose is then obtained by maximum a posteriori (MAP) estimation, equivalently by maximizing the reliability-weighted log-likelihood as follows:

$$(\hat{\mathbf{R}}, \hat{\mathbf{t}}) = \arg\max_{\mathbf{R},\mathbf{t}} \log P(\mathbf{R}, \mathbf{t} \mid \mathcal{D}) = \arg\max_{\mathbf{R},\mathbf{t}} \sum_{i=1}^{K} \sum_{j=1}^{K} \pi_{ij} \log P(\hat{\mathbf{R}}_{ij}, \hat{\mathbf{t}}_{ij} \mid \mathbf{R}, \mathbf{t}). \tag{8}$$

We solve Eq. 8 using iteratively reweighted pose averaging, as detailed in Appendix A.1.

3.3 Uncertainty-Guided Adaptive Verification

The Bayesian estimator in Sec. 3.2 provides not only a MAP pose estimate, but also a posterior distribution that characterizes how strongly the cross-view hypotheses support this estimate. While a concentrated posterior indicates strong agreement around the MAP solution, a diffuse posterior reveals ambiguity that may lead to unreliable camera-based pose estimation. We exploit this posterior uncertainty to determine when explicit geometric verification is necessary. Specifically, confident estimates are accepted directly, whereas uncertain ones trigger point-level geometric reliability refinement using the 3D pointmaps predicted by the foundation model. This adaptive mechanism retains the efficiency of direct MAP-based pose inference while reserving more expensive geometric recovery for ambiguous cases.

Local Posterior Uncertainty. Let $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = (\hat{\mathbf{R}}, \hat{\mathbf{t}})$ denote the MAP estimate from Eq. 8. Since $\mathbf{R} \in \mathrm{SO}(3)$ and $\mathbf{t} \in \mathbb{S}^2$, we characterize local pose variations around $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ by $\boldsymbol{\xi} = (\boldsymbol{\omega}, \mathbf{v}) \in \mathbb{R}^5$, where $\boldsymbol{\omega} \in \mathbb{R}^3$ parameterizes a local rotation perturbation and $\mathbf{v} \in \mathbb{R}^2$ lies in the tangent space of $\mathbb{S}^2$ at $\hat{\mathbf{t}}$. We denote the corresponding perturbed pose by $\boldsymbol{\theta}(\boldsymbol{\xi})$, with $\boldsymbol{\theta}(\mathbf{0}) = \hat{\boldsymbol{\theta}}_{\mathrm{MAP}}$. We approximate the posterior around the MAP solution using a Laplace approximation (Tierney & Kadane, 1986), where the Hessian $\mathbf{H}$ of the negative log-posterior characterizes the local posterior curvature, and its inverse determines the local covariance, yielding the following Gaussian approximation:

$$\mathbf{H} = \nabla^2_{\boldsymbol{\xi}} \left[-\log P(\boldsymbol{\theta}(\boldsymbol{\xi}) \mid \mathcal{D}) \right] \Big|_{\boldsymbol{\xi}=\mathbf{0}}, \quad P_u(\boldsymbol{\xi} \mid \mathcal{D}) \approx \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma} = (\mathbf{H} + \lambda \mathbf{I})^{-1}). \tag{9}$$

Here, $\boldsymbol{\Sigma}$ approximates the pose covariance, and λ is a small damping term for numerical stability. A concentrated local posterior therefore indicates low pose uncertainty, whereas a diffuse posterior reflects ambiguity among competing pose hypotheses.

Posterior Pose Confidence. Based on the local posterior in Eq. 9, we define pose confidence as the posterior probability mass within a prescribed neighborhood of the MAP estimate $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ (Tierney & Kadane, 1986), which we estimate using Monte Carlo sampling as follows:

$$q = \Pr_{\boldsymbol{\xi} \sim P_u(\boldsymbol{\xi} \mid \mathcal{D})} \left(\|\boldsymbol{\omega}\|_2 \le \epsilon_R,\ \|\mathbf{v}\|_2 \le \epsilon_t \right) \approx \frac{1}{N} \sum_{s=1}^{N} \mathbf{1} \left[\|\boldsymbol{\omega}^{(s)}\|_2 \le \epsilon_R \ \wedge \ \|\mathbf{v}^{(s)}\|_2 \le \epsilon_t \right], \tag{10}$$

where $\boldsymbol{\xi}^{(s)} = (\boldsymbol{\omega}^{(s)}, \mathbf{v}^{(s)}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ is the s-th posterior sample, N is the number of Monte Carlo samples, and ϵ_R and ϵ_t specify the tolerated rotation and translation-direction deviations, respectively. A high q indicates that the posterior is strongly concentrated within the tolerated neighborhood of the MAP estimate, providing high confidence in the estimated pose; conversely, a low q indicates substantial pose uncertainty and suggests that additional geometric recovery is needed.

Adaptive Geometric Verification. The posterior confidence q provides a principled criterion for determining whether the MAP pose estimate can be reliably accepted or requires further geometric refinement. Given a threshold η, we directly retain $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ when $q \ge \eta$. When $q < \eta$, we further exploit the dense 3D pointmaps, the byproduct of the foundation-model inference in Sec. 3.1, to refine the optimization weights $\pi_{ij} = r_i^P r_j^Q$ (*i.e.*, hypothesis reliability) in Eq. 8 for MAP estimation.

Our key insight is that *an accurate view pose should make the predicted 3D points geometrically consistent with their corresponding calibrated viewing rays*. Specifically, for the view-pose estimate $(\hat{\mathbf{R}}_i^P, \hat{\mathbf{t}}_i^P)$ of view $\mathbf{p}_i$, we evaluate its consistency with the pointmaps of all views in V_P. For a global 3D point $\mathbf{x}_{v,n}^P$ predicted at pixel n of view $\mathbf{p}_v$, its viewing direction induced by the estimated pose of $\mathbf{p}_i$ is $\hat{\mathbf{d}}_{i,v,n}^P = \mathcal{N}\Big(\mathbf{C}_v^\top \mathbf{C}_i \hat{\mathbf{R}}_i^P (\mathbf{x}_{v,n}^P - \hat{\mathbf{c}}_i^P)\Big)$. We compare this induced direction with the calibrated viewing ray $\mathbf{d}_{v,n}$ of the corresponding pixel and aggregate their angular discrepancies over all views

Table 1: **Comparison with state-of-the-art methods on Stanford2D3D (Armeni et al., 2016).** We report registration accuracy (ACC) and area under the recall curve (AUC) under different overlap ranges. AGV: Adaptive Geometric Verification.

Method	Overlap 0.1–0.3					Overlap 0.3–0.5					Overlap 0.5–1.0				
	ACC ↑		AUC ↑			ACC ↑		AUC ↑			ACC ↑		AUC ↑		
	@5	@10	@5	@10	@20	@5	@10	@5	@10	@20	@5	@10	@5	@10	@20
VGGT	0.29	1.21	0.12	0.40	1.70	0.65	2.98	0.27	0.97	3.41	0.39	1.75	0.11	0.61	2.68
π^3	0.12	0.96	0.08	0.27	1.20	0.14	1.23	0.08	0.41	2.07	1.04	3.90	0.35	1.29	4.35
TORB	2.00	3.79	0.61	1.85	4.00	4.67	10.83	2.22	4.53	10.07	36.52	48.17	19.61	31.34	42.36
TSIFT	1.01	1.64	0.52	0.92	1.69	7.55	11.98	4.13	7.01	11.15	41.75	49.06	25.20	35.64	44.15
SPHORB	3.50	7.04	1.31	3.36	7.08	11.70	23.20	5.00	11.39	20.38	40.45	53.31	21.56	34.72	46.78
SphereGlue	7.74	15.32	3.34	7.49	14.45	30.60	48.39	16.78	28.27	41.89	77.21	84.87	53.79	67.69	77.38
PanoSplatt3R	0.71	3.12	0.32	1.09	3.76	8.69	16.15	4.67	8.64	15.83	51.75	63.12	30.59	44.56	56.60
G2G	3.54	7.66	1.32	3.66	7.24	6.84	18.62	2.46	7.64	16.67	25.97	44.61	12.07	24.03	39.60
EDM	21.65	42.59	9.13	20.92	35.73	54.76	77.62	32.54	49.36	67.40	90.39	95.45	62.44	77.83	87.45
PanoREG	88.30	93.09	55.53	73.59	83.86	91.92	95.55	63.59	79.03	87.89	95.84	98.18	69.07	83.21	90.90
PanoREG+AGV	**89.18**	**93.63**	**56.61**	**74.40**	**84.41**	**92.95**	**95.62**	**64.63**	**79.71**	**88.31**	**96.17**	**98.38**	**69.80**	**83.66**	**91.17**

and points to obtain the point-level geometric reliability:

$$g_i^P = \exp\left[-\frac{1}{K}\sum_{v=1}^{K}\sum_{n} c_{v,n}^P \rho\Big(\angle(\hat{\mathbf{d}}_{i,v,n}^P, \mathbf{d}_{v,n})\Big)\right], \tag{11}$$

where $c_{v,n}^P$ denotes the point confidence predicted by the foundation model, and $\rho(\cdot)$ is the Cauchy robust penalty. The point-level reliability g_j^Q is obtained analogously for panorama Q. Finally, we refine the hypothesis reliability as $\tilde{\pi}_{ij} = \pi_{ij} g_i^P g_j^Q$ and use $\tilde{\pi}_{ij}$ in Eq. 8 for refined MAP estimation. We note that computing point-ray geometric reliability incurs additional computational cost. We therefore selectively invoke this refinement only for low-confidence cases thanks to this uncertainty-guided adaptive verification mechanism, while most samples already yield reliable MAP estimates and thus bypass the unnecessary point-level computation.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate PanoREG across three panoramic benchmarks: indoor datasets Stanford2D3D (Armeni et al., 2016) (overlap splits $[0.1, 0.3]$, $[0.3, 0.5]$, $[0.5, 1.0]$) and Matterport3D (Chang et al., 2017) ($[0.1, 0.3]$, $[0.3, 1.0]$), alongside outdoor PanoCity (Guo et al., 2026b) (10-frame intervals). Together, these datasets evaluate pose robustness under severe low-overlap conditions and test zero-shot generalization across diverse indoor-outdoor domains.

Baselines. We compare against classical spherical matching methods: *TORB* (Murrugarra-Llerena et al., 2022), *TSIFT* (Murrugarra-Llerena et al., 2022), *SPHORB* (Zhao et al., 2015), learned panoramic matching methods: *SphereGlue* (Gava et al., 2023), *EDM* (Jung et al., 2025), and recent foundation-model-based approaches: *G2G* (Wei et al., 2026), *PanoSplatt3R* (Ren et al., 2025), *PanoVGGT* (Guo et al., 2026b). We also include the pretrained 3D foundation models *VGGT* (Wang et al., 2025) and π^3 (Wang et al., 2026) as perspective-based baselines.

Implementation Details. We decompose each panorama into $K = 6$ views and use pretrained VGGT-1B for inference. We set reliability scales to $(\sigma_R, \sigma_c) = (8^\circ, 0.2)$, Cauchy scales to $(\tau_R, \tau_t) = (5^\circ, 15^\circ)$ with 5-iteration MAP averaging, and confidence tolerances to $(\epsilon_R, \epsilon_t) = (5^\circ, 15^\circ)$ over $N = 20{,}000$ Monte Carlo samples. The Hessian uses step size 10^{-3} with 10^{-6} damping, and adaptive verification adopts $\eta = 0.03$. All experiments run on a single NVIDIA L40 GPU.

4.2 Comparison with SOTA methods

Quantitative Comparison. We conduct zero-shot comparisons with state-of-the-art methods on Stanford2D3D (Armeni et al., 2016), Matterport3D (Chang et al., 2017), and PanoCity (Guo et al., 2026b), as reported in Tables 1 and 2. Under severe low-overlap conditions, correspondence-based methods (*e.g.*, SPHORB, TSIFT) degrade sharply, whereas PanoREG maintains high registration

Table 2: **Comparison with state-of-the-art methods on Matterport3D (Chang et al., 2017) and PanoCity (Guo et al., 2026b).** *AGV*: Adaptive Geometric Verification.

Method	Matterport3D										PanoCity				
	Overlap 0.1–0.3					Overlap 0.3–1.0					Frame Interval 10				
	ACC ↑		AUC ↑			ACC ↑		AUC ↑			ACC ↑		AUC ↑		
	@5	@10	@5	@10	@20	@5	@10	@5	@10	@20	@5	@10	@5	@10	@20
VGGT	0.36	1.64	0.11	0.54	1.94	1.09	3.28	0.35	1.30	3.68	19.49	30.26	9.73	17.45	27.80
π^3	0.18	1.15	0.06	0.36	1.62	0.93	2.59	0.25	1.00	3.33	13.46	25.15	5.94	12.80	23.42
TORB	9.65	12.65	5.15	8.30	11.07	21.28	28.34	12.95	19.30	25.13	60.72	72.98	36.56	52.21	65.02
TSIFT	16.34	18.48	10.98	14.28	16.78	38.16	42.01	26.97	33.68	38.42	75.32	82.15	54.73	67.14	75.64
SPHORB	9.61	15.19	5.25	8.89	14.17	20.32	37.95	8.64	19.24	33.95	56.60	71.22	31.91	48.68	62.44
SphereGlue	33.15	42.60	19.95	29.32	38.35	73.75	79.83	51.72	65.01	73.95	81.48	87.87	59.50	72.41	81.48
PanoSplatt3R	20.75	31.65	10.33	18.51	28.09	59.35	70.81	33.31	49.77	62.42	30.51	53.10	13.06	28.03	46.05
G2G	6.91	14.77	2.65	6.76	10.77	20.33	36.49	10.63	18.86	33.13	12.81	26.58	5.30	12.53	24.47
PanoVGGT	9.83	33.20	2.72	12.15	32.26	16.83	50.47	4.82	19.65	45.03	74.60	95.39	33.63	61.21	79.68
EDM	50.16	65.71	26.50	42.80	57.76	83.31	90.90	**52.25**	71.56	82.56	20.60	37.76	7.90	19.05	31.87
PanoREG	79.07	89.61	46.38	66.16	79.41	85.70	95.39	51.67	71.75	84.62	95.20	97.79	69.46	83.25	90.82
PanoREG+AGV	**79.69**	**89.95**	**46.72**	**66.49**	**79.61**	**86.32**	**95.92**	**52.25**	**72.24**	**85.00**	**95.23**	**97.82**	**69.47**	**83.26**	**90.84**

Table 3: **Ablation Studies and Downstream Evaluation.** (a-c) Ablations on Stanford2D3D; (d) downstream evaluation on 360DVO (Guo et al., 2026a); (e) generalization across foundation models.

ID	Method	AUC@5 ↑	AUC@10 ↑	AUC@20 ↑
	(a) Effect of Pose Aggregation Strategy			
v1	Forward-Facing View Only	24.25	37.76	48.89
v2	Uniform Pose Averaging	41.25	56.63	68.85
v3	Foundation Conf.-Weighted	44.18	59.80	71.66
v4	Reliability-based Averaging	50.24	68.64	80.22
	(b) Effect of Pose Hypothesis Reliability Cues			
v5	w/o Reliability	51.16	69.59	80.75
v6	Rotation Consistency	54.00	71.02	81.94
v7	Center Consistency	53.72	72.12	82.60
v8	Full Geometric Prior	**55.53**	**73.59**	**83.86**
	(c) Effect of Likelihood Function			
v9	Gaussian	50.27	68.70	80.32
v10	Cauchy (PanoREG)	**55.53**	**73.59**	**83.86**

(d) Panoramic Visual Odometry on 360DVO			
Method	**ATE ↓ [m]**	**RPE(t) ↓ [m/f]**	**RPE(r) ↓ [deg/f]**
TORB	19.42	7.83	6.06
TSIFT	29.88	10.99	8.25
SPHORB	28.69	7.78	8.19
PanoSplatt3R	29.65	7.73	3.89
G2G	33.15	9.11	8.20
SphereGlue	28.67	9.27	6.61
PanoVGGT	19.64	7.80	6.41
EDM	28.74	7.13	26.53
360DVO	26.44	8.69	16.67
Ours	**16.19**	**7.03**	**3.39**

(e) Generalization across 3D Foundation Models									
Foundation Model	**Stanford2D3D** (0.1–0.3)			**Matterport3D** (0.1–0.3)			**PanoCity**		
	AUC@5	**AUC@10**	**AUC@20**	**AUC@5**	**AUC@10**	**AUC@20**	**AUC@5**	**AUC@10**	**AUC@20**
π^3	73.26	84.01	89.99	62.68	76.30	84.29	87.02	92.65	95.78
VGGT-Ω	70.31	83.62	90.87	71.72	84.90	91.91	89.59	94.64	97.28

accuracy and yields substantial AUC gains on both Stanford2D3D and Matterport3D. Moreover, its strong performance on the outdoor PanoCity benchmark demonstrates robust cross-domain zero-shot generalization without scene-specific adaptation, while learning-based methods (*e.g.*, EDM and G2G) exhibit greater sensitivity to domain shifts.

Qualitative Comparison. Figure 3 further presents qualitative comparisons under challenging low-overlap cases from Matterport3D (Chang et al., 2017) and PanoCity (Guo et al., 2026b). While competing methods suffer severe pose drift or complete registration failures, PanoREG consistently produces geometrically coherent alignments across both indoor and outdoor environments, confirming the robustness of our view-level reasoning and zero-shot transferability.

4.3 Ablation Study and Analysis

Effect of key design choices. Table 3 (a–c) evaluates key design choices in PanoREG on Stanford2D3D (overlap 0.1–0.3). In (a), we compare different pose aggregation strategies. Uniform multi-view averaging (v2) improves over the single forward-view baseline (v1), while weighting by foundation-model confidence (v3) or our geometric reliability (v4) yields further gains, with reliability-based weighting performing best. In (b), we examine the reliability cues used to assess pose hypotheses. Rotation consistency (v6) and camera-center consistency (v7) each improve over the variant without reliability modeling (v5), while combining both cues (v8) achieves the best per-

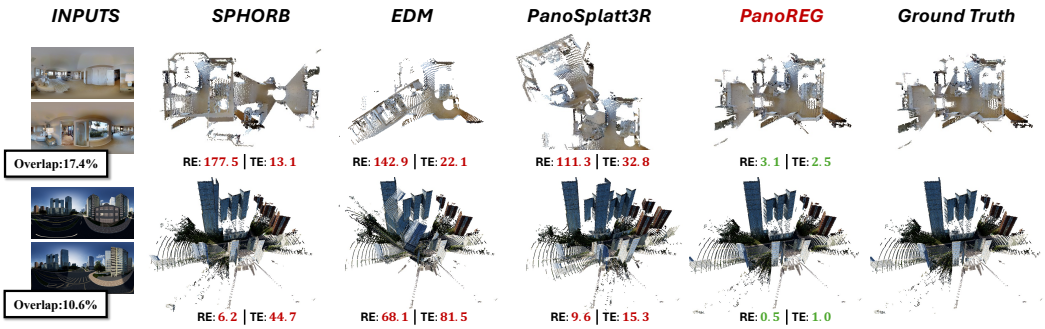
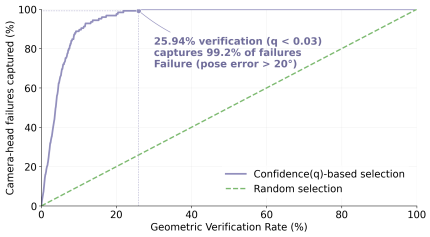


Figure 3: **Qualitative comparison under low overlap.** The first and second rows show examples from Matterport3D (Chang et al., 2017) and PanoCity (Guo et al., 2026b), respectively.

Table 4: **Analysis of uncertainty-guided adaptive verification.**



(a) Confidence-guided failure capture

q Range	#Pairs	AUC@5 ↑	ΔAUC@5 ↑	ACC@5 ↑	ΔACC@5 ↑
< 0.005	199	7.97 → **10.95**	**+2.98**	19.60 → **24.12**	**+4.52**
0.005–0.01	127	30.40 → **34.77**	**+4.37**	63.78 → **68.50**	**+4.72**
0.01–0.03	297	47.60 → **52.42**	**+4.82**	85.19 → **87.21**	**+2.02**
0.03–0.05	354	58.12 → **59.90**	**+1.78**	93.22 → **95.76**	**+2.54**
≥ 0.05	1425	65.57 → **66.05**	**+0.48**	**99.51** → 99.16	−0.35
η	**Time (s) ↓**	**AUC@5 ↑**	**ΔAUC@5 ↑**	**ACC@5 ↑**	**ΔACC@5 ↑**
0.01	1.97	55.53 → **56.01**	**+0.48**	88.30 → **88.93**	**+0.63**
0.03	2.03	55.53 → **56.61**	**+1.08**	88.30 → **89.18**	**+0.88**
0.05	2.10	55.53 → **56.87**	**+1.34**	88.30 → **89.55**	**+1.25**

(b) Point refinement across confidence levels

formance. In (c), we evaluate the likelihood design: replacing the Gaussian likelihood (v9) with the heavy-tailed Cauchy likelihood (v10) further improves robustness to inconsistent pose hypotheses.

Analysis of uncertainty-guided adaptive verification. Table 4(a) shows that posterior confidence q effectively identifies unreliable camera-head estimates, with $q < 0.03$ capturing 99.2% of severe failures. Table 4(b) shows that point-based refinement yields larger gains for low-confidence samples ($q < 0.03$; AUC@5 improves by 2.98-4.82 percentage points), compared with only 0.48 percentage points for high-confidence samples ($q \geq 0.05$). We therefore set $\eta = 0.03$, which achieves a favorable accuracy-efficiency trade-off with modest additional computation.

Generalization across 3D Foundation Models. We further evaluate PanoREG with different 3D foundation models. As shown in Table 3(e), PanoREG maintains strong performance with both π^3 and VGGT-Ω across multiple datasets, demonstrating that our view-pose reasoning framework is not tied to a specific foundation model and can readily leverage different pretrained geometric priors.

4.4 Downstream Evaluation: Large-Baseline Panoramic Visual Odometry

We further evaluate PanoREG on the hard sequences of 360DVO (Guo et al., 2026a) in Table 3(d), where pairwise relative poses between frames sampled at an interval of 10 are used for sequential camera-motion estimation. PanoREG consistently outperforms all competing methods, including the dedicated 360DVO approach (Guo et al., 2026a), demonstrating effective transfer of robust pairwise pose estimation to large-baseline panoramic visual odometry.

5 Conclusion

We presented PanoREG, a training-free framework that reformulates panoramic registration from conventional correspondence-centric match-then-RANSAC pipelines into a viewpose-then-Bayesian reasoning paradigm. By exploiting the calibrated geometry of panoramic sub-views, PanoREG establishes cross-view pose associations from frozen 3D foundation models, yielding multiple panorama-level pose hypotheses whose reliability is assessed through intrinsic intra-panorama geometric consistency. A heavy-tailed Bayesian estimator robustly consolidates these hypotheses, while posterior confidence selectively invokes point-level geometric refinement only when needed. Extensive experiments demonstrate strong robustness, cross-model generalization, and effective transfer to panoramic visual odometry.

AI USE STATEMENT

In this work, we used generative AI tools solely for language editing and polishing of the manuscript, including improvements to grammar, clarity, and readability. All AI-assisted revisions were carefully reviewed by the authors. The authors take full responsibility for the technical content, scientific claims, and final content of this work.

ETHICS STATEMENT

This work focuses on panoramic relative pose estimation and does not involve human-subject studies or the collection of personally identifiable information. All experiments are conducted on publicly available datasets following their respective licenses and usage policies. We are not aware of any significant ethical concerns arising from the methodology or experimental evaluation presented in this work.

REPRODUCIBILITY STATEMENT

We provide a complete description of the proposed method and evaluation protocol in the main paper. The supplementary material further documents implementation details, hyperparameters, dataset preprocessing, overlap computation, pair construction, and additional experimental settings and analyses. Together, these materials provide the information necessary to reproduce the reported experiments and results.

REFERENCES

Zafer Arican and Pascal Frossard. Scale-invariant features and polar descriptors in omnidirectional imaging. *IEEE Transactions on Image Processing*, 21(5):2412–2423, 2012.

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.

David Caruso, Jakob Engel, and Daniel Cremers. Large-scale direct slam for omnidirectional cameras. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 141–148. IEEE, 2015.

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.

Javier Cruz-Mota, Iva Bogdanova, Benoît Paquier, Michel Bierlaire, and Jean-Philippe Thiran. Scale invariant feature transform on the sphere: Theory and applications. *International journal of computer vision*, 98(2):217–241, 2012.

Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. Tangent images for mitigating spherical distortion. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12423–12431. IEEE, 2020.

Christiano Gava, Vishal Mukunda, Tewodros Habtegebrial, Federico Raue, Sebastian Palacio, and Andreas Dengel. Sphereglue: Learning keypoint matching on high resolution spherical images. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 6134–6144. IEEE, 2023.

Christiano Couto Gava, Jean-Marc Hengen, Bertram Taetz, and Didier Stricker. Keypoint detection and matching on high resolution spherical images. In *International Symposium on Visual Computing*, pp. 363–372. Springer, 2013.

Hao Guan and William AP Smith. Brisks: Binary features for spherical images on a geodesic grid. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4886–4894. IEEE, 2017.

Xiaopeng Guo, Yinzhe Xu, Huajian Huang, and Sai-Kit Yeung. 360dvo: Deep visual odometry for monocular 360-degree camera. *IEEE Robotics and Automation Letters*, 2026a.

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*, 2026b.

Peter Hansen, Peter Corke, Wageeh Boles, and Kostas Daniilidis. Scale-invariant features on the sphere. In *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8. IEEE, 2007.

Paul W Holland and Roy E Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics-theory and Methods*, 6(9):813–827, 1977.

Will Hutchcroft, Yuguang Li, Ivaylo Boyadzhiev, Zhiqiang Wan, Haiyan Wang, and Sing Bing Kang. Covispose: Co-visibility pose transformer for wide-baseline relative pose estimation in 360 indoor panoramas. In *European Conference on Computer Vision*, pp. 615–633. Springer, 2022.

San Jiang, Junhuan Liu, Yaxin Li, Duojie Weng, and Wu Chen. Reliable feature matching for spherical images via local geometric rectification and learned descriptor. *Remote Sensing*, 15 (20):4954, 2023.

Dongki Jung, Jaehoon Choi, Yonghan Lee, Somi Jeong, Taejae Lee, Dinesh Manocha, and Suyong Yeon. Edm: Equirectangular projection-oriented dense kernelized feature matching. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6337–6347. IEEE, 2025.

Junho Kim, Jiwon Jeong, and Young Min Kim. Fully geometric panoramic localization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20827–20837. IEEE, 2024.

Jeffri Murrugarra-Llerena, Thiago LT Da Silveira, and Claudio R Jung. Pose estimation for two-view panoramas based on keypoint matching: A comparative study and critical analysis. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 5198–5207. IEEE, 2022.

Jiahui Ren, Mochu Xiang, Jiajun Zhu, and Yuchao Dai. Panosplatt3r: Leveraging perspective pre-training for generalized unposed wide-baseline panorama reconstruction. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28959–28969. IEEE, 2025.

Shiming Song, Yongjun Zhang, Yuanze Wang, Mengzhu Wang, Yuetian Wang, Zhuojing Tian, Jinming Song, and Dianxi Shi. Pano3r: Training free panoramic 3d reconstruction. In *ECAI 2025: 28th European Conference on Artificial Intelligence 25-30 October 2025, Bologna, Italy-Including 14th Conference on Prestigious Applications of Intelligent Systems (PAIS 2025) Proceedings*, pp. 565–572. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 2025.

Jheng-Wei Su, Chi-Han Peng, Peter Wonka, and Hung-Kuo Chu. Gpr-net: Multi-view layout estimation via a geometry-aware panorama registration network. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 6469–6478. IEEE, 2023.

Luke Tierney and Joseph B Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association*, 81(393):82–86, 1986.

Diantao Tu, Hainan Cui, Xianwei Zheng, and Shuhan Shen. Panopose: Self-supervised relative pose estimation for panoramic images. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20009–20018. IEEE, 2024.

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.

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.

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, 2026.

Yufei Wei, Shuhao Ye, Chenxiao Hu, Yiyuan Pan, Dongyu Feng, Rong Xiong, Yue Wang, and Yanmei Jiao. G2g: Exploiting intra-group geometry for inter-group pose estimation. *arXiv preprint arXiv:2606.08284*, 2026.

Qiang Zhao, Wei Feng, Liang Wan, and Jiawan Zhang. Sphorb: A fast and robust binary feature on the sphere. *International journal of computer vision*, 113(2):143–159, 2015.

A Appendix

A.1 Detailed Optimization of Robust Bayesian Pose Estimation

We provide additional details on the optimization of the reliability-aware Bayesian pose estimator introduced in Sec. 3.2 of the main paper. Given the cross-view pose hypotheses $\mathcal{D} = \{(\hat{\mathbf{R}}_{ij}, \hat{\mathbf{t}}_{ij})\}_{i,j=1}^{K}$ and their geometric reliability weights $\pi_{ij} = r_i^P r_j^Q$, our goal is to estimate the panorama-level relative pose $\boldsymbol{\theta} = (\mathbf{R}, \mathbf{t}) \in SO(3) \times S^2$. Under the uniform pose prior adopted in the main paper, MAP estimation is equivalent to minimizing the negative reliability-weighted log-likelihood:

$$\begin{aligned}(\hat{\mathbf{R}}, \hat{\mathbf{t}}) &\in \arg\max_{\mathbf{R},\,\mathbf{t}} \log P(\mathbf{R}, \mathbf{t} \mid \mathcal{D}) \\ &= \arg\min_{\mathbf{R},\,\mathbf{t}} \sum_{i=1}^{K}\sum_{j=1}^{K} \pi_{ij} \left[\log\left(1 + \frac{d^2_{SO(3)}(\mathbf{R}, \hat{\mathbf{R}}_{ij})}{\tau_R^2}\right) + \log\left(1 + \frac{d^2_{S^2}(\mathbf{t}, \hat{\mathbf{t}}_{ij})}{\tau_t^2}\right)\right],\end{aligned} \tag{12}$$

where τ_R, τ_t denote the Cauchy scale parameters for rotation and translation direction, respectively.

Iteratively reweighted pose averaging. We approximately minimize Eq. 12 using iteratively reweighted least squares (IRLS) (Holland & Welsch, 1977) on $SO(3)$ and S^2. For clarity, we define the weighted intrinsic means as

$$\begin{aligned}\mathrm{RotMean}\left(\{\mathbf{R}_{ij}\}, \{w_{ij}\}\right) &\in \arg\min_{\mathbf{R}\in SO(3)} \sum_{i,j} w_{ij} d^2_{SO(3)}(\mathbf{R}, \mathbf{R}_{ij}), \\ \mathrm{DirMean}\left(\{\mathbf{t}_{ij}\}, \{w_{ij}\}\right) &\in \arg\min_{\mathbf{t}\in S^2} \sum_{i,j} w_{ij} d^2_{S^2}(\mathbf{t}, \mathbf{t}_{ij}).\end{aligned} \tag{13}$$

We initialize the hypothesis weights using their geometric reliability:

$$w_{ij}^{(0)} = \frac{\pi_{ij}}{\sum_{m=1}^{K}\sum_{n=1}^{K} \pi_{mn}}, \tag{14}$$

and obtain the initial pose as

$$\mathbf{R}^{(0)} = \mathrm{RotMean}\left(\{\hat{\mathbf{R}}_{ij}\}, \{w_{ij}^{(0)}\}\right), \quad \mathbf{t}^{(0)} = \mathrm{DirMean}\left(\{\hat{\mathbf{t}}_{ij}\}, \{w_{ij}^{(0)}\}\right). \tag{15}$$

At iteration s, we compute the disagreement between each hypothesis and the current estimate:

$$a_{ij}^{(s)} = d_{SO(3)}\left(\mathbf{R}^{(s)}, \hat{\mathbf{R}}_{ij}\right), \qquad b_{ij}^{(s)} = d_{S^2}\left(\mathbf{t}^{(s)}, \hat{\mathbf{t}}_{ij}\right). \tag{16}$$

The corresponding Cauchy consistency factors are

$$c_{ij}^{R,(s)} = \frac{1}{1 + \left(a_{ij}^{(s)}/\tau_R\right)^2}, \qquad c_{ij}^{t,(s)} = \frac{1}{1 + \left(b_{ij}^{(s)}/\tau_t\right)^2}. \tag{17}$$

Since the objective in Eq. 12 is additive in the rotation and translation-direction terms, the two components use separate reweighting coefficients. Specifically, the concavity of $\log(1 + u/\tau^2)$ in the squared residual u yields a quadratic majorizer with coefficients

$$\alpha_{ij}^{R,(s)} = \frac{\pi_{ij}}{\tau_R^2 + \left(a_{ij}^{(s)}\right)^2}, \qquad \alpha_{ij}^{t,(s)} = \frac{\pi_{ij}}{\tau_t^2 + \left(b_{ij}^{(s)}\right)^2}. \tag{18}$$

Normalizing these coefficients separately gives

$$w_{ij}^{R,(s+1)} = \frac{\pi_{ij} c_{ij}^{R,(s)}}{\sum_{m,n} \pi_{mn} c_{mn}^{R,(s)}}, \qquad w_{ij}^{t,(s+1)} = \frac{\pi_{ij} c_{ij}^{t,(s)}}{\sum_{m,n} \pi_{mn} c_{mn}^{t,(s)}}. \tag{19}$$

The component-wise constants $1/\tau_R^2$ and $1/\tau_t^2$ cancel during normalization. We then update the pose using the corresponding weighted intrinsic means:

$$\begin{aligned} \mathbf{R}^{(s+1)} &= \mathrm{RotMean}\left(\{\hat{\mathbf{R}}_{ij}\}, \{w_{ij}^{R,(s+1)}\}\right), \\ \mathbf{t}^{(s+1)} &= \mathrm{DirMean}\left(\{\hat{\mathbf{t}}_{ij}\}, \{w_{ij}^{t,(s+1)}\}\right). \end{aligned} \tag{20}$$

Exact minimization of these weighted subproblems ensures that the objective in Eq. 12 does not increase. When the subproblems are solved numerically, the same property holds provided that each update does not increase its corresponding quadratic surrogate.

After S reweighting iterations, we return $(\mathbf{R}^{(S)}, \mathbf{t}^{(S)})$ as the approximate MAP estimate. We use a fixed budget of $S = 5$ reweighting iterations in all experiments.

A.2 Derivation of Posterior Uncertainty and Confidence

We provide additional details on the posterior uncertainty estimation introduced in Sec. 3.3 of the main paper. Given the MAP panorama pose $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = (\hat{\mathbf{R}}, \hat{\mathbf{t}})$, with $\hat{\mathbf{R}} \in SO(3)$ and $\hat{\mathbf{t}} \in S^2$, we characterize the local posterior around this solution in a five-dimensional tangent space.

Local pose parameterization. A local perturbation around $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ is parameterized as

$$\boldsymbol{\xi} = (\boldsymbol{\omega}, \mathbf{v}) \in \mathbb{R}^3 \times \mathbb{R}^2, \tag{21}$$

where $\boldsymbol{\omega} \in \mathbb{R}^3$ represents a local rotation perturbation and $\mathbf{v} \in \mathbb{R}^2$ parameterizes a perturbation in the tangent space of S^2 at $\hat{\mathbf{t}}$. For rotation, we map the perturbation back to $SO(3)$ using the exponential map:

$$\mathbf{R}(\boldsymbol{\omega}) = \mathrm{Exp}\left([\boldsymbol{\omega}]_\times\right) \hat{\mathbf{R}}, \tag{22}$$

where $[\cdot]_\times$ denotes the skew-symmetric matrix representation. For translation direction, let $\mathbf{B} \in \mathbb{R}^{3\times 2}$ be an orthonormal basis of the tangent plane $T_{\hat{\mathbf{t}}} S^2$. The corresponding tangent perturbation is $\mathbf{B}\mathbf{v} \in \mathbb{R}^3$, and the perturbed translation direction is obtained through the spherical exponential map:

$$\mathbf{t}(\mathbf{v}) = \mathrm{Exp}_{\hat{\mathbf{t}}}(\mathbf{B}\mathbf{v}). \tag{23}$$

More explicitly, letting $\rho = \|\mathbf{v}\|_2$, this mapping is

$$\mathbf{t}(\mathbf{v}) = \cos(\rho)\hat{\mathbf{t}} + \frac{\sin(\rho)}{\rho}\mathbf{B}\mathbf{v}, \tag{24}$$

with $\mathbf{t}(\mathbf{0}) = \hat{\mathbf{t}}$. The resulting locally perturbed pose is therefore denoted as

$$\boldsymbol{\theta}(\boldsymbol{\xi}) = \left(\mathbf{R}(\boldsymbol{\omega}), \mathbf{t}(\mathbf{v})\right), \qquad \boldsymbol{\theta}(\mathbf{0}) = \hat{\boldsymbol{\theta}}_{\mathrm{MAP}}. \tag{25}$$

Laplace approximation. We approximate the posterior distribution locally around the MAP solution using a Laplace approximation. Specifically, we define the local negative log-posterior as

$$\mathcal{F}(\boldsymbol{\xi}) = -\log P\left(\boldsymbol{\theta}(\boldsymbol{\xi}) \mid \mathcal{D}\right). \tag{26}$$

Since $\boldsymbol{\xi} = \mathbf{0}$ corresponds to the MAP solution, the first-order derivative vanishes locally, and a second-order expansion gives

$$\mathcal{F}(\boldsymbol{\xi}) \approx \mathcal{F}(\mathbf{0}) + \frac{1}{2}\boldsymbol{\xi}^\top \mathbf{H} \boldsymbol{\xi}, \tag{27}$$

where

$$\mathbf{H} = \nabla^2_{\boldsymbol{\xi}} \left[-\log P\left(\boldsymbol{\theta}(\boldsymbol{\xi}) \mid \mathcal{D}\right) \right]\Big|_{\boldsymbol{\xi}=\mathbf{0}} \in \mathbb{R}^{5\times 5} \tag{28}$$

is the Hessian of the negative log-posterior at the MAP estimate. This yields the local Gaussian approximation

$$P_u(\boldsymbol{\xi} \mid \mathcal{D}) \approx \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right), \qquad \boldsymbol{\Sigma} = (\mathbf{H} + \lambda \mathbf{I})^{-1}, \tag{29}$$

where a small damping term λ is added for numerical stability. Intuitively, a large posterior curvature corresponds to a small covariance and thus a well-constrained pose estimate, whereas a flat posterior produces a larger covariance and indicates pose ambiguity.

Posterior pose confidence. We use the local posterior above to quantify how much probability mass is concentrated around the MAP estimate. Following the main paper, the posterior confidence is defined as

$$q = \Pr_{\boldsymbol{\xi} \sim P_u(\boldsymbol{\xi}|\mathcal{D})} \left(\|\boldsymbol{\omega}\|_2 \le \epsilon_R,\ \|\mathbf{v}\|_2 \le \epsilon_t \right), \tag{30}$$

where ϵ_R and ϵ_t denote the tolerated local deviations for rotation and translation direction, respectively. Under the exponential-map parameterization above, $\|\boldsymbol{\omega}\|_2$ directly corresponds to the local geodesic rotation deviation, while $\|\mathbf{v}\|_2$ corresponds to the angular displacement on S^2 from $\hat{\mathbf{t}}$. Therefore, q directly measures the posterior probability that the pose remains within the prescribed rotational and translation-direction neighborhood of the MAP estimate. We estimate this probability using Monte Carlo sampling. Specifically, we draw

$$\boldsymbol{\xi}^{(s)} = \left(\boldsymbol{\omega}^{(s)}, \mathbf{v}^{(s)}\right) \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right), \qquad s = 1, \ldots, N, \tag{31}$$

and compute

$$q \approx \frac{1}{N} \sum_{s=1}^{N} 1 \left[\|\boldsymbol{\omega}^{(s)}\|_2 \le \epsilon_R \ \wedge\ \|\mathbf{v}^{(s)}\|_2 \le \epsilon_t \right]. \tag{32}$$

A high value of q indicates that most posterior mass is concentrated close to the MAP estimate and therefore suggests a reliable pose estimate. Conversely, a low q indicates that substantial posterior mass lies outside the tolerated neighborhood, revealing ambiguity among the pose hypotheses. This confidence is subsequently used by the adaptive verification mechanism to determine whether additional point-level geometric refinement is required.

A.3 Details of Point-Ray Geometric Reliability

We provide additional details on the point-level geometric reliability used by the uncertainty-guided adaptive verification in Sec. 3.3 of the main paper. When the posterior confidence satisfies $q < \eta$, we further exploit the dense 3D pointmaps predicted by the frozen 3D foundation model to refine the view-pose reliability before re-estimating the panorama-level pose.

Point-ray consistency. Our key observation is that an accurate view pose should place the predicted 3D points along the calibrated viewing rays associated with their image pixels. For the predicted pose $(\hat{\mathbf{R}}_i^P, \hat{\mathbf{t}}_i^P)$ of view $\mathbf{p}_i$, its corresponding camera center is derived as $\hat{\mathbf{c}}_i^P = -(\hat{\mathbf{R}}_i^P)^\top \hat{\mathbf{t}}_i^P$. For a 3D point $\mathbf{x}_{v,n}^P$ predicted at pixel n of view $\mathbf{p}_v$, we compute the viewing direction induced by the candidate pose of view $\mathbf{p}_i$ as

$$\hat{\mathbf{d}}_{i,v,n}^P = \mathcal{N}\left((\mathbf{C}_v^P)^\top \mathbf{C}_i^P \hat{\mathbf{R}}_i^P \left(\mathbf{x}_{v,n}^P - \hat{\mathbf{c}}_i^P \right) \right), \tag{33}$$

where $\mathcal{N}(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|_2$ denotes ℓ_2 normalization, and $\mathbf{C}_i^P$ and $\mathbf{C}_v^P$ are the predefined view-to-panorama rotations associated with views $\mathbf{p}_i$ and $\mathbf{p}_v$.

We then compare this pose-induced direction $\hat{\mathbf{d}}_{i,v,n}^P$ with the calibrated viewing ray $\mathbf{d}_{v,n}$ corresponding to pixel n of view $\mathbf{p}_v$. Their angular discrepancy is

$$\alpha_{i,v,n}^P = \angle \left(\hat{\mathbf{d}}_{i,v,n}^P, \mathbf{d}_{v,n} \right). \tag{34}$$

A small angular discrepancy indicates that the candidate camera pose is geometrically compatible with the predicted 3D point, whereas a large discrepancy suggests an inconsistent view-pose estimate.

Point-level geometric reliability. To aggregate the point-ray discrepancies over all views and pixels, we weight each point using the confidence $c_{v,n}^P$ predicted by the foundation model and apply a Cauchy robust penalty $\rho(\cdot)$. Following the main paper, the point-level geometric reliability of view $\mathbf{p}_i$ is defined as

$$g_i^P = \exp\left[-\frac{1}{K}\sum_{v=1}^{K}\sum_{n} c_{v,n}^P \rho\left(\alpha_{i,v,n}^P\right)\right]. \tag{35}$$

The confidence weighting reduces the contribution of uncertain 3D points, while the robust penalty limits the influence of points with large angular inconsistency. Consequently, g_i^P is large when the candidate view pose is broadly consistent with the dense geometric predictions across the panorama and decreases as the point-ray disagreement increases. The point-level reliability g_j^Q for a view $\mathbf{q}_j$ in panorama Q is computed analogously.

Refined hypothesis reliability. The point-level reliability provides an additional geometric cue that is complementary to the intra-panorama reliability r_i^P and r_j^Q used in the initial Bayesian pose estimation. For a cross-panorama pose hypothesis induced by $(\mathbf{p}_i, \mathbf{q}_j)$, we refine its original reliability

$$\pi_{ij} = r_i^P r_j^Q \tag{36}$$

as

$$\tilde{\pi}_{ij} = \pi_{ij} g_i^P g_j^Q. \tag{37}$$

The refined reliability $\tilde{\pi}_{ij}$ is then substituted for π_{ij} in the robust Bayesian pose estimator to obtain the refined panorama-level MAP estimate.

A.4 Dataset Details

We evaluate panoramic relative pose estimation on three datasets: Stanford2D3D (Armeni et al., 2016), Matterport3D (Chang et al., 2017), and PanoCity (Guo et al., 2026b), covering indoor environments and outdoor aerial urban scenes. Panoramas are represented in equirectangular projection (ERP), and the input RGB images are resized to 1280×640 pixels.

Stanford2D3D. Stanford2D3D (Armeni et al., 2016) contains 1,413 panoramic RGB images from six indoor areas across three buildings, together with camera poses and per-pixel world-coordinate annotations. Following Jung et al. (2025), we construct candidate pairs by exhaustively pairing distinct camera viewpoints within the same room and area. Geometric overlap is computed from the provided global XYZ maps. The evaluation contains 5,403 image pairs, comprising 2,402 low-overlap (0.1-0.3), 1,461 medium-overlap (0.3-0.5), and 1,540 high-overlap (0.5-1.0) pairs. Ground-truth relative poses are derived from the provided camera extrinsics.

Matterport3D. The Matterport3D (Chang et al., 2017) test set contains 2,014 panoramas from 18 scenes. We convert the provided skybox images into ERP panoramas and exhaustively enumerate all within-scene panorama pairs, yielding 136,753 candidate pairs. Following the protocol of Jung et al. (2025), we compute pairwise overlap using ground-truth depth and camera poses. Retaining pairs with at least 0.1 overlap results in 9,432 evaluation pairs spanning 2,009 distinct panoramas, including 5,564 low-overlap, 3,104 medium-overlap, and 764 high-overlap pairs.

PanoCity. PanoCity (Guo et al., 2026b) is a panoramic dataset containing 119,537 RGB panoramas with corresponding depth maps and camera poses across Beijing, Jinan, and Ningbo. We use the standard test split containing 6,064 panoramas from 100 urban blocks, including 792 images from Beijing, 4,223 from Jinan, and 1,049 from Ningbo. Within each block, we pair frames whose indices differ by 10, requiring every intermediate frame to belong to the test split to avoid bridging discontinuities. This produces 3,714 evaluation pairs.

A.5 Additional Ablation Studies

Effect of the number of perspective views. We study the number of perspective views K used for panoramic decomposition. As shown in Table 5, increasing K from 4 substantially improves registration accuracy, as denser panoramic coverage provides more reliable view-pose evidence and cross-view hypotheses. The improvement, however, gradually saturates as K further increases: increasing K from 8 to 10 yields only marginal gains despite increasing the number of pose hypotheses

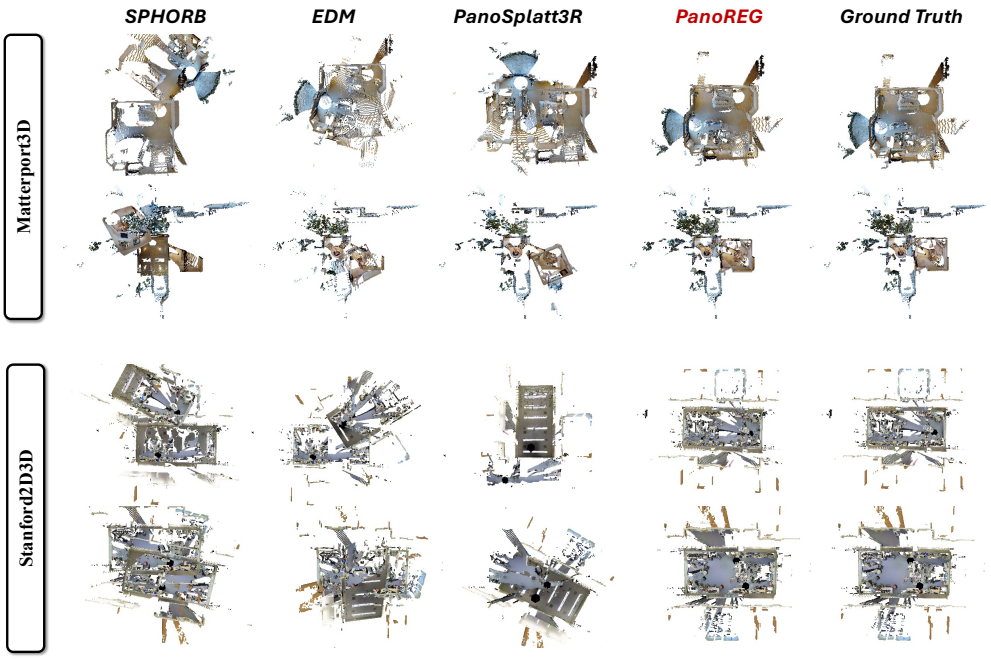


Figure 4: **Qualitative comparison under low overlap (0.1-0.3).** The first and second rows show examples from Matterport3D (Chang et al., 2017) and Stanford2D3D (Armeni et al., 2016)

from 64 to 100. Meanwhile, a larger K also incurs higher computational cost due to the increased number of input views and pose hypotheses. Considering this accuracy-efficiency trade-off, we adopt $K = 6$ as the default setting, which provides sufficient geometric coverage while maintaining moderate inference cost.

Table 5: Effect of the number of perspective views K. We vary the number of perspective views per panorama on the low-overlap (0.1–0.3) split of Stanford2D3D.

K	#Hyp.	AUC@5 ↑	AUC@10 ↑	AUC@20 ↑	Time ↓
4	16	37.21	59.16	73.37	1.63
6	36	55.53	73.59	83.86	2.03
8	64	59.60	76.79	86.34	2.91
10	100	**60.97**	**77.85**	**87.41**	3.94

Sensitivity to Cauchy scale parameters. Table 6 studies the sensitivity of PanoREG to the Cauchy scale parameters τ_R and τ_t on the low-overlap Stanford2D3D split. The performance remains stable over a broad range of parameter choices, indicating that the robust Bayesian estimator is not sensitive to precise scale tuning. Moderate scales perform slightly better overall, with $(\tau_R, \tau_t) = (5°, 15°)$ achieving the best AUC@10 and AUC@20.

Table 6: Sensitivity to the Cauchy scale parameters on the low-overlap (0.1–0.3) split of Stanford2D3D. AUC values are reported in percentages. Best results are highlighted in bold.

τ_R (°)	τ_t (°)	AUC@5↑	AUC@10↑	AUC@20↑
2	15	**55.56**	73.47	83.76
10	15	54.78	73.15	83.64
5	15	55.53	**73.59**	**83.86**
5	5	55.15	72.81	82.90
5	30	55.06	73.21	83.61

A.6 Additional Qualitative Results

Figure 4 qualitatively demonstrates the robustness of PanoREG under low overlap. Across both Matterport3D and Stanford2D3D, competing methods often exhibit large rotational deviations, structural misalignment, or fragmented reconstructions. In contrast, PanoREG preserves the global scene structure and achieves substantially better alignment with the ground truth, highlighting the robustness of view-pose reasoning when reliable pixel-level correspondences are scarce.

A.7 Failure Cases and Limitations

PanoREG substantially improves robustness over correspondence-based registration, particularly under low overlap where reliable pixel correspondences become sparse or ambiguous. Residual failures mainly arise in extreme cases where the underlying 3D foundation model produces systematically biased geometric predictions across most sub-views, causing multiple view-pose hypotheses to converge toward an incorrect consensus that cannot be fully resolved by robust Bayesian fusion. Such cases typically involve extremely limited overlap, severe visual ambiguity, or repetitive and textureless regions with intrinsically scarce geometric evidence. Our uncertainty-guided adaptive verification further mitigates these failures by introducing complementary point-level geometric cues, but in rare cases where both camera-pose and pointmap predictions are severely degraded, point-ray consistency may still provide insufficient evidence for further correction. These remaining failures therefore mainly correspond to extreme regimes in which both view-level and point-level geometric cues become unreliable. Furthermore, adaptive geometric verification introduces additional computation for low-confidence samples. Although the confidence-controlled design avoids this cost for most reliable cases, further improving the efficiency of foundation-model inference and point-level verification remains an important direction for future work.