SHARED RECOVERABLE EVIDENCE FOR GROUND-AERIAL 3D GAUSSIAN SPLATTING REGISTRATION AND FUSION

Anonymous authors
Paper under double-blind review

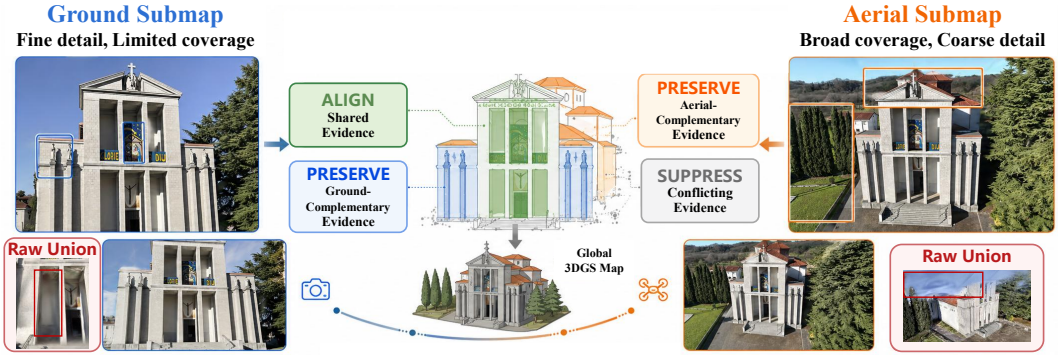

Figure 1: **Qualitative comparison of ground-aerial 3DGS registration and fusion.** Shared recoverable evidence aligns ground-aerial 3DGS submaps, while camera-conditioned fusion preserves complementary evidence and suppresses conflicting evidence.

ABSTRACT

Registration and fusion of independently reconstructed ground and aerial 3D Gaussian Splatting (3DGS) submaps are essential for large-scale scene reconstruction, while cross-platform viewpoint disparities make the task ill-posed. Existing methods overlook the view-dependent radiance of 3DGS primitives, yielding brittle geometric correspondences and severe photometric interference upon naive merging. We present a framework that couples Shared Recoverable Evidence (SRE)-guided registration with camera-conditioned fusion of ground-aerial submaps. SRE-guided registration derives submap-local support from associated training views of 3DGS submaps, then converts it into transform-consistent shared support for Sim(3) proposal refinement and selection. With the transformation fixed, camera-conditioned fusion learns per-Gaussian soft routing selection weights through cross-view map-teacher supervision from frozen submap renders, and modulates opacity to produce one global 3DGS map. Extensive evaluations on ground-aerial pairs from the TerraSky3D, Horizon-GS, and official ScanNet-GSReg benchmarks demonstrate that our approach outperforms existing state-of-the-art methods in both registration accuracy and cross-view rendering fidelity.

1 INTRODUCTION

High-fidelity visual mapping increasingly targets large-scale environments captured across heterogeneous, multi-platform setups. While neural radiance-field representations, including 3D Gaussian Splatting (3DGS), offer high-fidelity 3D scene representations (Mildenhall et al., 2020; Kerbl et al., 2023), scaling to large-scale environments requires spatial partitioning. Existing large-scale frameworks typically decompose global scenes into local submaps for independent optimization, necessitating post-hoc registration and fusion to form a globally unified representation (Tancik et al., 2022; Turki et al., 2022; Lin et al., 2024; Liu et al., 2024). However, seamlessly integrating independently optimized 3DGS submaps remains a critical bottleneck, particularly under heterogeneous ground-

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.

aerial captures. Because aerial drones and ground vehicles observe the scene from drastically distinct vantage points with non-uniform spatial coverages, registering and fusing these disparate submaps requires overcoming severe cross-perspective ambiguities beyond the assumptions of centralized or idealized spatial partitioning.

To address ground-aerial reconstruction, existing joint optimization approaches typically rely on end-to-end centralized training (Ham et al., 2024; Zhang et al., 2024; Jiang et al., 2025; Wang et al., 2026b). However, these methods strictly require centralizing all multi-source raw observations within a single coordinate system. For asynchronous or multi-session captures, such centralized optimization becomes computationally intractable and sacrifices system modularity. A more scalable alternative is post-hoc submap merging without joint retraining, yet this paradigm remains highly ill-posed (Cadena et al., 2016; Saeedi et al., 2016; Liu et al., 2025). Because aerial trajectories yield top-down macro-structural context whereas ground observations provide high-frequency street-level details and fine-grained textures, independently optimized 3DGS submaps exhibit severe geometric and appearance asymmetries. Consequently, registering and fusing such heterogeneous submaps without global re-optimization remains a formidable challenge.

At the core of this difficulty is the fundamental representation shift from discrete point clouds to anisotropic 3DGS primitives. Unlike passive geometric points, 3DGS primitives explicitly parameterize geometry and view-dependent appearance via anisotropic covariance, opacity, and spherical harmonics (Kerbl et al., 2023). Consequently, when observing the same physical surface, ground and aerial captures yield severe parameter divergence, wherein identical regions are encoded with drastically heterogeneous primitive scales, densities, and view-dependent attributes. This attribute heterogeneity renders traditional rigid overlap metrics invalid for 3DGS registration. While recent 3DGS registration methods can estimate Sim(3) transformations (Chang et al., 2024; Yuan et al., 2024), pure geometric alignment alone remains fundamentally insufficient. Directly concatenating aligned submaps leads to severe alpha-blending artifacts, depth-sorting errors, and visual inconsistency. Therefore, effective 3DGS integration requires coupling spatial registration with photometric attribute fusion.

To overcome these limitations, we present a novel framework explicitly designed for the robust registration and seamless fusion of independent 3DGS submaps. Rather than relying on purely geometric heuristics or direct submap overlays, our approach introduces Shared Recoverable Evidence (SRE) to evaluate local reconstruction reliability, enabling transformation-consistent correspondence estimation between heterogeneous ground and aerial submaps. Furthermore, we perform camera-conditioned fusion by dynamically modulating Gaussian contribution weights according to the query viewpoint, adaptively resolving cross-submap redundancy and rendering interference prior to rasterization. Our main contributions are summarized as follows:

- We formulate ground-aerial 3DGS integration as a unified framework combining evidence-driven spatial registration with camera-conditioned fusion, enabling robust post-hoc submap merging without joint retraining.
- We introduce SRE to identify reliable correspondences across heterogeneous submaps, and propose a camera-conditioned fusion strategy that adaptively modulates primitive contributions to preserve complementary details while eliminating inter-submap rendering artifacts.
- Extensive experiments demonstrate that our approach achieves superior registration accuracy over existing point-based and 3DGS baselines, while consistently improving view-dependent rendering quality across challenging ground-aerial environments.

2 Related Work

Large-Scale Scene Reconstruction. Block-NeRF (Tancik et al., 2022) and Mega-NeRF (Turki et al., 2022) decompose large captures into spatially specialized neural fields, while VastGaussian and CityGaussian (Lin et al., 2024; Liu et al., 2024) bring parallel partitioning and subsequent merging to large-scale 3DGS. Related 3DGS systems make the representation view-adaptive or control projected sampling (Lu et al., 2024; Yu et al., 2024). These frameworks coordinate partitioning within a unified reconstruction pipeline. Similarly, existing cross-platform ground-aerial methods exploit jointly available streams or cross-view priors: DRAGON (Ham et al., 2024) synthesizes intermediate elevations, Horizon-GS (Jiang et al., 2025) and HeteroArch-GS (Wang et al., 2026b)

optimize unified aerial-ground representations, and intermediate 3DGS views (Yu et al., 2026) support cross-platform image matching prior to joint Structure-from-Motion (SfM) (Sch"onberger & Frahm, 2016). Our problem begins after separate submap optimization and therefore requires post-hoc integration without a coordinated global reconstruction.

3DGS Registration and Fusion. Aligning disparate 3DGS models has recently gained significant traction. GaussReg (Chang et al., 2024) performs coarse alignment on point clouds extracted from Gaussian primitives followed by image-guided refinement, while PhotoReg (Yuan et al., 2024) combines 3D foundation-model geometry, scale consistency, and photometric refinement. SIREN (Shorinwa et al., 2025) uses semantic support for initialization-free multi-robot mapping, BEV-SplatReg (Leblond et al., 2026) aligns orthographic map projections, and RegGS (Cheng et al., 2025) jointly reconstructs unposed sparse views. GStitch (Shen et al., 2026) studies system-level appearance-aware stitching, whereas GSA (Amoyal et al., 2026) targets category-level cross-instance object alignment. Photometric radiance-field pose optimization jointly refines a scene and its cameras (Lin et al., 2021), while dense feed-forward geometry models provide cross-view pointmaps and camera priors (Wang et al., 2024; 2025). We use such geometry as proposal priors and complement it with SRE to score radiance support after alignment. Our setting jointly requires ground-aerial Sim(3) registration and a single camera-conditioned Gaussian map. Liu et al. (Liu et al., 2025) combine skeleton alignment and Gaussian-adaptive features with structure-aware primitive selection using skeleton, local-detail, and scene-center scores. Our fusion instead conditions source contributions on the query camera and preserves native-view appearance through frozen-submap map-teacher supervision, with geometry and appearance fixed.

Point Cloud Registration. Classical registration ranges from ICP-style local refinement to robust global estimation (Besl & McKay, 1992; Pomerleau et al., 2015; Yang et al., 2021). Learned methods then improve point descriptors, overlap prediction, and correspondence consistency, including RPM-Net (Yew & Lee, 2020), PREDATOR (Huang et al., 2021), PointDSC (Bai et al., 2021), CoFiNet (Yu et al., 2021), GeoTransformer (Qin et al., 2022), and Lepard (Li & Harada, 2022). Their support measures are defined over geometry or learned point features. SRE instead asks whether a frozen 3DGS primitive is recoverable in the associated training views that optimized its radiance parameters before treating it as correspondence evidence.

3 METHOD

3.1 PROBLEM FORMULATION

Let $\mathcal{M}^r = (\mathcal{G}^r, \mathcal{C}^r)$ denote an independently reconstructed 3DGS submap and its calibrated associated training camera poses, where $r \in \{g, a\}$ indexes ground and aerial acquisition. A primitive $g_i^r = (\boldsymbol{\mu}_i^r, \mathbf{q}_i^r, \mathbf{s}_i^r, \alpha_i^r, \mathbf{h}_i^r)$ comprises its center, unit orientation quaternion, positive anisotropic scale, activated opacity, and spherical-harmonic (SH) coefficients. Its covariance is $\boldsymbol{\Sigma}_i^r = \mathbf{R}(\mathbf{q}_i^r)\,\mathrm{diag}((\mathbf{s}_i^r)^2)\mathbf{R}(\mathbf{q}_i^r)^\top$.

The two submaps occupy independent similarity coordinate frames. We estimate $\mathbf{T}_{a\to g} = (s, \mathbf{R}, \mathbf{t}) \in \mathrm{Sim}(3)$, with $\mathbf{T}[\mathbf{x}] = s\mathbf{R}\mathbf{x} + \mathbf{t}$, and express the aerial submap in the ground reference frame. This leaves two coupled decisions: which primitives supply transform-consistent support, and which aligned opposite-origin 3DGS primitives should contribute at a query camera. We address them with *SRE-guided registration* and *camera-conditioned fusion*, respectively.

3.2 SHARED RECOVERABLE EVIDENCE CONSTRUCTION

We first measure whether a primitive is recoverable in the associated training views that optimized its radiance parameters. For g_i^r and C_v^r, let $\nu_{iv}^r \in \{0, 1\}$ be a valid-rasterization indicator. We normalize four cues $f_{iv}^{r,k} \in [0, 1]$: screen-space footprint scale, frozen-render gradient, normalized inverse-square-root depth cue, and activated opacity, indexed by $\mathcal{K} = \{\mathrm{footprint}, \mathrm{grad}, \mathrm{depth}, \mathrm{opacity}\}$. Their saturated multi-view aggregation is

$$e_i^r = 1 - \prod_{v \in \mathcal{C}^r} \left(1 - \nu_{iv}^r \prod_{k \in \mathcal{K}} \left[\delta_k + (1 - \delta_k) f_{iv}^{r,k}\right]\right), \tag{1}$$

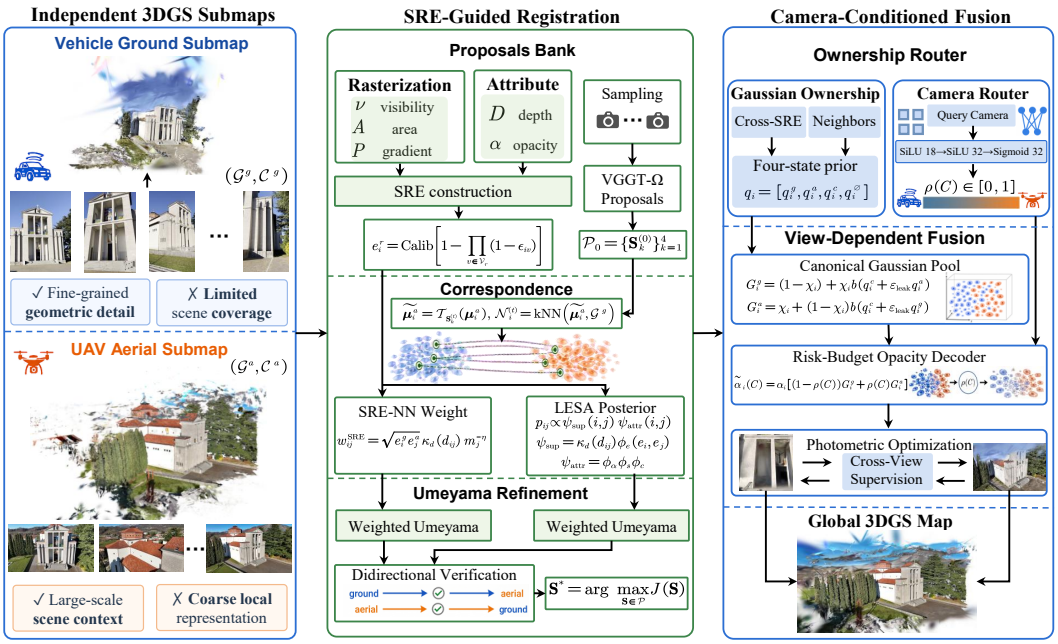

Figure 2: **Framework overview.** SRE-guided registration first identifies transform-consistent support from frozen submap renders and estimates a relative similarity transform. Camera-conditioned fusion then admits complementary registered Gaussians for each query camera before standard 3DGS rasterization, producing one renderable map.

where $\delta_k \in (0, 1)$ prevents one weak cue from nulling an otherwise valid observation. Let $q^r_{.99}$ be the 99th percentile of positive e^r_i in submap r; with $\epsilon > 0$, we use $\widehat{e}^r_i = \min\{e^r_i/(q^r_{.99}+\epsilon), 1\}$. Thus, e^r_i is *submap-level* recoverability. It becomes shared evidence only after a candidate transform pairs it in the other submap.

3.3 SRE-Guided Registration

Reference-frame proposals. To enter the small-overlap regime from a plausible similarity coordinate frame, we form four deterministic camera-subset banks $\mathcal{V}^r_b \subset \mathcal{C}^r$. VGGT-Ω (Wang et al., 2026a) processes each joint bank of frozen submap renders and predicts camera centers $\widehat{\mathbf{c}}^b_v$ in a common, arbitrary similarity coordinate frame Ω_b. We align each associated training-camera trajectory to this frame and compose the two alignments:

$$\begin{aligned} \mathbf{A}^b_r &= \arg\min_{\mathbf{A}\in \mathrm{Sim}(3)} \sum_{v\in\mathcal{V}^r_b} \left\| \mathbf{A}[\mathbf{c}^r_v] - \widehat{\mathbf{c}}^b_v \right\|_2^2, \\ \mathbf{T}^{(0)}_{a\to g,b} &= (\mathbf{A}^b_g)^{-1} \circ \mathbf{A}^b_a. \end{aligned} \tag{2}$$

Equation (2) converts shared-frame camera predictions into a relative Sim(3) initializer by composing the two submap alignments, so the estimate depends only on the ground-aerial relative configuration.

Recoverability-weighted refinement. Given $\mathbf{T}^{(t)}_{a\to g}$, we use Latent Effective Support Alignment (LESA) to assign a normalized posterior $p^{(t)}_{ij}$ over $j \in \mathcal{N}_k(\mathbf{T}^{(t)}_{a\to g}[\boldsymbol{\mu}^a_i])$ from spatial, recoverability, opacity, shape, and low-order appearance compatibility. Let $\mathcal{P}^{(t)}$ contain the posterior-selected pairs, and let $m^{(t)}_j$ count pairs assigned to target j. With $d^{(t)}_{ij} = \|\mathbf{T}^{(t)}_{a\to g}[\boldsymbol{\mu}^a_i] - \boldsymbol{\mu}^g_j\|_2$ and $\kappa_{\sigma_t}(d) = \exp[-d^2/(2\sigma_t^2)]$, we solve

$$\begin{aligned} w^{(t)}_{ij} &= p^{(t)}_{ij} \sqrt{\widehat{e}^a_i \widehat{e}^g_j}\, \kappa_{\sigma_t}(d^{(t)}_{ij})\, (m^{(t)}_j)^{-\eta}, \\ \mathbf{T}^{(t+1)}_{a\to g} &= \arg\min_{\mathbf{T}\in \mathrm{Sim}(3)} \sum_{(i,j)\in\mathcal{P}^{(t)}} w^{(t)}_{ij} \left\| \mathbf{T}[\boldsymbol{\mu}^a_i] - \boldsymbol{\mu}^g_j \right\|_2^2, \end{aligned} \tag{3}$$

which has the weighted Umeyama solution (Umeyama, 1991). $p_{ij}^{(t)}$ resolves ambiguity among local alternatives, whereas $\sqrt{\widehat{e}_i^a \widehat{e}_j^g}$ measures absolute recoverability; $\eta \geq 0$ prevents repeated target structure from dominating the estimate. Across the refined proposal set $\mathcal{H}$, we select the candidate maximizing bidirectional recoverable support after normalized residual and center-scale anchor regularization. The posterior affinity, pair-selection rule, and complete candidate score are specified in the supplementary material.

3.4 Camera-Conditioned Submap Fusion

We transform aerial 3DGS primitives and associated camera poses into the ground reference frame and form the disjoint union $\mathcal{G}^0 = \mathcal{G}^g \uplus \mathbf{T}^\star_{a\to g}(\mathcal{G}^a)$. The similarity transformation maps $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mapsto (s\mathbf{R}\boldsymbol{\mu} + \mathbf{t}, s^2\mathbf{R}\boldsymbol{\Sigma}\mathbf{R}^\top)$ and rotates SH coefficients in the real-SH basis (Blanco et al., 1997). This common coordinate representation is not yet a fused radiance model: full-opacity rendering retains competing explanations of coincident surfaces.

For each primitive in $\mathcal{G}^0$, let $o_i \in \{0, 1\}$ indicate aerial origin and let $\gamma_i \in [0, 1]$ be its learned soft routing selection weight for opposite-submap admission. A camera-role mixing coefficient $\rho_\theta(C) = \sigma(h_\theta(\phi(C)))$ maps a normalized query-camera pose encoding $\phi(C)$ to $[0, 1]$, with the endpoints denoting ground and aerial camera roles. We decode opacity before depth sorting and alpha compositing as

$$\begin{aligned} G_i^g &= (1 - o_i) + o_i\gamma_i, \qquad G_i^a = o_i + (1 - o_i)\gamma_i, \\ \widetilde{\alpha}_i(C) &= \alpha_i\big[(1 - \rho_\theta(C))G_i^g + \rho_\theta(C)G_i^a\big]. \end{aligned} \tag{4}$$

Thus, each submap-associated camera role retains its own 3DGS primitives while admitting only the learned complementary contribution from the other submap. The final camera-conditioned map $\mathcal{G}^\star = (\mathcal{G}^0, \boldsymbol{\gamma}, \theta)$ supplies query-dependent opacities to the standard 3DGS rasterizer.

3.5 Cross-View Map-Teacher Supervision

Fusion uses cross-view map-teacher supervision from frozen submap renders and calibrated camera poses. Define $(\bar{\mathcal{G}}^g, \bar{C}_v^g) = (\mathcal{G}^g, C_v^g)$ and $(\bar{\mathcal{G}}^a, \bar{C}_v^a) = (\mathbf{T}^\star_{a\to g}(\mathcal{G}^a), \mathbf{T}^\star_{a\to g}(C_v^a))$. The detached reference render and routed render are $Y_v^r = \mathrm{sg}[\mathcal{R}(\bar{\mathcal{G}}^r, \bar{C}_v^r)]$ and $F_v^r = \mathcal{R}(\mathcal{G}^0, \bar{C}_v^r; \boldsymbol{\gamma}, \theta)$, where sg denotes detachment. Let Ω_v^r retain reference-supported pixels, and let $y^g = 0, y^a = 1$ be the associated-view role labels. We optimize

$$\begin{aligned} \mathcal{L}_{\text{fuse}} = &\sum_{r\in\{g,a\}} \frac{1}{|\mathcal{C}^r|} \sum_{v\in\mathcal{C}^r} \mathcal{D}_{\Omega_v^r}(F_v^r, Y_v^r) \\ &+ \lambda_{\text{cls}} \sum_{r\in\{g,a\}} \frac{1}{|\mathcal{C}^r|} \sum_{v\in\mathcal{C}^r} \ell_{\text{BCE}}(\rho_\theta(\bar{C}_v^r), y^r) \\ &+ \lambda_{\text{SRE}}\mathcal{R}_{\text{SRE}}(\boldsymbol{\gamma}; \boldsymbol{\pi}, t), \end{aligned} \tag{5}$$

where $\mathcal{D}_\Omega$ is masked L_1-SSIM and $\mathcal{R}_{\text{SRE}}$ is an annealed SRE-prior regularizer. Only $\boldsymbol{\gamma}$ and θ are optimized while position, base opacity, appearance, and $\mathbf{T}^\star_{a\to g}$ remain fixed. Detailed loss, similarity transformation, and posterior definitions are provided in the supplementary material.

4 Experiments

4.1 Experimental Setup

We evaluate ground-aerial 3DGS submap integration on TerraSky3D (D'Urso et al., 2026), Horizon-GS (Jiang et al., 2025), and the public ScanNet-GSReg replay benchmark. TerraSky3D and Horizon-GS exhibit ground-aerial perspective disparity and non-uniform coverage, whereas ScanNet-GSReg provides dense indoor overlap. Ground-aerial point-cloud baselines receive ground-truth relative scale before rigid registration; our method estimates scale. Ground-truth rotation and translation are used for evaluation.



Figure 3: **Qualitative comparison of registration.** Gray and orange represent the ground and aerial submaps, respectively.

Table 1: **Quantitative evaluation of registration.** Per-scene columns report RRE. Avg. equally weights dataset means over 10, 2, and 79 common-output pairs, respectively. Best and second-best values are in bold and underlined, respectively.

Method	TerraSky3D			Horizon-GS		ScanNet-GSReg			Avg.		
	AG_0000	AG_0009	AG_0018	Road	Park	0802	0727	0726	RRE↓	RTE↓	T-RMSE/ q_{90} ↓
FGR + ICP	5.94	152.24	48.33	172.35	162.35	5.58	162.72	52.03	109.18	3.137	2.175
GeoTransformer	157.23	70.32	145.23	74.62	142.51	98.44	168.69	93.70	102.17	11.471	3.609
PREDATOR	15.40	34.07	39.56	35.11	179.49	20.45	168.20	83.99	87.76	4.839	2.458
GaussReg	118.25	142.51	152.01	162.39	79.86	1.81	3.28	1.74	82.74	75.704	1.914
PhotoReg	90.23	43.58	4.83	166.21	110.86	15.48	2.41	**0.11**	72.73	1.597	1.447
BEV-SplatReg	65.25	65.95	2.61	60.39	138.36	164.74	144.65	77.91	90.30	6.914	3.824
Ours	**0.19**	**1.79**	**0.21**	**2.43**	**16.91**	**0.88**	**1.80**	0.61	**4.16**	**0.125**	**0.102**

Datasets and protocols. We compare FGR+ICP, GeoTransformer (Qin et al., 2022), and PREDATOR (Huang et al., 2021) on sampled Gaussian centers with GaussReg (Chang et al., 2024), PhotoReg (Yuan et al., 2024), and scale-and-gravity-assisted BEV-SplatReg. T-RMSE measures source-center RMSE between estimated and reference transformations; q_{90} is the 90th-percentile target-center radius about its median. Table 1 averages common-output pairs within each dataset (10 TerraSky3D, 2 Horizon-GS, 79 ScanNet-GSReg), then equally weights dataset means for all metrics. The eight displayed pairs are examples; Section B.2 gives the aggregation formula.

Table 1 uses each pipeline's initializer: DUSt3R-based coarse alignment for PhotoReg, the released coarse-registration model for GaussReg, and VGGT-Ω proposals for our SRE-guided refinement and selection. The matched-initialization controls in Section F give ICP, GICP, robust ICP, and our refiners the same cached proposal under the stated scale policies.

Table 2 fixes registered submaps, cameras, geometry, SH appearance, and rasterization settings. Fusion trains on detached submap renders; PSNR/SSIM/LPIPS use real RGB at cameras withheld from

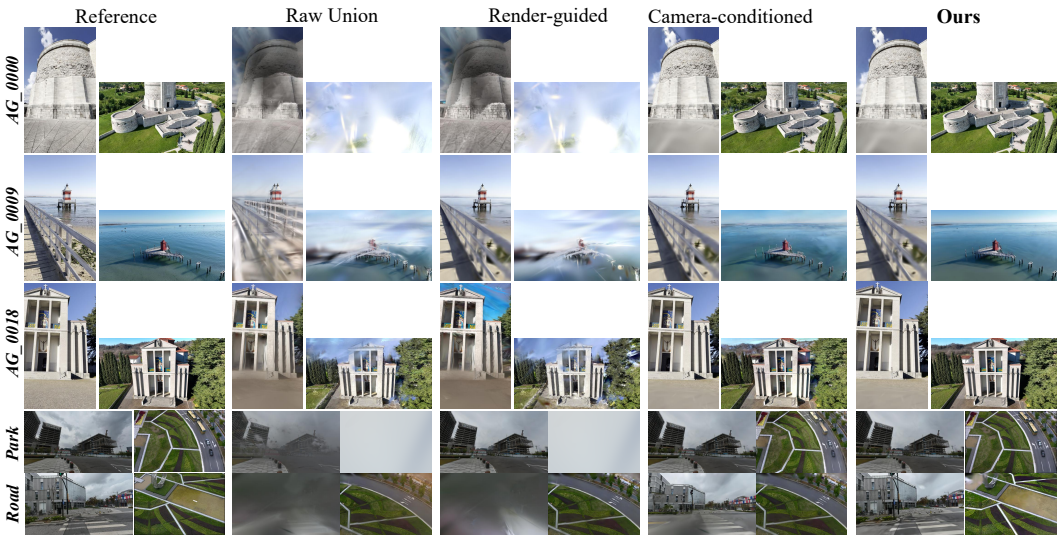

Figure 4: **Qualitative comparison of fixed-registration fusion.** Paired ground and aerial queries show viewpoint-dependent rendering differences.

Table 2: **Quantitative evaluation of fixed-registration fusion.** All variants use identical frozen registered submaps. Per-scene columns report PSNR. The final group reports PSNR, SSIM, and LPIPS averages.

Method	TerraSky3D			Horizon-GS		ScanNet-GSReg			Avg.		
	AG_0000	AG_0009	AG_0018	Road	Park	0802	0727	0726	PSNR↑	SSIM↑	LPIPS↓
Raw union	12.19	11.83	15.09	12.12	15.58	15.55	18.07	19.59	14.71	0.567	0.651
Render-guided selection	14.19	12.55	16.51	11.95	19.18	17.62	18.32	20.69	16.37	0.588	0.616
Camera-conditioned selection	21.71	19.40	23.13	19.83	22.05	18.33	18.79	20.65	20.61	0.687	0.467
Ours	**23.13**	**21.68**	**26.05**	**19.98**	**22.23**	**18.56**	**19.75**	**21.17**	**21.61**	**0.707**	**0.449**

fusion training. Metrics average views within scenes, scenes within datasets, then equally weight the three datasets (15 TerraSky3D, 2 Horizon-GS, 82 ScanNet-GSReg pairs); eight pairs are displayed.

The two selection baselines are internal controls. Render-guided selection ranks source renders by L_1–SSIM agreement with the raw-union render; camera-conditioned selection ranks sources by training-camera distance and direction compatibility. Both composite the secondary render through the primary render's residual alpha. We could not obtain runnable public implementations of SIREN (Shorinwa et al., 2025), Liu et al. (Liu et al., 2025), or GStitch (Shen et al., 2026) for evaluation. RegGS (Cheng et al., 2025) jointly reconstructs unposed sparse views; the appendix evaluates its mixture-Wasserstein term on frozen submaps as RegGS-MW2. Fusion costs appear in Section F.1.

4.2 Evaluation of Sim(3) Submap Registration

Tab. 1 reports the registration comparison across heterogeneous submap pairs. Ours achieves 4.16° RRE, 0.125 RTE, and 0.102 T-RMSE/q_{90}. PhotoReg, the strongest baseline, reaches 72.73° RRE and 1.597 RTE.

Point-cloud matchers rely on sampled centers whose density and local geometric descriptors differ across the two roles. Existing 3DGS registrars assume substantial spatial and appearance overlap, so photometric consistency becomes locally ambiguous under heterogeneous attributes and asymmetric coverage. SRE evaluates submap-level native-view recoverability, filters view-degenerated primitives, and selects transformation-consistent correspondence support. The selected support is determined by stable native-view recovery rather than point density or a single photometric threshold.

The scene-level results show the same pattern across distinct scene topologies. Ours remains below 2° on `AG_0000`, `AG_0009`, and `AG_0018`, whereas FGR+ICP changes from 5.94° on `AG_0000` to 152.24° on `AG_0009`. On Horizon-GS `Park`, where structural anchors are sparse and vegetation

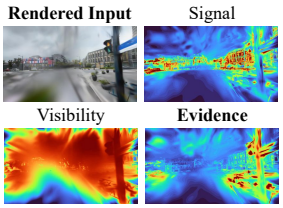

Figure 5: **Qualitative ablation of SRE support.** Rendered input, visibility, signal, and selected-camera evidence.

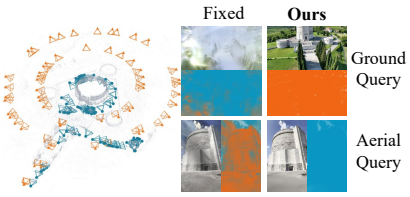

Figure 6: **Qualitative ablation of camera-conditioned routing.** Native camera frusta and matched ground and aerial queries.

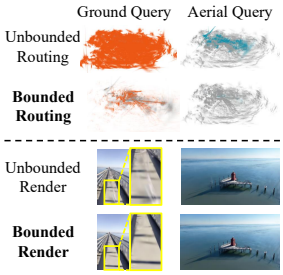

Figure 7: **Qualitative ablation of opacity budget.** Common-view opacity transport and matched query renders.

Table 3: **Quantitative ablation of SRE-guided registration and camera-conditioned opacity routing.** (a) Sim(3) errors after one registration component replacement. (b) Fusion uses fixed registered submaps and estimated Sim(3). Each row changes only the named component.

(a) Evidence guided registration

Setting	RRE (°)↓	RTE↓	RSE↓
Full SRE + LESA	0.49	0.0242	**0.0116**
Uniform evidence	**0.42**	0.0332	0.0154
No multiplicity correction	0.97	0.0373	0.0297
Hard 1-NN support	0.64	**0.0235**	0.0132
No SRE refinement	1.44	0.0656	0.0424

(b) Camera-conditioned opacity routing

Setting	PSNR↑	Ground↑	Aerial↑	Min.↑
Raw union	12.62	15.20	10.47	10.47
Camera-independent mixture	13.67	16.01	11.73	11.69
No opacity budget	22.06	20.76	23.27	20.65
Ours	**23.36**	**22.41**	**24.20**	**22.18**

is dense, Ours obtains 16.91° compared with 79.86° for GaussReg. This contrast separates dense, repeated outdoor structure from the sparse anchors in `Park`. On dense-overlap ScanNet-GSReg scenes, Ours remains competitive with dedicated indoor 3DGS registrars.

Fig. 3 shows the corresponding failure modes. Point-cloud overlays exhibit rotational orientation drift, including tilted aerial submaps, while 3DGS registrars show spatial displacement along facades and disconnected road networks. Ours preserves global orientation and structural co-location of the ground and aerial submaps, particularly at the facade and road structures that define the scene layout.

4.3 Evaluation of Fixed-Registration Submap Fusion

Tab. 2 evaluates fixed-registration fusion under identical frozen submaps. *Raw union* attains 14.71 dB PSNR, 0.567 SSIM, and 0.651 LPIPS, showing that direct Gaussian superposition retains substantial inter-submap interference. *Render-guided selection* reaches 16.37 dB but only 11.95 dB on `Road`.

Camera-conditioned selection improves to 20.61 dB, a 4.24 dB increase over render-guided selection, and reaches 19.83 dB on `Road`. The query camera therefore provides the primary selection cue for submap-level rendering ambiguity. Ours further reaches 21.61 dB, 0.707 SSIM, and 0.449 LPIPS. Its PSNR gains over *Raw union* are 10.94 dB, 9.85 dB, and 10.96 dB on the three TerraSky3D scenes. The remaining improvement compares the complete routing and composition rules, which change both source selection and opacity treatment.

Fig. 4 links these gains to the composition rule. *Raw union* produces ghosting and boundary corruption because opposing primitives compete during alpha blending and depth sorting. *Render-guided selection* removes global duplicate surfaces but leaves spatial seams.

Camera-conditioned selection prioritizes the submap compatible with the query viewpoint and suppresses macroscopic ghosting. The opacity router modulates the registered Gaussians before joint rasterization, while the selection baseline composites separately rendered source images. The paired queries show the resulting differences in facade texture, roof boundaries, and background appearance under ground and aerial viewpoints.

4.4 COMPONENT ABLATIONS

We conduct component-wise ablations on TerraSky3D. Tab. 3(a) replaces one registration component, and Tab. 3(b) replaces camera-conditioned routing or the opacity budget. Each row preserves the submaps, cameras, and evaluation protocol. For estimated and reference similarities $(s, \mathbf{R}, \mathbf{t})$ and $(s^\star, \mathbf{R}^\star, \mathbf{t}^\star)$, RRE measures angular error in degrees, $\mathrm{RTE} = \|\mathbf{t} - \mathbf{t}^\star\|_2 / \|\mathbf{t}^\star\|_2$, and $\mathrm{RSE} = |s/s^\star - 1|$. Fusion scores average ten TerraSky3D pairs, with Min. computed as the scene mean of the lower ground-view or aerial-view PSNR.

Full SRE + LESA yields registration errors of 0.49° RRE, 0.0242 RTE, and 0.0116 RSE. *Uniform evidence* lowers RRE to 0.42° but raises RTE and RSE by 37% and 33%. This discrepancy shows that rotation alone does not characterize similarity estimation quality. Uniform weighting removes the recoverability distinction between primitives; its lower rotation error accompanies higher translation and scale errors in this comparison. Hard 1-NN yields the lowest RTE (0.0235), while the full model yields the lowest RSE among the five variants.

Without *Multiplicity correction*, RRE and RTE rise to 0.97° and 0.0373 because repeated clusters on ground planes and facades dominate the alignment. *Hard 1-NN support* gives 0.64° RRE, and removing *SRE refinement* gives 1.44° RRE and 0.0656 RTE. Multiplicity damping limits repeated support, whereas LESA distributes support over compatible neighborhoods instead of committing to one nearest primitive. Fig. 5 shows how visibility and rasterization signal isolate selected-camera structural anchors for similarity estimation. The rendered input alone contains many visible surfaces, but the evidence panel retains only the support used by the similarity estimate.

Camera-independent mixture reaches 13.67 dB overall and 11.69 dB worst-case PSNR, only slightly above *Raw union* at 12.62 dB and 10.47 dB. Static mixing cannot represent view-dependent occlusion and perspective changes, so contradictory primitives continue to conflict during rasterization. The low minimum-role PSNR shows that a single global composition cannot preserve both native explanations.

Camera-conditioned routing without the opacity budget reaches 22.06 dB overall and 20.65 dB worst-case PSNR, an 8.39 dB gain over static mixing. Fig. 6 shows that source assignment changes with the query camera. The camera frusta locate the two native trajectories in the shared scene, and the paired ground and aerial queries show the corresponding change in source contribution. Ground queries activate ground-origin primitives and suppress aerial clutter, while aerial queries reverse this selection.

Adding the *Opacity budget* gives 23.36 dB overall and 22.18 dB worst-case PSNR, gains of 1.30 dB and 1.53 dB. Fig. 7 separates the common-view opacity transport from the matched query renders. The unbounded model admits more cross-submap opacity, whereas the budget reduces transport and the visible color-bleeding and alpha-blending artifacts in the matched renders.

The controls distinguish correspondence support from rendering contribution. Visibility and rasterization signal select structural anchors for registration, while camera conditioning selects the source appropriate to each query. The opacity-budget comparison then measures the effect of limiting cross-submap transport at a fixed routing stage. In Fig. 7, this transport is concentrated around occlusion boundaries and background regions, where the unbounded variant exhibits color bleeding and haze.

Across matched ground and aerial queries, camera-conditioned routing improves PSNR from 16.01 to 20.76 dB and from 11.73 to 23.27 dB, respectively, over the camera-independent mixture. The larger aerial-view gain shows the unequal effect of camera conditioning across the two acquisition trajectories in Fig. 6. Role-resolved scores complement overall PSNR by separating these changes from the relative numbers of ground and aerial evaluation views.

5 CONCLUSION

We present SRE for ground and aerial 3DGS submap registration and camera conditioned fusion. It estimates Sim(3) from transform consistent support and routes camera compatible Gaussian contributions before rasterization. Across outdoor protocols and ScanNet-GSReg, it improves registration and fixed registration cross-view rendering while fusion uses frozen submap renders rather than dataset RGB.

USE OF GENERATIVE AI

Generative AI tools assisted literature search, hypothesis refinement, experimental design, code implementation, result interpretation, figure and table preparation, and manuscript editing. The authors are responsible for the accuracy, originality, and evidentiary support of the final submission.

REFERENCES

Roy Amoyal, Oren Freifeld, and Chaim Baskin. Cross-instance gaussian splatting registration via geometry-aware feature-guided alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4993–5002, 2026.

Xuyang Bai, Zixin Luo, Lei Zhou, Hongkai Chen, Lei Li, Zeyu Hu, Hongbo Fu, and Chiew-Lan Tai. Pointdsc: Robust point cloud registration using deep spatial consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15859–15869, 2021.

Paul J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992. doi: 10.1109/34.121791.

Miguel A. Blanco, M. Flórez, and M. Bermejo. Evaluation of the rotation matrices in the basis of real spherical harmonics. *Journal of Molecular Structure: THEOCHEM*, 419(1–3):19–27, 1997.

Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J. Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 32(6):1309–1332, 2016.

Jiahao Chang, Yinglin Xu, Yihao Li, Yuantao Chen, Wensen Feng, and Xiaoguang Han. Gaussreg: Fast 3d registration with gaussian splatting. In *European Conference on Computer Vision*, 2024.

Chong Cheng, Yu Hu, Sicheng Yu, Beizhen Zhao, Zijian Wang, and Hao Wang. Reggs: Unposed sparse views gaussian splatting with 3dgs registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8100–8109, 2025.

Mattia D'Urso, Yuxi Hu, Christian Sormann, Mattia Rossi, and Friedrich Fraundorfer. Terrasky3d: Multi-view reconstructions of european landmarks in 4k. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2026.

Yujin Ham, Mateusz Michalkiewicz, and Guha Balakrishnan. DRAGON: Drone and ground gaussian splatting for 3d building reconstruction. In *IEEE International Conference on Computational Photography*, 2024.

Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. PREDATOR: Registration of 3d point clouds with low overlap. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4267–4276, 2021.

Lihan Jiang, Kerui Ren, Mulin Yu, Linning Xu, Junting Dong, Tao Lu, Feng Zhao, Dahua Lin, and Bo Dai. Horizon-GS: Unified 3d gaussian splatting for large-scale aerial-to-ground scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26789–26799, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimk"uhler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. In *ACM SIGGRAPH Conference Proceedings*, 2023.

H. Leblond, G. Simon, R. Martins, C. Demonceaux, and M.-O. Berger. Gaussian splatting map registration with orthographic bird's-eye-view renderings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6080–6089, 2026.

Yang Li and Tatsuya Harada. Lepard: Learning partial point cloud matching in rigid and deformable scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. BARF: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5741–5751, 2021.

Jiaqi Lin, Zhihao Li, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Jiayue Liu, Yangdi Lu, Xiaofei Wu, Songcen Xu, Youliang Yan, and Wenming Yang. Vastgaussian: Vast 3d gaussians for large scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5166–5175, 2024.

Shiyang Liu, Dianyi Yang, Yu Gao, Bohan Ren, Yi Yang, and Mengyin Fu. Automated 3d-gs registration and fusion via skeleton alignment and gaussian-adaptive features, 2025.

Yang Liu, Chuanchen Luo, Lue Fan, Naiyan Wang, Junran Peng, and Zhaoxiang Zhang. Citygaussian: Real-time high-quality large-scale scene rendering with gaussians. In *European Conference on Computer Vision*, 2024.

Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20654–20664, 2024.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pp. 405–421, 2020.

François Pomerleau, Francis Colas, and Roland Siegwart. A review of point cloud registration algorithms for mobile robotics. *Foundations and Trends in Robotics*, 4(1):1–104, 2015. doi: 10.1561/2300000035.

Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11143–11152, 2022.

Sajad Saeedi, Michael Trentini, Mae Seto, and Howard Li. Multiple-robot simultaneous localization and mapping: A review. *Journal of Field Robotics*, 33(1):3–46, 2016.

Johannes L. Sch"onberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4104–4113, 2016.

Licheng Shen, Ho Ngai Chow, Lingyun Wang, Tong Zhang, Mengqiu Wang, Xiaoxin Li, Yibo Li, and Yuxing Han. GStitch: Spatial-temporal fusion of 3d gaussian splattings for scalable 3d reconstruction. *Pattern Recognition*, 173:112826, 2026. doi: 10.1016/j.patcog.2025.112826.

Olao Shorinwa, Jiankai Sun, Mac Schwager, and Anirudha Majumdar. SIREN: Semantic, initialization-free registration of multi-robot gaussian splatting maps. In *Proceedings of the Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 3804–3827, 2025.

Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P. Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-NeRF: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8248–8258, 2022.

Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-NeRF: Scalable construction of large-scale NeRFs for virtual fly-throughs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12922–12931, 2022.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991.

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*, pp. 5294–5306, 2025.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω, 2026a.

Junfan Wang, Han Hu, Zhihao Jia, Yang Jia, Bo Xiang, Jiwei Deng, and Qing Zhu. HeteroArch-GS: Aerial-ground mesh-guided gaussian splatting for heterogeneous architectural landmarks with a real-world dataset. https://vrlab.org.cn/~hanhu/projects/heteroarch-gs/, 2026b. Preprint.

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*, pp. 20697–20709, 2024.

Heng Yang, Jingnan Shi, and Luca Carlone. TEASER: Fast and certifiable point cloud registration. *IEEE Transactions on Robotics*, 37(2):314–333, 2021.

Zi Jian Yew and Gim Hee Lee. RPM-Net: Robust point matching using learned features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11824–11833, 2020.

Hao Yu, Fu Li, Mahdi Saleh, Benjamin Busam, and Slobodan Ilic. Cofinet: Reliable coarse-to-fine correspondences for robust point cloud registration. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Jiangxue Yu, Hui Wang, San Jiang, Xing Zhang, Dejin Zhang, and Qingquan Li. Aerial-ground image feature matching via 3d gaussian splatting-based intermediate view rendering. *Geo-spatial Information Science*, 2026. doi: 10.1080/10095020.2026.2658226.

Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19447–19456, 2024.

Ziwen Yuan, Tianyi Zhang, Matthew Johnson-Roberson, and Weiming Zhi. Photoreg: Photometrically registering 3d gaussian splatting models, 2024.

Saining Zhang, Baijun Ye, Xiaoxue Chen, Yuantao Chen, Zongzheng Zhang, Cheng Peng, Yongliang Shi, and Hao Zhao. Drone-assisted road gaussian splatting with cross-view uncertainty. In *Proceedings of the British Machine Vision Conference*, 2024.

APPENDIX

This appendix specifies the SRE-guided registration and camera-conditioned fusion objectives, evaluation protocols, and metrics. It provides benchmark leaderboards, scene-level results, and controlled ablations. Main-table results and the four-scene fixed-checkpoint controls have separate evaluation populations.

A IMPLEMENTATION DETAILS

This section specifies registration, the admission decoder, and map-teacher supervision. The router-only variant is a separate diagnostic control.

A.1 SUBMAP TRANSFORMATION AND SRE CONSTRUCTION

We map the aerial submap and its associated camera poses into the ground reference frame before fusion. For a Gaussian covariance $\mathbf{\Sigma} = \mathbf{R}(\mathbf{q})\,\mathrm{diag}(\mathbf{s}^2)\mathbf{R}(\mathbf{q})^\top$ and $\mathbf{T} = (s, \mathbf{R}, \mathbf{t})$, the implementation applies

$$\boldsymbol{\mu}' = s\mathbf{R}\boldsymbol{\mu} + \mathbf{t}, \qquad \mathbf{\Sigma}' = s^2\mathbf{R}\mathbf{\Sigma}\mathbf{R}^\top, \qquad \mathbf{h}' = \mathbf{D}_{\mathrm{SH}}(\mathbf{R})\mathbf{h}. \tag{6}$$

The camera center and camera-to-world orientation are transformed by the same $s\mathbf{R}$ and $\mathbf{R}$, respectively. For degree-three real SH, we instantiate $\mathbf{D}_{\mathrm{SH}}(\mathbf{R})$ by fitting the 16×16 real-SH basis transform on 512 deterministic unit directions and apply it independently to the three color channels.

For each visible pair, let $r_{iv,x}, r_{iv,y}$ be screen-space radii, z_{iv} depth, and π_{iv} the projected mean.

$$\begin{aligned}
x_{iv}^{\mathrm{fp}} &= \log(1 + \sqrt{r_{iv,x} r_{iv,y}}),\\
x_{iv}^{\mathrm{dep}} &= (\max\{z_{iv}, 10^{-4}\})^{-1/2},\\
x_{iv}^{\mathrm{sig}} &= \|\nabla \widehat{I}_v(\pi_{iv})\|_2,\\
x_{iv}^{\mathrm{opa}} &= \alpha_i,\\
u_{iv}^k &= \mathrm{clip}_{[0,1]}\left(\frac{x_{iv}^k - Q_5(x_v^k)}{Q_{95}(x_v^k) - Q_5(x_v^k) + \epsilon}\right), \quad k \in \mathcal{K},\\
a_{iv}^k &= \delta_k + (1 - \delta_k) u_{iv}^k,\\
b_{iv} &= \min\left\{0.95, \nu_{iv} \prod_{k\in\mathcal{K}} a_{iv}^k\right\},\\
e_i &= 1 - \prod_{v\in\mathcal{C}} (1 - b_{iv}).
\end{aligned} \tag{7}$$

where $\widehat{I}_v$ is the associated-view image and its signal is a finite-difference luminance gradient. Historical SRE fields use associated dataset images, whereas the rendered-input variant uses frozen submap renders. The map-teacher controls use frozen submap renders. We subsequently use $\widehat{e}_i = \min\{e_i/(Q_{.99}(\{e_j : e_j > 0\}) + \epsilon), 1\}$.

SRE extraction settings. We evaluate every associated training view at render factor one. Each cue is normalized independently within a view by clipping from Q_5 to Q_{95}, with floors $(\delta_{\mathrm{fp}}, \delta_{\mathrm{sig}}, \delta_{\mathrm{dep}}, \delta_{\mathrm{opa}}) = (0.10, 0.10, 0.10, 0.20)$. Per-view evidence is capped at 0.95 before noisy-OR aggregation; the rasterizer uses near and far planes of 0.01 and 10^{10}, respectively.

A.2 PROPOSAL BANK AND SRE-GUIDED REGISTRATION

VGGT-Ω provides four image-based relative-pose initializers. For bank b, it predicts camera centers $\widetilde{\mathbf{c}}_{C,b}^r$ in a shared similarity coordinate frame. We fit each known submap trajectory to that frame and compose the two fits:

$$\begin{aligned}
\mathbf{A}_b^r &= \arg\min_{\mathbf{A}\in\mathrm{Sim}(3)} \sum_{C\in\mathcal{V}_b^r} \left\|\mathbf{A}[\mathbf{c}_C^r] - \widetilde{\mathbf{c}}_{C,b}^r\right\|_2^2,\\
\mathbf{T}_b^{(0)} &= (\mathbf{A}_b^g)^{-1} \circ \mathbf{A}_b^a.
\end{aligned} \tag{8}$$

The four banks are coverage-20, uniform-4 with phase 0.5, coverage-6, and uniform-8 with phase 0.25, where each number is the number of associated images per submap. Each VGGT input is resized to a maximum side length of 1024.

LESA instantiates the local posterior used in Eq. (3) over $k = 3$ neighbors. With $\kappa_\sigma(d) = \exp[-d^2/(2\sigma^2)]$, $\phi_s = \exp[-\|\operatorname{sort}(\log \mathbf{s}_i) - \operatorname{sort}(\log \mathbf{s}_j)\|_2^2/(2\sigma_s^2)]$, and $\phi_h = \exp[-\|\mathbf{h}_i^{(0)} - \mathbf{h}_j^{(0)}\|_2^2/(2\sigma_h^2)]$, the implemented posterior and update weight are

$$\begin{aligned} \psi_{ij}^{(t)} &= \kappa_{\sigma_t}(d_{ij}^{(t)}) \left(\sqrt{\widehat{e}_i^a \widehat{e}_j^g}\right)^{1.2} \left(\sqrt{\alpha_i^a \alpha_j^g}\right)^{0.5} \\ &\quad \cdot \phi_s(i,j)^{0.3} \phi_h(i,j)^{0.2}, \\ p_{ij}^{(t)} &= \frac{\psi_{ij}^{(t)}}{\sum_{\ell \in \mathcal{N}_k(i)} \psi_{i\ell}^{(t)}}, \\ w_i^{(t)} &= \psi_{ij_i}^{(t)} \left[1 + 0.5 \left(p_{ij_i}^{(t)} - p_{i,2\mathrm{nd}}^{(t)}\right)\right]. \end{aligned} \tag{9}$$

We retain $d_{ij_i}^{(t)} \le 2\sigma_t$ and the top 35% of weights, with a 128-correspondence fallback, then use the weighted Umeyama update, where j_i and $p_{i,2\mathrm{nd}}^{(t)}$ denote the highest- and second-highest-posterior neighbors, respectively. The SRE-only branch instead uses the nearest-neighbor weight defined before Eq. (3), including $m_j^{-0.75}$ to downweight repeated assignments to one target primitive.

Let S_{bi} be the smaller directed collision-corrected support, r_{bi} the larger directed weighted residual, and $\mathbf{T}_0 = (s_0, \mathbf{R}_0, \mathbf{t}_0)$ the center-scale initializer. We rank the refined bank by

$$\begin{aligned} J(\mathbf{T}) = S_{\mathrm{bi}} &- 0.10 \frac{r_{\mathrm{bi}}}{\delta_{\mathrm{reg}}} + 0.015 \log(1+n) \\ &- 0.05 \frac{d_R(\mathbf{R}, \mathbf{R}_0)}{180} \\ &- 0.04 \frac{\|\mathbf{t} - \mathbf{t}_0\|_2}{q_{90}} - 0.06 \left|\log \frac{s}{s_0}\right|, \end{aligned} \tag{10}$$

where n is the smaller directed support count and q_{90} is the 90th-percentile radius about the target-center median. Candidate-bank dispersion is retained as a reliability diagnostic; it never modifies the selected transform.

Registration schedule. For each submap, we draw $70{,}000$ uniform and $70{,}000$ SRE-weighted Gaussian centers, and set $\delta_{\mathrm{reg}} = 0.035 q_{90}$ in the target submap. Each proposal independently initializes ten SRE-Umeyama updates with σ_t decreasing from $3\delta_{\mathrm{reg}}$ to δ_{reg}, a $2\sigma_t$ distance gate, and multiplicity exponent 0.75, and eight LESA updates over $k = 3$ neighbors using the same scale schedule, $\sigma_s \ge 0.35$, and $\sigma_h = 0.25$. We retain the top 35% of LESA weights, with a minimum of 128 correspondences. Both branches enter candidate selection. The matched-initialization control separately tests SRE followed by LESA on the same single proposal.

A.3 Camera-Conditioned Fusion

The main formulation combines the camera encoding in Eq. (14), the admission decoder below, and frozen-render map-teacher supervision. A separate router-only control fixes opposite-source admission to zero and optimizes the 1697-parameter camera router using

$$\mathcal{L}_{\mathrm{role}} = \frac{1}{2} \sum_{r \in \{g,a\}} \frac{1}{|\mathcal{C}^r|} \sum_{v \in \mathcal{C}^r} \ell_{\mathrm{BCE}}(\rho_\theta(\bar{C}_v^r), y^r). \tag{11}$$

This diagnostic control uses 3000 Adam updates at learning rate 10^{-3} and sampled BCE weight 0.01, with frozen Gaussian attributes and registration. It is distinct from the 30,000-step map-teacher model in the main text.

A.4 Admission Decoder

The following admission and map-teacher formulation specifies the decoder used in the main formulation and qualitative figures.

The registered full map carries an origin indicator $\zeta_i \in \{0,1\}$ (1 for an aerial Gaussian). Let $u_i = \widehat{e}_i^{\text{own}}$ and let $v_i = \widehat{e}_{\nu(i)}^{\text{cross}} \exp[-d_i^2/(2r^2)]$ be the cross-submap support of its nearest opposite-submap Gaussian. The local radius is $r = \max\{0.0025L, 0.03\}$, where L is the combined-map bounding-box diagonal; nearest-neighbor search uses at most 500,000 opposite-submap centers. The four-state SRE prior is

$$\begin{aligned} m_i &= \sqrt{u_i v_i}, \qquad \omega_i = \sigma((u_i - v_i)/0.08), \\ \overline{\boldsymbol{\pi}}_i &= 0.01\mathbf{1}_4 + m_i[\omega_i \mathbf{e}_{r_i} + (1-\omega_i)\mathbf{e}_{\bar{r}_i}] \\ &\quad + u_i(1-v_i)\mathbf{e}_{\mathrm{c}} + (1-u_i)(1-0.5v_i)\mathbf{e}_{\perp}, \\ \boldsymbol{\pi}_i &= \frac{\overline{\boldsymbol{\pi}}_i}{\|\overline{\boldsymbol{\pi}}_i\|_1}. \end{aligned} \tag{12}$$

where $r_i \in \{g, a\}$ is the origin role, $\bar{r}_i$ the opposite role, and $(\mathbf{e}_g, \mathbf{e}_a, \mathbf{e}_{\mathrm{c}}, \mathbf{e}_{\perp})$ denote the ground-owned, aerial-owned, complementary, and reject states.

The map-teacher model uses the two-state decoder: the first three prior masses are collapsed into the origin-role state and the reject mass is retained. Thus, with $\tau = 0.7$,

$$\begin{aligned} \boldsymbol{\pi}_i^{(2)} &= \big(\pi_{i,g} + \pi_{i,a} + \pi_{i,\mathrm{c}},\, \pi_{i,\perp}\big), \\ \mathbf{q}_i &= \mathrm{softmax}(\mathcal{M}_{r_i,\perp}(\mathbf{l}_i)/\tau)\,, \end{aligned} \tag{13}$$

where $\mathcal{M}_{r_i,\perp}$ masks the two inactive logits.

For camera pose C, define the normalized center $\bar{\mathbf{c}} = (\mathbf{c} - \mathrm{mean}(\mathcal{C}))/q_{90}(\mathcal{C})$ and forward axis $\mathbf{f}$. The camera-role mixing coefficient is

$$\begin{aligned} \boldsymbol{\phi}(C) &= [\bar{\mathbf{c}}, \mathbf{f}, \sin(\pi\bar{\mathbf{c}}), \cos(\pi\bar{\mathbf{c}}), \\ &\qquad \sin(2\pi\bar{\mathbf{c}}), \cos(2\pi\bar{\mathbf{c}})] \in \mathbb{R}^{18}, \\ \rho_\theta(C) &= \sigma(h_\theta(\boldsymbol{\phi}(C))), \qquad h_\theta : 18 \to 32 \to 32 \to 1. \end{aligned} \tag{14}$$

Both hidden layers use SiLU. With $q_{i,\mathrm{own}}$ denoting the active origin-role mass, source admission is

$$\begin{aligned} b_i &= b\varepsilon q_{i,\mathrm{own}}, \qquad b = 0.02, \quad \varepsilon = 0.02, \\ G_i^g &= (1-\zeta_i) + \zeta_i b_i, \qquad G_i^a = \zeta_i + (1-\zeta_i) b_i, \\ \gamma_i(C,t) &= (1-\lambda_t) G_i^{y_C} \\ &\quad + \lambda_t\big[(1-\rho_\theta(C))G_i^g + \rho_\theta(C) G_i^a\big], \\ \widetilde{\alpha}_i(C,t) &= \alpha_i \gamma_i(C,t), \\ \lambda_t &= \mathrm{clip}_{[0,1]}\, \frac{t - 0.2T}{0.8T}. \end{aligned} \tag{15}$$

During training, y_C is the associated-view role; after the first 20% of iterations, the exact role gate is linearly replaced by the camera-pose prediction. At inference, the final single map uses only $\rho_\theta(C)$ before standard depth sorting and alpha compositing.

A.5 Map-Teacher Supervision

For an associated training view C of role r, the reference is a detached render $Y_C^r = \mathrm{sg}[\mathcal{R}(\overline{\mathcal{G}}^r, C)]$ of the frozen role-matched submap and $\Omega_C^r = \{p : \alpha_{Y_C^r}(p) \geq 0.05\}$. For the routed full-map render F_C, the complete optimization objective is

$$\begin{aligned} \ell_1(C) &= \mathrm{mean}_{p \in \Omega_C^r} \|F_C(p) - Y_C^r(p)\|_1, \\ \ell_{\mathrm{ssim}}(C) &= 1 - \mathrm{SSIM}_{\Omega_C^r}(F_C, Y_C^r), \\ \mathcal{D}(C) &= 0.8\ell_1(C) + 0.2\ell_{\mathrm{ssim}}(C), \\ \mathcal{R}_t &= 10^{-3}\, \mathrm{mean}_{i \in \mathcal{I}_t}\, \mathrm{KL}(\boldsymbol{\pi}_i^{(2)} \| \mathbf{q}_i) \\ &\quad + 10^{-4}\, \mathrm{mean}_{i \in \mathcal{I}_t}\, \mathcal{H}(\mathbf{q}_i), \\ \mathcal{L}_t &= \mathcal{D}(C) + \beta_t \mathcal{R}_t + 10^{-2}\, \mathrm{BCE}(\rho_\theta(C), y_C), \\ \beta_t &= \max\{0, 1 - t/(0.8T)\}, \qquad T = 30{,}000. \end{aligned} \tag{16}$$

The KL and entropy terms are averaged over a random visible subset of at most 131,072 Gaussians. Reference alpha and completion losses are set to zero.

Optimization updates the soft routing weights and camera-role function; registered geometry, appearance, and reference rendering remain fixed.

Fusion optimization settings. We use Adam with learning rate 10^{-3} for the ownership logits and camera-role MLP, and train for 30,000 iterations with batch size one, role-balanced sampling, and 512-pixel patches. The detached reference support uses frozen-submap opacity at least 0.05, with loss weights specified in Eq. (16). We disable densification, opacity reset, reference-alpha loss, and completion loss, so the optimization consumes only frozen submaps and calibrated camera poses.

A.6 Native-Teacher Objective and Source Suppression

Consider the opacity rule in Eq. (4) with $\boldsymbol{\gamma} = \mathbf{0}$ and a correctly predicted acquisition role $\rho_\theta(\bar{C}_v^r) = y^r$. Ground queries then retain only ground primitives, and aerial queries retain only aerial primitives. With identical camera transforms, backgrounds, and rasterization settings, the routed render equals its detached source teacher:

$$F_v^r(\mathbf{0}, \rho_\theta = y^r) = Y_v^r, \qquad \mathcal{D}_{\Omega_v^r}(F_v^r, Y_v^r) = 0. \tag{17}$$

Since the photometric discrepancy is nonnegative, this is a global minimum of the photometric term. A sigmoid router approaches the role endpoints as its logits increase in magnitude, so the role BCE also approaches zero when the training roles are separable in the pose encoding. The equality requires matched rendering conventions and exact endpoint gates; it does not assert that finite-logit optimization attains them exactly.

The SRE-prior regularizer can oppose this solution while its coefficient is positive, but it does not provide a persistent constraint: the implemented β_t reaches zero at $0.8T$. Entropy regularization encourages decisive states without requiring a complementary state. Consequently, the native-teacher objective preserves source appearance but does not identify nonzero cross-source admission or guarantee complementary reconstruction. A nonzero parameter alone is insufficient: it must produce a measurable opposite-source contribution to the rendered image.

In the two-state decoder, opposite-source opacity is bounded by $\alpha_i b\varepsilon \leq 0.0004$. At an exact native-role endpoint, its projected alpha is below the rasterizer's $1/255$ cutoff. This explains why different admission policies can produce identical endpoint renders; the bound does not apply to intermediate camera mixtures, where both source gates can be appreciable. Table 15 tests the resulting equivalence, and Section F.2 evaluates a separate cross-source supervision control. Neither the prior nor the opacity budget constitutes a proof that the zero-admission solution is avoided.

A.7 Algorithmic Workflow

Algorithm 8 summarizes the complete execution order. The registration stage searches the proposal bank with SRE-supported correspondence updates, and the fusion stage then trains admission weights and the camera-role router using the selected transformation.

B Experimental Protocols and Metrics

B.1 Ground-Aerial Benchmark Protocols

Each ground and aerial submap is independently reconstructed for 30,000 steps from its respective image set. TerraSky3D is the primary ground-aerial benchmark: point-registration results use its 15-pair protocol, whereas 3DGS-registration uses the fixed 10-pair method-intersection protocol. Horizon-GS retains two ground-aerial pairs and is reported in a separate benchmark-specific table. BEV-SplatReg follows its released scale-and-gravity-assisted configuration and is reported in the benchmark-specific block that preserves those input assumptions.

Table 2 averages 15 TerraSky3D, two Horizon-GS, and 82 ScanNet-GSReg pairs, with equal scene weights within each dataset and equal dataset weights. The fixed-checkpoint diagnostic controls use the four retained checkpoints for AG_0015, AG_0018, Park, and Road, with 43, 34, 800, and 315

Figure 8: **Algorithmic workflow.** SRE-guided registration is completed before camera-conditioned fusion. All functions and objective terms are defined in the preceding implementation subsections.

Require: Frozen $\mathcal{M}^g = (\mathcal{G}^g, \mathcal{C}^g)$ and $\mathcal{M}^a = (\mathcal{G}^a, \mathcal{C}^a)$; associated-view images for registration; camera banks $\{\mathcal{V}_b^g, \mathcal{V}_b^a\}_{b=1}^B$; stage counts $L_{\mathrm{SRE}}, L_{\mathrm{LESA}}$; fusion horizon T
Ensure: Relative transformation $\mathbf{T}_{a\to g}^\star$ and a single renderable map $\mathcal{G}^\star$
for $r \in \{g, a\}$ **do**
  $\widehat{\mathbf{e}}^{\,r} \leftarrow \mathrm{SRE}(\mathcal{G}^r, \mathcal{C}^r)$ using Eq. (7).
end for
$\mathcal{H} \leftarrow \varnothing$ and $J^\star \leftarrow -\infty$
for $b = 1, \ldots, B$ **do**
  Obtain $\mathbf{T}_b^{(0)}$ from Eq. (8)
  for $u \in \{\mathrm{SRE}, \mathrm{LESA}\}$ **do**
    $\mathbf{T} \leftarrow \mathbf{T}_b^{(0)}$
    Set $(L_u, \mathrm{Support}_u) \leftarrow (L_{\mathrm{SRE}}, \mathrm{SRESupport})$ if $u = \mathrm{SRE}$; otherwise set $(L_{\mathrm{LESA}}, \mathrm{LESASupport})$
    for $\ell = 1, \ldots, L_u$ **do**
      $(\mathcal{P}, \mathbf{w}) \leftarrow \mathrm{Support}_u(\mathbf{T}, \widehat{\mathbf{e}}^{\,g}, \widehat{\mathbf{e}}^{\,a})$
      if $|\mathcal{P}| < 3$ or $\sum \mathbf{w} \le 0$ **then**
        break
      else
        $\mathbf{T} \leftarrow \mathrm{WeightedUmeyama}(\mathcal{P}, \mathbf{w})$
      end if
    end for
    $\mathcal{H} \leftarrow \mathcal{H} \cup \{\mathbf{T}\}$
    if $J(\mathbf{T}) > J^\star$ by Eq. (10) **then**
      $(\mathbf{T}_{a\to g}^\star, J^\star) \leftarrow (\mathbf{T}, J(\mathbf{T}))$
    end if
  end for
end for
$\mathcal{G}^0 \leftarrow \mathcal{G}^g \uplus \mathbf{T}_{a\to g}^\star(\mathcal{G}^a)$; transform $\mathcal{C}^a$ and initialize admission logits and θ
for $t = 1, \ldots, T$ **do**
  Sample a role-balanced associated camera (r, C)
  Render the frozen source teacher and routed map; update admission logits and θ using Eq. (16)
end for
Output Registered Gaussians, source identities, admission weights, normalization, and θ for Eq. (4)

validation views. Within each pair, every composition rule uses the same frozen submaps, estimated Sim(3), and transformed associated cameras. Ground-aerial point-registration baselines use ground-truth relative scale, while the matched-initialization controls use the estimated proposal scale. Fusion optimization uses associated training-view poses, while evaluation uses camera poses held out from fusion optimization. Because each frozen submap retains the image set used for its independent reconstruction, this fixed-reconstruction protocol isolates native-view preservation for the camera-conditioned fusion objective while holding the independent submap reconstructions fixed. All fusion comparisons use the same undistorted full-resolution RGB targets, black background, and estimated-registration camera convention. The aerial camera centers and orientations are transformed with the aerial submap. Intrinsics follow the COLMAP parser's undistortion and calibrated image size. The RGB targets enter metric computation only. Scene means give equal weight to scenes rather than pooling their unequal numbers of views.

B.2 Registration and Rendering Metrics

Let $\mathbf{T} = (s, \mathbf{R}, \mathbf{t})$ and $\mathbf{T}^\star = (s^\star, \mathbf{R}^\star, \mathbf{t}^\star)$ denote an estimate and reference Sim(3). We report the standard relative rotation, translation, and scale errors

$$\begin{aligned} \mathrm{RRE} &= d_R(\mathbf{R}, \mathbf{R}^\star), \\ \mathrm{RTE} &= \frac{\|\mathbf{t} - \mathbf{t}^\star\|_2}{\|\mathbf{t}^\star\|_2}, \\ \mathrm{RSE} &= \left| \frac{s}{s^\star} - 1 \right|. \end{aligned} \tag{18}$$

Table 4: **Registration leaderboard.** Methods are grouped by their input representation, while every benchmark remains a separate block. Each value averages the common numeric scene records of the methods in its representation group under that benchmark's native input assumptions. RRE is in degrees and T-RMSE/q_{90} measures the complete estimated transformation.

Method	TerraSky3D		Horizon-GS		ScanNet-GSReg	
	RRE (deg)↓	T-RMSE/q_{90} ↓	RRE (deg)↓	T-RMSE/q_{90} ↓	RRE (deg)↓	T-RMSE/q_{90} ↓
Point-cloud methods						
FGR + ICP	89.04	8.033	167.35	0.807	88.76	0.799
GeoTransformer	113.13	9.719	108.57	0.776	90.65	0.932
PREDATOR	88.17	6.869	107.30	0.634	80.82	0.730
Gaussian-submap methods						
GaussReg	124.36	4.932	121.13	0.769	2.73	0.041
PhotoReg	65.54	3.385	138.54	0.731	14.13	0.224
BEV-SplatReg	47.82	3.346	99.37	1.298	123.70	6.828
Ours	0.49	0.049	9.67	0.209	2.33	0.048

RRE is reported in degrees. To measure the complete transformation on the source Gaussian centers, let q_{90} be the 90th percentile of $\|\boldsymbol{\mu}_j^g - \mathrm{median}(\mathcal{G}^g)\|_2$ over the target submap and report

$$\frac{\text{T-RMSE}}{q_{90}} = \frac{\left[\frac{1}{N}\sum_i \|\mathbf{T}\boldsymbol{\mu}_i^a - \mathbf{T}^\star\boldsymbol{\mu}_i^a\|_2^2\right]^{1/2}}{q_{90}}. \tag{19}$$

where N is the number of aerial Gaussian centers. Proposal and matched-initialization controls instead report $\Delta t/q_{90}$, with $\Delta t = \|\mathbf{t}-\mathbf{t}^\star\|_2$. This radius-normalized translation error differs from both RTE and transformed-center RMSE. Within each dataset, registration averages give equal weight to scenes. The Avg. column in Table 1 equally weights the three dataset means over its stated common-output populations.

Dataset-balanced aggregation. For a method m, metric E, and the common-output set S_d of dataset d, Table 1 reports

$$\overline{E}_m = \frac{1}{3}\sum_{d=1}^{3}\frac{1}{|S_d|}\sum_{s\in S_d} E_{m,d,s}, \qquad (|S_1|,|S_2|,|S_3|) = (10,2,79). \tag{20}$$

The displayed eight pairs are a subset of these populations. Their arithmetic mean differs from both Eq. (20) and the pooled 91-pair mean. Each pair has equal weight within its dataset, and each dataset has weight $1/3$. The same aggregation applies to all methods and all three aggregate error columns. The ScanNet-GSReg intersection contains 79 of the 82 official pairs; PhotoReg has no output on scene0782_00, scene0724_00, and scene0719_00. These aggregate errors describe the common-output population.

PSNR, SSIM, and LPIPS are computed from standard rendered RGB images; ground-view and aerial-view values average their corresponding evaluation cameras separately.

B.3 Public ScanNet-GSReg Replay

The public-view replay uses all 82 official test pairs and both released 10,000-iteration 3DGS submaps per pair. The continuous registration comparison uses the fixed 79-pair method-intersection protocol. The matched-initialization experiment retains all 82 pairs and three shared Gaussian-sampling seeds.

C Main Leaderboards

C.1 Registration

Point-cloud and Gaussian-submap methods use distinct map representations and correspondence mechanisms. The leaderboard groups these families by row and reports each dataset in a separate benchmark block with its own average.

On the primary TerraSky3D Gaussian-submap endpoint, the complete method reaches 0.49° RRE and 0.049 normalized transformed-center RMSE on the ten-scene common protocol. The detailed



Figure 9: **Extended qualitative registration results.** Top: ground–aerial scenes. Bottom: room-level ScanNet-GSReg scenes.

TerraSky3D and ScanNet-GSReg ledgers below expose the corresponding scene-level estimates under the same fixed protocols.

C.2 Fusion

Table 6 reports the separate router-only composition controls by scene. Table 5 reports role-resolved interventions on the retained map-teacher checkpoints using one fixed-router evaluator. Hard selection and the learned decoder have matching rounded role means, while both improve over raw union. Section F separates opposite-source admission from camera-role selection on these four scenes.

Scene	Method	Ground	Aerial	Overall
AG_0015	Raw union	18.832	9.183	13.671
AG_0015	Camera-conditioned hard selection	26.305	24.003	25.074
AG_0015	Historical decoder	26.305	24.003	25.074
AG_0018	Raw union	17.355	12.540	15.089
AG_0018	Camera-conditioned hard selection	26.335	25.734	26.052
AG_0018	Historical decoder	26.335	25.734	26.052
Park	Raw union	16.782	8.763	15.559
Park	Camera-conditioned hard selection	22.499	20.709	22.226
Park	Historical decoder	22.499	20.709	22.226
Road	Raw union	11.379	12.608	12.136
Road	Camera-conditioned hard selection	21.397	19.102	19.983
Road	Historical decoder	21.397	19.102	19.983

Table 5: **Quantitative evaluation of source selection.** PSNR in dB with identical inputs within each scene. Overall averages all views, not the two role means.

Composition	AG_0015	AG_0018	Park	Road
Raw union	13.671	15.089	15.559	12.136
Static mixture	14.375	15.894	16.496	13.160
Nearest-camera selection	25.074	26.052	22.226	19.983
Distance-weighted opacity routing	21.667	22.437	22.222	19.983
Distance-weighted image blending	24.329	25.613	22.234	19.986
Camera-conditioned hard selection	25.074	26.052	22.226	19.983
Same-router image blending	25.074	26.052	22.226	19.984
Router-only opacity gating	25.074	26.052	22.224	19.984

Table 6: **Quantitative evaluation of fixed-registration fusion by scene.** PSNR in dB. Learned rows use the three BCE-only pose routers, with the same router shared across their composition controls.

Table 7: **Quantitative evaluation on the TerraSky3D benchmark.** Scene-wise estimated-Sim(3) 3DGS registration comparison on the fixed 10-scene TerraSky3D protocol. The table reports the exact common-output scene set; protocol assumptions and coverage are reported in the text and supplementary material.

Method	Metric	AG_0012	AG_0001	AG_0000	AG_0009	AG_0016	AG_0017	AG_0004	AG_0020	AG_0018	AG_0025	Avg.
GaussReg	RRE (deg)↓	89.74	159.88	118.25	142.51	83.44	152.92	92.34	133.85	152.01	118.70	124.36
PhotoReg	RRE (deg)↓	66.52	8.29	90.23	43.58	46.61	10.06	171.18	105.95	4.83	108.13	65.54
Ours	RRE (deg)↓	**0.29**	**0.09**	**0.19**	**1.79**	**0.68**	**0.09**	**0.62**	**0.36**	**0.21**	**0.55**	0.49
GaussReg	RTE↓	175.107	161.175	237.293	10.564	818.805	3.677	597.100	11.439	186.315	57.366	225.884
PhotoReg	RTE↓	2.693	0.900	1.391	2.123	2.189	3.053	2.130	2.563	0.122	8.926	2.609
Ours	RTE↓	**0.026**	**0.011**	**0.007**	**0.024**	**0.020**	**0.003**	**0.052**	**0.075**	**0.004**	**0.021**	0.024
GaussReg	RSE↓	0.979	0.986	0.903	0.942	0.982	0.582	0.987	0.709	0.992	0.954	0.902
PhotoReg	RSE↓	0.387	0.256	0.417	0.163	0.054	0.642	0.262	0.084	0.165	0.603	0.303
Ours	RSE↓	**0.008**	**0.005**	**0.007**	**0.005**	**0.017**	**0.005**	**0.034**	**0.026**	**0.004**	**0.004**	0.012
GaussReg	T-RMSE/q_{90} ↓	20.543	1.907	2.117	1.992	1.151	1.214	1.188	1.715	14.918	2.576	4.932
PhotoReg	T-RMSE/q_{90} ↓	17.950	0.519	2.181	1.635	0.948	1.178	1.058	2.732	2.647	2.999	3.385
Ours	T-RMSE/q_{90} ↓	**0.194**	**0.010**	**0.014**	**0.062**	**0.030**	**0.008**	**0.023**	**0.049**	**0.075**	**0.021**	0.049

D Detailed Main-Protocol Evidence

The scene-level records below resolve the common-output registration comparison.

D.1 Common-Output Registration

D.2 Cross-Benchmark Representative Scenes

The following registration table reports three TerraSky3D scenes, both available Horizon-GS scenes, and the six fixed ScanNet-GSReg pairs 0707, 0709, 0726, 0727, 0787, and 0802.

D.3 Room-Level ScanNet-GSReg Scene Ledger

The all-pair macro averages establish the room-level benchmark result. We next resolve six fixed public pairs to make the estimated transformations directly inspectable. The registration ledger keeps both RRE and transformed-center error: Ours averages 1.06° RRE and 0.030 T-RMSE/q_{90} across these pairs, compared with 2.97° and 0.048 for GaussReg.

Table 8: **Representative scene-level registration.** RRE (degrees) on three TerraSky3D scenes, both Horizon-GS scenes, and six fixed public ScanNet-GSReg pairs. The selected rows are the compact cross-benchmark display set; benchmark-wide macro averages remain in Table 4.

Method	TerraSky3D			Horizon-GS		ScanNet-GSReg					
	AG_0000	AG_0009	AG_0018	Road	Park	0707	0709	0726	0727	0787	0802
Point-cloud methods											
FGR + ICP	5.94	152.24	48.33	172.35	162.35	73.30	23.98	52.03	162.72	74.10	5.58
GeoTransformer	157.23	70.32	145.23	74.62	142.51	14.61	17.89	93.70	168.69	140.53	98.44
PREDATOR	15.40	34.07	39.56	35.11	179.49	174.77	81.66	83.99	168.20	153.35	20.45
Gaussian-submap methods											
GaussReg	118.25	142.51	152.01	162.39	79.86	1.38	0.70	1.74	3.28	8.90	1.81
PhotoReg	90.23	43.58	4.83	166.21	110.86	0.74	0.15	0.11	2.41	165.18	15.48
BEV-SplatReg	65.25	65.95	2.61	60.39	138.36	134.50	49.47	77.91	144.65	12.97	164.74
Ours	0.19	1.79	0.21	2.43	16.91	0.50	0.19	0.61	1.80	2.36	0.88

Table 9: **ScanNet-GSReg six-scene registration ledger.** Continuous estimated-Sim(3) errors for the six fixed public room-level pairs. Each metric block uses the same scene order; lower is better. The benchmark-wide result remains the all-pair common-record average in Table 4.

Metric	Method	0707	0709	0726	0727	0787	0802	Avg.
RRE (deg)↓	GaussReg	1.38	0.70	1.74	3.28	8.90	1.81	2.97
RRE (deg)↓	PhotoReg	0.74	0.15	0.11	2.41	165.18	15.48	30.68
RRE (deg)↓	BEV-SplatReg	134.50	49.47	77.91	144.65	12.97	164.74	97.37
RRE (deg)↓	**Ours**	0.50	0.19	0.61	1.80	2.36	0.88	1.06
T-RMSE/q_{90} ↓	GaussReg	0.010	0.007	0.024	0.050	0.178	0.020	0.048
T-RMSE/q_{90} ↓	PhotoReg	0.029	0.013	0.006	0.043	2.097	0.582	0.462
T-RMSE/q_{90} ↓	BEV-SplatReg	7.670	6.062	12.782	1.469	1.694	16.423	7.683
T-RMSE/q_{90} ↓	**Ours**	0.011	0.005	0.024	0.060	0.067	0.015	0.030

E ABLATIONS

Registration ablations separate proposal initialization, recoverability weighting, and local posterior estimation. Fusion controls hold the aligned submaps fixed to evaluate camera conditioning and, for the historical decoder, bounded source admission. We report continuous transformation errors and role-resolved rendering metrics so that each replacement is evaluated at the endpoint it directly affects.

Controlled replacements. Every registration variant reuses the same independently reconstructed submaps, associated training-camera poses, Gaussian sampling budget, proposal inputs, and benchmark split. The replacements alter only the evidence weights, the local correspondence posterior, or the number of transformation proposals. This isolates three questions in sequence: whether recoverability is more informative than uniform Gaussian support, whether soft local association is needed under repeated structures, and whether multiple initial proposals reduce the tail of the estimated transformation error. The historical fusion interventions reuse the same checkpoint, estimated transformation and learned camera router. They alter only the source-admission rule.

Endpoint assignment. The registration tables pair RRE with T-RMSE/q_{90} to measure two aspects of the same estimated transformation: RRE isolates orientation, whereas T-RMSE/q_{90} evaluates the complete transformed Gaussian centers after scale, rotation, and translation have been applied. Median and mean RRE are both retained for calibration controls; their separation reveals whether a design changes the typical estimate or suppresses a small number of high-error hypotheses. For fusion, aggregate PSNR, SSIM, and LPIPS are accompanied by ground-view and aerial-view PSNR.

Evidence chain. Component, calibration, and individual-cue ablations resolve successive stages of registration. Fusion controls separate native-role fidelity, source admission, and continuous-camera interpolation.

E.1 SRE-GUIDED REGISTRATION

Table 10 averages 15 TerraSky3D, two Horizon-GS, and 82 ScanNet-GSReg pairs within their respective blocks. The main-paper Table 3(a) uses a ten-pair TerraSky3D subset of these records, yielding 0.49° rather than the full 15-pair mean of 0.51°. The registration ablation separates cal-

Table 10: **Registration ablation.** Each row replaces one component of SRE-guided registration while retaining the same proposal, sampling, and evaluation protocol. Continuous errors are reported separately for every benchmark block.

	TerraSky3D		Horizon-GS		ScanNet-GSReg	
Configuration	RRE (deg)↓	T-RMSE/q_{90} ↓	RRE (deg)↓	T-RMSE/q_{90} ↓	RRE (deg)↓	T-RMSE/q_{90} ↓
Full SRE + LESA	0.51	0.051	9.67	0.209	2.30	0.225
Uniform evidence	0.46	0.051	10.14	0.211	2.68	0.256
Hard 1-NN correspondence	0.71	0.090	9.52	0.203	2.88	0.265
Without multiplicity correction	0.79	0.125	9.46	0.211	2.64	0.294
Single proposal	0.76	0.071	5.43	0.209	3.05	0.266
Without SRE refinement	1.34	0.302	21.98	0.210	4.16	0.310

ibrated recoverability, local correspondence weighting, and proposal refinement. On TerraSky3D, removing the SRE refinement raises the rotation error from 0.51° to 1.34° and the transformed-center error from 0.051 to 0.302. This replacement measures SRE refinement within an estimator that also includes LESA and the residual schedule. The same replacement raises Horizon-GS RRE from 9.67° to 21.98° and ScanNet-GSReg RRE from 2.30° to 4.16°.

The hard-correspondence and no-multiplicity variants test distinct choices. LESA selects correspondences using local attribute compatibility, while multiplicity correction downweights repeated target assignments in SRE support. Table 10 keeps these component replacements, while Table 11 reports calibration and correspondence controls. Hard 1-NN assignment raises T-RMSE/q_{90} from 0.051 to 0.090 on TerraSky3D and from 0.225 to 0.265 on ScanNet-GSReg; removing multiplicity correction raises the same errors to 0.125 and 0.294.

The aggregate replacements identify the estimator stages. Table 12 resolves their internal evidence construction: it isolates the four rasterization and primitive cues, the view-aggregation rule, and the LESA compatibility terms. Reporting TerraSky3D together with the public room-level replay distinguishes the local-posterior behavior from a single outdoor scene configuration.

Recoverability replacements. Uniform evidence preserves the proposal set and transformed neighborhoods but removes the relation between a Gaussian and the images that support it. The visibility-only control retains rasterization validity while discarding the screen-space, associated-image, normalized-depth, and opacity cues. The single-cue removals keep the calibrated product intact except for one factor, whereas the view-aggregation controls replace how observations from associated training cameras are summarized. These rows therefore distinguish cue redundancy from the broader question of whether registration should be weighted by recoverable image evidence at all. Their primary endpoint is RRE, with T-RMSE/q_{90} verifying that a favorable orientation also maps the Gaussian centers consistently. Maximum and mean view aggregation raise TerraSky3D RRE to 0.59° and 0.63°, and ScanNet-GSReg RRE to 3.13° and 3.15°, respectively, compared with 0.51° and 2.30° for the full field.

Local-posterior replacements. The LESA controls begin after a proposal has mapped the aerial submap into a target neighborhood. A single latent neighbor restricts that neighborhood to one candidate. Equal correspondence weights preserve the selected pairs and their total weight but remove relative weighting between pairs. Removing opacity, scale-shape, or low-frequency color compatibility retains the soft neighborhood but reduces the information used to distribute probability within it. Distinctiveness weighting uses the margin between the two largest posterior entries, whereas multiplicity correction counts source assignments to a target primitive. These changes affect the center pairs and weights used by the similarity update. A single latent neighbor increases TerraSky3D T-RMSE/q_{90} from 0.051 to 0.073, while equal correspondence weights increase the ScanNet-GSReg value from 0.225 to 0.236.

E.2 Camera-Conditioned Fusion

The inference controls preserve every Gaussian, the estimated registration, and the learned camera router. Zero admission suppresses all opposite-source endpoint weights. Constant admission replaces them by their source-wise mean, and shuffled admission permutes them within each source. Their matching native-view scores in Tab. 15 identify camera-conditioned source selection as the measured effect. The following controls test this interpretation at native-role queries and during camera motion.

Table 11: **Quantitative ablation of SRE calibration and correspondence.** All rows use the locked 15-pair TerraSky3D protocol and change one SRE or LESA choice while retaining the remaining procedure.

Configuration	Median RRE (deg)↓	Mean RRE (deg)↓	$\Delta t/q_{90}$ ↓	RSE↓
SRE calibration and correspondence controls				
Lower SRE floors	0.287	0.520	0.0131	0.0123
Higher SRE floors	0.381	0.441	0.0119	0.0105
Rank calibration	0.265	0.615	0.0124	0.0105
LESA, $k = 1$	0.549	0.621	0.0121	0.0095
LESA, $k = 5$	0.349	0.484	0.0116	0.0103
Without multiplicity correction	0.397	0.787	0.0241	0.0236
Full multiplicity correction	0.385	0.513	0.0130	0.0113

Table 12: **Evidence-field and LESA controls.** Continuous registration errors are reported on the primary TerraSky3D and public ScanNet-GSReg protocols. Each row changes one cue, aggregation rule, or local-posterior term while retaining the remaining SRE-guided registration procedure.

Configuration	TerraSky3D		ScanNet-GSReg	
	RRE (deg)↓	T-RMSE/q_{90} ↓	RRE (deg)↓	T-RMSE/q_{90} ↓
Evidence-field construction				
Full evidence field	0.51	0.051	2.30	0.225
Visibility-only evidence	0.47	0.046	2.40	0.225
Without screen-space footprint scale	0.48	0.043	2.20	0.211
Without rendered-image gradient	0.47	0.052	2.52	0.233
Without normalized depth cue	0.46	0.048	2.24	0.209
Without opacity cue	0.43	0.051	2.36	0.224
Arithmetic-mean cue aggregation	0.52	0.047	2.33	0.217
Maximum view aggregation	0.59	0.071	3.13	0.332
Mean view aggregation	0.63	0.081	3.15	0.336
Without submap-wise calibration	0.51	0.050	2.29	0.222
LESA local posterior				
Full SRE + LESA	0.51	0.051	2.30	0.225
Without opacity compatibility	0.51	0.055	2.31	0.211
Without scale-shape compatibility	0.53	0.051	2.33	0.233
Without low-frequency color compatibility	0.50	0.058	2.33	0.222
Single latent neighbor	0.62	0.073	2.40	0.224
Without distinctiveness weighting	0.52	0.051	2.30	0.225
Equal correspondence weights	0.54	0.056	2.30	0.236

F Initialization and Admission Controls

Shared initialization and scale. Table 13 evaluates all 82 ScanNet-GSReg pairs using the first cached VGGT proposal and 70000 uniformly sampled Gaussian centers per submap. Each sampling seed uses identical centers for all refiners. The three seeds measure sampling variation. Fixed-scale refinement holds the proposal scale constant, while joint similarity refinement estimates scale, rotation, and translation. In this table, translation error is divided by the target radius q_{90}, rather than by the reference translation magnitude used in the main leaderboard. ICP and GICP use 80 iterations. SRE uses 10 updates, followed by eight LESA updates where applicable. Robust ICP uses distance-weighted Umeyama updates with uniform evidence. These are matched-input comparisons rather than equal-runtime comparisons. The mean RRE of SRE followed by LESA is comparable to ICP under joint similarity refinement, while its mean normalized translation error is higher. This separates refinement performance from the larger end-to-end differences involving initialization and candidate selection.

Complete outdoor matched-initialization control. Table 14 includes all 15 TerraSky3D and two Horizon-GS pairs. Each scene has 33 records across three sampling seeds and the two scale policies. All submaps use freshly extracted rendered-RGB SRE. Retained and reconstructed replacement submaps form a separate cohort from the historical registration leaderboard. The first archived proposal is shared by every refiner, and SRE followed by LESA is a sequential control rather than the complete method's parallel candidate selection.

On TerraSky3D, sequential SRE and LESA reduces joint-refinement RRE from 1.447° to 0.746°, while ICP reaches 0.518°. Relative to ICP, rotation error is lower on three pairs and higher on twelve. The mean paired difference is $+0.228^\circ$, with a scene-bootstrap interval of $[0.051, 0.440]^\circ$. Translation and scale errors also favor ICP. On Horizon-GS, SRE followed by LESA lowers trans-

Table 13: **Quantitative evaluation of matched-initialization refinement.** All 82 ScanNet-GSReg pairs use three shared sampling seeds. RRE is in degrees. Fixed scale uses the proposal scale.

Refiner	Mean RRE↓	Median RRE↓	$\Delta t/q_{90}$ ↓	RSE↓
Fixed proposal scale				
VGGT proposal	4.844	3.670	0.286	0.037
ICP	2.378	1.395	0.191	0.037
GICP	2.250	1.152	0.188	0.037
Robust ICP	3.764	2.860	0.234	0.037
SRE	3.427	2.327	0.218	0.037
SRE + LESA	2.531	1.429	0.182	0.037
Joint $\mathrm{Sim}(3)$				
VGGT proposal	4.844	3.670	0.286	0.037
ICP	2.493	1.375	0.171	0.035
Robust ICP	3.842	3.001	0.279	0.057
SRE	3.427	2.346	0.203	0.033
SRE + LESA	2.528	1.365	0.182	0.041

Refiner	TerraSky3D (15)			Horizon-GS (2)		
	RRE	$\Delta t/q_{90}$	RSE	RRE	$\Delta t/q_{90}$	RSE
Fixed proposal scale						
VGGT proposal	1.447	0.041	0.049	9.473	0.205	0.384
ICP	1.073	0.047	0.049	4.700	0.227	0.384
GICP	1.230	0.051	0.049	5.551	0.246	0.384
Robust ICP	1.177	0.046	0.049	5.470	0.225	0.384
SRE	1.071	0.039	0.049	5.546	0.227	0.384
SRE + LESA	1.084	0.046	0.049	5.141	0.223	0.384
Joint $\mathrm{Sim}(3)$						
VGGT proposal	1.447	0.041	0.049	9.473	0.205	0.384
ICP	0.518	0.009	0.009	4.927	0.260	0.443
Robust ICP	1.123	0.027	0.018	5.413	0.241	0.412
SRE	0.883	0.017	0.012	5.545	0.233	0.394
SRE + LESA	0.746	0.016	0.014	5.090	0.242	0.415

Table 14: **Quantitative evaluation of matched-initialization outdoor refinement.** All 17 pairs use the first archived proposal, shared Gaussian samples, and rendered-RGB SRE. Values average three sampling seeds within each scene, then scenes within each dataset. RRE is in degrees. Reconstructed replacement submaps define a separate cohort from Table 1.

lation and scale errors on both pairs, while rotation error increases. These comparisons distinguish improvement from the proposal from improvement over a conventional refiner.

External Gaussian-distribution control. We adapt the official RegGS mixture-Wasserstein term (Cheng et al., 2025) to frozen submap refinement, denoted RegGS-MW2. The control retains its Gaussian-mixture compression and loss, using 1000 components and 100 Adam updates from the same first proposal on `AG_0018`. Fixed-scale and joint Sim(3) refinement give $5.775°$ and $5.719°$ RRE, respectively. This is a term-level adaptation, excluding RegGS's image and depth objectives and its own reconstruction stage. Liu et al. (Liu et al., 2025) instead combine skeleton support, learned GA-KPConv detail features, and scene-center scores for primitive selection. The executable feature pipeline and weights were unavailable for our evaluation. We discuss this method as related work.

Historical fixed-router intervention. We retain all Gaussian primitives and freeze the query-camera router while replacing opposite-source admission. The 1192 validation views comprise 34 from `AG_0018`, 43 from `AG_0015`, 800 from `Park`, and 315 from `Road`. These four scenes retain the complete fusion checkpoints needed for replay. Hard selection uses the same router output and retains its selected submap. Mean PSNR for learned admission and hard selection agrees within 10^{-5} dB in each scene. Zero admission, source-wise constant admission, and source-wise shuffled admission give the same native-role behavior. These inference-time interventions leave the trained parameters unchanged. The two-state endpoint cap of 0.0004 lies below the rasterizer's $1/255$ alpha cutoff, explaining why opposite-source primitives have no rendered contribution at those endpoints.

Composition and camera-router controls. The comparison uses the four checkpoints, 1192 native validation views, and four 180-frame paths. Nearest-camera selection uses the closest training-

Admission	PSNR↑	SSIM↑	LPIPS↓
Zero	23.334	0.712	0.378
Constant	23.334	0.712	0.378
Shuffled	23.334	0.712	0.378
Learned	23.334	0.712	0.378

Table 15: **Quantitative ablation of opposite-source admission.** Fixed-checkpoint interventions on four scenes and 1192 validation views, with the learned camera router unchanged.

Input	Teacher	AG_0015	AG_0018	Park	Road	Avg.
Position	No	25.074 ± 0.000	26.052 ± 0.000	22.185 ± 0.055	19.984 ± 0.000	23.324 ± 0.014
Position + direction	No	25.074 ± 0.000	26.052 ± 0.000	22.224 ± 0.002	19.984 ± 0.000	23.333 ± 0.001
Position	Yes	25.074 ± 0.000	26.052 ± 0.000	22.207 ± 0.015	19.984 ± 0.000	23.329 ± 0.004
Position + direction	Yes	25.074 ± 0.000	26.052 ± 0.000	22.174 ± 0.081	19.984 ± 0.000	23.321 ± 0.020

Table 16: **Quantitative ablation of router-only training.** PSNR mean and standard deviation over three seeds. Avg. equally weights four scenes within each seed. Every row uses role BCE and zero per-Gaussian admission.

Composition	AG_0015	AG_0018	Park	Road
Camera-conditioned hard selection	0.3215	0.1792	0.1351	0.2816
Same-router image blending	0.0535	0.0391	0.0065	0.0772
Router-only opacity gating	0.0431	0.0386	0.0074	0.0853

Table 17: **Quantitative analysis of camera-router transitions.** Adjacent-frame RGB MAE at each BCE-only router's hard-switch frames, averaged over three seeds. Compositions share frames within each scene and seed. Paths have no reference RGB.

Composition	AG_0015	AG_0018	Park	Road
Camera-conditioned hard selection	0.3293	0.1703	0.1296	0.2892
Same-router image blending	0.0639	0.0460	0.0098	0.0865
Same-router opacity routing	0.0451	0.0390	0.0109	0.0820
Historical decoder	0.0451	0.0390	0.0109	0.0820

Table 18: **Quantitative analysis of source transitions.** Adjacent-frame RGB MAE at the same learned-router hard switch. These paths have no reference RGB.

camera center across both submaps. Distance weighting uses $\rho = (d_g + 10^{-8})/(d_g + d_a + 2 \cdot 10^{-8})$, where d_g and d_a are the nearest training-camera distances for each source. Image blending interpolates two separately rendered RGB images. Opacity routing applies the same coefficient before joint rasterization. These controls do not use validation RGB to set weights.

Router-only training crosses position-only and position-plus-direction inputs with role BCE and optional map-teacher supervision. Each of the 48 runs uses 3000 Adam updates at learning rate 10^{-3} and a shared role-balanced camera schedule for each seed. Both input conditions retain the same 18–32–32–1 network, with the three direction features set to zero in the position-only condition. All Gaussian attributes and registration remain frozen. The native role label selects the map teacher, and only rendered RGB enters its supported L1 and SSIM loss. The loss weights are 0.01 for BCE, 0.8 for L1, and 0.2 for SSIM. Random calibrated crops have at most 512×512 pixels, with no image resizing or Gaussian subsampling. Evaluation uses full-resolution native views and the original continuous paths. The final update is fixed before evaluation.

The following transition table reports the historical decoder and its matched router controls. Tab. 17 reports the BCE-only routers.

F.1 Fusion Computational Cost

The archived F2 map-teacher training records contain completed 30,000-step runs for 15 TerraSky3D and two Horizon-GS pairs. Over these 17 runs, mean recorded wall time is 13.27 minutes and mean peak PyTorch-allocated GPU memory is 4.29 GiB. These are retrospective run-level measurements, rather than an isolated equal-hardware throughput comparison. They exclude the independent submap reconstruction and registration stages.

Measurement	Time	Allocated GPU memory
30,000-step fusion (17-pair mean)	13.27 min	4.29 GiB peak
AG_0018 rendering (mean)	15.66 ms	5.05 + 1.10 GiB

Table 19: **Quantitative evaluation of map-teacher fusion cost.** Training reports archived run-level measurements; inference uses a separate full-resolution A100 benchmark.

Router	Composition	Native (ms)	Transition (ms)
Original	Raw union	9.64	7.37
Original	Nearest-camera selection	6.51	5.18
Original	Hard selection	6.66	5.40
Original	Image blending	12.43	9.94
Original	Opacity routing	6.93	6.04
Original	Full model	6.94	6.09
Router-only BCE	Raw union	9.72	7.33
Router-only BCE	Nearest-camera selection	6.46	5.13
Router-only BCE	Hard selection	6.59	5.43
Router-only BCE	Image blending	12.27	9.83
Router-only BCE	Opacity routing	6.92	6.85

Table 20: **Quantitative evaluation of composition latency.** Mean per-query median over 16 native and 16 transition cameras on an A100, with three warm-ups and ten repetitions. Source selection and image blending use cached full submap tensors.

For AG_0018, a separate inference benchmark uses 9,794,092 Gaussians at 3839×2159 resolution on one NVIDIA A100-SXM4-80GB. After 25 warm-up iterations, 100 timed iterations give a mean latency of 15.66 ms and a 95th-percentile latency of 15.71 ms. Timing includes camera routing, opacity decoding, and packed RGB rasterization, excluding checkpoint loading. Resident tensors occupy 5.05 GiB, with an additional peak allocation of 1.10 GiB during rendering, for approximately 6.15 GiB total allocated memory.

Fusion stores four ownership logits per Gaussian, of which two are active in the two-state decoder, together with 1697 camera-router parameters. For N Gaussians, this state scales as $4N+1697$ stored parameters; Gaussian geometry and appearance remain frozen. These measurements characterize the fusion stage. End-to-end cost also includes SRE construction, proposal inference, and registration; a hardware-matched stage-level profile and a centralized-retraining comparison are not available in these records.

Measured composition cost. Tab. 20 uses four evenly spaced native validation cameras per scene and path frames 60, 80, 100, and 120 at 1280×720. Each query has three warm-ups and ten timed repetitions with CUDA synchronization. Timing includes query weights and rendering, but excludes input loading. Hard selection and image blending use cached tensors containing all primitives of each selected source. Timing on the measured GPU excludes concurrent registration jobs. Peak allocation measures the evaluation harness's memory footprint, including cached joint and split tensors. With the BCE-only router, opacity routing takes 6.85 ms on the sampled transition queries, versus 9.83 ms for image blending and 5.43 ms for hard selection. This measures a rendering-cost trade-off, separately from the no-reference transition diagnostic.

A separate diagnostic measurement of eight compositions, including static mixture and both distance-weighted controls, uses the same query selection and timing protocol. It gives 6.78 ms for opacity routing and 9.78 ms for image blending. The BCE-only pose router's fixed 3000-update training loop averages 6.15 seconds over twelve runs. This timing covers the optimization loop and excludes data loading and initialization.

Camera-reference consistency. Fixed-registration comparisons express both submaps and their associated cameras using the same estimated transformation. Evaluating the fused model with ground-truth-transformed aerial cameras is a different experiment. Holding the `AG_0018` checkpoint, evaluator, 34 image IDs, and reference RGB bytes fixed, changing only to the latter cameras reduces aerial PSNR from 25.734 to 15.784 dB. Ground PSNR remains 26.335 dB. The historical aerial value of 17.385 dB comes from a separate historical run whose exact checkpoint is unavailable. Both historical and current records average the same 34 image IDs with 18 ground and 16 aerial views, giving 22.122 and 26.052 dB overall. The discrepancy is therefore within a scene,

Control	Full PSNR	Region PSNR
Hard selection	26.05	25.66
Per-Gaussian, native + cross	22.09 ± 0.04	11.05 ± 0.04
Source scalar, native + cross	16.40 ± 0.06	10.67 ± 0.03
Per-Gaussian, native only	25.30 ± 0.01	22.91 ± 0.04

Table 21: **Quantitative analysis of separate-region supervision.** AG_0018 development experiment with 1000 updates and three camera-order seeds. Region denotes the fixed alpha-defined mask, not missing-surface ground truth.

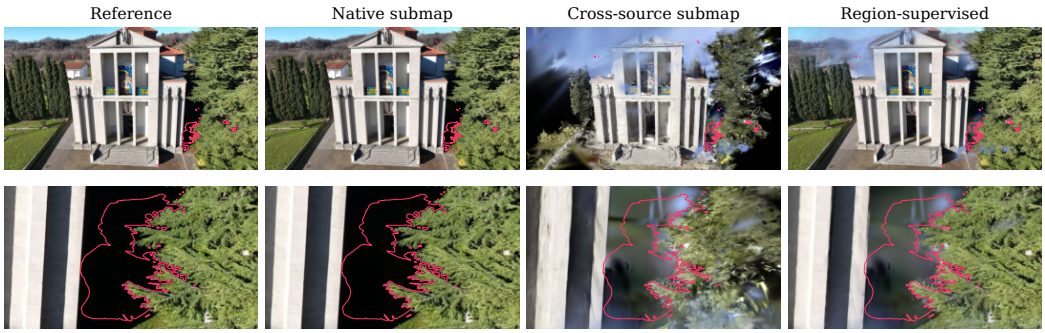

Figure 10: **Qualitative diagnosis of alpha-based cross-submap supervision.** The first aerial validation query of AG_0018 is shown with the supervision mask outlined. The lower row enlarges its largest connected component. Native and cross-source renders share the same camera. The trained variant uses separate-region supervision with seed zero.

rather than a change in dataset averaging. The camera intervention establishes sensitivity to the evaluation frame; its 1.601 dB aerial difference from the historical run remains unresolved.

F.2 Diagnostic Study of Alpha-Based Cross-Submap Supervision

Table 21 tests a development variant on `AG_0018`. Geometry, appearance, registration, and the router remain fixed. Cross-source admission starts at 0.1 and is optimized for 1000 updates using three camera-order seeds. Native teacher loss uses pixels with native alpha above 0.5. Cross-source teacher loss uses pixels with native alpha below 0.2 and opposite-source alpha above 0.6, with each loss normalized by its own region size. These masks cover 0.111% of evaluation pixels and are nonempty in 26 of 34 views.

Per-Gaussian admission reaches 22.09 dB, compared with 16.40 dB for two source-wise scalars under the same objective. Both remain below hard selection at 26.05 dB. Native-only supervision reaches 25.30 dB with the same initialization and budget. In the inspected aerial query, the alpha-defined region includes accurately rendered dark surfaces. The native and cross-source teachers reach 39.64 and 9.11 dB in this region, respectively. The less accurate cross-source target identifies a mask failure in this query; the full-image scores above cover all evaluation views.

Continuous-camera diagnostics. The supplementary trajectories interpolate camera position and orientation over 180 frames using the same poses for all composition controls. The interpolated cameras have no reference RGB, so these videos provide a visual comparison rather than a PSNR benchmark. Camera-router continuity and image continuity are distinct because alpha compositing also depends on visibility and depth ordering.

Table 22 and Fig. 11 compare hard selection, the historical decoder, and soft routing with zero opposite-source admission. Each path is selected by camera position and orientation, without RGB-based selection. At the hard-selection switch, the historical decoder reduces adjacent-frame RGB MAE from 0.3293 to 0.0451 on AG_0015, from 0.1703 to 0.0390 on AG_0018, from 0.1296 to 0.0109 on Park, and from 0.2892 to 0.0820 on Road. Zero-admission soft routing gives the same values at four-decimal precision. Its same-camera difference from the historical decoder averages between 1.07×10^{-5} and 6.94×10^{-5} in RGB MAE across these paths. The switching behavior is

Scene	Hard selection	Historical	Zero admission	Historical–zero MAE
AG_0015	0.3293	0.0451	0.0451	1.69e-05
AG_0018	0.1703	0.0390	0.0390	1.10e-05
Park	0.1296	0.0109	0.0109	1.07e-05
Road	0.2892	0.0820	0.0820	6.94e-05

Table 22: **Quantitative analysis of historical source transitions.** The first three columns give adjacent-frame RGB MAE at hard-selection switches. The last averages same-camera historical–zero RGB MAE over all 180 frames. Lower motion-induced change alone does not establish rendering fidelity.

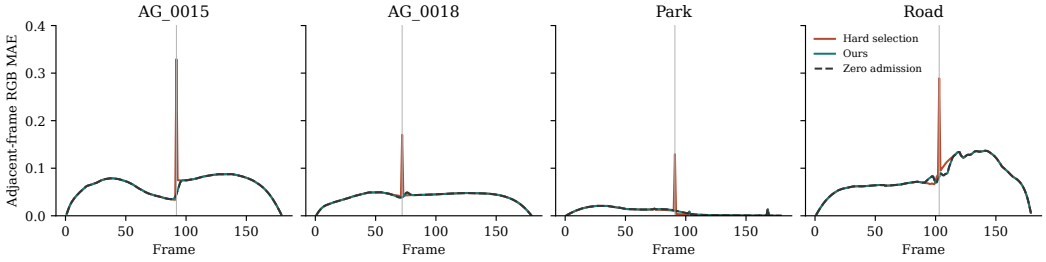

Figure 11: **Quantitative analysis of camera-conditioned source transitions.** This archived diagnostic uses the historical decoder, labeled Ours in the plot. Vertical markers locate its hard-selection switch.

therefore largely reproduced by camera-level interpolation. These unwarped differences also contain camera-induced motion and visibility changes, and do not measure temporal fidelity.