Supplementary Material for SCCM: Spherically Consistent Coarse Matching for ERP Dense Feature Correspondence

Gyeonggwan Lee[1,2], Eunsoo Im[1], Seunghwan Hong[1], and Junghun Suh[1]

[1] Kakao Mobility Corp., Seongnam, Republic of Korea
[2] Korea University, Seoul, Republic of Korea
{gandan.lee,soo.hd,logan.sh,jude.suh}@kakaomobility.com

A Implementation Details

Reproducibility. Tab. 1 summarizes the shared training protocol used for all models retrained on Matterport3D [2] (the Holo360D training protocol is in Sec. P); SCCM changes only the coarse-matcher modules, leaving the loss, encoder, and refiner architecture unchanged.
Numerical TPB details. For a ray $\mathbf{r} = (\cos\varphi \sin\lambda, \sin\varphi, \cos\varphi\cos\lambda)$, the orthonormal tangent frame $\mathbf{B_r} = [\mathbf{e}_\varphi\ \mathbf{e}_\lambda]$ (Sec. 4.1, main) is

$$\mathbf{e}_\varphi = \frac{\partial \mathbf{r}}{\partial \varphi} = (-\sin\varphi\sin\lambda,\ \cos\varphi,\ -\sin\varphi\cos\lambda), \qquad \mathbf{e}_\lambda = \mathbf{e}_\varphi \times \mathbf{r}. \tag{1}$$

$\mathbf{e}_\varphi$ has unit norm for every (φ, λ), including the poles; $\mathbf{e}_\lambda$ is automatically unit and orthogonal since $\mathbf{r} \perp \mathbf{e}_\varphi$, and is aligned with the $+\lambda$ direction. We read $(\sin\lambda, \cos\lambda)$ off the ray as $x/\cos\varphi$, $z/\cos\varphi$ with $\cos\varphi = \sqrt{\max(1-y^2, \epsilon_\mathrm{b})}$ ($\epsilon_\mathrm{b} = 10^{-7}$, distinct from the LAC clamp $\epsilon = 10^{-3}$ in Eq. (5), main); these ratios stay finite because $x, z \to 0$ together at the poles, so the frame is well-defined at every coarse-grid location (grid token centers never sample the exact poles, so this clamp is effectively inactive). We then evaluate the spherical log-map (Eq. (3), main) in a stable form with an inner-product clamp (keeping $\theta < \pi$) and a norm clamp (removing the 0/0 pole form), so the resulting bias is finite for every token pair. Because TPB forms an all-to-all ($N \times N$) table on the 32×64 coarse grid, exact antipodal pairs occur; these are only $O(N)$ of the $O(N^2)$ entries, and the norm clamp assigns them a bounded, deterministic bias value, so $\boldsymbol{\delta}_{q,k}$ is geometrically exact away from this sparse set.

Encoding and MLP. Before Fourier encoding, the offset is scaled per query by $\cos^\alpha\varphi_q$ with a single learnable scalar α (initialized at 0, i.e. the identity; the released Matterport3D checkpoint learns $\alpha = -0.13$), which lets the bias adapt its latitude conditioning. The scaled offset is encoded with 8 geometric frequencies $(1, 2, \ldots, 2^7)$ per tangent coordinate as (sin, cos) pairs (32 dimensions) and mapped by a $32 \to 32 \to 1$ MLP (GELU, zero-initialized output; 1,089 parameters); α and this MLP are the only learned parts of TPB.

Table 1: Shared training protocol for all models retrained on Matterport3D, used for the fixed-scaffold ablations.

Setting	Value
Optimizer	AdamW (β_1=0.9, β_2=0.999, wd 0.01); non-zero LR only on trainable modules
LR schedule	base 1×10^{-4}, constant, $\times0.1$ at step 112,500 (90% of the 125K steps)
Budget	1 M samples (≈125K steps); all rows identical
Loss	RoMa [5] dense regression + certainty BCE + coarse-grid CE, unmodified (SCCM adds none)
GT corresp.	spherical warp from per-pixel depth + relative pose (intrinsics-free ERP)
MP3D split	official 61/11/18 scenes = 54,015/4,625/15,682 pairs; native 1024×512
Augmentation	pair-shared horizontal flip ($p = 0.5$) and independent per-frame yaw $\theta_A, \theta_B \in [0, 2\pi)$ (EDM [6] convention), since a shared yaw would be a near no-op for yaw-equivariant architectures; no color/crop
Encoder	DINOv2-Large (patch 14), frozen; $448 \times 896 \to 32 \times 64$ tokens (N=2048)
Resolution	inputs resized to 448×896; metrics on the dense 448×896 refiner warp
Hardware	8×RTX 4090 (24 GB), batch 1/GPU, bf16 forward; ≈22 h/run

B Stanford2D3D Test Protocol

Frame source. Stanford2D3D [1] provides 1,413 equirectangular panoramas at 4096×2048 across 6 indoor areas (area 5 is split into $5a/5b$ as separately scanned wings). We drop one truncated panorama in area 3, leaving **1,412** usable panoramas (area 1: 190, area 2: 299, area 3: 84, area 4: 258, area 5a: 143, area 5b: 230, area 6: 208).

Coordinate harmonization. The dataset stores camera poses as 3×4 world-to-camera matrices in a $+y$-down image-frame convention and depth as 16-bit PNGs with scale 1/512 m. Two normalizations match our runtime ERP convention (Sec. 3, main): *(i)* a $\mathrm{diag}(1, -1, 1)$ y-flip on each pose's rotation block, so image-top corresponds to $+y$ (up); *(ii)* the /512 depth scaling (verified against the dataset documentation; the alternative /1000 yields depths $\approx 2\times$ too small and breaks the dense GT warp).

Pair generation. Within each area we enumerate all uni-directional pairs (i, j), $i < j$, and compute an EDM-style overlap score:

$$\mathrm{ov}(i, j) \;=\; \frac{|\{\, p \in V_i \;:\; |d_{i\to j}(p) - d_j(\Pi(p))| \,/\, d_j(\Pi(p)) < 0.1 \,\}|}{|V_i|},$$

where V_i is the set of valid-depth pixels of I_i, Π is the depth-and-pose-driven ERP reprojection from I_i to I_j, and the relative-depth consistency threshold 0.1 follows EDM [6]. We retain pairs with $\mathrm{ov}(i,j) \in [0.30, 0.80]$.
Resulting pair count.

Area	Frames	All pairs ($i<j$)	Retained ($0.30 \le \mathrm{ov} \le 0.80$)
area 1	190	17,955	1,057
area 2	299	44,551	3,403
area 3	84	3,486	233
area 4	258	33,153	1,178
area 5a	143	10,153	620
area 5b	230	26,335	1,118
area 6	208	21,528	1,135
Total	1,412	157,161	**8,744**

All 8,744 pairs are used at evaluation time; we do not subsample.
Difference from EDM's protocol. EDM [6] retains pairs with ov $>$ 0.50 (3,460 pairs). Our $[0.30, 0.80]$ band is chosen to probe generalization rather than to reproduce EDM's setting, and is not a superset of EDM's ov$>$0.5 set: the lower bound admits harder pairs (low covisibility, large yaw) that probe cross-dataset generalization, while the upper bound excludes highly overlapping near-duplicate pairs to focus the evaluation on more challenging viewpoints.

C Downstream Evaluation

The main paper summarizes both downstream tasks (relative pose and 3D reconstruction) in Sec. 5.4 under one unified protocol. This section provides the full protocol, the overlap-selected pose subset (Tab. 2), and top-down reconstructions (Fig. 1).
Pose. For each Matterport3D test pair we sample 2,000 matches by certainty-weighted multinomial sampling (RoMa [5]), unproject them to unit-ray pairs, and estimate the essential matrix with the same 8-point + RANSAC solver for every method, decomposing it to $(\mathbf{R}, \mathbf{t})$ by cheirality; the pose AUC of Tab. 3 (main) is aggregated over the 11,574-pair subset defined under Reconstruction below. To control for pair difficulty, we additionally report an overlap-selected subset (ov$>$0.5, 9,268 pairs; Tab. 2) under the same pipeline. All methods use this single evaluator, so differences reflect the matches rather than the estimator; this is *not* an EDM-official reproduction.
Reconstruction. For each test pair we triangulate stride-4 pixels of image A against their predicted matches under the GT pose (two-ray intersection, cheirality-filtered), using a shared per-pixel certainty gate $\tau = 0.5$—the same certainty signal the pose pipeline samples from (RoMa [5], DKM [3]). All methods are scored on the common 11,574-pair subset where every method yields $\geq$200 confident matches; the triangulated cloud is compared to the GT depth cloud for accuracy, completeness, and F-score at 5/10/20 cm. The ungated variant (no gate) is discussed in Sec. N.1.

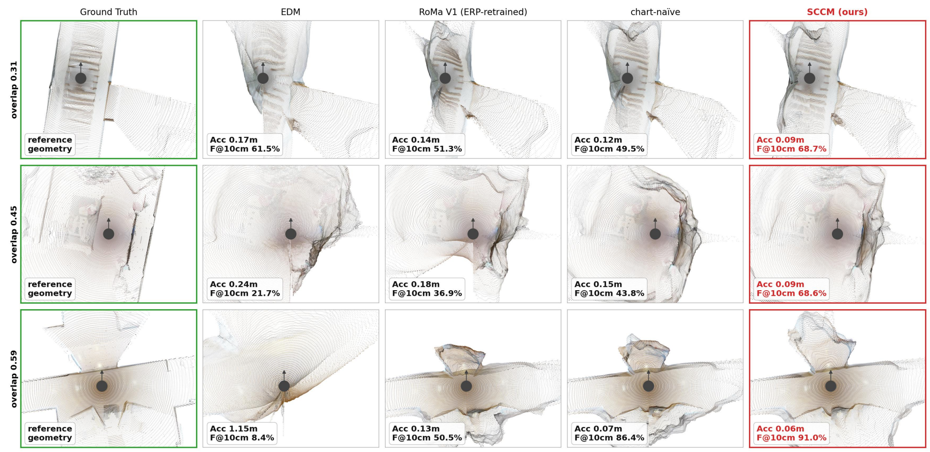

Fig. 1: Top-down reconstruction comparison on three representative test pairs (qualitative; aggregate reconstruction metrics are in Tab. 3, main). Each panel projects the triangulated cloud onto the X–Z plane (camera A at origin). On these examples the baselines (EDM, RoMa V1 (ERP-retrained)) and our chart-naïve scaffold produce distorted room contours, whereas SCCM's cloud (red) aligns more closely with the GT reference (green).

Table 2: Pose on the overlap-selected subset (ov > 0.5, 9,268 pairs) under our unified ray-essential evaluator. We report Pose-AUC using $\max(\text{rotation}, \text{translation})$ error. This is *not* an EDM-official reproduction. Best **bold**, second <u>underlined</u>.

Method	**AUC@5°**	**AUC@10°**	**AUC@20°**
EDM [6]	11.52	24.55	38.17
RoMa V1 (ERP-retrained) [5]	14.69	33.60	54.28
chart-naïve	<u>17.56</u>	<u>37.68</u>	<u>57.97</u>
SCCM	**22.06**	**42.36**	**61.17**

D Statistical Robustness: Bootstrap Confidence Intervals

The main ablation trains each configuration once (training-seed variance of its endpoints is bounded separately in Sec. R); here we quantify *evaluation-side* uncertainty (not training-seed variance) with 10,000 paired bootstrap resamples of the test pairs. Tab. 3 reports the SCCM − chart-naïve difference Δ, its 95% CI, and the win rate: every CI excludes zero and SCCM wins all 10,000 resamples on both Matterport3D and zero-shot Stanford2D3D. The same bootstrap confirms the scaffold gain (chart-naïve exceeds ERP-retrained RoMa V1 by +3.1 pp PCK@1°, no resample reversing it); the SCCM−chart-naïve gain itself is larger on zero-shot Stanford2D3D (+5.0 pp, 95% CI [+4.7, +5.4]).

Table 3: Pair-level bootstrap (10,000 paired resamples) on Matterport3D test (15,682 pairs) and zero-shot Stanford2D3D (8,744 pairs). PCK is fraction-of-pixels (↑); MAE/median in degrees (↓). Every Δ (SCCM − chart-naïve) 95% CI excludes zero, and SCCM wins all 10,000/10,000 resamples for every metric on both datasets. Best in **bold**. Values are computed at full precision from per-pair records collected in a separate evaluation pass; entries and Δs may differ from the main-paper tables in the last reported digits on both datasets (aggregation and evaluation-pass differences).

Set	**Method**	PCK@1°↑	PCK@3°↑	PCK@5°↑	MAE↓	Med↓
MP3D	chart-naïve	0.2299	0.6086	0.7674	5.667	2.225
	SCCM	**0.2748**	**0.6653**	**0.8059**	**5.360**	**1.875**
	Δ	+.045	+.057	+.039	−.307	−.350
	95% CI	[+.042, +.048]	[+.053, +.061]	[+.035, +.042]	[−.42, −.19]	[−.40, −.30]
Stanford2D3D	chart-naïve	0.1789	0.4746	0.6267	18.86	3.250
	SCCM	**0.2291**	**0.5637**	**0.6974**	**17.09**	**2.450**
	Δ	+.050	+.089	+.071	−1.77	−.800
	95% CI	[+.047, +.054]	[+.083, +.095]	[+.065, +.077]	[−2.40, −1.13]	[−.90, −.70]

E Sphere-Area-Weighted Metric

Because SCCM's largest gains are at high latitude (Fig. 4a, main), and a pixel-uniform PCK over-weights precisely those latitudes, one could suspect the aggregate gain is a polar over-counting artifact. To rule this out, we recompute every metric with per-pixel weights proportional to $\cos\varphi$ (sphere area). Table 4 shows the chart-naïve → SCCM gain persists on every metric under area weighting, on both Matterport3D and zero-shot Stanford2D3D, so the improvement is not caused by polar over-counting.

F Off-Gravity Behavior Beyond Upright ERP

SCCM assumes upright ERP. To probe this assumption we synthesize off-gravity captures from the test set: for a fixed pitch rotation R we resample both panoramas by R and conjugate the relative pose $T' = GTG^{-1}$ ($G = \mathrm{blkdiag}(R, 1)$), which preserves the GT correspondences up to resampling artifacts while tipping the chart off the gravity horizon. Table 5 reports PCK@1° for SCCM, its chart-naïve scaffold, EDM, and ERP-retrained RoMa V1 on the full test split under 10°/20°/30° pitch. All ERP matchers degrade off-gravity; SCCM is not tilt-equivariant, but owing to its larger upright margin it retains the highest *absolute* PCK@1° at every tested tilt. Moderate real-world tilt is a different regime: Holo360D pairs have a median relative tilt of 12°, and after training on such data SCCM leads there (Sec. P). Large synthetic pitch is outside SCCM's intended scope and can in practice be mitigated by IMU- or horizon-based rectification; a natively tilt-equivariant matcher remains future work.

Table 4: Sphere-area-weighted metrics: each pixel weighted by $\cos\varphi$ vs the pixel-uniform grid, on Matterport3D test and zero-shot Stanford2D3D. SCCM leads all baselines (EDM, ERP-retrained RoMa V1, the chart-naïve scaffold) on every metric under both pixel-uniform and sphere-area-weighted aggregation, on both datasets. The pixel-uniform columns reproduce the main-paper tables; the area-weighted columns re-weight the same test pixels by $\cos\varphi$.

		PCK@1°↑	PCK@3°↑	PCK@5°↑	MAE↓	Med↓
Matterport3D test						
EDM	pixel-uniform	0.163	0.400	0.509	16.78	4.79
	area-weighted	0.186	0.419	0.518	17.16	4.55
RoMa V1 (ERP-retrained)	pixel-uniform	0.198	0.538	0.711	6.60	2.76
	area-weighted	0.238	0.565	0.716	6.84	2.44
chart-naïve	pixel-uniform	0.229	0.609	0.769	5.68	2.23
	area-weighted	0.264	0.622	0.764	5.98	2.06
SCCM	pixel-uniform	**0.275**	**0.665**	**0.806**	**5.36**	**1.88**
	area-weighted	**0.303**	**0.667**	**0.795**	**5.74**	**1.78**
Zero-shot Stanford2D3D						
EDM	pixel-uniform	0.104	0.262	0.355	35.70	10.69
	area-weighted	0.136	0.302	0.387	35.11	9.68
RoMa V1 (ERP-retrained)	pixel-uniform	0.167	0.452	0.628	17.80	3.45
	area-weighted	0.215	0.494	0.643	17.27	3.06
chart-naïve	pixel-uniform	0.179	0.475	0.627	18.86	3.26
	area-weighted	0.221	0.504	0.638	17.99	2.96
SCCM	pixel-uniform	**0.229**	**0.564**	**0.697**	**17.09**	**2.45**
	area-weighted	**0.274**	**0.588**	**0.710**	**16.25**	**2.20**

G Rationale for Backbone and Baseline Training Choices

Why not V2. RoMa V2 [4] does not provide a publicly reproducible training recipe comparable to V1 (loss, data loaders, schedule, 10-dataset mixture), so V2-based variants cannot be reproducibly retrained on ERP under our controlled 1 M-sample protocol. Moreover, V2's intertwined positional changes (normalized-grid RoPE, fixed frequency $\omega = 1$, multi-view Transformer with alternating attention) interact directly with the attention-side modules of SPA, so re-applying SPA on V2 would confound "what the V2 positional changes provide" with "what our sphere-aware prior provides."

V1 is the appropriate scaffold for controlled ablation. On the fixed chart-naïve attention scaffold (cross-attention + dual-softmax, no PE) that replaces V1's GP coarse stage, we add our yaw-periodic RoPE, tangent-plane bias, and AAC cumulatively; this is the scaffold used for the cumulative ablation (Tab. 2, main) and the factorial decomposition (Sec. M). V2 is retained as a zero-shot perspective baseline in Tab. 1 (main), where its released inference checkpoint suffices.

EDM baseline. EDM's [6] public release provides inference code and a checkpoint trained on Matterport3D; the training code, loss, and data pipeline are not released. Retraining EDM under our 1 M-sample protocol would therefore

Table 5: Off-gravity tilt robustness on Matterport3D test (15,682 pairs). A fixed SO(3) pitch is applied to both views and the relative pose is conjugated accordingly, so the ground truth remains consistent up to resampling artifacts. We report PCK@1° under increasing pitch. All ERP matchers degrade off-gravity; SCCM is not tilt-equivariant but retains the highest absolute accuracy at every tested angle. Best per column in **bold**.

Method	upright (resampled)	pitch 10°	pitch 20°	pitch 30°
EDM [6]	0.163	0.105	0.044	0.018
RoMa V1 (ERP-retrained) [5]	0.198	0.096	0.042	0.022
chart-naïve	0.230	0.099	0.043	0.023
SCCM (ours)	**0.273**	**0.142**	**0.064**	**0.034**

require re-implementing its full training stack, with the same reproduction confound noted for V2 above. We instead evaluate EDM from its released in-domain checkpoint (trained by its authors on the same Matterport3D data under their own recipe and budget), and separately evaluate its input-side absolute spherical PE as a mechanism-only control retrained under the identical protocol (Tab. 2, main; Sec. J).

Portability to other matchers. SPA and AAC attach to two generic interfaces rather than to RoMa-specific components: SPA to the pairwise attention logits of the coarse matcher, and AAC to its per-pixel covisibility logits. LoFTR-style matchers [12] (cross-attention + dual-softmax) expose both; our chart-naïve scaffold R1 *is* such a pipeline, so the cumulative ablation (Tab. 2, main) demonstrates both modules on it. GP-based coarse stages (DKM [3], original RoMa V1 [5]) have no coarse attention logits, so only AAC applies there; SPA would first require replacing the GP stage with an attention-based one, which is exactly the R1 substitution.

H Yaw-Periodic RoPE: Frequency Choice

Restricting the longitudinal RoPE frequencies to integers ($\omega_i^\lambda \in \mathbb{Z}$, Sec. 4.1 (i) main) is the only choice consistent with 2π-periodicity at the chart seam: any fractional ω^λ shifts the cross-seam relative phase by a non-multiple of 2π and breaks the cyclicity SPA enforces. This does not limit resolution at our operating grid: at coarse grid width $W = 64$ the longitudinal Nyquist limit is 32 cycles per 360°, and our $\omega_{\max} = d_h/4 = 16$ (per-head dimension $d_h = 64$) sits at half Nyquist—so the finest encoded period (22.5°) spans about four token spacings ($360°/64 \approx 5.6°$), comfortably above the sampling limit. The integer constraint also does not cost accuracy in practice: the geometric-frequency control (Tab. 2, main) shows that non-integer frequencies underperform the integer-constrained design. Only the longitudinal half of each head is restricted; the latitude half keeps the standard geometric schedule ($\omega_i^\varphi = \theta_0^{-2i/(d_h/2)}$, $\theta_0 = 10000$), so fine latitude discrimination is preserved as in the latitude subspace of standard RoPE.

I Chart-naïve Coarse Matcher (R1 Backbone)

The chart-naïve backbone (row R1) keeps RoMa V1's [5] frozen DINOv2-Large [9] encoder and the ConvRefiner *architecture* (the refiner weights, like all non-encoder modules, are trained from scratch), and replaces *only* V1's Gaussian-process coarse matcher with a standard attention + dual-softmax coarse stage. It carries *no* sphere-aware component—no yaw-periodic RoPE, tangent-plane bias, EDM abs. PE, or log-area correction—and every ablation row, the EDM abs. PE control, and SCCM is this same backbone augmented with one or more such terms, which is what makes the comparison controlled.
Coarse stage. Coarse features $\mathbf{x}_A, \mathbf{x}_B \in \mathbb{R}^{N\times d}$ ($d = 512$, $N = H \times W$ the coarse grid) pass through a cascade of $n_{\text{blocks}} = 4$ Cross $\to$ Self attention blocks (8 heads, each with a feed-forward sub-layer) using *no* positional encoding and *no* relative-position bias. Assignment is by dual-softmax

$$\mathbf{P} = \text{softmax}_{\text{row}}(\mathbf{S}/\tau) \odot \text{softmax}_{\text{col}}(\mathbf{S}/\tau), \quad \mathbf{S} = \mathbf{x}_A \mathbf{x}_B^\top, \quad \tau = 0.1, \tag{2}$$

and the soft correspondence (the expected target coordinate under the row-normalized $\mathbf{P}$) is passed to the RoMa V1 ConvRefiner.
Covisibility gate. A per-query head produces a covisibility logit $\ell_p = \text{MLP}(\mathbf{x}_p)$ (128-d hidden) mapped to a gate $\mu_p = \sigma(\ell_p)$ that down-weights non-covisible queries. In R1 this is the bare MLP; AAC (Sec. 4.2, main) adds the single log-area term on top of this same logit. All components except the frozen encoder are randomly initialized and trained from scratch under the protocol of Sec. A.

J EDM Positional Embedding: Re-implementation

Scope of this probe. The purpose here is to isolate a single design axis—an *absolute, input-side* sphere-coordinate positional encoding (as in EDM) vs. our *relative, intra-attention* bias (SPA)—under one identical training protocol, data split, and evaluator. To do so we place EDM's positional embedding (EDM Eq. (1) and Eq. (9)) on the same R1 backbone (cross-attention + dual-softmax) used by SCCM. We implement EDM's published PE equations rather than using EDM's full system. Any conclusion drawn here therefore bears on the *absolute-PE mechanism itself*, not on EDM as a whole system (which also has a distinct backbone and a geodesic-flow refiner).
Formulation. For each coarse-grid token at latitude/longitude (φ, λ) we form the 3D unit ray (EDM Eq. (1))

$$\mathbf{r}(\varphi, \lambda) = (\sin\lambda\cos\varphi,\ \sin\varphi,\ \cos\lambda\cos\varphi), \tag{3}$$

the same unit ray $\mathbf{r}(\varphi, \lambda)$ used by our TPB (Sec. 4.1, main). The embedding (EDM Eq. (9)) is

$$\boldsymbol{\chi} = \cos(\mathbf{W}_{\text{PE}}\,\mathbf{r} + \mathbf{b}), \qquad \mathbf{W}_{\text{PE}} \in \mathbb{R}^{D\times 3},\ \mathbf{b} \in \mathbb{R}^{D}, \tag{4}$$

with $\mathbf{W}_{\text{PE}}$ a learned 1×1 projection $(3\to D)$, $\mathbf{b}$ a learned bias, and cos element-wise. $\boldsymbol{\chi}$ is added to *both* token sets *before* the attention cascade ($\mathbf{x}\leftarrow\mathbf{x}+\boldsymbol{\chi}$, $\mathbf{y}\leftarrow\mathbf{y}+\boldsymbol{\chi}$). It is purely absolute (per-token, independent of the query–key pair): there is no relative-phase mechanism, which is precisely the property that distinguishes it from SPA.

Identity initialization (fair start). We zero-initialize $\mathbf{W}_{\text{PE}}$ and set $\mathbf{b}=\frac{\pi}{2}$, so $\boldsymbol{\chi}=\cos(\frac{\pi}{2})=\mathbf{0}$ at step 0: the model is functionally equivalent to R1 at initialization, matching the identity-at-start contract of TPB (zero-initialized); LAC instead starts at the analytic area value $\alpha_{\text{LAC}}=1$, not at identity. Setting $\mathbf{b}=\frac{\pi}{2}$ (rather than 0) keeps the gradient live ($\partial\cos/\partial(\cdot)=-\sin(\frac{\pi}{2})=-1$), so the projection trains from the first step rather than from a dead point.

Reproducibility. We implement only the EDM absolute sphere-coordinate embedding (unit ray Eq. (1) $\to$ cos of a learned $3\to D$ projection Eq. (9) $\to$ added to both token grids before the R1 coarse matcher); all other EDM components are not used.

K AAC/LAC Numerical Sanity Checks

Clamp sensitivity. The numerical safety clamp $\epsilon=10^{-3}$ in Eq. (5) (AAC) of the main paper is active only in an extremely thin numerical pole band ($|\varphi|>89.94°$, $\sim$0.06% of latitudes); all remaining pixels see the exact unclamped $\cos\varphi$. It acts on the non-trainable $\cos\varphi$ buffer, so the learnable α_{LAC} (entering only as $\alpha_{\text{LAC}}\log[\cdot]$) retains a well-defined gradient dominated by the unclamped majority. Tab. 6 shows that varying ϵ over $[10^{-4},10^{-2}]$ leaves every metric unchanged.

Table 6: Sensitivity to the numerical safety clamp ϵ used in Eq. (5) of the main paper, on the full Matterport3D *test* split (15,682 pairs) with the full SCCM model. Varying ϵ over $[10^{-4},10^{-2}]$ leaves all metrics unchanged at reported precision, confirming that the clamp is not driving the result. PCK ↑; MAE/median in degrees ↓.

ϵ	PCK@1°↑	PCK@3°↑	PCK@5°↑	MAE°↓	Med°↓
10^{-4}	0.2748	0.6653	0.8059	5.360	1.875
10^{-3} (used)	0.2748	0.6653	0.8059	5.360	1.875
10^{-2}	0.2748	0.6653	0.8059	5.360	1.875

LAC strength. The learned LAC strengths remain near the analytic area value ($\alpha^{A}_{\text{LAC}}=0.978$, $\alpha^{B}_{\text{LAC}}=1.028$; init $\alpha=1$). Overriding them to $\alpha=1$ at inference time leaves the metrics unchanged to within 10^{-4} (Tab. 7), supporting that LAC's contribution comes from the pre-sigmoid area term itself rather than scalar tuning. (This isolates the *final* scalar value; fully excluding any effect of keeping α learnable *during training* would require a separate fixed-α run.)

Table 7: LAC strength sanity check on the full Matterport3D test split. The trained SCCM checkpoint is evaluated with its *learned* per-view LAC strengths $(\alpha^A_{\text{LAC}}, \alpha^B_{\text{LAC}})$ and with α *fixed to* 1 *at inference time*. The two settings yield metrics identical to within 10^{-4}, supporting that LAC's contribution comes from placing the analytic log-area term before the sigmoid rather than from scalar tuning.

Model	$\alpha^A_{\text{LAC}}/\alpha^B_{\text{LAC}}$	PCK@1°↑	PCK@3°↑	PCK@5°↑	MAE°↓
SCCM (learned α)	0.978/1.028	0.2748	0.6653	0.8059	5.360
SCCM ($\alpha=1$, infer.)	1.000/1.000	0.2748	0.6654	0.8059	5.360

L Computational Complexity Analysis

Tab. 8 reports trainable parameters and FLOPs. SCCM adds only a parameter-free yaw-periodic RoPE, an $\approx$1.1K-parameter TPB MLP, and one LAC scalar per covisibility side; the TPB bias costs $\approx$8.9 GFLOPs ($\approx$1.2%) when recomputed every forward and nothing when cached. Tab. 9 reports latency: uncached TPB recomputation is memory-bound, but caching the content-independent TPB table reduces the deployed overhead to $\sim$5%. We detail the FLOPs bound and latency protocol below.

Table 8: Computational footprint relative to the R1 chart-naïve baseline: trainable-parameter and FLOP additions at the medium ERP resolution (448×896). Inference latency is reported separately in Tab. 9.

Configuration	**Trainable params**	**GFLOPs**
(R1) chart-naïve (cross-attn+DS)	137.17 M	751.15
+ SPA (RoPE + TPB)	$+\approx$1.1 k	$\approx$9.0 ($\approx$0.1 cached)
+ AAC (LAC)	+2	$\approx$0.0
SCCM total	137.17 M	$\approx$760.2 ($\approx$751.3 cached)
Increase vs. baseline	$<$0.002%	$\approx$1.2% ($\approx$0.01% cached)

L.1 FLOPs

The chart-naïve baseline forward pass measures **751.15** GFLOPs at the medium ERP resolution 448×896 (`thop`, RTX 4090). The SCCM additions are bounded analytically. RoPE, a per-channel rotation of queries and keys in each attention layer, costs about 0.1 GFLOPs in total, and AAC is an N-element pointwise term with negligible cost. TPB evaluates its $32\to32\to1$ MLP on all $N^2 = 2048^2$ token pairs, $\approx$4.4 GMACs $\approx$8.9 GFLOPs, i.e. $\approx$1.2% of the 751 GFLOPs baseline when the bias is recomputed every forward, and 0 when the content-independent bias is cached (Sec. L.2). The wall-clock overhead of the uncached path (Tab. 9) exceeds this compute share because it is memory-bound (an $N\times N\times32$ intermediate) with kernel-launch overhead; caching eliminates it, leaving a deployed $\sim$5%.

L.2 Inference Latency Protocol

The latency values in Tab. 9 are wall-clock times on an RTX 4090, single image pair at the medium ERP resolution, mean over 100 forward passes after a 10-iteration warm-up. The pairwise tangent log-map buffer (Sec. 4.1 (ii), main) is precomputed and excluded from the per-forward timing, consistent with deployment.

TPB bias caching. The TPB bias is a fixed grid-geometry prior at inference: it depends only on the token positions, not on image content, so it can be precomputed once and reused for every pair; the cache is behavior-preserving and disabled during training. Tab. 9 reports a back-to-back A/B on a single RTX 4090: recomputing the bias each forward costs +20.9%, whereas caching it reduces SCCM's overhead to +5.0%. We therefore regard SCCM's deployed runtime cost as the cached ∼5%, the larger figure being an artifact of redundant recomputation rather than added computation.

Table 9: TPB bias caching latency. Wall-clock latency on a single RTX 4090 at medium ERP resolution, averaged over 100 timed forwards after warm-up. The cached variant precomputes the content-independent TPB bias for the fixed ERP grid and reuses it at inference time.

Method	**Latency (ms)**	**Overhead**
chart-naïve	115.76	—
SCCM (TPB uncached)	139.95	+20.90%
SCCM (TPB cached)	121.60	**+5.04%**

M Per-Band Ablation Decomposition

This section adds the per-latitude and per-longitude decomposition of each module's contribution (Fig. 2), complementing the body ablation (Tab. 2, main; Sec. 5.3).

From Fig. 2, the three priors split cleanly by axis. *Yaw-periodic RoPE* provides the dominant PCK@1° gain (+3.8 pp over R1) and stays positive across all longitudes, including the seam, and toward the pole; this is specific to its 2π-periodic integer-frequency design—substituting standard geometric-frequency RoPE (Tab. 2, main, [§]) yields only +0.7 pp and *degrades* MAE ($5.68° \to 6.79°$, vs. a +0.19° regression for yaw-periodic RoPE). *TPB* mainly reduces MAE ($5.87° \to 5.58°$) with little change in PCK@1° or median, consistent with metric calibration rather than improving near-threshold precision in isolation. In the cumulative R2b→R3 step, *LAC (AAC)* adds a smaller but consistent improvement across all three PCK thresholds and the error metrics (reattributed to the TPB×LAC interaction in the factorial decomposition below), holding across latitude in MAE and in PCK@1° except a slight dip in the 85–90° band, with

α_{LAC} settling near the analytic area value (0.978/1.028 per view)—supporting its role as a pre-sigmoid area-correction term rather than scalar tuning (Sec. K).

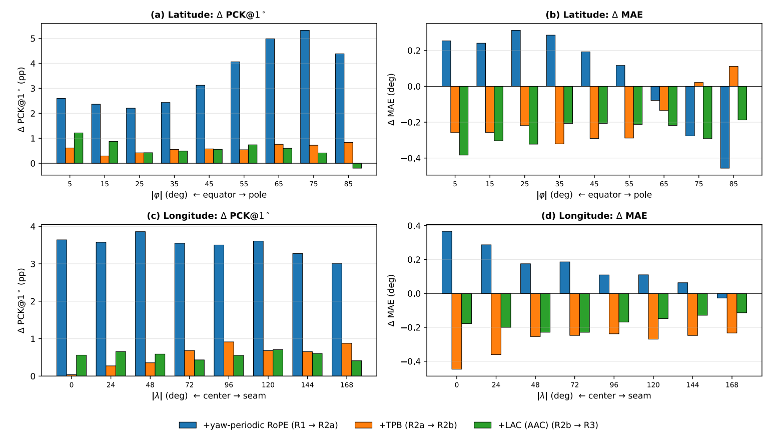

Fig. 2: Band-wise marginal decomposition of R1→R2a→R2b→R3 (Tab. 2, main). Bars show each module's marginal effect: blue +yaw-periodic RoPE (R2a−R1); orange +TPB (R2b−R2a); green +LAC/AAC (R3−R2b). Left column: ΔPCK@1° (pp); right column: ΔMAE (deg). Top row: latitude bands $|\varphi|$; bottom row: longitude bands $|\lambda|$. Positive ΔPCK and negative ΔMAE indicate improvement.

TPB error-tail quantiles. To verify that TPB acts on the angular-error tail rather than strict threshold precision, we compare pixel-level angular-error quantiles of R2a and R2b on the full Matterport3D test set (Tab. 10). TPB leaves PCK@1° and the median (P50) essentially unchanged, but reduces the high-error quantiles (P90/P95/P99 by 0.41°/1.80°/4.96°). This confirms the interpretation in Tab. 2 (main) that TPB calibrates the pairwise metric and suppresses large-error tails rather than improving near-threshold precision.

Factorial decomposition: the two priors interact. The cumulative ladder of Tab. 2 (main) credits each gain to the last added term. To separate the roles of TPB and LAC we train the missing combination (yaw-periodic RoPE + LAC, no TPB) under the identical protocol and re-evaluate all four cells in a single evaluation pass, completing a 2×2 factorial over the R2a base (Tab. 11). In isolation, neither prior moves PCK@1° (-0.17 pp, 95% CI $[-0.37, +0.04]$; -0.14 pp, $[-0.35, +0.07]$), yet jointly they add $+0.70$ pp—a positive interaction of $+1.00$ pp (95% CI $[+0.72, +1.28]$), positive in all 10,000 paired resamples. MAE behaves oppositely: each prior lowers it independently ($-0.21°/-0.26°$, both CIs excluding zero) with no detectable interaction ($+0.05°$, $[-0.10, +0.20]$); the conditional and marginal MAE effects nearly coincide ($-0.17°/-0.22°$ vs. $-0.21°/-0.26°$), confirming additivity. Under this fixed scaffold, the strict-PCK@1° gain thus emerges only when TPB and LAC are combined, consistent with their complementary roles: LAC corrects *how much* matching mass survives the area-aware

Table 10: TPB error-tail quantiles on Matterport3D test (full split, ∼3.0B valid pixels). R2a is yaw-periodic RoPE only; R2b adds TPB (full SPA). Angular-error quantiles Pk are in degrees; in the Δ row, PCK@1° is in percentage points and all other columns in degrees. The ΔPCK@1° of −0.05 pp is computed from full-precision values before rounding (the body Tab. 2 rounds the two rows to 0.267 and 0.266, i.e. −0.1 pp). TPB leaves strict precision (PCK@1°) and the median (P50) nearly unchanged but lowers the high-error quantiles, confirming its role as tail calibration.

Row	PCK@1° ↑	MAE↓	P50↓	P75↓	P90↓	P95↓	P99↓
R2a (RoPE)	0.2667	5.87	1.92	4.18	9.29	19.49	94.23
R2b (full SPA)	0.2662	5.58	1.94	4.13	8.88	17.69	89.27
Δ	−0.05	−0.29	+0.02	−0.05	−0.41	−1.80	−4.96

gate, TPB corrects *where* the surviving mass ranks under the spherical metric; fixing either alone leaves the coarse argmax dominated by the remaining distortion. The per-latitude decomposition supports this reading: the PCK@1° interaction is positive in every $|\varphi|$ band and largest at high latitudes (+1.25/+1.16 pp at 60–75°/75–90° vs. +0.8–1.0 pp below 60°), where both chart distortions are strongest. This also explains why the cumulative ladder reads TPB's PCK@1° marginal as ≈0 (R2a→R2b): the joint gain is folded into the final row.

N Analysis of Remaining Errors

SCCM's residual errors fall into two groups. **(a) Geometry-specific failures** occur outside the intended upright-ERP setting: a large camera pitch/roll breaks the gravity-aligned prior and accuracy drops sharply (PCK@1° 0.273→0.034 at a 30° tilt; Fig. 3, Sec. F). Within upright ERP, the numerical clamps (Secs. A and K) introduce no approximation error, since no coarse-grid token lies in their pole band. **(b) Generic dense-matching failures** remain in texture-less, reflective/HDR, low-overlap, and depth-discontinuity regions; these are shared by perspective and ERP matchers alike (cf. RoMa [5], EDM [6], LoFTR [12]), are driven by visual ambiguity rather than chart geometry, and are orthogonal to the spherical priors—requiring upstream changes (a stronger encoder, photometric-invariant features, occlusion handling).

N.1 Certainty-gated vs. ungated reconstruction

The ungated variant is not our main protocol because all evaluated dense matchers produce certainty estimates, and the pose pipeline already uses the same certainty signal for sampling. When all covisible pixels are triangulated without the shared certainty gate, SCCM's mean accuracy is affected by a rare low-certainty tail, while its median accuracy and majority-pair performance remain better than chart-naïve. The shared $\tau = 0.5$ gate removes this tail for every method under the same rule, which is why the main table reports the certainty-gated protocol.

Table 11: TPB×LAC factorial on Matterport3D test (15,682 pairs): all four combinations over the R2a (yaw-periodic RoPE) base under the shared protocol of Tab. 1. Effects are paired-bootstrap estimates (10,000 resamples). All four cells are evaluated in a single evaluation pass; the body table and Tab. 10 stem from earlier passes, and repeated evaluation of the same checkpoint can shift the last reported digits through floating-point nondeterminism (observed up to $\sim 10^{-3}$ in PCK@1° and $\sim 0.1°$ in MAE). The body table's final-row marginal (+0.9 pp) corresponds to LAC given TPB here (+0.86 pp), i.e., the interaction plus LAC's solo effect. Neither prior improves strict precision alone; their combination does.

Cell	TPB	LAC	PCK@1°↑	PCK@3°↑	PCK@5°↑	MAE↓
R2a (RoPE)	✗	✗	0.2679	0.6558	0.7987	5.789
R2b (+TPB)	✓	✗	0.2662	0.6521	0.7989	5.576
+LAC (no TPB)	✗	✓	0.2664	0.6583	0.8017	5.527
R3 (SCCM)	✓	✓	**0.2748**	**0.6653**	**0.8059**	**5.360**

Effect (95% CI)	ΔPCK@1° (pp)	ΔMAE (deg)
TPB alone (R2b−R2a)	−0.17 [−0.37, +0.04]	−0.21 [−0.33, −0.10]
LAC alone	−0.14 [−0.35, +0.07]	−0.26 [−0.37, −0.16]
TPB given LAC	+0.83 [+0.62, +1.05]	−0.17 [−0.28, −0.05]
LAC given TPB (R3−R2b)	+0.86 [+0.66, +1.06]	−0.22 [−0.32, −0.11]
Interaction	**+1.00** [+0.72, +1.28]	+0.05 [−0.10, +0.20] (n.s.)

O Refinement-Stage Diagnosis

To locate where the mechanism-isolated SCCM gain enters the pipeline, we measure PCK@1° after the coarse stage and after each refinement stage of the shared cascade, on the full Matterport3D test split (no masking or object filtering; ∼3.0B valid pixels). Tab. 12 shows that SCCM's advantage over the chart-naïve scaffold is already present at the coarse-anchor stage (+4.1 pp) and remains essentially unchanged through refinement (+4.6 pp at the final dense output), with small non-monotonic fluctuation across intermediate stages (+4.1 to +5.4 pp). The architecturally unchanged RoMa V1 refiner provides a similar coarse-to-final lift for both models (+10.7 pp for chart-naïve, +11.2 pp for SCCM), preserving and slightly amplifying the coarse-stage advantage rather than creating it. This supports our interpretation that the ERP-specific gain enters primarily through coarse-anchor formation rather than being created by the refiner.

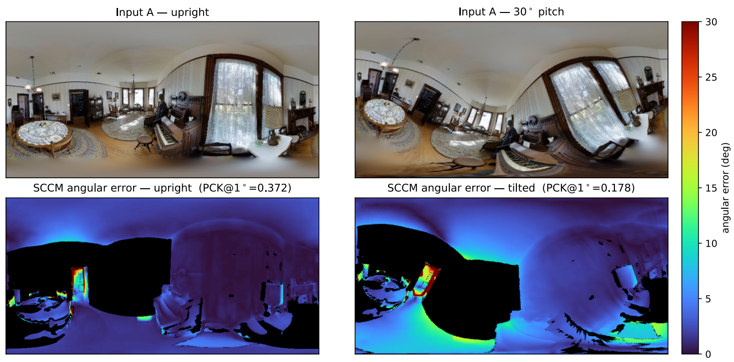

Fig. 3: SCCM's off-gravity failure mode, shown on a single Matterport3D test pair (the *same* pair throughout). *Top*: input view A upright vs. under a synthetic 30° off-gravity pitch (Sec. F), which tips the ERP chart off the gravity horizon. *Bottom*: SCCM's per-pixel angular error for the warp $A \to B$ (black = non-covisible; colorbar 0° to $\geq 30^\circ$). Upright, SCCM attains PCK@1° = 0.37 on this pair (mostly low error); under tilt the gravity-aligned prior no longer matches the tilted chart and accuracy drops markedly (PCK@1° = 0.18), with error concentrated where the chart distortion is largest. Holding the pair fixed isolates tilt—not scene content—as the cause; full-set tilt averages are in Tab. 5. This is a geometry-specific failure outside the intended upright-ERP setting.

Table 12: Refinement-stage PCK@1° diagnostic on the full Matterport3D test split (no masking/object filtering). PCK@1° after the coarse stage and after each refinement stage. SCCM's advantage is present before refinement and preserved through the architecturally unchanged RoMa V1 refiner; the similar coarse-to-final lift for both models indicates that the gain enters primarily through coarse-anchor formation rather than being created by the refiner.

Stage	chart-naïve	SCCM	Δ (pp)
coarse anchor (1/16)	0.122	0.163	+4.1
refiner stage (1/8)	0.133	0.187	+5.4
refiner stage (1/4)	0.188	0.234	+4.6
refiner stage (1/2)	0.219	0.263	+4.4
final dense	**0.229**	**0.275**	+4.6

P Outdoor Evaluation on Holo360D

This section details the outdoor experiment summarized in Tab. 1(c) of the main paper. Holo360D [10] is an in-the-wild dataset captured with a handheld LiDAR scanner coupled to a 360° camera along continuous trajectories. Its outdoor sequences are far deeper and sparser in supervision than Matterport3D (median scene depth 4.3 m vs. 2.0 m; 54% vs. 95% of ERP pixels carry valid depth) and are not upright (median relative tilt between paired views 12°; 79% of pairs exceed 5°).

Protocol. The three retrained matchers of Tab. 1 (main)—the ERP-retrained RoMa V1 (V1-GP), the chart-naïve scaffold (R1), and SCCM—are initialized from their Matterport3D checkpoints and trained on Holo360D under one identical protocol (25k optimizer steps, learning-rate scale 0.1, same augmentation and loss), on scene-disjoint splits of 5 training, 3 validation, and 4 test scenes, and score 8,000 held-out test pairs (2,000 per test scene; overlap 0.3–0.8) with the angular evaluator of the main paper. One component differs for the SCCM row only: during its Holo360D training the RoPE rotation angles are multiplied by a gain $\gamma \sim U[0,1]$ drawn at every forward pass, a stochastic regularizer on the RoPE module; at evaluation $\gamma = 1$, which is bit-identical to the yaw-periodic RoPE of the main paper. The two baselines carry no RoPE, so the term has no counterpart there, and it is not used for any Matterport3D or Stanford2D3D result.

Table 13: Outdoor training on Holo360D (8,000 test pairs). All three matchers start from their Matterport3D checkpoints and are trained under one protocol; SCCM leads on every metric, including the two error statistics. Best in **bold**.

Method	PCK@0.35°↑	PCK@0.5°↑	PCK@1°↑	PCK@3°↑	PCK@5°↑	MAE°↓	Med.°↓
RoMa V1 (ERP-retrained)	0.123	0.182	0.322	0.592	0.726	5.66	2.12
chart-naïve (R1)	0.119	0.180	0.331	0.626	0.761	5.02	1.93
SCCM	**0.137**	**0.202**	**0.357**	**0.645**	**0.776**	**4.94**	**1.76**

Result. SCCM leads on all seven metrics (Tab. 13), including the two stricter thresholds not shown in the main table (+1.8 pp at 0.35°); the +2.6 pp margin over R1 is more than four times the seed standard deviation of Sec. R.

Zero-shot transfer. Applying the same Holo360D-trained models *unchanged* to a second outdoor corpus (Mapillary Metropolis [8]; 2,484 street-level pairs resampled to match the rotation distribution of the Holo360D test set) tests transfer without any adaptation. SCCM retains its lead at tight thresholds (+1.9 pp PCK@0.35°, +0.5 pp PCK@1° over R1) but not at 3°–5° or in mean angular error (10.08° vs. 9.39° for R1), i.e. its zero-shot advantage is confined to precision. We therefore claim outdoor gains under in-domain training and report the zero-shot outdoor limit in the main paper (Sec. 6).

Q Additional Zero-Shot Baselines on ERP

Tab. 14 extends Tab. 1 of the main paper with four further published matchers, evaluated zero-shot on ERP under the same angular evaluator, each released checkpoint run at its own operating resolution with the ground truth on that grid (as for EDM in Sec. R). All four fall well below EDM, the strongest released ERP-native baseline, confirming that perspective-trained dense matchers and

3D foundation models do not transfer to the ERP chart without sphere-aware design.

Table 14: Additional zero-shot baselines (PCK@1°). †Sparse matchers scored on their own matches only. EDM from Tab. 1 for reference.

Method	Matterport3D	Stanford2D3D
DKM [3] (CVPR'23)	0.035	0.026
MASt3R† [7] (ECCV'24)	0.057	0.028
LoFTR† [12] (CVPR'21)	0.105	0.058
VGGT [13] (CVPR'25)	0.000	0.000
EDM [6] (reference)	0.163	0.104

R Additional Controls: Seeds, Resolution, and Pose Solver

Training-seed variance. The main ablation trains each configuration once. To bound training-side variance we retrained the two endpoints of the ablation, chart-naïve (R1) and SCCM, with two additional seeds each under the identical protocol and re-scored the full Matterport3D test split (Tab. 15). The R1→SCCM margin reproduces on every seed (+4.6/+5.3/+4.9 pp PCK@1°); the per-configuration seed standard deviation is 0.4–0.6 pp, so the margin sits an order of magnitude above seed noise, whereas the +0.9 pp LAC step of Tab. 2 is only about twice it and, being measured from single runs, should be read as indicative on the training side (its evaluation-side stability is established in Tab. 11).

Table 15: Multi-seed retraining of the ablation endpoints (Matterport3D test, PCK@1°). Paper seed first.

Configuration	seed 1 (paper)	seed 2	seed 3	s.d. (pp)
chart-naïve (R1)	0.229	0.230	0.223	0.4
SCCM	**0.275**	**0.283**	**0.272**	0.6
margin (pp)	+4.6	+5.3	+4.9	

EDM resolution control. EDM [6] is evaluated at its native training resolution (320×640); its ground truth and metrics are computed on that grid. Because no re-scaling of SCCM to EDM's resolution is canonical, we bracket it: (i) down-sampling SCCM's input to 320×640 and back (matching input information only) gives PCK@1° 0.273; (ii) matching input, ground truth, *and* scored pixels to EDM's grid (the same 1.47 B supervised pixels) gives 0.279. Both straddle the reported 0.275: re-scaling moves SCCM by ±0.4 pp in no consistent direction while its gap over EDM (0.163) stays at 11.0–11.6 pp. Training both methods at one resolution is not possible because EDM's public release is inference-only (Sec. G).

Pose solver. The downstream pose numbers of Tab. 3 use one shared ray-space 8-point+RANSAC evaluator for all methods, whereas EDM's own paper reports a different pipeline (the Robust 360-8PA solver of Solarte *et al*. [11] and its own pair selection). Swapping our essential-matrix estimator for Robust 360-8PA on the same matches and inliers changes AUC@5° by at most 0.01 pp for every method (Tab. 16; the ≥ 200-confident-matches filter is applied over these four methods only, leaving 15,112 pairs rather than the 11,574 of Tab. 3, which also intersects the perspective and sparse baselines, hence the slightly higher absolute values), so the ranking, and the gap to the numbers reported in the EDM paper, are properties of the evaluation protocol (pair selection, sampling budget) rather than of the solver.

Table 16: Solver swap on 15,112 Matterport3D test pairs: Pose AUC@5° with our 8-point+RANSAC estimator vs. Robust 360-8PA on identical matches.

Method	8-pt+RANSAC	Robust 360-8PA
EDM [6]	8.28	8.29
RoMa V1 (ERP-retrained)	11.69	11.68
chart-naïve (R1)	14.17	14.16
SCCM	**17.96**	**17.96**

References

1. Armeni, I., Sax, S., Zamir, A.R., Savarese, S.: Joint 2D-3D-Semantic data for indoor scene understanding. arXiv preprint arXiv:1702.01105 (2017)
2. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3D: Learning from RGB-D data in indoor environments. In: 3DV (2017)
3. Edstedt, J., Athanasiadis, I., Wadenbäck, M., Felsberg, M.: DKM: Dense kernelized feature matching for geometry estimation. In: CVPR (2023)
4. Edstedt, J., Nordström, D., Zhang, Y., Bökman, G., Astermark, J., Larsson, V., Heyden, A., Kahl, F., Wadenbäck, M., Felsberg, M.: RoMa v2: Harder better faster denser feature matching. In: ECCV (2026)
5. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: RoMa: Robust dense feature matching. In: CVPR (2024)

6. Jung, D., Choi, J., Lee, Y., Jeong, S., Lee, T., Manocha, D., Yeon, S.: EDM: Equirectangular projection-oriented dense kernelized feature matching. In: CVPR (2025)
7. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3D with MASt3R. In: ECCV (2024)
8. Mapillary: Mapillary Metropolis dataset. `https://www.mapillary.com/dataset/metropolis`, accessed September 2026
9. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
10. Ou, J., Cao, Z., Ren, Y., Li, Z., Zhu, J., Hua, T., Zhang, S., Xiong, H., Zhao, W.: Holo360D: A large-scale real-world dataset with continuous trajectories for advancing panoramic 3D reconstruction and beyond. arXiv preprint arXiv:2604.22482 (2026)
11. Solarte, B., Wu, C.H., Lu, K.W., Sun, M., Chiu, W.C., Tsai, Y.H.: Robust 360-8PA: Redesigning the normalized 8-point algorithm for 360-FoV images. In: ICRA (2021)
12. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-free local feature matching with transformers. In: CVPR (2021)
13. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: VGGT: Visual geometry grounded transformer. In: CVPR (2025)