WHAT A 3D RECONSTRUCTION BENCHMARK CAN AND CANNOT TELL ABOUT GEOMETRIC FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Benchmarks for 3D geometric foundation models rank models by surface metrics such as Chamfer distance, without testing whether those metrics can tell a reconstruction of the right scene from one of the wrong scene. This paper tests it for fourteen models on six domains by scoring two substitutes through the unmodified pipeline: another scene's ground truth, planted on the target's cameras, and the target's own ground truth flattened onto a plane. Where the ground has relief, the wrong scene is rejected on 98% of cells. On low-relief ground, rendered lunar terrain and rover imagery of Mount Etna and Saharan sand alike, the flattened ground truth outscores every model on all 21 windows, and on about half of them even when every submission receives the same similarity fit. A closed-form bound explains why: a plane's Chamfer cannot exceed the target's mean deviation from its own plane, so the plane wins wherever the best model errs by more. The wrong scene also outscores every model on most low-relief cells, but on the real rover data only through placement, which the shared similarity fit removes. The ground-level tables here and those the field publishes are too small to separate their leading models once every pair is tested. The controls and the evaluation code will be released, with a checklist for benchmarks on low-relief terrain.

1 INTRODUCTION

Feed-forward 3D geometric foundation models such as DUSt3R (Wang et al., 2024), VGGT (Wang et al., 2025a) and MapAnything (Keetha et al., 2026) reconstruct scene geometry and camera poses from a handful of unposed images (Wang et al., 2026b), and their evaluation has converged on surface metrics against ground-truth geometry, principally Chamfer distance and threshold F-scores, aggregated over a scene set into a ranking.

This paper asks not whether the models are good but whether the measurement is informative. The test is a control: the ground truth of a *different scene* is submitted to the evaluator as though it were a prediction and scored through the unmodified pipeline. Such a prediction is geometrically valid, correctly scaled and of the wrong place, so every metric must score it worse than a genuine reconstruction of the target, and a metric that does not is not measuring reconstruction.

On low-relief ground the metric fails this test. On rendered lunar terrain (LuSNAR), a rover's images of Mount Etna (AMEDS) and Saharan sand at a Mars-analogue site (Erfoud), the ground truth of a different patch of the same kind of ground outscores every model on 77% of cells planted on the target's cameras and on 51% of pairs planted on a model's own, and the target's own ground truth flattened onto its best-fit plane does so on all 21 windows (Figure 1). Where the ground has relief, on ETH3D's laser scans and TartanGround's simulated environments, the wrong scene is rejected on 98% of cells, and a donor from another domain is rejected on lunar targets too, so what decides the outcome is the ground truth, the target's deviation from its own plane and, for the wrong scene, the donor's gap to it, not a dataset or a kind of imagery.

Two failure modes produce the result, and the paper separates them. The first is the ground's small deviation from its own plane: a plane's Chamfer distance cannot exceed the target's mean deviation from that plane, $\mathrm{dev}(h_t)$, so wherever the models err by more than the ground varies, the flattened target lies closer to the target than their reconstructions do, and on most such cells a different place

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

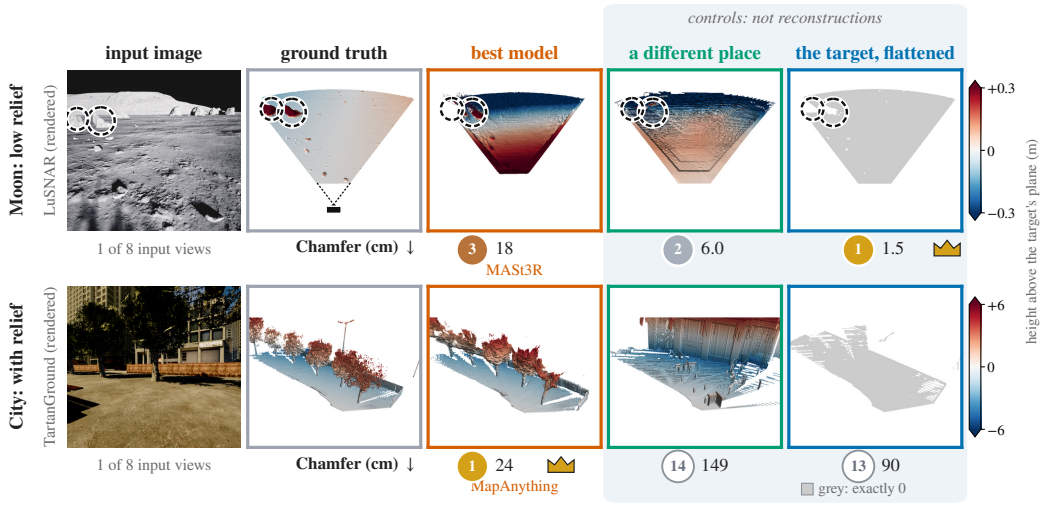


Figure 1: Chamfer (lower is better) on two targets; medals rank each tile among every model and both substitutes, and the crown marks the metric's choice. Each cloud is the one scored, seen from behind the input cameras and coloured by height above the target's least-squares plane; the flattened target is exactly zero, drawn grey, and its rings hold only the gaps the boulders hid. On rendered lunar ground both substitutes outscore the best model, which recovers the ringed boulders that neither substitute has; the different place's seams are where its frames, each planted on a target camera, meet. On a rendered city block all 12 models beat both. On the 21 low-relief audit windows the flattened target beats every model on all 21, and a different place on 24 of their 31 cells, rendered lunar ground and real rover imagery of Etna and the Sahara alike (Table 2); under a similarity fit shared by every submission the flattened target still wins this cell, as on 7 of LuSNAR's 11, and the different place no longer does. The rows' colour scales differ twentyfold.

does too. The plane wins wherever the best model errs by more than that deviation, and the model won all six benchmark windows the bound leaves open, so the rule needs only the ground truth (Section 4). The second is placement: on the two real rover domains the models' surfaces are misplaced as a whole, and once every submission, the substitutes included, receives the same similarity fit to the geometry, the wrong scene wins none of their cells; on rendered lunar ground it still wins 4 of 11, and the plane keeps 7 of LuSNAR's 11 windows and 4 of the rover domains' 10.

Passing the control and agreeing with a rover's navigation are different properties, and Chamfer's agreement is that of its completeness term: relief correlation, a metric built from the bound to pass the control, rejects both substitutes on nearly every cell and, in rank composite with completeness, keeps those rejections while its agreement with navigation coverage stays within 0.04 of Chamfer's or above it (Section 5, Table 3). Rankings are also less firm than they look: no ground-level domain large enough to test separates its two leading models once every pair is tested, published tables carry too few datasets for such a test to separate any pair, and unstated scoring defaults push rank agreement below 0.90 in more than half of the comparisons that can move (Section 6).

Contributions.

1. **A closed-form bound on when a surface metric must stop discriminating**, a consequence of orthogonal projection rather than a finding, and the rule it supports, both computed from the target's ground truth alone (Section 4).
2. **Two controls that test it**, a wrong scene and the flattened ground truth, with variants that separate relief from placement and from a donor's pose advantage, to be released as a function taking any scoring pipeline (Section 4); and *relief correlation*, which no plane passes and a wrong place rarely does, with the completeness it needs to rank (Section 5).
3. **An account of what such a benchmark can establish**: agreement with navigation (Section 5), statistical resolution against the smallest p-value a design admits, on published

tables too (Section 6.1), sensitivity to eighteen protocol axes (Section 6.2), and a low-relief benchmark checklist (Section 7); Appendix K separates the new from the transferred.

2 Related work

Feed-forward 3D reconstruction. Beginning with pairwise pointmap regression (Wang et al., 2024; Leroy et al., 2024) and extended to many views (Wang et al., 2025a; Yang et al., 2025; Keetha et al., 2026; Wang et al., 2026c), these models predict scene geometry and camera parameters directly from unposed images. Later members add streaming operation (Wang et al., 2025b; Wang & Agapito, 2025; Cabon et al., 2025; Zhuo et al., 2026), larger view budgets (Tang et al., 2025), training-efficient scaling with dynamic-scene support (Wang et al., 2026a) and joint depth and pose estimation (Lin et al., 2026; Burzio et al., 2026), and test-time self-calibration adapts a pretrained model to a target scene (Lu et al., 2025).

Benchmarks and surface metrics. Benchmarks rank these models by surface and depth metrics (Cong et al., 2025; Peng et al., 2026), also under aerial and degraded capture (Yang et al., 2026; Wu et al., 2025b; Liu et al., 2026) and on lunar terrain, for which models are adapted and distilled (Grethen et al., 2025; Grethen, 2026; Grethen et al., 2026) and lunar monocular depth has a benchmark of its own (Sekmen et al., 2026). Their metrics are the field's standard ones: accuracy and completeness against a reference scan (Seitz et al., 2006), the F-score at a distance threshold (Knapitsch et al., 2017; Schöps et al., 2017) and Chamfer distance, recently reweighted for density as a completion metric (Wu et al., 2021) and made hyperbolic (Lin et al., 2023) or geodesic (Alonso et al., 2025) as a training loss, where its gradient collapses in a way no local regulariser prevents (Song & Hyde, 2026). Those variants change how nearest-neighbour matching weighs density, outliers and topology; the failure studied here is one of Chamfer as an evaluator, where the signal is absent rather than mis-weighted, and Equation 2 gives the condition under which Chamfer can still tell a surface from its absence: a model erring by less than half the target's deviation from its own plane is guaranteed to beat the flattened target. A trivial retrieval baseline has been shown statistically indistinguishable from learned single-view reconstruction under such metrics (Tatarchenko et al., 2019), and concurrent benchmarks ship calibration, falsification and wrong-scene controls (Cui et al., 2026; Shi & Gong, 2026; Paul et al., 2026). What is contributed here is the terrain class on which a flat plane must win, bounded in closed form; Appendix K covers the alternative metrics and audits of metrics.

Benchmark validity and statistical resolution. The control applies the logic of sanity checks for saliency maps (Adebayo et al., 2018), that a diagnostic must change its verdict when what it detects is replaced by something it should reject, to a surface metric, and scoring metrics against navigation follows the argument that depth metrics should be judged by their downstream use (Li et al., 2025). Rank-based comparison over several datasets (Demšar, 2006) is standard in parts of machine learning but essentially absent from 3D vision benchmarks; Appendix K places both analyses among the benchmark critiques of other fields.

3 Benchmark construction

Domains. Six domains are used, so that the study is not about one capture style (Table 1): rendered lunar rover imagery from LuSNAR (Liu et al., 2024); a rover's real imagery of volcanic ground on Mount Etna, the standard terrestrial lunar analogue, from AMEDS (Suter et al., 2023); a real rover traverse of flat Saharan sand, the one low-relief domain not chosen to resemble the Moon, from Erfoud (Lacroix et al., 2020); twenty simulated environments of TartanGround (Patel et al., 2025), built on TartanAir (Wang et al., 2020); real laser-scanned scenes from ETH3D (Schöps et al., 2017); and a nadir descent camera over the lunar south pole from StereoLunar (Grethen et al., 2025). LuSNAR, AMEDS and Erfoud form the low-relief class.

Models. Fourteen feed-forward models are evaluated, each one's outputs and exclusions in Table B.1. Three output no depth map on any domain, so the depth metric is reported over the other eleven, nine on ETH3D and TartanGround. Every surface and coverage number is reported over every model its domain carries, except where the evaluator refuses to score: a prediction whose

Table 1: Domains and the analyses each enters. Cells count zero-shot evaluations after deduplication, restriction to the frozen protocol and removal of control rows, 1611 over 59 scenes; StereoLunar's row is its first altitude band. Validity needs a rover's navigation task, which a descent camera lacks. Resolution means one planned comparison can reach $p < 0.05$ under the exact signed-rank test on scene-level scores, which takes six or more scenes; for every pair, ETH3D's eleven suffice only on its nine-model depth metric (Section 6.1). The protocol study evaluates every model at every level of each scoring axis on one low-relief and one high-relief domain, the input axes on subsets (Table H.1) and ETH3D on its ground-truth extent axis alone.

Domain	Imagery	Ground truth	Scenes	Models	Cells	Control	Validity	Resolution	Protocol
LuSNAR	rendered lunar rover	rendered depth	5	14	280	yes	yes	no	yes
AMEDS	real rover, Etna	LiDAR	4	14	63	yes	yes	no	no
Erfoud	real rover, Sahara	LiDAR	5	14	70	yes	yes	no	no
TartanGround	simulated, terrestrial	rendered depth	20	12	478	yes	yes	yes	yes
ETH3D	real, terrestrial	laser scan	11	12	132	yes	yes	yes	one axis
StereoLunar	rendered lunar descent	rendered depth	14	14	588	plane only	no	yes	no

alignment leaves it covering under 1% of the ground truth has failed in alignment or scale, and is not given a score. That happens nine times, to three models on four AMEDS windows and to two on one TartanGround window, which is why AMEDS has 63 cells rather than 70 and TartanGround 478 rather than 480 (Appendices B and C.1). Training data is read from each model's released code and checkpoints, and ETH3D and TartanGround each run twelve of the fourteen: three are excluded from one or both for a documented training-data overlap or an authors' contamination notice. MASt3R, CUT3R and MapAnything also train on TartanAir and are kept, so TartanGround's rankings carry that overlap (Appendix B).

Protocol. Every model receives the same eight images per cell, selected by a frozen rule computed once from ground truth so selection cannot vary by model. Predictions are aligned to camera poses and never to the geometry, and under the frozen protocol no scale is fitted on test data: a model not used at its own metric scale receives a scale calibrated once on a held-out validation split, and what is then fitted on the test cell is a rigid $SE(3)$ transform, not a free similarity; the oracle scale level of Section 6.2 relaxes this as a diagnostic.

Criteria and notation. Surface metrics compare a prediction P with the ground truth G through unsquared Euclidean nearest-neighbour distances $d(x, Y) = \min_{y \in Y} \|x - y\|_2$, in metres, each direction averaged over its own points. Chamfer is the mean of accuracy, the first term, and completeness, the second,

$$\mathrm{CD}(P, G) \;=\; \tfrac{1}{2}\Big(\tfrac{1}{|P|}\textstyle\sum_{p \in P} d(p, G) + \tfrac{1}{|G|}\sum_{g \in G} d(g, P)\Big), \tag{1}$$

and the F-score at tolerance τ is the harmonic mean of the shares of P and of G lying within τ of the other cloud. Per-view depth metrics (on AMEDS and Erfoud, against their LiDAR projected into each view) are computed before any global alignment, against ground truth resized to the prediction. Navigation metrics rasterise the reconstruction onto a gravity-aligned grid and measure terrain fields, traversability and path feasibility; the primary one, coverage, is the share of the ground truth's occupied grid cells that the reconstruction also fills. A *cell* is one model or control scored on one frame window, and $\mathrm{dev}(h_t)$ is the mean absolute distance of the target's ground truth h_t from its least-squares plane, onto which it projects as h_{plane}; Appendix A collects the other terms.

4 When a surface metric stops discriminating

Two controls. The first submits the ground truth of a different scene to the evaluator as though it were a prediction: a real surface, correctly scaled and of the wrong place, which every metric must score worse than a genuine reconstruction of the target. It complements the oracle control, the target's own ground truth, which a metric keyed to scene-agnostic statistics such as point density would pass. The donor's per-frame geometry is planted on the target's ground-truth poses (LOCAL), so the alignment is an identity and only the terrain is wrong, though the donor then has an exact camera track no model has; planted on a real model's estimated poses for the same cell (LOCAL-MODEL) it carries that model's pose error instead, and keeping the donor's own poses (WORLD) is

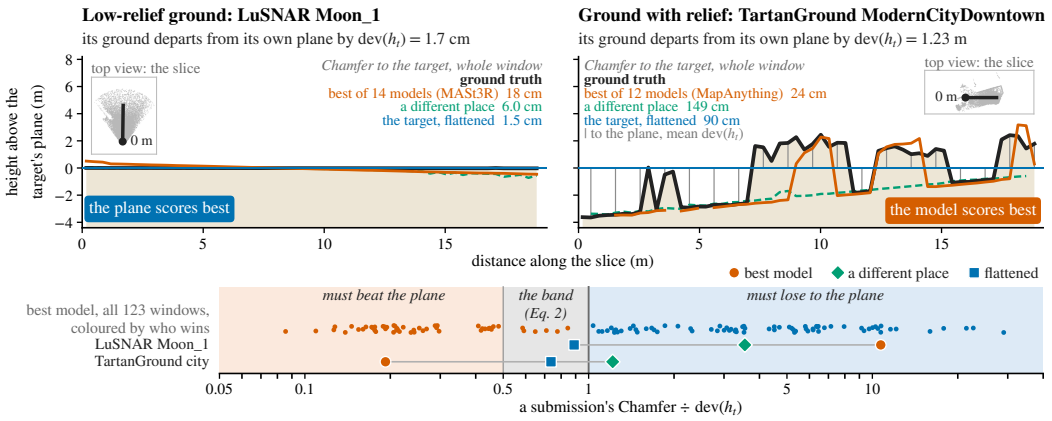


Figure 2: The bound on real cross-sections of Figure 1's targets, on one scale; each curve is one slice (inset), while every Chamfer is over the whole window. The grey sticks join the ground truth to its least-squares plane; their mean, $\mathrm{dev}(h_t)$, is the flattened target's completeness, so by Equation 2 its Chamfer lies in the band from $\frac{1}{2}\,\mathrm{dev}(h_t)$ to $\mathrm{dev}(h_t)$. Below, each submission's Chamfer over $\mathrm{dev}(h_t)$ (circle: best model; diamond: a different place; square: flattened), with the best model of all 123 frame windows, blue where the plane beats it.

reported only for continuity (Appendix D). The second control, registered before it was run, projects the target's ground truth onto its least-squares plane and scores it as a prediction: it carries the target's footprint and point density but no relief, and needs no donor or inference.

A closed-form bound. Every point of the flattened target is the orthogonal foot of one ground-truth point on the plane. The flattened point nearest a ground-truth point therefore lies at exactly that point's distance from the plane, so the flattened target's completeness is $\mathrm{dev}(h_t)$; and the ground-truth point nearest a flattened point lies no farther than the point it came from, so its accuracy lies between 0 and $\mathrm{dev}(h_t)$. Chamfer as defined in Equation 1, the mean of the two, therefore satisfies

$$\tfrac{1}{2}\,\mathrm{dev}(h_t) \;\le\; \mathrm{CD}(h_{\mathrm{plane}}, h_t) \;\le\; \mathrm{dev}(h_t). \tag{2}$$

A model whose Chamfer exceeds $\mathrm{dev}(h_t)$ must lose to the plane and one below half of it must beat the plane; between the two, the plane's own accuracy decides, and every quantity is a statistic of the target's ground truth. The flattened target carries one point per ground-truth point, so it matches the target's density and footprint exactly and the band holds whatever the sampling. No relief sweep is needed to apply it: of 48 definitions of local relief tested, each keeps LuSNAR's error-to-relief ratio above ETH3D's and TartanGround's, but none yields a fixed threshold, which $\mathrm{dev}(h_t)$ does (Appendices D and F). Appendix F derives the band and a general form for any flat submission (Equation 4), and bounds any surface that replaces the target's heights the same way: a bowl of the target's amplitude and the wrong shape beats every model on 17 of the 21 low-relief windows.

Result. The outcome divides by terrain rather than by dataset or kind of imagery (Table 2, Figure 1). Where the ground has relief, on ETH3D and TartanGround, the wrong scene is rejected on all but one of the same-domain and unseen cells, including environments the study never ranked models on. On low-relief ground it wins most cells of every domain, rendered or real, 77% of them in all (scene-bootstrap interval 59 to 93%), and the plane wins every one. Nine further LuSNAR cells and ten on a second Erfoud drive bring it to 40 of 50 (Appendix D.1). Across all 123 frame windows of the benchmark, StereoLunar's descent windows included, the plane wins exactly when the best model's Chamfer exceeds $\mathrm{dev}(h_t)$; Equation 2 guarantees this only outside its band, and the 6 windows inside the band all go to the model (Figure 2; Appendix F), and where the plane wins the wrong scene mostly wins too. Three checks rule out simpler explanations. A donor from another domain is rejected on all but two cells, lunar targets included, so the wrong scene does not win by resembling a different kind of place. Planted on a model's own estimated poses it still beats that model on nearly every LuSNAR pair, but against the best model it wins only 214 of 420 low-relief pairs, reproducing the verdict on LuSNAR and Erfoud and not on AMEDS, so part of the LOCAL

Table 2: The wrong-scene control, LOCAL construction, whose donor has the target's exact poses. Margin is the best zero-shot model's Chamfer on the identical scene and frame window divided by the control's, so a margin above 1 means the wrong scene scored better than every model; rows in **green** are those where it did so on the median cell. A scene is won when its median margin over its windows exceeds 1. Equal fit counts the cells still won when every submission receives the same similarity fit to the geometry (Table D.3); n/a, not fitted. Figure D.1 plots both controls against $\mathrm{dev}(h_t)$. Donors are chosen by hand; donors retrieved by rule win on the same 18 of the 21 low-relief windows as the best hand-picked ones, and on 6 of TartanGround's 54 windows against 1 (Appendix D.1).

Domain	Donor	Cells	Median margin	Range	Cells won	Scenes won	Equal fit
LuSNAR	regolith	11	**3.01×**	[0.97, 12.07]	**10 / 11**	**5 / 5**	4 / 11
	cross-domain	10	0.34×	[0.19, 0.61]	0 / 10	0 / 5	n/a
AMEDS	volcanic	10	**1.20×**	[0.26, 4.01]	**7 / 10**	**2 / 4**	0 / 10
Erfoud	sand	10	**2.59×**	[0.48, 5.95]	**7 / 10**	**4 / 5**	0 / 10
ETH3D	within-domain	8	0.43×	[0.05, 0.54]	0 / 8	0 / 8	n/a
	cross-domain	10	0.43×	[0.05, 1.06]	1 / 10	1 / 10	n/a
TartanGround	within-domain	22	0.38×	[0.08, 1.41]	1 / 22	1 / 20	0 / 22
	unseen scenes	35	0.30×	[0.03, 0.99]	0 / 35	0 / 16	n/a
	cross-domain	10	0.35×	[0.07, 1.34]	1 / 10	1 / 10	n/a

rate is a pose advantage. COLMAP (Schönberger & Frahm, 2016; Schönberger et al., 2016) loses to the plane on all but one of the low-relief windows it reconstructs and on about half of the high-relief ones, so the failure is not peculiar to learned models.

Two failure modes. The two substitutes win for different reasons, and the paper keeps them apart. The plane carries no misplacement: the best model errs by a median of 6.1, 3.5 and 4.5 times $\mathrm{dev}(h_t)$ on LuSNAR, AMEDS and Erfoud (Appendix F); on LuSNAR most of that error is a tilt of the whole ground, and close to half of it survives even the best similarity fit. Given every submission, the substitutes included, the same similarity fit to the geometry, as a benchmark aligning on geometry would, the wrong scene still beats every model on 4 of LuSNAR's 11 cells and on none of AMEDS's or Erfoud's, and the plane on 7 of LuSNAR's 11 windows, 1 of AMEDS's 5 and 3 of Erfoud's 5 (Appendix D): on the real rover domains the wrong scene wins by placement, and so, on four of AMEDS's five windows, does the plane. A benchmark on such terrain asks whether its models' surfaces are right and whether they are in the right place, and once the models err by more than the ground varies a surface metric answers neither.

5 WHICH METRIC SHOULD BE REPORTED

Two probes, one table. Table 3 sets two probes side by side. The first asks whether a metric rejects the wrong scene on the cells of Table 2, in the direction registered for it in advance, per domain and never pooled: pooled, Chamfer rejects on 71 of 96 cells, a rate that hides its 1 of 11 on lunar terrain. The second measures agreement, within each scene so that scene difficulty cannot drive it, between a metric's model ranking and a navigation outcome's; its verdict holds at each of 60 rover capability profiles and with the hazard map or planned paths as the outcome (Appendix B).

What the probes show. Chamfer is the case the control was built for: it rejects the wrong scene and predicts navigation on TartanGround and ETH3D ($\rho = +0.80$ and $+0.51$) and does neither on lunar terrain, at $+0.20$ to $+0.44$ across the four navigation outcomes, every one of them consistent with zero under model resampling (Appendix G). The two properties are nonetheless distinct. Per-view depth rejects the wrong scene on every cell of every domain carrying depth but one of Erfoud's, AMEDS and Erfoud included once their LiDAR is projected into the views, and predicts coverage better than Chamfer on LuSNAR and Erfoud as a point estimate (Appendix G); against the target's plane ray-cast into its cameras, though, it loses on 4 of Erfoud's 5 windows (Appendix E). Completeness predicts navigation everywhere despite failing the lunar control, plausibly because both it

Table 3: Metric verdicts per domain. Rejects counts the same-domain and unseen cells of Table 2 where the wrong scene scores worse than the model best on that column's metric; ρ cov. is the mean within-scene Spearman correlation between the metric's and the navigation coverage's model rankings, the registered primary of four outcomes, the other three being in Table G.3. Relief corr. is Equation 3, and Composite its rank composite with completeness; the principal other metrics are in Table E.1.

Domain	Chamfer		Completeness		Depth AbsRel		Relief corr.		Composite	
	Rej.	ρ cov.	Rej.	ρ cov.	Rej.	ρ cov.	Rej.	ρ cov.	Rej.	ρ cov.
LuSNAR	1 / 11	$+0.20$	2 / 11	$+0.75$	11 / 11	$+0.66$	9 / 11	$+0.08$	9 / 11	$+0.63$
AMEDS	3 / 10	$+0.84$	4 / 10	$+0.85$	10 / 10	$+0.77$	10 / 10	$+0.52$	10 / 10	$+0.80$
Erfoud	3 / 10	$+0.52$	2 / 10	$+0.66$	9 / 10	$+0.65$	10 / 10	$+0.41$	9 / 10	$+0.66$
ETH3D	8 / 8	$+0.51$	7 / 8	$+0.51$	8 / 8	$+0.11$	6 / 8	$+0.33$	8 / 8	$+0.47$
TartanGround	56 / 57	$+0.80$	54 / 57	$+0.85$	57 / 57	$+0.24$	57 / 57	$+0.56$	56 / 57	$+0.79$

and coverage measure footprint support, and on AMEDS and Erfoud Chamfer's ranking still agrees with navigation ($+0.84$ and $+0.52$). The control tests the top of a leaderboard; agreement over a whole ranking is a different property.

Does a better metric fix it? Not one that measures distance. Density-aware Chamfer (Wu et al., 2021), a proposed replacement, at any of four temperatures, point-to-plane distance, density-equalised Chamfer, the F-score at every tolerance from 0.01 to 0.5 m, HyperCD (Lin et al., 2023) and a geodesic Chamfer after GeoCD (Alonso et al., 2025) each accept the wrong scene or the flattened target on most low-relief cells they were run on, and earth mover's distance (Rubner et al., 2000; Fan et al., 2017) accepts both there and even on cells Chamfer rejects. Criteria that score orientation or appearance rather than position, such as normal consistency, slope correlation and photometric consistency, reject the wrong scene on most lunar cells (Table E.1, which lists the principal metrics tested).

Relief correlation. Equation 2 constrains every metric of how far a submission's points lie from the target's, not one of where the relief lies. *Relief correlation* fixes one raster over the ground truth's footprint, in its principal frame with 0.2 m cells chosen before any result was seen, takes each cloud's mean height $h_X(c)$ in every cell c, removes each raster's own least-squares plane π_X, and takes the Pearson correlation of what remains over the cells $\mathcal{C}$ that both clouds occupy:

$$\mathrm{RC}(P, G) \;=\; \mathrm{corr}_{c \in \mathcal{C}}\big(h_P(c) - \pi_P(c),\; h_G(c) - \pi_G(c)\big). \tag{3}$$

Three properties follow from the construction (Figure E.1). A flat submission has no residual and scores 0 by convention, so the plane control cannot pass; a global vertical offset, tilt or positive vertical scale leaves the score unchanged, so a reconstruction is not charged for being misplaced vertically as a whole; and a different place scores only for relief lying where the target's does.

What relief correlation establishes. It passes the control where Chamfer fails it. Against the model best on it, relief correlation rejects the wrong scene on every AMEDS, Erfoud and TartanGround cell and on 9 of 11 LuSNAR cells (Table 3), and the plane on every cell of every domain, as any flat submission must; it loses two ETH3D cells, both small scenes whose walls a height raster represents poorly, and as a height-field metric it cannot score an overhang or a second storey. The verdict survives other raster sizes, a robust or quadratic detrending, and height noise, heavy-tailed or not, far above the lunar relief, the robust fit and the noise tested without Erfoud (Appendix E). Alone it ranks the models below Chamfer on navigation coverage on all five domains, $+0.08$ against $+0.20$ on LuSNAR and $+0.56$ against $+0.80$ on TartanGround, but Chamfer's agreement is its completeness's: with completeness held fixed, Chamfer's four-outcome agreement lies between -0.15 and $+0.19$ on the five domains and relief correlation's between $+0.06$ and $+0.41$, above Chamfer's on every one (Appendix E). The outcomes reward where the ground is covered, which completeness measures and relief correlation, scored on the cells both clouds occupy, does not. Beside completeness it keeps both properties: the rank composite of the two, the mean of their within-window ranks with nothing tuned, rejects both substitutes on 9 of 11 LuSNAR, 10 of 10 AMEDS and 9 of 10 Erfoud cells, where Chamfer rejects none, and on navigation coverage it is

within 0.04 of Chamfer's or above it on every domain (Table 3), parity rather than an improvement. Completeness was chosen as the partner after that decomposition was measured, and a grid of $9{,}453$ formulas finds none that does better and transfers (Appendix E); on cells added after the pairing was chosen it rejects both substitutes on 17 of 19 (Appendix D.1).

6 What a ranking can and cannot resolve

A ranking is only as firm as its sample and its protocol, and this section measures both.

6.1 Statistical resolution

Models are compared with the Friedman test (Friedman, 1937) and the Nemenyi post-hoc procedure (Nemenyi, 1963) as Demšar (2006) recommends, with the Iman and Davenport correction (Iman & Davenport, 1980) and scenes as blocks. Scenes order the models with Kendall's W from 0.28 to 0.98 (Table G.4), and Nemenyi separates no neighbouring pair anywhere: its critical difference of 2.69 to 8.87 ranks would need 145 to 394 scenes even under perfect agreement.

What a pairwise test separates. Nemenyi's verdict for one pair depends on how many other models share the diagram (Benavoli et al., 2016), so every pair is also tested with an exact Wilcoxon signed-rank test and Holm's step-down correction (Holm, 1979). On TartanGround's twenty scenes it separates 33 of 66 pairs on Chamfer and 41 of 66 on coverage, almost all between models far apart in rank; the one adjacent pair that separates on each metric lies at the bottom. On the two ground-level datasets that admit the test at all, TartanGround on every metric and ETH3D on per-view depth, the first-ranked model is not separated from the second once every pair is tested. That is a limit of these designs, not evidence that the two models perform alike.

The descent camera is the exception. On StereoLunar's 14 scenes the test separates 89 of 91 pairs on Chamfer under the frozen scale, 11 of 13 adjacent ones, though that scale is unstable there and a per-scene fit separates fewer (Appendix G). Elsewhere the limit is the number of scenes: an exact signed-rank test over n blocks cannot return a two-sided p below $2/2^n$, nor, with Holm's correction over m pairs, below $m \times 2/2^n$ (Tarone, 1990). The floor binds exact rank tests over blocks, scenes or datasets, and not a hierarchical model over windows nested in scenes (Corani et al., 2017). It is a diagnostic of sample size, not a verdict on the models: it is 0.064 for ETH3D's surface metrics and 0.035 on its depth subset, and 1 for LuSNAR's, AMEDS's and Erfoud's five or fewer scenes.

The Bayesian signed-rank test of Benavoli et al. (2017), with a 5% region of practical equivalence, makes the distinction explicit: across the nine ground-level designs, three metrics each on LuSNAR, ETH3D and TartanGround, 504 of 573 model pairs are *undetermined*, the posterior settling neither on a winner nor inside the region. TartanGround's top two on Chamfer separate in 3% of resamples at its 20 scenes and 16% at four times as many (Table G.5, Appendix G).

The same test on published tables. The same test applies to tables the field has already published, since a table of per-dataset results for several models is a models-by-columns block design and its attainable floor follows from its shape. Sixteen metric rows transcribed verbatim from five tables of E3D-Bench (Cong et al., 2025) separate no pair, for a reason that precedes their contents: the floor $m \times 2/2^n$ lies between 0.43 and 1.0 in every table, and a pair could separate only with 10 to 13 columns, so a table of this shape cannot separate any pair once every pair is tested, whatever its cells contain (Table G.2, Appendix G). A single pre-planned comparison is a family of one and has a lower floor, which seven of the sixteen rows clear; none of the sixteen separates its own leading pair even then.

6.2 Protocol sensitivity

Eight input axes and ten scoring axes, the latter re-scoring stored predictions, were measured (Table H.1). The scoring axes are where agreement, Spearman's ρ between a metric's model ranking under a changed setting and under its reference setting, is weakest: on LuSNAR 150 of 266 comparisons fall below $\rho = 0.90$, and on TartanGround 148 of 268, and 85 and 86 of them respectively stay below 0.90 under a scene-level bootstrap, excluding comparisons that cannot move by construction.

Rescaling at all, frozen or oracle, and the alignment move rankings most, then confidence filtering, and depth scaling moves the depth ranking in all 8 LuSNAR comparisons but 4 of 8 on TartanGround (Appendix H). The claim is degraded agreement rather than reversal: ρ's interval lies wholly below zero in 13 of 266 and 2 of 268, although the first-ranked model changes in 127 and 96 of them.

Input corruption. Seven corruptions at three severities are applied to the input images only, following common-corruption methodology (Hendrycks & Dietterich, 2019). On LuSNAR, degrading the input *improves* Chamfer in 17 of 21 conditions (Table H.4), and not because the prediction thins: over the 233 of 420 lunar cells where it improves, the point count is unchanged (median ratio 1.00) and accuracy and completeness improve together in 171 (Appendix H).

Input crop moves a reported ranking. Five models' loaders silently crop a square frame to 4:3; uncropped, MASt3R, the best lunar model on Chamfer here, moves from a mean rank of 1.80 to 7.25, worse in 19 of 20 lunar windows yet better in 24 of 40 terrestrial ones (Table H.2).

7 A CHECKLIST FOR BENCHMARKS ON LOW-RELIEF TERRAIN

The results reduce to nine steps, in the order an organiser would run them (Appendix L).

1. Compute $\mathrm{dev}(h_t)$ for every target: the bound needs it, and no relief sweep replaces it.
2. Run the plane control: a model erring by more than $\mathrm{dev}(h_t)$ loses to it (Equation 2), and a plane that beats the field means the metric is not ranking fidelity to the scene. Run it on every metric reported; for per-view depth, ray-cast the plane into the target's cameras.
3. Run the wrong-scene control on the target's own poses (LOCAL) and on each model's estimated poses (LOCAL-MODEL); report both margins, not a verdict.
4. Where a control wins, give every submission, the controls included, the same similarity fit to the geometry: a win that vanishes is one of placement, one that survives is one of surface.
5. Where a control beats the best model, withdraw the surface metric on those cells rather than offsetting it; elsewhere print the control's score as a floor.
6. Express surface error in units of $\mathrm{dev}(h_t)$, not in bare metres.
7. Report relief correlation (Equation 3) beside completeness, or rank by their composite.
8. Score the primary result under one frozen protocol, with alignment, scale track, filtering and grid named, and report other settings as sensitivity.
9. Before reading significance into a table, compute the smallest p-value its design admits.

8 DISCUSSION

Limitations. The failure is shown on three low-relief domains and repeats on NASA's POLAR testbed (Hansen et al., 2024), added after every rule was fixed (Appendix C.3); snow, ice, seafloor and indoor corridors are untested. On the two real rover domains the wrong scene wins by placement, which a fit shared by every submission removes (Section 4), and none of the four metrics run on the target moved by 1 m (Chamfer, completeness, relief correlation and their composite) certifies placement at that scale (Appendix D.1). The navigation outcomes are a ground vehicle's, and the low-relief domains carry 14 scenes, too few for pairwise tests (Section 6.1).

Conclusion. Reconstruction benchmarks assume rather than test that their metrics discriminate. Tested, the assumption holds where the ground has relief and fails where it does not: under the frozen protocol a wrong place outscores every model on most low-relief cells and the plane on all 21 windows, and on 6 of 8 lunar cells even after a 3.6-fold fine-tuning gain (Appendix F). Where the best model errs by more than the ground's relief, the plane's win is Chamfer correctly reporting that the models have negative skill against the terrain's climatological forecast (Appendix K). What fails is discrimination: at that error a reconstruction of the right place scores no better than a wrong place, so the benchmark cannot certify that its leader reconstructs the ground, even where Chamfer's ranking still agrees with navigation, as on AMEDS and Erfoud (Section 5).

Ethics statement

This study uses only datasets and model checkpoints released by their authors (LuSNAR, AMEDS, Erfoud, TartanGround, ETH3D, StereoLunar and POLAR), each under its own terms of use; it involves no human subjects and no personal data. The controls and code that will be released are evaluation tools: by exposing where a benchmark cannot tell reconstructions apart, they are meant to prevent misleading model selection in planetary and field robotics, and we are not aware of harmful applications or conflicts of interest.

Reproducibility statement

Section 3 and Appendix B specify the models, domains, frame selection and the frozen evaluation protocol; Appendix C documents how the ground truth of the real domains and of POLAR was curated and checked; Appendix A defines the terms, and Appendix I lists the checks applied to the evaluation outputs. Every number is computed from our evaluation outputs except values quoted from cited work, of which only the E3D-Bench rows of Table G.2 and the corruption levels of Table M.10 enter an analysis, each marked as transcribed. The code, the controls (also as a standalone function that takes any scoring pipeline and a set of target and donor pairs), the evaluation outputs and the scripts that rebuild every audit cell from each dataset's own release will be released.

Use of large language models

In this work we used generative AI tools to give feedback on the experimental design, proposing ablations and controls and critiquing those we had chosen, and to help implement the analysis code. We also used them to search for related work, datasets and published result tables, to produce figures and tables from our evaluation outputs, and to draft and edit parts of the text. We reviewed all AI-assisted work: every cited work and transcribed number was checked against its primary source, and the analysis code and its results were checked by the authors, including a recomputation of every number computed in this study from the evaluation outputs. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 9525–9536, 2018.

Pedro Alonso, Tianrui Li, and Chongshou Li. GeoCD: A Differential Local Approximation for Geodesic Chamfer Distance. *arXiv preprint arXiv:2506.23478*, 2025.

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. MET3R: Measuring Multi-View Consistency in Generated Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6034–6044. Computer Vision Foundation / IEEE, 2025.

Alessio Benavoli, Giorgio Corani, and Francesca Mangili. Should We Really Use Post-Hoc Tests Based on Mean-Ranks? *Journal of Machine Learning Research*, 17:5:1–5:10, 2016.

Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a Change: a Tutorial for Comparing Multiple Classifiers Through Bayesian Analysis. *Journal of Machine Learning Research*, 18:77:1–77:36, 2017.

Varun Burde, Assia Benbihi, Pavel Burget, and Torsten Sattler. Comparative Evaluation of 3D Reconstruction Methods for Object Pose Estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7669–7681. IEEE, 2025.

Varun Burde, Pavel Burget, and Torsten Sattler. Benchmarking the Effects of Object Pose Estimation and Reconstruction on Robotic Grasping Success. *arXiv preprint arXiv:2602.17101*, 2026.

Alessandro Burzio, Tobias Fischer, Sven Elflein, Qunjie Zhou, Riccardo de Lutio, Jiawei Ren, Jiahui Huang, Shengyu Huang, Marc Pollefeys, Laura Leal-Taixé, Zan Gojcic, and Haithem

Turki. Déjà View: Looping Transformers for Multi-View 3D Reconstruction. *arXiv preprint arXiv:2605.30215*, 2026.

Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jérôme Revaud, and Vincent Leroy. MUSt3R: Multi-view Network for Stereo 3D Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1050–1060. Computer Vision Foundation / IEEE, 2025.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With Little Power Comes Great Responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9263–9274. Association for Computational Linguistics, 2020.

Yang Chen and Gérard G. Medioni. Object Modelling by Registration of Multiple Range Images. *Image and Vision Computing*, 10(3):145–155, 1992.

Wenyan Cong, Yiqing Liang, Yancheng Zhang, Ziyi Yang, Yan Wang, Boris Ivanovic, Marco Pavone, Chen Chen, Zhangyang Wang, and Zhiwen Fan. E3D-Bench: A Benchmark for End-to-End 3D Geometric Foundation Models. *arXiv preprint arXiv:2506.01933*, 2025.

Giorgio Corani, Alessio Benavoli, Janez Demšar, Francesca Mangili, and Marco Zaffalon. Statistical Comparison of Classifiers through Bayesian Hierarchical Modelling. *Machine Learning*, 106(11): 1817–1837, 2017.

Benjamin Coupry, Baptiste Brument, Antoine Laurent, Jean Mélou, Yvain Quéau, and Jean-Denis Durou. Assessing the Quality of 3D Reconstruction in the Absence of Ground Truth: Application to a Multimodal Archaeological Dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3457–3466. IEEE, 2025.

Yangzhi Cui, Feng Qiao, and Nathan Jacobs. StereoGenBench: A Synthetic Multi-Camera Benchmark for Stereo Generation under Controlled Baseline Regimes. *arXiv preprint arXiv:2605.23237*, 2026.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The Benchmark Lottery. *arXiv preprint arXiv:2107.07002*, 2021.

Janez Demšar. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Nanxing Nick Deng, Qing Cheng, Niclas Zeller, and Daniel Cremers. A Calibration Audit of Confidence in Feed-Forward 3D Reconstruction. *arXiv preprint arXiv:2608.29705*, 2026.

Jingming Dong and Stefano Soatto. Domain-Size Pooling in Local Descriptors: DSP-SIFT. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5097–5106. IEEE Computer Society, 2015.

Yanwei Du, Shiyu Feng, Carlton G. Cort, and Patricio A. Vela. Task-driven SLAM Benchmarking for Robot Navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7147–7153. IEEE, 2025.

Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A Point Set Generation Network for 3D Object Reconstruction from a Single Image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2463–2471. IEEE Computer Society, 2017.

Milton Friedman. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937.

Kyle Gorman and Steven Bedrick. We Need to Talk about Standard Splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2786–2791. Association for Computational Linguistics, 2019.

Yvette Graham and Timothy Baldwin. Testing for Significance of Increased Correlation with Human Judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 172–176. Association for Computational Linguistics, 2014.

Clémentine Grethen. Geometric Foundation Model Distillation for Efficient Lunar 3D Reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 17032 of *Lecture Notes in Computer Science*, pp. 529–545. Springer, 2026.

Clémentine Grethen, Simone Gasparini, Géraldine Morin, Jérémy Lebreton, Lucas Marti, and Manuel Sanchez-Gestido. Adapting Stereo Vision from Objects to 3D Lunar Surface Reconstruction with the StereoLunar Dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 3810–3819. IEEE, 2025.

Clémentine Grethen, Yuang Shi, Simone Gasparini, and Géraldine Morin. MoonAnything: A Vision Benchmark with Large-Scale Lunar Supervised Data. In *Proceedings of the ACM Multimedia Systems Conference (MMSys)*, pp. 457–463. ACM, 2026.

Margaret Hansen, Uland Wong, and Terrence Fong. The POLAR Traverse Dataset: A Dataset of Stereo Camera Images Simulating Traverses across Lunar Polar Terrain under Extreme Lighting Conditions. *arXiv preprint arXiv:2403.12194*, 2024.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2019.

Nicolai Hermann, Jorge Condor, and Piotr Didyk. Puzzle Similarity: A Perceptually-Guided Cross-Reference Metric for Artifact Detection in 3D Scene Reconstructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28881–28891. IEEE, 2025.

Antti Hietanen, Jyrki Latokartano, Alessandro Foi, Roel Pieters, Ville Kyrki, Minna Lanz, and Joni-Kristian Kämäräinen. Benchmarking Pose Estimation for Robot Manipulation. *Robotics and Autonomous Systems*, 143:103810, 2021.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

Yubin Hu, Kairui Wen, Heng Zhou, Xiaoyang Guo, and Yong-Jin Liu. SS3DM: Benchmarking Street-View Surface Reconstruction with a Synthetic 3D Mesh Dataset. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, pp. 106649–106666, 2024.

Ronald L. Iman and James M. Davenport. Approximations of the Critical Region of the Friedman Statistic. *Communications in Statistics – Theory and Methods*, 9(6):571–595, 1980.

Nikhil Varma Keetha, Norman Müller, Johannes L. Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian A. Scherer, and Peter Kontschieder. MapAnything: Universal Feed-Forward Metric 3D Reconstruction. In *International Conference on 3D Vision (3DV)*, pp. 499–509. IEEE, 2026.

Ihor Kendiukhov. Quantifying Ranking Instability Across Evaluation Protocol Axes in Gene Regulatory Network Benchmarking. *arXiv preprint arXiv:2603.03493*, 2026.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023.

R. L. Kirk, E. Howington-Kraus, M. R. Rosiek, J. A. Anderson, B. A. Archinal, K. J. Becker, D. A. Cook, D. M. Galuszka, P. E. Geissler, T. M. Hare, I. M. Holmberg, L. P. Keszthelyi, B. L. Redding, W. A. Delamere, D. Gallagher, J. D. Chapel, E. M. Eliason, R. King, and A. S. McEwen. Ultrahigh resolution topographic mapping of Mars with MRO HiRISE stereo images: Meter-scale slopes of candidate Phoenix landing sites. *Journal of Geophysical Research: Planets*, 113(E3): E00A24, 2008.

Randolph L. Kirk, David P. Mayer, Robin L. Fergason, Bonnie L. Redding, Donna M. Galuszka, Trent M. Hare, and Klaus Gwinner. Evaluating Stereo Digital Terrain Model Quality at Mars Rover Landing Sites with HRSC, CTX, and HiRISE Images. *Remote Sensing*, 13(17):3511, 2021.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction. *ACM Transactions on Graphics*, 36(4):78:1–78:13, 2017.

Anany Kotawala. Resolution Diagnostics for Paired LLM Evaluation. In *ICML 2026 Workshop on Hypothesis Testing*, 2026.

Simon Lacroix, Andrea De Maio, Quentin Labourey, Ellon Paiva Mendes, Pierre Narvor, Vincent Bissonette, Clément Bazerque, Fabrice Souvannavong, Raphaël Viards, and Martin Azkarate. The Erfoud dataset: a comprehensive multi-camera and Lidar data collection for planetary exploration. In *15th Symposium on Advanced Space Technologies in Robotics and Automation (ASTRA)*, 2020.

Jack Langerman, Denys Rozumnyi, Yuzhong Huang, and Dmytro Mishkin. Explaining Human Preferences via Metrics for Structured 3D Reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26944–26953. IEEE, 2025.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding Image Matching in 3D with MASt3R. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 15130 of *Lecture Notes in Computer Science*, pp. 71–91. Springer, 2024.

Zhenyu Li, Haotong Lin, Jiashi Feng, Peter Wonka, and Bingyi Kang. BenchDepth: Are We on the Right Way to Evaluate Depth Foundation Models? *arXiv preprint arXiv:2507.15321*, 2025.

Fangzhou Lin, Yun Yue, Songlin Hou, Xuechu Yu, Yajun Xu, Kazunori D. Yamada, and Ziming Zhang. Hyperbolic Chamfer Distance for Point Cloud Completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14549–14560. IEEE, 2023.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the Visual Space from Any Views. In *International Conference on Learning Representations (ICLR)*, pp. 141261–141285, 2026.

Jiayi Liu, Qianyu Zhang, Xue Wan, Shengyang Zhang, Yaolin Tian, Haodong Han, Yutao Zhao, Baichuan Liu, Zeyuan Zhao, and Xubo Luo. LuSNAR: A Lunar Segmentation, Navigation and Reconstruction Dataset based on Muti-sensor for Autonomous Exploration. *arXiv preprint arXiv:2407.06512*, 2024.

Shuhong Liu, Chenyu Bao, Ziteng Cui, Yun Liu, Xuangeng Chu, Lin Gu, Marcos V. Conde, Ryo Umagami, Tomohiro Hashimoto, Zijian Hu, Tianhan Xu, Yuan Gan, Yusuke Kurose, and Tatsuya Harada. RealX3D: A Physically-Degraded 3D Benchmark for Multi-view Visual Restoration and Reconstruction. *International Journal of Computer Vision*, 134(8):382, 2026.

Ziqi Lu, Heng Yang, Danfei Xu, Boyi Li, Boris Ivanovic, Marco Pavone, and Yue Wang. LoRA3D: Low-Rank Self-Calibration of 3D Geometric Foundation Models. In *International Conference on Learning Representations (ICLR)*, pp. 44356–44378. OpenReview.net, 2025.

Lars M. Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy Networks: Learning 3D Reconstruction in Function Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4460–4470. Computer Vision Foundation / IEEE, 2019.

Mateusz Michalkiewicz, Anekha Sokhal, Tadeusz Michalkiewicz, Piotr Pawlikowski, Mahsa Baktashmotlagh, Varun Jampani, and Guha Balakrishnan. GIQ: Benchmarking 3D Geometric Reasoning of Vision Foundation Models with Simulated and Real Polyhedra. In *International Conference on Learning Representations (ICLR)*, pp. 15468–15481, 2026.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 12346 of *Lecture Notes in Computer Science*, pp. 405–421. Springer, 2020.

Allan H. Murphy. Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient. *Monthly Weather Review*, 116(12):2417–2424, 1988.

Kevin Musgrave, Serge J. Belongie, and Ser-Nam Lim. A Metric Learning Reality Check. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 12370 of *Lecture Notes in Computer Science*, pp. 681–699. Springer, 2020.

Peter B. Nemenyi. *Distribution-free Multiple Comparisons*. PhD thesis, Princeton University, 1963.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Manthan Patel, Fan Yang, Yuheng Qiu, Cesar Cadena, Sebastian A. Scherer, Marco Hutter, and Wenshan Wang. TartanGround: A Large-Scale Dataset for Ground Robot Perception and Navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 20524–20531. IEEE, 2025.

Soumava Paul, Prakhar Kaushik, and Alan Yuille. Can These Views Be One Scene? Evaluating Multiview 3D Consistency when 3D Foundation Models Hallucinate. *arXiv preprint arXiv:2605.18754*, 2026.

Haosong Peng, Hao Li, Jiaqi Chen, Yuhao Pan, Runmao Yao, Yalun Dai, Fushuo Huo, Fangzhou Hong, Zhaoxi Chen, Haozhao Wang, Dingwen Zhang, Ziwei Liu, and Wenchao Xu. SpatialBench: Is Your Spatial Foundation Model an All-Round Player? *arXiv preprint arXiv:2605.27367*, 2026.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet Classifiers Generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400. PMLR, 2019.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4104–4113. IEEE Computer Society, 2016.

Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise View Selection for Unstructured Multi-View Stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 9907 of *Lecture Notes in Computer Science*, pp. 501–518. Springer, 2016.

Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A Multi-view Stereo Benchmark with High-Resolution Images and Multi-camera Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2538–2547. IEEE Computer Society, 2017.

Steven M. Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pp. 519–528. IEEE Computer Society, 2006.

Aytac Sekmen, Fatih Gunes, Furkan Horoz, Umut Isik, Alp Ozaydin, Altay Topaloglu, Umutcan Ustundas, Alp Yeni, Ersin Soken, Erol Sahin, Gokberk Cinbis, and Sinan Kalkan. LuMon: A Comprehensive Benchmark and Development Suite with Novel Datasets for Lunar Monocular Depth Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 10368–10378, 2026.

Yukun Shi and Minglun Gong. Future Rendering $\neq$ Future Surface: A Benchmark and Dataset for Dynamic Surface Reconstruction Beyond the Observed Window. *arXiv preprint arXiv:2607.21471*, 2026.

Chang-Yong Song and David Hyde. On the Structural Failure of Chamfer Distance in 3D Shape Optimization. *arXiv preprint arXiv:2603.09925*, 2026.

Willem Suter, Luca Fornarelli, Emiel Den Exter, Aaron Pereira, Benoit Putzeys, Edmundo Ferreira, Levin Gerdes, and Thomas Krueger. The ARCHES Mount Etna Dataset (AMEDS): A Planetary Rover Data Collection in a Lunar Analogue Environment. In *17th Symposium on Advanced Space Technologies in Robotics and Automation (ASTRA)*, 2023.

Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, Justin Kerr, and Angjoo Kanazawa. Nerfstudio: A Modular Framework for Neural Radiance Field Development. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 72:1–72:12. ACM, 2023.

Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander G. Schwing, and Zhicheng Yan. MV-DUSt3R+: Single-Stage Scene Reconstruction from Sparse Views In 2 Seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5283–5293. Computer Vision Foundation / IEEE, 2025.

Yifu Tao, Miguel Ángel Muñoz-Bañón, Lintong Zhang, Jiahao Wang, Lanke Frank Tarimo Fu, and Maurice Fallon. The Oxford Spires Dataset: Benchmarking Large-Scale LiDAR-Visual Localisation, Reconstruction and Radiance Field Methods. *The International Journal of Robotics Research*, 45(6):839–857, 2026.

Robert E. Tarone. A Modified Bonferroni Method for Discrete Data. *Biometrics*, 46(2):515–522, 1990.

Maxim Tatarchenko, Stephan R. Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What Do Single-View 3D Reconstruction Networks Learn? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3405–3414. Computer Vision Foundation / IEEE, 2019.

Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun D. Preece. Sanity Checks for Saliency Metrics. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 6021–6029. AAAI Press, 2020.

Antonio Torralba and Alexei A. Efros. Unbiased Look at Dataset Bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1521–1528. IEEE Computer Society, 2011.

Julián Urbano. Stop Using the Wilcoxon Test: Myth, Misconception and Misuse in IR Research. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 2758–2768. ACM, 2026.

Hengyi Wang and Lourdes Agapito. 3D Reconstruction with Spatial Memory. In *International Conference on 3D Vision (3DV)*, pp. 78–89. IEEE, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotný. VGGT: Visual Geometry Grounded Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. Computer Vision Foundation / IEEE, 2025a.

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

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D Perception Model with Persistent State. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522. Computer Vision Foundation / IEEE, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3D Vision Made Easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709. IEEE, 2024.

Weijie Wang, Qihang Cao, Sensen Gao, Donny Y. Chen, Haofei Xu, Wenjing Bian, Songyou Peng, Tat-Jen Cham, Chuanxia Zheng, Andreas Geiger, Jianfei Cai, Jia-Wang Bian, and Bohan Zhuang. Feed-Forward 3D Scene Modeling: A Problem-Driven Perspective. *arXiv preprint arXiv:2604.14025*, 2026b.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian A. Scherer. TartanAir: A Dataset to Push the Limits of Visual SLAM. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4909–4916. IEEE, 2020.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-Equivariant Visual Geometry Learning. In *International Conference on Learning Representations (ICLR)*, pp. 10481–10497. proceedings.iclr.cc, 2026c.

Zengyan Wang, Sirshapan Mitra, Rajat Modi, Hui Lim, and Yogesh Rawat. Sky2Ground: A Benchmark for Site Modeling under Varying Altitude. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12227–12236, 2026d.

Siyang Wu, Jack Nugent, Willow Yang, and Jia Deng. Toward A Better Understanding of Monocular Depth Evaluation. *arXiv preprint arXiv:2510.19814*, 2025a.

Tong Wu, Liang Pan, Junzhe Zhang, Tai Wang, Ziwei Liu, and Dahua Lin. Balanced Chamfer Distance as a Comprehensive Metric for Point Cloud Completion. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 29088–29100, 2021.

Xinyi Wu, Steven Landgraf, Markus Ulrich, and Rongjun Qin. An Evaluation of DUSt3R/MASt3R/VGGT 3D Reconstruction on Photogrammetric Aerial Blocks. *Geo-spatial Information Science*, pp. 1–19, 2025b.

Butian Xiong, Nanjun Zheng, Junhua Liu, and Zhen Li. GauU-Scene V2: Assessing the Reliability of Image-Based Metrics with Expansive Lidar Image Dataset Using 3DGS and NeRF. *arXiv preprint arXiv:2404.04880*, 2024.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3R: Towards 3D Reconstruction of 1000+ Images in One Forward Pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21924–21935. Computer Vision Foundation / IEEE, 2025.

Xiang Yang, Yongli Wang, HaiFeng Li, and Yunsheng Zhang. UAVFF3D: A Geometry-Aware Benchmark for Feed-Forward UAV 3D Reconstruction. *arXiv preprint arXiv:2605.17942*, 2026.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming Visual Geometry Transformer. In *International Conference on Learning Representations (ICLR)*, pp. 88055–88072, 2026.

GUIDE TO THE APPENDIX

The appendix follows the order of the paper. The questions below are answered across more than one of its sections; its last section, Appendix M, holds the full versions of analyses summarised earlier in a sentence or a table row.

A reader's question	Where it is answered
Where is each term defined, and which population of cells does each count belong to?	Appendix A; Appendix D, *Which population a count belongs to*.
Does the wrong scene win because of the donor, the poses or the alignment?	Appendix D, with the equal fit of Table D.3; donors retrieved by rule in Appendix D.1.
Is the failure specific to lunar ground, or to rendered imagery?	Appendices C.1, C.2 and J.
Do more views, fine-tuning or a classical pipeline remove it?	Appendix D.1; Appendix F; Appendix D, *A classical pipeline loses to the plane as well*.
Do the rules hold on data added after they were fixed?	POLAR, Appendix C.3; the extension windows, Appendix D.1.
Does any other metric reject both controls?	Table E.1; Appendices M.1 and M.2.
Would per-view depth, an image-based criterion or a learned metric avoid the failure?	Appendix E, *The control per-view depth cannot face*; Appendix J.
How many scenes does a ranking need, and what does the floor assume?	Appendix G, Tables G.4 and G.5; Appendix M.3.
Which protocol choices move a ranking, and which cannot?	Appendix H and Table H.1; Appendix M.4.
How were COLMAP, the per-scene baselines and fine-tuning run, and what does each model cost?	Appendix M.

A TERMINOLOGY

The constructs the analyses are stated over, collected here because they are defined where they are first needed and used throughout.

Cell One scored evaluation: a single model, on a single scene, over a single frame window, under one protocol. It is the unit of every model-level count and the unit the run log stores one row per. Other counts name their own unit (windows, scenes, pairs, profiles), and the navigation grid's cells are a different object, called grid cells throughout. A *control cell* (Appendix D) is a target window, a donor and the model pool scored on it, and a *raster cell* (Section 5) is one bin of relief correlation's height grid.

Deduplication The run log keeps every attempt at a cell; before any aggregation a cell's newest successful row supersedes the earlier ones.

Relief-threshold twin A control row re-scored in the relief-scaled threshold sweep. It carries the same Chamfer as its frozen-protocol original, since a threshold changes an F-score and not a distance, and the frozen-protocol filter removes it so that no cell is counted twice.

Frame window The eight images a cell is evaluated on, identified by the index of the first, so *Moon_1 at* 120 means frames 120 to 148 at stride four. A model's score moves more between windows of one scene than between models, so a window is never pooled across.

Scene A place, holding one or more frame windows (one on ETH3D and Erfoud). Scenes are the blocks in every statistical test on this study's own designs, since a hard scene depresses every model at once; a published table's blocks are its columns (Appendix G).

Domain One of the six datasets: LuSNAR (lunar rover, rendered), AMEDS (real rover imagery of volcanic lunar-analogue ground on Mount Etna), Erfoud (real rover imagery of Saharan interdune sand), TartanGround (terrestrial simulation), ETH3D (real, laser-scanned) and StereoLunar (lunar descent camera). The first five are rover domains and carry a navigation task; StereoLunar does not. ETH3D's *electro* sequence, whose control is run against every other ETH3D scene as donor (Appendix J), is reported as its own group beside ETH3D. Results are reported per domain; a joint rate over several domains or groups is stated only with the per-domain counts beside it. As a label on results, *terrestrial* means TartanGround,

ETH3D and electro; Erfoud is terrestrial ground but is reported with the low-relief domains, LuSNAR, AMEDS and Erfoud. POLAR (Appendix C.3) is a held-out testbed reported apart from the six; it is one scene, whose ten windows form one block.

Constructions and populations LOCAL, LOCAL-MODEL and WORLD, the three ways of planting a donor, are defined in Section 4 and compared in Table D.1; control cells, audit cells, plane-control windows and retrieval windows are defined in Appendix D.

The band The interval from $\frac{1}{2}\mathrm{dev}(h_t)$ to $\mathrm{dev}(h_t)$ in which Equation 2 leaves the plane's verdict to its own accuracy.

Relief correlation Equation 3: the correlation of two clouds' raster heights after each raster's own least-squares plane is removed (Section 5).

Frozen protocol The one combination of frame selection, alignment, scale, region of interest and thresholds that the reported numbers are computed under. Its alignment is a frozen per-model scale, for every model not used at its own metric scale, followed by a rigid $SE(3)$ fit to camera poses, so the composite is a similarity whose scale is never fitted on test data. Rows from any other protocol are excluded by reading each row's recorded protocol, never by parsing its identifier.

Zero-shot row A cell scored from a released checkpoint under the frozen protocol, with no adaptation. These are the rows a leaderboard would report.

Control row A cell whose "prediction" is ground truth rather than a model output. The *discriminative* control submits a different scene's ground truth and must be rejected; the *oracle* control submits the target's own and must be accepted; the *plane* control submits the target's own ground truth projected onto its least-squares plane and must be rejected. Control rows are excluded from every model comparison.

Donor and target In a discriminative control, the donor is the scene the substituted ground truth comes from and the target is the scene being scored.

Margin A reference's Chamfer on a cell divided by the control's Chamfer on that same cell. The reference is the best zero-shot model unless the text names another, such as the model whose poses a LOCAL-MODEL control borrows. Above 1 means the control scored better than the reference, which is the control being accepted.

Relief The scene's local terrain variation: the median, over 1 m columns of the gravity-aligned ground-truth cloud, of the within-column height spread. A property of the scene, or on AMEDS and Erfoud of the scored window, fixed before any model is scored, and read from a pinned manifest that model code cannot reach.

Resolution What a comparison can establish, as distinct from what it reports: whether the design separates two models at all, given how many scenes it has. Reported alongside every ranking.

Separation A pair of models distinguished by an exact Wilcoxon signed-rank test over scenes with Holm correction across the family of pairwise tests, at $\alpha = 0.05$. Where the Nemenyi test is reported instead, a pair separates when its mean-rank gap exceeds the critical difference.

Attainable floor The smallest p-value a design can return whatever the data. An exact two-sided signed-rank test over n scenes cannot go below $2/2^n$, the probability that all n differences share a sign by chance, and Holm's correction multiplies the first of m tests by m; when $m \times 2/2^n$ exceeds α, no pair can separate, however different the models are.

Smallest detectable correlation The smallest $|\rho|$ that a test at $\alpha = 0.05$ over the given number of points could declare significant. A correlation that is not significant does not show that the association is weaker than this bound; it leaves open any association up to its upper confidence limit.

Rank agreement Spearman's ρ between the models' ranking on one metric under a changed protocol setting and under the axis's reference setting, which is the frozen protocol's except on the scale track and depth scaling (Table H.1), each model's value being its mean over the cells both share; its interval comes from resampling scenes.

Calibrated-scale track The frozen protocol's treatment of a model not used at its own metric scale: one scale per model and domain, fitted on the domain's held-out validation scenes (two on ETH3D, three on StereoLunar, one elsewhere) and applied unchanged to every test cell.

Navigation coverage Printed as coverage in tables: the fraction of the ground truth's occupied grid cells that a reconstruction also fills. It reads where a reconstruction lies in plan and not its height.

Track B The two per-scene optimisation baselines, reported apart from the fourteen feed-forward models, which form the benchmark's main track (Appendix G).

Axis One protocol choice varied deliberately while everything else is held fixed, such as input crop or ground-truth extent. Axis rows are scored from stored predictions where no re-inference is needed, so two levels of an axis differ only by that axis.

B BENCHMARK DETAIL

Counting the composition. The cells of Table 1 are counted after deduplication, after restriction to the frozen protocol and after removal of control rows (Appendix A). StereoLunar's row is the first altitude band, and the second band's eight scenes are not counted. Models is the roster a domain carries: ETH3D and TartanGround each omit two, individually, for a documented training-data overlap or, for VGGT-Ω on ETH3D, its authors' contamination notice. AMEDS's fourteen is a maximum, since the evaluator's coverage guard refuses seven model cells over four of its windows (Appendix C.1), and TartanGround's 478 cells are 480 less two refused on one ConstructionSite window. The unseen-scene row of Table 2 adds zero-shot rows on windows of 16 further TartanGround environments, which enter that table alone and are never ranked.

The models at a glance. Table B.1 names the fourteen models in one place with what each produces and what it is withheld from. They are DUSt3R (Wang et al., 2024), MASt3R (Leroy et al., 2024), MUSt3R (Cabon et al., 2025), MV-DUSt3R+ (Tang et al., 2025), Spann3R (Wang & Agapito, 2025), CUT3R (Wang et al., 2025b), Fast3R (Yang et al., 2025), VGGT (Wang et al., 2025a), VGGT-Ω (Wang et al., 2026a), StreamVGGT (Zhuo et al., 2026), Pi3 (Wang et al., 2026c), MapAnything (Keetha et al., 2026), Depth Anything 3 (Lin et al., 2026) and DVLT, the released name of Déjà View (Burzio et al., 2026). Two per-scene optimisation baselines, a neural radiance field (Mildenhall et al., 2020) and a Gaussian splatting variant (Kerbl et al., 2023), Nerfstudio's Nerfacto and Splatfacto (Tancik et al., 2023), are evaluated separately as a second track, since they optimise the test scene itself from ground-truth camera poses. What it does not say is why these fourteen and not others, which is a separate question from withholding and is answered in Appendix K, where the inclusion rule, the date the pool was closed and the nearest eligible models it does not contain are stated. Three return no depth map on any domain, so the depth metric is reported over the other eleven less those a domain withholds, which is nine on ETH3D (Pi3 and VGGT-Ω withheld) and nine on TartanGround (DVLT and VGGT-Ω); one, MapAnything, is used at its own metric scale and is therefore the only model not rescaled by the frozen per-model factor of the calibrated-scale track. CUT3R and MASt3R also describe metric-scale output (Wang et al., 2025b; Leroy et al., 2024); MASt3R's global alignment does not preserve it, and both are placed on the calibrated-scale track with the other models, so a scale error either makes is corrected by a factor fitted on validation scenes rather than charged to it. The withheld column is derived from the domains a model has rows on rather than from a flag, and three models are held out of one or more. Which models were adapted is a separate question, answered in Appendix H. Pi3 is absent from ETH3D because its own artefacts disagree: its paper lists ETH3D as evaluation-only while its official training branch declares it among the datasets fed to the training loss, and an exposure that cannot be settled is treated as exposure. VGGT-Ω is absent from ETH3D because its authors' notice of 18 August 2026 asks that the released model's numbers on a benchmark list that includes ETH3D not be relied on, and absent from TartanGround because it trains on TartanAir V2 and on TartanGround itself; it is admitted elsewhere because the other domains are neither in its training data nor benchmarks that notice names. DVLT is absent from TartanGround because it trains on TartanAir V1 and V2, from whose environments TartanGround is rendered, and is scored on ETH3D, which VGGT-Ω withholds. MASt3R, CUT3R and MapAnything also train on TartanAir (Leroy et al., 2024; Wang et al., 2025b; Keetha et al., 2026), MapAnything on TartanAir V2 under a per-scene split, and all three were kept on TartanGround; its ranking and resolution results, among them its leading pair of MapAnything and MASt3R, should be read with that overlap in mind, while the low-relief results do not rest on TartanGround's ranking.

Table B.1: The fourteen evaluated models. Depth map, and whether the pipeline uses a model at its own metric scale, are read from each model's own stored prediction; domains counts the six of Table 1 on which it has zero-shot rows; withheld names what it is held out of, and why is given above. The spread of runtime and memory across models on TartanGround is in Table M.7.

Model	Depth map	Metric scale	Domains	Withheld from
CUT3R	yes	no	6	none
Depth Anything 3	yes	no	6	none
DUSt3R	yes	no	6	none
DVLT	yes	no	5	TartanGround
Fast3R	no	no	6	none
MapAnything	no	yes	6	none
MASt3R	yes	no	6	none
MUSt3R	yes	no	6	none
MV-DUSt3R+	yes	no	6	none
Pi3	yes	no	5	ETH3D
Spann3R	no	no	6	none
StreamVGGT	yes	no	6	none
VGGT	yes	no	6	none
VGGT-Ω	yes	no	4	ETH3D, TartanGround

Contamination screening in detail. Each model's training data is taken from its released code and checkpoints, which record the datasets a training run was given, and for VGGT-Ω, which releases no training code, from its paper's list of training datasets. Fourteen models are evaluated overall; ETH3D and TartanGround each run twelve of them, for three separate, model-specific reasons rather than any shared one. One model lists ETH3D as evaluation-only in its paper while its official training branch feeds ETH3D to the training loss, and an exposure that cannot be settled is treated as exposure, so it is excluded from ETH3D only. A second model trains on the TartanAir environments TartanGround is rendered from and is excluded from TartanGround but retained on ETH3D, which is one concrete argument for carrying a second terrestrial domain. A third model trains on TartanGround itself, which excludes it from TartanGround, and separately carries its own authors' contamination notice naming ETH3D among the benchmarks its released checkpoint should not be scored against, which excludes it from ETH3D; it is evaluated on LuSNAR, AMEDS, Erfoud and StereoLunar, and on the POLAR testbed. No model is excluded for any reason other than a documented training-data overlap with the domain it is missing from or, in that one case, a published contamination notice; the converse does not hold, since three retained models also train on TartanAir (above).

StereoLunar. The nadir subset is used within one selection band, 3.0 to 5.0 km, because a non-metric model's frozen scale tracks scene depth, which varies about ninefold over the full set; the scenes the band yields lie between 3.5 and 4.4 km, where one pixel spans five to seven metres of ground. Even inside it a non-metric model's frozen scale changes by a median factor of 18 between calibration subsets, twice the ninefold variation the band was chosen to bound, so the restriction does not make the frozen scale stable and the per-scene fit is the more reliable reading of this domain (Appendix G). A scene is a cluster of four bursts whose centres lie within 30% of one footprint of a seed burst and within 10% of its altitude, each burst seen from two viewpoints under three illuminations that repeat the same camera poses, so a scene carries three windows, one per illumination. Scene centres are at least one footprint apart, which does not by itself keep footprints apart, since a burst may lie 30% of a footprint from its seed. Measured on the scenes built, the nearest bursts of two scored scenes are at least 1.08 footprints apart for 90 of the 91 pairs, so their ground does not overlap, and 0.77 apart for the remaining pair, t07 and t12, which therefore share a strip of up to a quarter of a footprint; two scored scenes lie within one footprint of a validation scene, at 0.86 and 0.97. Three scenes fix the frozen scales and fourteen are scored. Twelve were drawn first. The band then held no other scene one footprint from every accepted one except on ground near a cluster of bursts held out of this study, so a second, append-only pass took the two such scenes without reusing those bursts. A second band of eight scenes at 6.5 to 6.9 km, drawn by the same rule, is not in Table 1 and is not reported, because it could be scored only with band one's frozen scales, fitted at little more than half its altitude, which the depth dependence above rules out. F-score

thresholds are one, two and four ground sample distances (6, 12 and 24 m), and the depth files store planar depth, contrary to the dataset's documentation. At those thresholds F-score is exactly zero in 92 to 93% of cells, because errors there run to hundreds of metres, so it is not used to rank models on this domain.

Alignment. Predictions are placed in the ground-truth frame by a similarity transform whose rotation and translation are fitted to camera poses only, never to the geometry, and whose scale is the frozen per-model factor, so the alignment cannot absorb geometric error. Models not used at their own metric scale receive a scale fitted once on a validation split and frozen; no scale is fitted on test data. The resolved absolute transform is recorded with every run. A protocol that fits one similarity to the dense geometry and applies it to cameras and points alike, in the manner of UAVFF3D's shared global alignment (Yang et al., 2026), would remove the tilt of the ground that dominates the best LuSNAR models' height error (Appendix F), but not bring the typical model's error under the terrain's relief: the residual after the best similarity fit is still a median 21.8 times it on LuSNAR over all models at each scene's window 120, 70 cells (Table D.5).

Region of interest. Both clouds are trimmed to within a fixed range of the camera path before scoring, under both protocols and by the same criterion, so that a model is not charged for geometry far outside what the cameras observed. That range is 20 m from the nearest camera centre on every ground-level domain, and 4700 m on StereoLunar, whose camera looks down from kilometres up, both fixed in the benchmark's configuration. The trim is applied as a criterion rather than as a bounding box, since a box in a gravity-aligned frame is not the same region as a distance from the trajectory.

Depth. Depth is evaluated before any global alignment, against ground truth resized to the prediction, under a primary alignment fixed in advance, one median scale per view and no shift, with the other strategies of the depth-scaling axis reported as sensitivity (Table H.1). Alignment destroys the units in which depth was predicted, so evaluating depth after it would measure the alignment. Each view is compared with that camera's own ground-truth depth image, so no frustum or occlusion test is needed: a pixel is what that camera saw. The ground truth is first centre-cropped to the prediction's aspect ratio when a model's input pipeline cropped the image, then resized to the prediction's raster by nearest neighbour, so the prediction is never resampled and no depth is interpolated across an edge. A pixel is scored when both values are finite and the ground truth lies between 1 mm and the domain's evaluation range (50 m on LuSNAR, 30 m on AMEDS, 20 m on Erfoud, 60 m on ETH3D, 100 m on TartanGround and 8000 m on StereoLunar), which also removes LuSNAR's sky. AMEDS's and Erfoud's depth is sparse, their LiDAR projected into each view (Appendices C.1 and C.2). The rule is identical for every model; the pixel count differs only through each model's output raster. The reported depth error is that primary one.

Poses. ETH3D camera parameters are the dataset's released structure-from-motion reconstruction (Schönberger & Frahm, 2016), whose pose convention is world-to-camera with a scalar-first quaternion and is inverted on read. The convention was verified against the released sparse points rather than assumed, since it differs from the other domains' conventions.

Navigation. Reconstructions are rasterised onto a gravity-aligned grid, and terrain fields, traversability masks and path feasibility are computed under a set of capability profiles. The planner arm reaches 6 of ETH3D's 11 scenes. Its ground truth is culled to the union of eight wide-baseline frusta, and a union of wedges is fragmented by construction, so on the other 5 no traversable component spans the required fraction of the grid. That count itself depends on the scan registration that Appendix I checks: measured against unregistered scans the planner reached no scene at all, and courtyard went from 0 feasible goals to 50 once the scans were placed in the image frame. Lowering the connectivity threshold for one domain would manufacture a number by varying the protocol per domain, which is the failure this paper is about, so the threshold is unchanged and any ETH3D path number names the 6 scenes it comes from.

Navigation robustness. Navigation coverage, the outcome Section 5 correlates against, depends only on which grid cells a reconstruction fills, not on any traversability threshold. The outcomes that do depend on the thresholds were tested at each of the 60 capability profiles (5 slope, 4 step and 3

roughness limits) separately: the traversability F1, which scores the predicted hazard map grid cell by grid cell, and the safe-path rate, the fraction of feasible start and goal pairs whose path planned on the reconstruction stays on truly traversable grid cells. Section 5's comparison holds at every profile, which is a statement about stability across profiles and not a test of the difference. On LuSNAR per-view depth agrees with the hazard map at $+0.73$ to $+0.86$ against Chamfer's $+0.43$ to $+0.50$, and with the safe-path rate at $+0.60$ to $+0.81$ against $+0.28$ to $+0.44$; on TartanGround Chamfer agrees at $+0.70$ to $+0.76$ and $+0.49$ to $+0.66$ against depth's $+0.13$ to $+0.30$ and $+0.12$ to $+0.30$, and on ETH3D Chamfer's agreement with the hazard map exceeds depth's at every profile. Of these comparisons and depth's advantage on lunar coverage, the TartanGround hazard-map comparison and that coverage advantage reach $p < 0.05$ under resampling of the model pool before correction, the lunar hazard-map and safe-path gaps and the ETH3D one do not, the TartanGround safe-path gap is not among the eleven comparisons of Appendix G, and none of those eleven survives Holm's correction. The rate at which untraversable grid cells are called traversable is not used as an outcome: unobserved grid cells count as untraversable, so a sparser reconstruction can lower that rate without being better. The grid's own parameters were varied separately, by re-scoring the stored predictions at every level of the four of them Section 6.2 counts, grid cell size from 0.05 to 0.5 m, the minimum points per grid cell and the height quantile and plane window that build the terrain fields, and of the region radius, which trims both clouds and is counted as its own axis. Those re-scorings use a six-profile subset of the capabilities, so the frozen grid is recomputed from them rather than read from Section 5, and models are compared only on the frame windows they share. On TartanGround the within-scene agreement between Chamfer and navigation coverage is $+0.80$ at the frozen grid and stays between $+0.78$ and $+0.82$ at all 19 settings, the 14 levels of Table H.1 and the five axes' reference settings, positive on all 20 scenes at 18 of them and on 19 at the other. The lunar re-scoring covers fewer models, five on three of its five scenes, so its value is not the one Section 5 reports, but it stays between -0.46 and -0.26 at every level, so the gap between the domains holds at every setting. ETH3D was not re-scored along these axes after its scans were registered, the planner has no obstacle-inflation radius to vary, and the 2.5 cm voxelisation that precedes rasterisation was held fixed.

C THE REAL DOMAINS AND THE POLAR TESTBED: CURATION AND CHECKS

AMEDS and Erfoud are the two real rover domains, and neither ships in a form a reconstruction benchmark can score directly: their scenes, frame windows, scale calibration, sensor conventions and per-view depth were built for this study, the same way for both. Scenes are intervals of a traverse with RTK-corrected navigation. The ground truth is the LiDAR, placed in the world by the navigation pose and accumulated over the scored window. Windows are chosen by the ground-truth camera path alone, and each model's frozen scale is fitted on a scene that is never scored. Every sensor convention is tested against the data, by competing hypotheses and a margin, rather than assumed. Per-view depth is the accumulated cloud projected into each frame through its ground-truth pose, nearest return per pixel. Appendices C.1 and C.2 give what is specific to each domain, AMEDS's image timestamps and Erfoud's intrinsics and camera pitch among it, and the checks run on the result. POLAR, a laboratory testbed of lunar polar terrain, was added after every rule in this paper was fixed, to test those rules rather than to inform them; Appendix C.3 gives how it was prepared and what it replicates.

C.1 AMEDS: THE LUNAR RESULT ON A REAL LUNAR ANALOGUE

AMEDS (Suter et al., 2023) was recorded by ESA's Interact rover on the slopes of Mount Etna, the standard terrestrial lunar analogue: a scientific camera on the rover's arm, a VLP-32C LiDAR, and an inertial navigation system with RTK correction. It is included to answer the one objection nothing inside LuSNAR can, that the lunar result rests on a single rendered dataset.

Preparation. Scenes are maximal intervals of traverses 2, 4 and 5 over which the navigation solution is RTK-fixed and no sensor stream has a gap. Images are taken at 2 Hz, so that an eight-frame window covers about the ground a LuSNAR window does, and are resampled into a 1440×576 pinhole camera that excludes the rover's own arm. The ground truth is thinned to one point per 3 cm voxel and culled to the scored cameras' view.

Conventions. The world frame, the attitude order and direction, the LiDAR mount and the camera extrinsic agree on all three traverses. One correction was needed. The image timestamps lag the navigation clock: photometric agreement between frames, warped through the LiDAR, peaks at -0.20, -0.30 and -0.25 s on the three traverses, each an interior maximum, and -0.25 s is applied. Left at zero, it would misplace every camera by 0.06 to 0.21 m, the rover moving at 0.26 to 0.84 m/s over the scored windows.

Windows and calibration. Windows are chosen by the ground-truth camera path alone, at least 2 m of travel over eight frames taken every fourth record of the 2 Hz image stream, so that a window spans 14 s, because the rover stops and starts: one segment is excluded outright, having moved 0.3 m in 76 s. Four scenes yield five scored windows (Table C.1). The frozen scales are fitted on a fifth, held-out scene, on three disjoint blocks of frames chosen the same way. That scene is never scored, but on AMEDS nearly every recorded frame, the validation frames included, views ground some scored window also sees; a scale fitted on shared ground can only help the models, so the plane's win on AMEDS is conservative. Four models' scales vary by more than a factor of two across those blocks. On seven of the seventy model cells the evaluator refuses the aligned prediction for overlapping less than one percent of the region scored, a scale failure rather than a score: Fast3R on four windows, and MV-DUSt3R+ on two and Spann3R on one, all among those four. Between 11 and 14 models are scored on every window.

Table C.1: AMEDS per scored window. Path is the ground-truth camera travel over the window (m); relief is the median height spread in one-metre columns (m); the best model is the lowest Chamfer (m) among the zero-shot models scored on the window, and ratio is that Chamfer over the relief; depth ρ is the median over models of the Spearman correlation between a model's own per-view depth and the LiDAR projected into the same pixels without the z-buffer, the registration check's value (keeping only the nearest return, as the stored depth maps do, raises it by up to 0.12); wrong counts the cells, one per donor, on which the wrong scene beats that best model, and plane the same cells for the flattened scene, whose verdict does not depend on the donor and so is counted once per donor cell.

Window	Path	Relief	Best model	Chamfer	Ratio	Depth ρ	Wrong	Plane
T2b@148	8.5	0.0975	MapAnything	1.051	10.78	0.91	2 / 2	2 / 2
T2c@33	11.8	0.1176	MapAnything	1.053	8.95	0.94	2 / 2	2 / 2
T2c@78	11.6	0.1106	Pi3	1.204	10.89	0.92	2 / 2	2 / 2
T4a@89	9.0	0.0977	MapAnything	0.434	4.44	0.89	1 / 2	2 / 2
T5a@171	3.6	0.0836	MUSt3R	0.348	4.16	0.73	0 / 2	2 / 2

Two checks on the result. The ground truth and the planted donor are both LiDAR scans with the same ring pattern around the camera, while the models' clouds are dense, and a distance can reward matching sampling. Thinned to one point per voxel, which removes density as a signal, the wrong scene still beats every model on 6 of 10 cells at 0.1 m and 5 of 10 at 0.2 m; only margins of 1.32 or less change sides. If the ground-truth cameras were misplaced, the models' clouds would be misregistered while the donor, planted on the same cameras, stayed self-consistent. Per-view depth does not depend on the multi-view poses, and it agrees with the projected LiDAR at the median ρ of Table C.1: 0.89 to 0.94 on every window where the wrong scene wins a cell, and lowest, 0.73, on the one window where it wins none.

Relief within AMEDS. The error to relief ordering of Table D.2 also holds window by window within AMEDS: the wrong scene wins both cells of each of the three windows whose ratio is 8.95 to 10.89, and one of the four cells of the two windows at 4.16 and 4.44.

Per-view depth on real lunar-analogue imagery. AMEDS ships no depth map, so a sparse depth ground truth is built as for both rover domains (Appendix C), through the virtual pinhole. It is not the same object as LuSNAR's dense rendered depth: between 17% and 38% of pixels carry a return, a median of 32% per window, and the coverage falls from the top of the frame to the bottom, so the metric is read on the pixels the LiDAR reached. The build is checked rather than assumed: every stored map is bit-identical to a fresh projection, and scoring the same points without the z-buffer reproduces the registration check's median Spearman on every window exactly, 0.912, 0.940,

0.917, 0.886 and 0.730. Keeping only the nearest return then raises those correlations by up to 0.12, which is the z-buffer removing points the surface occludes; the vertical-flip control collapses as it should, while the mirror control cannot discriminate on this geometry, because depth here is nearly a function of image row, and is reported for that reason rather than as a pass.

Scored this way, per-view depth rejects the wrong scene on all 10 AMEDS control cells, where Chamfer rejects 3: the best model's median absolute relative error runs from 0.063 to 0.181 against the placed wrong scene's 0.20 to 0.61. The recommendation of Section 5, to report per-view depth beside a surface metric, rests elsewhere on rendered depth on LuSNAR and TartanGround and on measured depth on ETH3D and Erfoud; it also holds on real rover imagery of lunar-analogue ground scored against LiDAR.

C.2 Erfoud: low relief at a Mars analogue, not a lunar one

LuSNAR is rendered lunar terrain and AMEDS is a lunar analogue: volcanic ground on Mount Etna, photographed because it resembles the Moon. Both leave the same objection open, that the failure follows the choice of a Moon-like place. Erfoud answers it. It is a rover dataset recorded by LAAS-CNRS at the Merzouga site in the Moroccan Sahara (Lacroix et al., 2020), one of three sites its authors describe as Mars-like planetary analogues rather than lunar ones, and it is the flattest real terrain in this study.

Preparation. Trajectory 21-1 carries an HDL-32 LiDAR and a 1920×1200 monochrome stereo pair at 2 Hz behind one hardware trigger, with RTK-corrected inertial navigation. Scenes are intervals of that traverse; the images are the left camera's raw frames undistorted to the published pinhole calibration, because the rectified images the archive also ships were produced with intrinsics it does not publish. Six windows were chosen by ground-truth camera path alone, one per scene, each covering 4.38 to 4.74 m of travel and 100 to 106 scans, with a minimum pairwise footprint separation of 45.3 m so that no two windows are the same ground. One of the six, M21d, is the validation scene on which the frozen per-model scale is fitted, so five are scored and M21d never enters a scored row in any role.

Conventions. Every convention was settled independently at three places on the traverse (camera frames 600, 1700 and 3400), with every verdict reproducing at all three. The point clouds are in the sensor frame: composing the robot and sensor poses reproduces the archive's own site-frame bounding boxes to 4.8×10^{-2} m against 1.34 m for the next hypothesis and 190 m for leaving the cloud untransformed. The attitude order is $R_z R_y R_x$ and the pose direction is robot-to-world by a margin of 6.0 to 9.8 times; the camera basis is the OpenCV optical convention by 1.35 to 2.9 times, with the reversed axis putting no points in the image at all. Gravity is $[0, 0, -1]$, the ground plane's normal lying 0.34 to 0.65 degrees from $+z$. The accumulated map is 0.0087 to 0.0110 m thick. Unlike AMEDS, no camera time offset is needed: the cross-correlation peak sits at exactly 0.00 s, in the interior of a ± 1.5 s grid, at all three places, which is what the archive's single hardware clock implies. An independent metric-scale check from the stereo pair, whose 0.2703 m baseline is not used anywhere else in the pipeline, puts block-matched range at 1.001 to 1.007 times the LiDAR's.

What the checks could not settle. On ground this flat, sharpness of the accumulated map does not resolve the LiDAR mount rotation, a 5% range scale or the mount yaw: the objective moves by at most 1.02, 1.05 and nothing interpretable across those grids, and the bounding-box and stereo checks settle them instead. One residual is recorded rather than corrected. The camera-rotation grid prefers a pitch 1 to 2 degrees below the published extrinsic, which is 0.17 to 0.35 m at a range of 10 m. Adopting it would be fitting the calibration to the evaluation, so the published extrinsic is used and the disagreement is stated here; every model was also re-scored with the camera pitched down by 1, 1.5 and 2 degrees, reproducing its logged Chamfer exactly at no offset, and the wrong scene then beats every model on 8, 8 and 9 of the 10 cells and the plane on all 10, against 7 and 10 at the published extrinsic, at a median margin falling from 2.59 to 1.46. The residual does not produce the result.

Relief. Measured as everywhere else in this paper, the median within-column height spread over 1 m columns, the five scored windows have a relief of 0.0287, 0.0349, 0.0369, 0.0534 and 0.0610 m, a median of 0.0369 m. The validation scene M21d, at 0.0335 m, is never scored and is not in that list;

over all six windows the median is 0.0359 m. That is between LuSNAR's 0.0151 m and AMEDS's 0.0977 m and far below ETH3D's 0.7531 m and TartanGround's 0.5952 m. The rover crosses flat interdune sand rather than climbing the dunes, and the figure is a property of the site rather than of the window selection: two further trajectories over the same ground give medians of 0.021 and 0.025 m.

Models. All fourteen models were run zero-shot on all five windows, 70 cells, under the frozen protocol and the calibrated-scale track. The best model on a window errs by 0.1737 to 0.4385 m, which is 4.59 to 7.19 times that window's relief and a median of 5.76 times. In those units, which are the units of Table D.2 throughout, the order is LuSNAR 16.69, AMEDS 8.95, Erfoud 5.76, ETH3D 1.12 and TartanGround 0.63. Measured instead against the ground truth's deviation from its own plane, the other normalisation this paper uses and one the two are easily confused for, it is LuSNAR 6.1, Erfoud 4.5, AMEDS 3.5, ETH3D 0.29 and TartanGround 0.21: the same except that Erfoud and AMEDS exchange places, because relief within a one-metre column ignores undulation broader than a column, which the deviation from a fitted plane includes.

Both controls fail here too. The wrong-scene control uses donors from the same site: the same sand, the same camera and the same illumination, minutes apart along one traverse, at least 45 m away. Reconstructing a different stretch of that ground outscores every one of the fourteen models on 7 of the 10 cells, at a median margin of 2.59, which is the rate AMEDS gives. Section 4's plane control gives the same answer on all 10 cells, and on all 5 of its windows in Appendix F's window study, where the inequality decides it as it decides every other window of the benchmark.

Per-view depth. Erfoud ships no per-frame depth either, so its maps are built as AMEDS's are (Appendix C), through the published pinhole. The maps are sparse, with a return on 11.5% to 13.3% of the pixels of a scored window and none inside the 2.18 m of range the scanner's lowest ring never reaches. They are checked against the models' own predicted depth through the evaluator's crop and resize: the eleven depth models agree with them at a median Spearman ρ of 0.966 to 0.997 per window, and the same maps turned upside down reverse the sign for every model and window, as a camera pitched down at open ground requires. Scored this way, per-view depth rejects the wrong scene on 9 of the 10 Erfoud control cells, where Chamfer rejects 3; the one it accepts is a pair whose best model already errs by 0.096 in relative depth.

The conclusion this domain supports is narrow. The plane, which carries no misplacement, beats every model on all 10 cells under the frozen protocol and on 6 of them when every submission receives the same similarity fit to the geometry (Appendix D): that failure follows the terrain rather than lunar imagery, rendering or a Moon-like site, and it appears on real desert photographed for other reasons. The wrong scene's win, 7 of 10 cells frozen and none after the same fit, is instead one of placement.

C.3 POLAR: A HELD-OUT REPLICATION ON A LUNAR POLAR TESTBED

Every rule in this paper was fixed on the domains above before POLAR was prepared: Equation 2 and its band, the two controls, relief correlation at 0.2 m against each raster's own plane, its rank composite with completeness, and the navigation outcomes. POLAR Traverse (Hansen et al., 2024) was then added to test them rather than to inform them. It is a NASA Ames testbed of regolith simulant shaped as lunar south polar terrain, lit to simulate the lunar poles, scanned by a FARO laser scanner, and photographed by a stereo bar moved along the bed in 24 straight traverses with a stereo pair every metre: two directions and two sun angles, crossed with two camera heights (1.3 and 0.65 m) and three pitches (14, 25 and 35 degrees). The archive numbers a traverse's direction and sun angle as its view i and its height and pitch as j, written ViTj below. It is one place. Every traverse crosses the same bed, about 19 by 5 m.

Preparation. The archive gives each traverse's camera poses in a frame of its own and the scan in another, and states no transform between them, so the transform was estimated. Each location's stereo pair was rectified with the published calibration, matched by semi-global block matching and triangulated over 0.8 to 3.5 m in front of the rig; the union was registered rigidly, without scale, to the scan by point-to-plane ICP from a grid of starts, and then refined on the near field, within 2.5 m of a camera and inside the scan. The pose convention was measured rather than assumed: read as

camera-to-traverse in the OpenCV basis, the optical axis points 0.66 to 1.69 degrees further down than the pitch the archive states on the accepted traverses, and the offset from the left to the right camera is 0.392 to 0.398 m along the camera's $+x$. A traverse is accepted if its refined residual has a median below 2.5 cm and at least half its points within 3 cm. That rule replaced a first one after the first stage's residuals had been seen, and before any model was run on POLAR. Eight of the 24 traverses pass, at medians of 0.55 to 2.46 cm, and the other sixteen lie at 2.77 to 21.61 cm. The three accepted traverses of one direction and sun angle fix the frozen per-model scale. They cross the scored bed, so on POLAR the scale is fitted on the ground it is scored on; that can only help the models, so the plane's win there is conservative. The other five are scored, each at two windows of eight consecutive locations (0 to 7 and 3 to 10, 7 m of travel), so the ten windows overlap in pairs and cover one bed from five rig configurations. The images are the left camera's, undistorted to the published pinhole, at the 100 ms exposure, the middle of fifteen, fixed beforehand. The ground truth is the scan thinned to 1 cm and cut to each scored camera's frustum out to 10 m. No visibility test is applied, so points hidden behind relief count against every submission alike. The preparation code will be released; the data are not redistributed.

Relief and the bound. Measured as everywhere else in this paper, the ten windows have a relief of 0.0366 to 0.0886 m, a median of 0.0513 m, between Erfoud's 0.0369 m and AMEDS's 0.0977 m, and a $\mathrm{dev}(h_t)$ of 0.072 to 0.075 m. All fourteen models were run zero-shot on all ten windows, 140 cells, under the frozen protocol. The best model on a window errs by 0.137 to 0.329 m, which is 2.07 to 6.43 times the window's relief (a median of 4.01) and 1.88 to 4.48 times its $\mathrm{dev}(h_t)$. Every window therefore lies outside the band, on the side where Equation 2 says the flattened target wins, and it wins on all ten, at 0.054 to 0.056 m.

The wrong scene on a single bed. The local construction plants a donor's camera-frame ground truth on the target's cameras, so every donor is matched on the rig's height and pitch. A first cell that ignored this, a 35 degree donor on a 14 degree target, was planted tilted by 21 degrees, scored 1.7 m and is discarded. A donor from the target's own traverse always overlaps the scored locations, since eleven locations admit no two disjoint windows, and the evaluator refuses it as a partial oracle. The first eight cells took the donor from another accepted traverse of the same height and pitch; only ground truth passes from a donor, so the calibration traverses serve as donors too. On 4 of them the donor runs the same way along the bed, and taken three locations on, it is the target's own ground three metres along: its completeness is 0.024 to 0.142 m, and its relief correlation with the target, 0.45 to 0.53, is above every model's on the same window. A 19 m bed has no second place, so every window was then given a donor that runs the other way and sees the bed from its opposite end. These are ten cells, from the accepted traverse at 35 degrees of the other sun angle and from four traverses that serve only as donors (V3T1, V4T2, V3T5 and V4T3), each admitted after its camera height above the scan and its pitch were checked against the accepted traverses of its class. They were added after the first eight had been seen, and both sets are reported. The wrong scene beats every model on 4 of the 4 same-direction cells and on 5 of the 14 reverse-direction cells (Table C.2).

The metrics. Chamfer rejects both substitutes on none of the 18 cells, since the flattened target beats the best model on all of them. On the 14 reverse-direction cells, relief correlation rejects both on 12, its rank composite with completeness on 13, and completeness alone on 2. On the four same-direction cells none of them does. There the metric reads the height field correctly: the displaced ground agrees with the target's heights more closely than any model does, and a score of shape ranks it first. What relief correlation cannot do is tell a reconstruction from a copy of the same ground displaced along its own broad relief, and Chamfer, which also ranks that copy above every model, cannot either. Averaged over the ten windows and the four navigation outcomes, the within-window rank agreement with navigation is 0.50 for completeness, 0.41 for the composite, 0.37 for Chamfer and 0.26 for relief correlation. Beyond completeness it is -0.12 for Chamfer, $+0.11$ for relief correlation and $+0.08$ for the composite: here, as on the other domains, what Chamfer shares with navigation is completeness's.

What POLAR does and does not add. The bound and the plane control reproduce on a domain added after the rules were fixed, on real imagery of a lunar analogue with a laser-scanned ground truth. Relief correlation's rejection of both substitutes reproduces on 12 of the 14 cells whose donor runs the other way. POLAR is one bed, its ten windows overlap in pairs, and three of its choices

were made after something had been seen: the registration rule, after the first stage's residuals and before any model ran; the first control cell, discarded after it scored 1.7 m; and the reverse-direction donors, four of them traverses admitted only as donors, after the first eight controls.

Table C.2: POLAR per scored window: relief (median within-column spread over 1 m columns), $\mathrm{dev}(h_t)$, the flattened target's Chamfer and the best model's (m); wrong-scene cells won against every model, by whether the donor runs the same way along the bed as the target or the other way; and relief correlation for the best model on it and for each donor, the same-direction donor first; a window without a same-direction donor lists its two reverse donors. n/a marks a window with no same-direction donor.

				Best model		Wrong scene wins		Relief corr.	
Window	Relief	$\mathrm{dev}(h_t)$	Plane	Model	Chamfer	same	reverse	best model	wrong
V1T1 s0	0.0512	0.073	0.055	DUSt3R	0.329	n/a	1 / 1	+0.20	+0.31
V1T1 s3	0.0886	0.074	0.056	DUSt3R	0.319	n/a	1 / 1	+0.22	+0.31
V1T2 s0	0.0462	0.073	0.055	MASt3R	0.272	1 / 1	1 / 1	+0.15	+0.47, -0.02
V1T2 s3	0.0513	0.073	0.055	Depth Anything 3	0.238	1 / 1	0 / 1	+0.38	+0.53, -0.04
V1T3 s0	0.0513	0.075	0.056	Pi3	0.146	n/a	0 / 2	+0.52	+0.25, +0.03
V1T3 s3	0.0513	0.072	0.054	VGGT-Ω	0.142	n/a	0 / 2	+0.46	+0.18, +0.11
V1T5 s0	0.0457	0.073	0.055	CUT3R	0.180	1 / 1	0 / 1	+0.33	+0.45, +0.13
V1T5 s3	0.0660	0.073	0.055	CUT3R	0.137	1 / 1	0 / 1	+0.31	+0.51, +0.13
V3T3 s0	0.0371	0.073	0.054	VGGT-Ω	0.152	n/a	2 / 2	+0.74	+0.30, +0.19
V3T3 s3	0.0366	0.074	0.055	VGGT-Ω	0.218	n/a	0 / 2	+0.76	+0.19, +0.03

D CONTROL DETAIL

Which population a count belongs to. Several quantities in this paper are reported over populations that differ and that a reader cannot tell apart from a bare count. They are named once here, with their sizes, and referred to by name afterwards.

The *control cells* are the wrong-scene pairs of Table 2, a target window, a donor and the model pool scored on it: 126 in all, of which the 96 with a same-domain or unseen donor (11 LuSNAR, 10 AMEDS, 10 Erfoud, 8 ETH3D and 57 TartanGround) carry every per-domain count, and the 30 cross-domain ones are reported apart; Table 2 counts them after the protocol filter, which removes the relief-threshold twins that carry a duplicate Chamfer. The *audit cells* are the wrong-scene cells whose clouds were saved and reproduce their logged control Chamfer, which the metric studies of Appendices E and F rescore: 106, the 96 together with the ten cells of ETH3D's electro sequence, which Table 2 does not list. Studies that cover only the 96 say so where they are reported. The *plane-control windows* are the 123 scored frame windows of the benchmark, 20 LuSNAR, 5 AMEDS, 5 Erfoud, 11 ETH3D, 40 TartanGround and StereoLunar's 42, one per scene and illumination, which the plane reaches because it needs no donor and no inference; StereoLunar's 42 wrong-scene cells are reported apart from Table 2, below, for their altitude confound. The *retrieval windows* are the 84 target windows of Appendix J's retrieved-donor study: 54 TartanGround, since TartanGround's 57 cells sit on 54 windows, three of them carrying two hand-picked donors; 9 ETH3D, its 8 and electro's one window, which carries ten donors; 11 LuSNAR; and 5 each for AMEDS and Erfoud, whose windows carry two donors each.

Why three constructions. The WORLD construction hands the evaluator the donor's own ground-truth camera poses, so the alignment fits two exact camera tracks. No model achieves that, so the margin it produces is not a like-for-like comparison. The LOCAL construction removes the donor's own track by removing the fit: the donor's geometry is taken per frame in that frame's own camera coordinates and re-planted on the target's ground-truth poses, so prediction and ground-truth poses coincide and the alignment is an identity. That hands the donor the target's exact poses instead, which no model has either, so LOCAL is an oracle-pose construction, and it is the one Table 2 reports. The LOCAL-MODEL construction plants the same wrong terrain on a real model's estimated poses for the identical cell, matching the pose error exactly. It is run on every audit cell with every roster model that has a stored prediction there, 1320 pairs over the 106, and compared both with the

best model on the cell and with the model whose poses it borrows, so pose quality is held equal in the second comparison. Table D.1 sets the three side by side.

Table D.1: The three wrong-scene constructions. All three submit a real surface of the wrong place, scored against the target's ground truth by the same evaluator; they differ only in whose camera poses carry the donor's geometry.

Construction	Donor's geometry planted on	Pose advantage
WORLD	the donor's own poses, fitted to the target's	none; the fit handicaps it
LOCAL	the target's ground-truth poses	the target's exact poses
LOCAL-MODEL	a model's estimated poses for the cell	none; that model's pose error

The like-for-like comparison strengthens at the larger sample. Against the model whose poses it borrows, the wrong terrain wins 150 of 154 pairs on LuSNAR, 107 of 126 on AMEDS and 125 of 140 on Erfoud, at median margins of 2.71, 2.89 and 2.81: a reconstruction of a different place, carrying that model's own pose error, scores about three times better than the model does. It wins on every one of the 11 LuSNAR, 10 AMEDS and 10 Erfoud cells for at least one model's poses.

Against the best model on the cell, over all fourteen models the wrong terrain beats the best model on 78 of 154 LuSNAR pairs, 85 of 140 Erfoud pairs and only 51 of 126 AMEDS pairs, because a control planted on a weak model's poses inherits that pose error and then loses to the strongest reconstruction on the cell. Restricted to four of the strongest models it wins 28 of 44 LuSNAR pairs and 27 of 40 Erfoud pairs but 19 of 40 AMEDS pairs, so the pose-matched comparison against the best model reproduces Table 2's verdict on LuSNAR and Erfoud and not on AMEDS, at either roster size.

The negative side is unchanged at the larger sample: against the best model the wrong terrain wins 0 of 96 ETH3D pairs, 0 of 120 electro pairs and 3 of 684 TartanGround pairs.

What each construction removes. Moving from WORLD to LOCAL removes the fit to the donor's own track and makes the control score better rather than worse, so that fit was penalising it. LOCAL still plants the donor on exact poses, the target's own, and only LOCAL-MODEL removes that advantage; against the best model it keeps the verdict on LuSNAR and Erfoud and loses it on AMEDS, as above.

Where the two controls disagree. The plane and the wrong scene disagree on 15 of the 106 cells. On 14 a plane wins and a different place does not, 7 of them TartanGround cells, 3 AMEDS, 3 Erfoud and one LuSNAR; on the remaining one a different place wins where a plane does not. Being one surface at the scale of the models' error is therefore where the failure lives without being the whole of it.

Scenes never ranked. The unseen-scene row of Table 2 covers 16 TartanGround environments the study never ranked models on, each held out entirely, and the control is rejected on all 35 of its cells (median margin 0.30), so the diagnostic carries to terrain the benchmark does not rank models on, where the models are scored only to be compared with the control.

Descent imagery and altitude. Planted in a nadir camera's frames, a donor keeps its own distance below the camera, so on StereoLunar the wrong scene's error follows the altitude gap between donor and target rather than its terrain. A height-matched construction removes that gap by moving each donor frame along the optical axis until its median depth equals the target frame's, a vertical shift that leaves the relief untouched; it is the altitude normalisation a descent-camera control needs, and it should be the default there. It lowers the control's Chamfer on all 42 cells, its median from 226 to 151 m, and raises its wins from 12 to 16 of 42 under the frozen scale and from 17 to 24 under a per-scene similarity fit. It does not equalise ground sample distance, since each cloud keeps its own point spacing, which is one reason the descent camera is reported apart from the rover domains.

Cropping everything to the support it shares. A donor ground truth is dense and complete over the whole footprint while a model has holes where it failed, and Chamfer and completeness both notice that difference rather than the difference in place. To remove it, all three clouds on a cell,

the target, the substitute and the best model, are cropped to the grid cells that all three occupy at the 0.10 m tolerance, and rescored by the evaluator's own Chamfer; the unmasked values reproduce the logged plane-control scores to a relative 10^{-6} first.

The plane control is unmoved, which is the part that carries the mechanism. On shared support the flattened ground truth still beats the best model on 11 of 11 LuSNAR, 10 of 10 AMEDS and 10 of 10 Erfoud cells, on none of ETH3D's 8 or electro's 10, and on 6 of TartanGround's 57 against 7 unmasked. Its median margin rises on LuSNAR, from 6.88 to 7.00, and falls on AMEDS and Erfoud, from 5.16 to 3.89 and from 5.99 to 2.15, but stays above one on every low-relief cell.

The wrong scene strengthens on regolith and sand and weakens on Etna. LuSNAR goes from 10 of 11 to 11 of 11, the one change being a cell that flips toward the substitute, and Erfoud from 7 of 10 to 9; TartanGround goes from 1 of 57 to none. AMEDS falls from 7 of 10 to 5, two cells flipping to the model, but the mask keeps a median of 93% of the ground truth on LuSNAR and 86% on Erfoud and only 54% on AMEDS, so the AMEDS figure is computed on roughly half the support the other domains keep and is a different measurement from the one beside it. On LuSNAR and Erfoud, where the crop removes little, the control wins more, so its wins do not come from carrying a footprint the models lack.

A classical pipeline loses to the plane as well. COLMAP (Schönberger & Frahm, 2016), given the dataset's own intrinsics and every departure from its defaults in its favour, reconstructs Erfoud well and beats the best network on 2 of its 5 windows, yet the flattened ground truth beats it on 28 of the 29 low-relief surfaces it produces, at a median margin of 28.0, and on 7 of its 13 high-relief ones at 1.40: a pipeline with no learned prior fails the plane control on low relief as well (Appendix M.1).

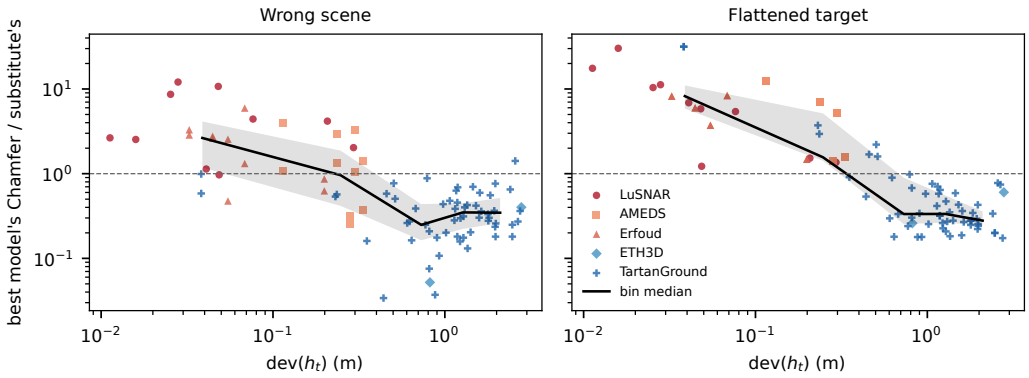


Figure D.1: Both controls' margins against the target's relief. One point per audit cell, at the best zero-shot model's Chamfer over the substitute's, so a point above the dashed line is a cell the substitute won; the line and band are the median and interquartile range in five bins of $\mathrm{dev}(h_t)$. The wrong scene's margin and the plane's fall together as the relief grows, and cross 1 between $\mathrm{dev}(h_t) \approx 0.2$ and 0.5 m, where the models' errors lie.

Would a fit at evaluation time rank the models better? The same fit that removes the wrong scene's placement advantage could be adopted as the protocol, so it was checked against the task: every zero-shot model's own fits (Table D.4) ranked within each window against the navigation outcomes, which are computed from the model's unfitted map because that is the map a rover would plan on. A similarity fit lowers Chamfer's agreement with coverage on LuSNAR from $+0.16$ on the models this comparison covers, against Table 3's $+0.20$ over every model, to -0.37, on AMEDS from $+0.84$ to $+0.64$ and on TartanGround from $+0.80$ to $+0.57$; a rigid fit raises it only on the two real rover domains, to $+0.90$ on AMEDS and $+0.85$ on Erfoud. A fitted score measures a surface the rover never had, so the fit is a diagnostic of placement, as step 4 of the checklist uses it, and not a better protocol.

How much of the error a global fit can remove, on four domains. The answer is not the same on all four domains (Table D.4). On LuSNAR and TartanGround about 45% of Chamfer survives

Table D.2: Terrain relief and model error. τ is the 0.10 m F-score tolerance. Each median names its own population. Relief is the median over every scene of the domain that carries a relief descriptor, which on LuSNAR and TartanGround is more than the scored pool of Table 1, since relief is a property of the terrain and is measured whether or not a scene is scored; several of the additional scenes serve as the wrong-scene control's donors. The final column, the best available model's Chamfer divided by that relief, is a median over the scored pool only. AMEDS and Erfoud are scored by window, so both of their medians are over their five scored windows; Erfoud's held-out validation window, M21d, is not among them.

Domain	Relief over	Relief (m)	Relief / τ	Ratio over	Best error / relief
LuSNAR (lunar)	9 scenes	0.0151	0.15×	5 scenes	16.69
AMEDS (real, lunar analogue)	5 windows	0.0977	0.98×	5 windows	8.95
Erfoud (real, desert sand)	5 windows	0.0369	0.37×	5 windows	5.76
ETH3D (real, laser)	11 scenes	0.7531	7.53×	11 scenes	1.12
TartanGround (simulation)	36 scenes	0.5952	5.95×	20 scenes	0.63

the best 7-DoF similarity fit, which is the result Section 4 reports: the error is not only a global misplacement. On AMEDS and Erfoud the residual share is 0.104 and 0.141, so on those two real rover domains most of the error *is* removable by a global scale, rotation and offset, as AMEDS's gross scale failures (Appendix C.1) lead one to expect, and the error is irreducible by a global fit on LuSNAR and TartanGround only. The control is scored under the frozen protocol's alignment, which fixes its number but not its meaning. The control carries none of the global misplacement a similarity removes: it is a correctly scaled surface planted on the target's own ground-truth poses, so its alignment is an identity. The decomposition already stores each model's Chamfer after the best per-cell 7-DoF similarity fit, and the same fit, through the same code on the same ground truth, was given to both substitutes on all 106 audit cells, so that every submission is aligned the same way, by its geometry, as a benchmark that fits a similarity to the point clouds would align it (Table D.3). Every cell first reproduces its logged wrong-scene and plane Chamfer from the stored clouds, and every comparison is made on the matched $200{,}000$-point subsamples the models' fits use, since a subsampled Chamfer set against a full-cloud one favours the full cloud; on that footing the models' fit alone leaves the plane 6 of Erfoud's cells and 2 of AMEDS's, and the frozen protocol reproduces Table 2. With every submission fitted, the wrong scene beats every model on 4 of LuSNAR's 11 cells and on none of AMEDS's or Erfoud's, and the plane on 7 of LuSNAR's, 2 of AMEDS's, one window counted once per donor, and 6 of Erfoud's, on three windows; on the 22 TartanGround cells the decomposition covers the wrong scene wins none and the plane one. On the two real rover domains the wrong scene therefore wins because the models are misplaced, not because their surfaces are worse, and so does the plane on four of AMEDS's five windows, while on LuSNAR both failures, and on three of Erfoud's five windows and one of AMEDS's the plane's, survive a fit that removes every global misplacement. The fit is an oracle, since it reads the ground truth, and the frozen protocol forbids it for the primary score precisely because it would hide scale error, so this is a diagnosis of what Chamfer fails to separate rather than a better protocol. On the high-relief cells whose models were not decomposed a fitted substitute can only be compared with the best model's frozen Chamfer, which a fit could only lower, so it is rejected by that bound or left undecided, never counted as a win.

Table D.3: Both controls under three alignments, on the audit cells whose models carry the 7-DoF decomposition: the frozen protocol, a per-cell similarity fit to the geometry given to the models alone, and the same fit given to every submission, the substitutes included. Each entry counts the cells on which the substitute beats every model, all on matched $200{,}000$-point subsamples; every cell first reproduces its logged Chamfers.

		Frozen protocol		Models fitted		Every submission fitted	
Domain	Cells	Wrong	Plane	Wrong	Plane	Wrong	Plane
LuSNAR	11	10	11	3	6	4	7
AMEDS	10	7	10	0	2	0	2
Erfoud	10	7	10	0	6	0	6
TartanGround	22	1	1	0	0	0	1

The relief-normalised columns of Table D.5 cannot be extended the same way: AMEDS and Erfoud are scored over frame windows and carry a per-window relief rather than the scene-level descriptor those columns divide by.

Table D.4: The share of Chamfer that survives the best 7-DoF similarity fit, median over every decomposed cell of the domain. A larger share means less of the error is a global misplacement. Cells here are every decomposition that completed, on every window, which is why the two counts exceed those of Table D.5.

Domain	Cells	Residual share
LuSNAR (lunar)	280	0.448
AMEDS (real, lunar analogue)	63	0.104
Erfoud (real, desert sand)	70	0.141
TartanGround (simulation)	478	0.434

Table D.5: Error decomposition at window 120 of each scene, the window each scene's relief is attached to: LuSNAR's five scenes and the fifteen TartanGround scenes that carry a relief descriptor, 250 cells. Residual is the Chamfer surviving the best 7-DoF similarity fit of prediction to ground truth, so it is the part no global correction removes, and residual share is that residual as a fraction of the protocol Chamfer; medians over models and scenes. Table D.4 gives the shares over every window.

Domain	Cells	Error / relief	Residual / relief	Residual share
LuSNAR (lunar)	70	45.20	21.76	0.503
TartanGround (simulation)	180	1.78	0.82	0.457

Relief under other definitions. The separation does not depend on how relief is defined, since over column footprints of 0.25 to 2 m, three within-column spread estimators and four ground-truth sampling strides, lunar error stays above 5 times relief and above ETH3D's and TartanGround's ratios at all 48 settings. Relief as pinned is the median, over one-metre columns of the gravity-aligned ground truth read at stride two, of the within-column standard deviation of height. Each of those three choices was varied independently: column footprints of 0.25, 0.5, 1 and 2 m; the standard deviation, the median absolute deviation scaled to be comparable with it, and the interquartile range; and ground-truth strides of 1, 2, 4 and 8. Table D.6 gives the best model's error over relief under the twelve footprint and estimator settings at stride two, computed as the pinned figures are. Relief itself moves with the definition, since a wider column admits more slope, but the lunar ratio never falls below 5 and stays above ETH3D's and TartanGround's at every one of the 48 settings. ETH3D's own ratio does not stay near one: the median absolute deviation reads a column that is mostly a flat facade with a few points off it as nearly flat, placing ETH3D at up to 49 times relief. The mechanism is therefore stated as a separation between domains, which every definition preserves, and not as a fixed threshold at one, which none of them makes invariant.

Table D.6: Relief sensitivity. Median over scenes of the best zero-shot model's Chamfer divided by the scene's relief, at ground-truth stride two, under each column footprint and within-column spread estimator. The pinned definition is the one-metre standard-deviation cell. Over strides one to eight the LuSNAR ratio ranges from 5.2 upward and never falls to ETH3D's at the same setting, nor to TartanGround's; neither the other strides nor TartanGround's ratios are shown.

Column (m)	LuSNAR			ETH3D		
	std	MAD	IQR	std	MAD	IQR
0.25	54.77	51.83	38.77	8.91	48.84	15.61
0.5	28.79	27.80	20.67	2.59	19.58	3.29
1	16.69	15.62	10.61	1.12	11.01	0.85
2	12.50	11.30	7.74	0.71	3.25	0.47

Other aggregations and tolerances. Density-aware Chamfer at four temperatures, point-to-plane distance, density-equalised Chamfer and dense F-score ladders up to 0.5 m accept the wrong scene on the lunar cells as Chamfer does, and TartanGround answers as ETH3D does (Table E.1 and Appendix M.1).

D.1 BEYOND THE AUDIT

The audit's cells were fixed before its results were read. The tests below were run afterwards, to ask whether the audit's size, its hand-picked donors, the placement it never probed, its averaging over vehicles and its one domain for more views hide anything. None changes a number reported for the audit; each is reported here as it came out.

More windows. The audit fixed its cells before any was scored, and only LuSNAR leaves windows it did not use. There all fourteen models were scored at windows 360, 600 and 840 of Moon_3, Moon_5 and Moon_7, where no wrong-scene control had run; with the donors those scenes' audit cells use, the wrong scene beats every model on 6 of the 9, at margins of 1.57 to 11.66 where it wins, against 10 of 11 in the audit. The flattened target beats every model on 8 of the 9 windows, the other lying inside the band, a result already counted among the 123 windows of Section 4. AMEDS and Erfoud have no further window under their own rules: on AMEDS every scene's at most two longest non-overlapping windows with 2 m of camera path are the ones scored, and Erfoud's trajectory 21-1 revisits its own sand, so no further window of it lies 45 m from every chosen window and from the validation scene's ground. Of three further drives of the same rover at the same site, prepared by the same code with the convention checks of Appendix C.2 rerun on each, trajectory 22-3 was set aside before any model saw it, because a second rover drives a few metres ahead of the camera throughout, and trajectory 22-1's LiDAR ran at 1 Hz and only over its first 465 m, so no window of it meets the LiDAR rule. Trajectory 22-4 does. Its images run at 4 Hz, so its windows take every eighth frame to span the same 14 s as 21-1's, and one camera row recording a different mount is excluded. Every scan inside a scored window's accumulation interval, placed as the evaluator places it, matches its documented box to a median of 0.6 to 1.1 cm per window (21-1's windows: 0.7 to 4.7 cm), and each window's ground truth agrees with itself, its even-numbered scans against its odd-numbered ones, to a median of 1.7 to 2.1 cm (21-1's: 1.9 to 2.4 cm). Chosen by the same rule against every window of both drives and the validation ground, it adds 5 windows at least 46 m from every other, with a relief of 0.0381 to 0.0672 m. With each window's two nearest other new windows as donors, the wrong scene beats every model on all 10 cells and the flattened target on all 5 windows. Pooled with the audit, the wrong scene beats every model on 40 of the 50 low-relief cells, 80% (scene-bootstrap interval 67 to 92%), against 24 of 31 in the audit alone, and on every new window the flattened target wins exactly where Equation 2 says it must.

The proposed metric on new ground. Relief correlation and its composite were chosen on the audit's windows. On the extension's 14 windows, with both probes applied as on POLAR to every zero-shot model's aligned cloud and every extension wrong-scene cell, relief correlation rejects both substitutes on 9 of 9 LuSNAR cells and 10 of 10 Erfoud cells, its composite with completeness on 9 of 9 and 8 of 10, and Chamfer on 1 of 9 and 0 of 10. Beyond completeness, relief correlation's agreement with the four navigation outcomes is $+0.05$ on LuSNAR and $+0.25$ on Erfoud, against Chamfer's -0.08 and -0.04. On ground it was not chosen on, the metric keeps the separation it was chosen for.

Donors retrieved by rule. Table 2's donors were chosen by hand. With the donor retrieved instead by the rule of Appendix J, the nearest scene-window over ten terrain descriptors, fixed before any score was read, the wrong scene beats every model on 18 of the 21 low-relief windows, the same windows the best hand-picked donor wins, and on 6 of TartanGround's 54, against 1 for the hand-picked donors: retrieval makes the control harder for the models, not easier.

How far must the right ground move? On POLAR a donor that was the target's own bed three metres along beat every model under the four metrics Appendix C.3 reports. Every domain can be asked that question without a donor. The target's own ground truth, moved rigidly by d in its ground plane along each of its two principal axes and in both directions, is scored against the target under Chamfer, completeness, relief correlation and their composite, and a window counts as rejecting the

copy at d when all four copies score worse than the model best on that metric (Table D.7). On none of the 21 low-relief windows does any of these four metrics reject a copy moved by 1 m, so at least one of its four copies beats the best model there. At 3 m relief correlation rejects it on 7 and at 5 m on 14, where Chamfer rejects it on 18. On TartanGround a copy moved by 2 m is rejected by Chamfer on 39 of 54 windows and by relief correlation on 18 of the 22 it covers. Moved by less than the 2 to 5 m at which these counts rise, and by less than the models' own error, a copy of the right ground still agrees with the target's heights more closely than any model does. Relief correlation rejects a different place, and Table 2's donors are other scenes, or on Erfoud ground at least 45 m away; it does not certify that a reconstruction is placed to within a metre. Per-view depth compares along each camera ray instead of between clouds, so it is the natural check of placement, and it was scored on the same copies: the moved ground truth rendered into each target camera, nearest point per pixel, through the evaluator's own per-frame depth score. On LuSNAR, whose ground truth is dense, the unmoved copy renders onto a median 67% of the target's pixels at an error of 0.0036, against 0.0154 for the best model, so the test is sound there, and depth rejects a copy moved by 1 m on 1 of the 11 windows, by 2 m on 3 and by 5 m on 7: it is no finer a check of placement than the surface metrics. On AMEDS and Erfoud the ground truth is sparse LiDAR, and even the unmoved copy renders onto only 28% and 13% of the target's pixels, so the test cannot be made there with this ground truth. On this terrain nothing tested here certifies placement to within a metre.

Table D.7: The target's own ground truth moved rigidly by d in its ground plane, along both principal axes and in both directions. For each metric, the smallest d in metres, of 0.1, 0.25, 0.5, 1, 2, 3 and 5 m, at which some window rejects the copy, all four copies scoring worse than the model best on that metric, and in brackets the windows that reject it at 5 m; none, no window by 5 m. The composite ranks relief correlation and completeness over the models and the copy. Completeness, relief correlation and the composite come from the metric grid, which covers 22 of TartanGround's 54 windows.

Domain (windows)	Chamfer	Completeness	Relief corr.	Composite
LuSNAR (11)	2 (11)	2 (6)	2 (9)	2 (7)
AMEDS (5)	5 (2)	5 (2)	5 (1)	none (0)
Erfoud (5)	3 (5)	none (0)	3 (4)	3 (3)
TartanGround (54; 22)	0.25 (49)	0.5 (20)	0.5 (21)	0.5 (20)

Two vehicles. The navigation outcomes average sixty vehicle capabilities. Split into two vehicles that share none, a cautious one (maximum slope 10 or 15 degrees, step 0.05 or 0.10 m, 12 capabilities) and a capable one (25 or 30 degrees, 0.20 or 0.30 m, 12), with the three outcomes each capability carries (traversability, reachability, safe paths), the agreement beyond completeness keeps its pattern on four of the five domains for both (Table D.8): relief correlation's lies between $+0.22$ and $+0.38$ on AMEDS, Erfoud and TartanGround and above Chamfer's, which lies between -0.23 and $+0.18$, and on LuSNAR every metric's is at most 0.16 in magnitude for both vehicles. On ETH3D it holds for the cautious vehicle only: for the capable one Chamfer's $+0.10$ exceeds relief correlation's $+0.01$. The result is not a property of one vehicle's thresholds; it remains a ground vehicle's, not a manipulator's or a flying platform's.

Table D.8: Within-window rank agreement with navigation beyond completeness (the partial correlation with completeness held fixed), averaged over windows and over traversability, reachability and safe paths, for two vehicles that share no capability.

	Cautious vehicle			Capable vehicle		
Domain	Chamfer	Relief corr.	Composite	Chamfer	Relief corr.	Composite
LuSNAR	-0.16	-0.07	-0.05	-0.07	-0.04	-0.04
AMEDS	-0.23	+0.22	+0.25	-0.22	+0.38	+0.35
Erfoud	+0.18	+0.23	+0.22	-0.09	+0.25	+0.24
ETH3D	0.00	+0.05	+0.04	+0.10	+0.01	0.00
TartanGround	+0.05	+0.26	+0.21	+0.15	+0.32	+0.28

More views on the real rover domains. The view study was run on LuSNAR (Appendix F). On every audit window of AMEDS and Erfoud the same eight models were rerun on the window's own 28-frame span at 15 and 29 views, every frame the span holds, so more views shorten the baseline and add no ground; the ground truth, the plane and $\mathrm{dev}(h_t)$ are recomputed for each count. The flattened target beats the best of the eight on 5, 5 and 5 of AMEDS's five windows at 8, 15 and 29 views and on 5, 5 and 5 of Erfoud's, with the best model's error a median 3.5, 3.1 and 2.6 times $\mathrm{dev}(h_t)$ on AMEDS and 4.5, 4.5 and 4.2 on Erfoud. On AMEDS's T2b window at 8 views seven of the eight models' rows exist.

E THE PRINCIPAL METRICS TESTED AGAINST THE CONTROLS, AND RELIEF CORRELATION

Section 4 predicts that any metric measuring how far points are will accept a surface at about the right height when the target is a single surface at the scale of the models' error. Table E.1 collects the principal metrics this paper tested against the two controls, one row each (Table G.1 adds the other statistics of each family), with the rule its best model is chosen by and the table that reports it in full: the distance families confirm the prediction on every low-relief domain they were run on, accuracy alone apart, which rejects the wrong scene on most AMEDS and Erfoud cells (Table G.1), and the metrics that compare where the relief is, rather than how far the points are, are the ones that reject both substitutes. The rest of this appendix tests that converse: whether such a metric rejects the wrong scene where Chamfer cannot, and whether it then ranks the models better.

Definition. The ground truth fixes one frame, its principal axes, and one raster over its footprint (the central 98% in each in-plane direction). Every cloud is rasterised by mean height per cell, and each raster has its own least-squares plane removed. Relief correlation is the Pearson correlation between a prediction's residual heights and the ground truth's over the cells both occupy, over at least 50 shared cells. It needs only the ground truth and the submitted cloud, so it also scores the three models that output no depth map. A flat surface has no residual and scores 0; a different place scores only for relief that happens to lie where the target's does; a global vertical offset, tilt or positive vertical scale does not change it, although an in-plane shift does (Table D.7). Pearson is undefined at zero variance rather than zero there, so the 0 is a convention and needs a threshold: a prediction whose residual has a standard deviation of 10^{-9} m or less is scored 0, and a cell whose *ground truth* residual has no variance is dropped rather than scored. The raster cell was fixed at 0.2 m before any result was seen, and 0.1 and 0.5 m are reported as sensitivity.

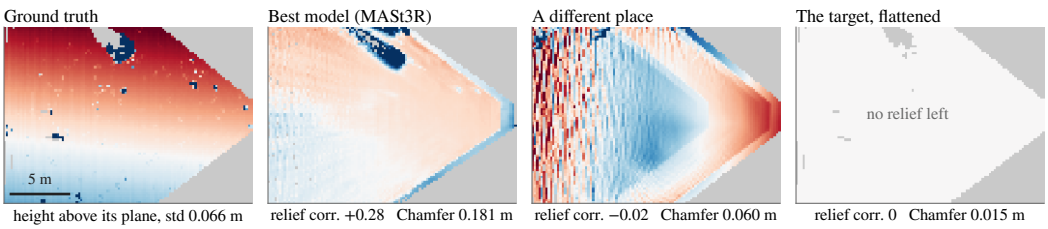


Figure E.1: What relief correlation compares, on the LuSNAR cell of Figure 1: unlike there, each cloud's height above (red) or below (blue) its own least-squares plane, on the metric's 0.2 m raster and each on its own symmetric colour scale, saturating at that panel's 98th percentile of absolute height; the std printed in a panel is its raster's standard deviation of that height, which weighs the boulders far more than the mean absolute deviation $\mathrm{dev}(h_t)$ of Figure 2 does. The stripes of the different place are the seams where its planted frames meet, as in Figure 1. Chamfer ranks the flattened target first and the model last; relief correlation ranks the model first and gives the flattened target exactly 0.

Why it passes the control and loses the task. The three invariances that make relief correlation reject a wrong place are the same three that detach it from navigation, which is worth stating because Section 5 reports the outcome without the mechanism. Removing each cloud's own plane deletes

Table E.1: The principal metrics tested against the controls: on how many cells each rejects the wrong scene and the flattened target, the substitute scoring worse than the best model. Best model is the model best on the row's own metric, or the model best on Chamfer, whose cloud the check was run on. The flattened target is reported on the three low-relief domains, where it is the substitute that matters, and on TartanGround, where it beats Chamfer's best model on 7 of the 57 cells against the wrong scene's 1; n/a marks a metric not scored on that domain or against that substitute, or, for the reference-free photometric row, the flattened target it is blind to by construction, scoring it exactly as the target's own ground truth, so that a count there would test the ground truth rather than the plane (Appendix K). Per-view depth faces the flattened target through the plane ray-cast into the target's cameras (below), and its entries there count windows. Density-aware Chamfer spans its four temperatures and was run on 15 of TartanGround's cells. The relief composites count TartanGround out of 56 because on the last cell one model has no calibrated alignment and the search needs every criterion for every model of a cell (Appendix M.2); Table 3, which keeps that cell, counts the composite's wrong-scene rejections there as 56 of 57. The last column names where the row is reported in full, this table where no other does.

Metric	Best model	Wrong scene rejected					Flattened target rejected				Detail
		LuSNAR	AMEDS	Erfoud	ETH3D	TartanGround	LuSNAR	AMEDS	Erfoud	TartanGround	
Distance to the nearest point											
Chamfer	own	1 / 11	3 / 10	3 / 10	8 / 8	56 / 57	0 / 11	0 / 10	0 / 10	50 / 57	Table 3
F-score at 0.1 m	own	1 / 11	3 / 10	1 / 10	8 / 8	51 / 57	n/a	n/a	n/a	n/a	Table G.1
Completeness	own	2 / 11	4 / 10	2 / 10	7 / 8	54 / 57	3 / 11	2 / 10	0 / 10	54 / 57	Table 3
Point-to-plane, mean	own	1 / 11	n/a	n/a	8 / 8	56 / 57	n/a	n/a	n/a	n/a	Table M.1
Density-equalised Chamfer	own	4 / 11	n/a	n/a	8 / 8	56 / 57	n/a	n/a	n/a	n/a	Table M.1
Density-aware Chamfer, four α	own	0 to 1 / 11	n/a	n/a	n/a	12 to 13 / 15	n/a	n/a	n/a	n/a	this table
Non-local matching											
Earth mover's distance	Chamfer's	1 / 11	0 / 10	0 / 10	6 / 8	39 / 57	0 / 11	0 / 10	0 / 10	15 / 57	this table
HyperCD	Chamfer's	1 / 11	3 / 10	3 / 10	8 / 8	53 / 57	0 / 11	0 / 10	0 / 10	49 / 57	this table
Geodesic Chamfer	Chamfer's	0 / 11	2 / 10	0 / 10	7 / 8	50 / 57	0 / 11	0 / 10	0 / 10	50 / 57	this table
Orientation											
Normal consistency	own	8 / 11	n/a	n/a	8 / 8	54 / 57	n/a	n/a	n/a	n/a	Table M.1
Normal consistency, $k = 30$	Chamfer's	5 / 11	8 / 10	10 / 10	8 / 8	52 / 57	1 / 11	2 / 10	2 / 10	51 / 57	Appendix E
Where the relief lies											
Relief correlation	own	9 / 11	10 / 10	10 / 10	6 / 8	57 / 57	11 / 11	10 / 10	10 / 10	57 / 57	Table 3
Slope correlation	own	10 / 11	10 / 10	5 / 10	8 / 8	57 / 57	11 / 11	10 / 10	10 / 10	57 / 57	Appendix E
Per pixel											
Per-view depth AbsRel	own	11 / 11	10 / 10	9 / 10	8 / 8	57 / 57	10 / 11	5 / 5	1 / 5	n/a	Table 3
Reference-free											
Photometric consistency	Chamfer's	11 / 11	3 / 10	n/a	8 / 8	45 / 57	n/a	n/a	n/a	n/a	this table
Camera-matrix reprojection	Chamfer's	4 / 11	4 / 10	4 / 10	7 / 8	51 / 57	3 / 11	4 / 10	2 / 10	54 / 57	this table
Learned, on depth renders											
LPIPS	Chamfer's	5 / 11	0 / 10	0 / 10	8 / 8	56 / 57	0 / 11	0 / 10	0 / 10	52 / 57	this table
DINOv2 class token	Chamfer's	3 / 11	0 / 10	0 / 10	6 / 8	56 / 57	5 / 11	0 / 10	0 / 10	52 / 57	this table
DINOv2 patch tokens	Chamfer's	5 / 11	0 / 10	0 / 10	8 / 8	57 / 57	1 / 11	0 / 10	0 / 10	52 / 57	this table
Rank composites											
Relief correlation with Chamfer	own	8 / 11	10 / 10	9 / 10	8 / 8	55 / 56	10 / 11	10 / 10	10 / 10	56 / 56	Table M.5
Slope correlation with Chamfer	own	6 / 11	8 / 10	4 / 10	8 / 8	55 / 57	11 / 11	10 / 10	10 / 10	57 / 57	Table M.5
Relief correlation with completeness	own	9 / 11	10 / 10	9 / 10	8 / 8	55 / 56	11 / 11	10 / 10	10 / 10	56 / 56	Appendix E

tilt, and tilt carries a median of 74% of the best model's mean squared height error on LuSNAR; on a gravity-aligned grid an unmodelled tilt is exactly what becomes an impassable slope, so the metric discards the error the planner cares about most. Pearson is invariant to a positive vertical scale, so a reconstruction that overstates every obstacle tenfold scores the same as one that gets the heights right. And correlating over the cells both clouds occupy charges nothing for extent a prediction never covered, where navigation coverage charges for exactly that. A criterion built to be blind to offset, tilt and scale cannot also be a criterion for a task whose costs are offsets, tilts and scales.

What the navigation outcomes reward. The last of those invariances is most of the loss, and it can be measured. Coverage is the share of the ground truth's navigation grid a reconstruction fills and completeness the mean distance from the ground truth to the reconstruction, so both reward footprint, and a metric can agree with navigation because it tracks footprint rather than the surface. Holding completeness fixed separates the two: the Spearman partial correlation of a metric's within-scene ranking with an outcome's, given completeness's, on the pool of Table 3 (Table E.2). Chamfer's agreement is its completeness's: beyond completeness its four-outcome agreement lies between -0.15 and $+0.19$ on the five domains, and its accuracy half, the distance from the prediction to the ground truth, agrees negatively on all three low-relief domains even before completeness is held fixed. Relief correlation's lies between $+0.06$ and $+0.41$, above Chamfer's on every domain; the difference is resolved on LuSNAR's coverage, $+0.41$, interval $+0.25$ to $+0.58$ over scenes, and

on Erfoud's four-outcome mean, $+0.40$, interval $+0.06$ to $+0.80$. Per-view depth, which scores every pixel of every view, keeps the most agreement beyond completeness on LuSNAR and Erfoud, $+0.54$ and $+0.54$. The outcomes therefore measure two properties, where the ground is covered and whether the surface there is right: completeness measures the first and fails the control, relief correlation measures the second and passes it, and Chamfer measures a mixture whose agreement with the task comes from the first. On LuSNAR relief correlation's own agreement beyond completeness is only $+0.06$, so its lead there reflects Chamfer's negative partial rather than an agreement of its own.

Table E.2: Agreement with navigation beyond completeness. Each entry is a metric's mean within-scene Spearman agreement with the four navigation outcomes, raw and, after the slash, as the partial correlation given completeness, on the pool of Table 3 (scenes with five models that carry completeness). The last row is relief correlation's partial agreement with coverage minus Chamfer's, with its 95% interval over scenes.

Metric	LuSNAR	AMEDS	Erfoud	ETH3D	TartanGround
Chamfer	+0.36 / -0.15	+0.68 / -0.03	+0.51 / -0.12	+0.46 / +0.19	+0.67 / +0.11
Completeness	+0.68 / n/a	+0.70 / n/a	+0.58 / n/a	+0.41 / n/a	+0.69 / n/a
Accuracy	-0.12 / -0.16	-0.17 / -0.43	-0.15 / -0.35	+0.25 / +0.22	+0.43 / +0.15
F-score at 0.1 m	+0.31 / -0.27	-0.02 / -0.23	-0.13 / -0.40	+0.25 / -0.01	+0.51 / +0.10
Per-view depth AbsRel	+0.74 / +0.54	+0.37 / +0.14	+0.73 / +0.54	+0.32 / +0.30	+0.17 / -0.03
Relief correlation	+0.18 / +0.06	+0.57 / +0.41	+0.43 / +0.28	+0.49 / +0.40	+0.57 / +0.23
Relief minus Chamfer, coverage, beyond completeness	+0.41 [+0.25, +0.58]	+0.28 [-0.13, +0.56]	+0.69 [+0.39, +0.98]	+0.05 [-0.32, +0.39]	-0.02 [-0.18, +0.15]

On the control cells. Relief correlation rejects the wrong scene (Table M.3) on most cells of all three low-relief domains, where Chamfer rejects at most three, at every raster size, and keeps Chamfer's result on TartanGround; it loses two ETH3D cells, both small scenes whose walls a height raster represents poorly. Relief correlation is a height-field metric, one height per raster cell over the target's ground plane, so a wall face, an overhang or a second storey is outside what it scores; where a scene is not a terrain it does not apply. On electro, which the table omits, it and Chamfer both reject the wrong scene on all 10 cells. The two variants that penalise amplitude rather than placement do worse than Chamfer on the high-relief domains: the models' clouds carry smooth warps several times larger than lunar relief, which a skill score charges in full and a correlation ignores. Keeping only the relief below about one metre, as the high-pass variant does, discards what the models get right, which is the placement of craters and rises rather than centimetre regolith. The last block is the limit: relief correlation's model ranking agrees with navigation coverage less than Chamfer's on every domain, although averaged over the four outcomes it leads Chamfer on ETH3D (Table M.5). A metric can be built to pass the control without becoming a better ranking: telling the scene apart and ranking the models are different properties, as Section 5 finds for per-view depth and completeness.

A per-pixel error does not help. Keeping per-view depth's pixel pairing in 3D, under the global alignment Chamfer uses, loses every lunar cell measured, so per-view depth's advantage is its per-view alignment.

Robustness to how the plane is removed. The plane fit is half the metric, so three detrendings were compared on the identical rasters of the 96 audit cells of the pre-extension pool (11 LuSNAR, 10 AMEDS, 8 ETH3D, 10 electro, 57 TartanGround; Erfoud's ten were saved later): the published least-squares plane, a least-squares quadratic, and a robust plane by iteratively reweighted least squares with Huber weights; outliers in these rasters are real geometry, a boulder or a crater rim, so a robust fit is a different choice rather than a better one.

The verdict does not depend on the choice. Counting the cells where the model best on Chamfer beats both substitutes, as Table M.3 does, the published fit gives 9 of 11 on LuSNAR, 8 of 10 on AMEDS, 16 of 18 on ETH3D with electro and 56 of 57 on TartanGround, and the robust fit gives 8, 8, 16 and 53. The 95% cell bootstrap intervals of the two fits overlap on all four, so on no domain does the choice reverse the verdict. TartanGround is the one place the choice is visible at all: the robust fit's interval, 49 to 56, overlaps the published fit's 54 to 57, but its point estimate of 53 lies just below it, making it the weaker of the two there. The quadratic moves AMEDS to 6 of 10 and TartanGround to 55 of 57, and both of its intervals overlap the plane's; on the fuller pool of Table M.3 it also takes Erfoud from 8 to 6 of 10. The published choice is therefore not load-bearing where the paper's argument lives.

Robustness to noise and to missing coverage. Beyond the raster cell swept in Table M.3, sensor noise on the scored heights and missing extent were run on the 96 pre-extension audit cells of the detrending comparison above, three seeds each, with the recomputation reproducing the published relief correlations exactly on all 288 model, donor and plane values before anything was perturbed.

Added zero-mean height noise barely moves the verdict. From $\sigma = 0.005$ m to $\sigma = 0.5$ m the best model still beats the wrong scene on 9 of LuSNAR's 11 cells, 8 of AMEDS's 10, 6 of ETH3D's 8 and 10 of electro's 10, unchanged at every level, and TartanGround falls only from 56 of 57 to 55.3. Half a metre of noise is more than thirty times LuSNAR's own relief, so the criterion is not resting on the fine structure the noise destroys. Only at $\sigma = 1$ and 2 m does LuSNAR move, to 8.3 and 8.0.

Heavy tails and uneven noise cost little more, and only at the largest scale. With Student-t noise of two degrees of freedom and scale σ, Gaussian noise carrying five percent of ten-sigma outliers, or noise growing across the window from none to 2σ as stereo error grows with range, each added per cell to every scored cloud on the same cells, rasters and seeds, AMEDS, ETH3D and electro keep 8, 6 and 10 at every level up to 0.5 m and LuSNAR keeps 9 up to 0.2 m; at 0.5 m the least is 8.7 of LuSNAR's 11 and 53.3 of TartanGround's 57, against 9 and 55.3 under Gaussian noise. On the model alone, against a clean donor, Student-t noise of 0.5 m takes LuSNAR to 6.7 and AMEDS to 5, where Gaussian noise takes them to 8 and 8.

Removing coverage costs more, as it should, and the metric degrades rather than breaking. Deleting contiguous blocks of the model's raster takes LuSNAR from 9 to 8.7, 8.3, 8.0, 7.3 and 7.3 of 11 as 10, 25, 50, 75 and 90% of its cells go: the coverage insensitivity named above, and a reason to report relief correlation beside a coverage measure rather than instead of one.

The pairing the leaderboard points to. All 91 one- and two-criterion combinations of thirteen bases, Chamfer, relief correlation and eleven terrain metrics, were scored on the 106 audit cells (Table M.5). A combination passes when, on every domain, it rejects both substitutes on at least as many cells as Chamfer and its four-outcome agreement is no more than 0.10 below Chamfer's. Six pass, and relief correlation with Chamfer rejects both on 8 of 11 LuSNAR, 10 of 10 AMEDS and 9 of 10 Erfoud cells (Appendix M.2).

Is a composite's agreement different from Chamfer's? Tested as a difference of dependent correlations over 2000 resamples of the models (Graham & Baldwin, 2014), the composite of relief correlation with Chamfer is at parity with Chamfer as a ranking, apart from a resolved loss on TartanGround's coverage, its interval -0.186 to -0.040; the composite with completeness was not tested this way, and its differences from Chamfer below are point estimates.

Completeness as the partner. The decomposition above says what the partner of relief correlation should measure: the footprint the outcomes reward, which completeness measures directly and Chamfer only through its completeness half. Completeness was therefore added to the thirteen bases and scored by the same code on the same 106 cells, its values for the wrong scene taken from the control rows and for the flattened target computed from the stored clouds as the plane's Chamfer is (which it reproduces to 10^{-14}). Alone it rejects both substitutes on 4 of the 31 low-relief cells, and its mean within-cell agreement with the four navigation outcomes over these audit cells (Table M.5) is $+0.60$ on LuSNAR, where Chamfer's is $+0.28$; that is a different pool and unit from Table E.2's within-scene $+0.36$. In rank composite with relief correlation it rejects both on 9 of 11 LuSNAR, 10 of 10 AMEDS and 9 of 10 Erfoud cells, 8 of 8 ETH3D, 10 of 10 electro and 55 of 56 TartanGround cells, one more low-relief cell than the composite with Chamfer, and its within-cell agreement differs from Chamfer's by -0.01 on electro to $+0.19$ on LuSNAR, untested point estimates. The pairing was chosen after the decomposition was measured, not before, which is why the grid search below, run over formulas and partners alike, is reported beside it.

Chosen without seeing the domain it is judged on. Every one of the 91 combinations above saw every one of the 106 cells, so the search selects on the data the verdict is read from. The test is to leave one domain out: run the whole search on the other five, apply a fixed rule that never reads the held-out domain, and score its winner there. The rule is to take, among the combinations never worse than Chamfer on either axis on all five, the one with the largest mean gain in rejection rate.

Across the six folds the rule selects three different winners, and every one pairs relief correlation with a terrain criterion: Pearson slope correlation in four folds, AMEDS, electro, Erfoud and ETH3D, slope error with LuSNAR held out and normal error with TartanGround held out. Every pick transfers. On each held-out domain it rejects both substitutes on at least as many cells as Chamfer, on 8 of 11 LuSNAR cells, all 10 of AMEDS's and 9 of Erfoud's 10 where Chamfer rejects none, and it ranks within the search's materiality margin of 0.1 of Chamfer, better on ETH3D and electro and at most 0.089 lower elsewhere. On LuSNAR the pick was chosen without LuSNAR in the selection set at all. What this licenses is a claim about pairings with relief correlation, of which the modal one contains no Chamfer, and not a claim that searching finds a single winner.

Two things limit this claim. The passing pairings were found by searching 91 combinations, and although the rejection gains are far too large to be sampling noise, the near-ties on the ranking axis are exactly the kind of margin a search of that size can manufacture, which the held-out test and the bootstrap bound but do not remove. And the pairings that contain Chamfer augment it rather than replace it, even though what a leaderboard would print changes. Slope correlation alone is the better discriminator on LuSNAR, rejecting both substitutes on 10 of 11 cells when each criterion is allowed its own best model, and trails Chamfer on ranking on AMEDS, Erfoud and TartanGround, by more than the materiality margin only on AMEDS, where Chamfer leads it by 0.216 in the four-outcome mean. That count is 7 of 11 under the stricter rule that holds the compared model fixed at the one Chamfer chose; the two rules answer different questions and neither is a correction of the other.

A grid search over terrain-field formulas. The thirteen bases above are one family, so a wider grid was scored to see whether a formula outside it does better: every combination of five raster cell sizes, from 0.05 to 0.8 m; seven trend removals, none, the ground truth's own plane, each raster's own plane or quadric, and three Gaussian high-pass filters; seven fields of the height, the height itself, slope, Laplacian, local roughness, topographic position, terrain ruggedness and vector ruggedness; and twelve comparisons, Pearson, Spearman, concordance, skill, mean, root mean square, robust and 95th-percentile error, a global structural similarity, normalised mutual information, the amplitude ratio and the error left after the best affine map; with gradient, power-spectrum and slope-distribution comparisons, and each in rank composite with Chamfer and with completeness, $9{,}453$ candidates in all. Each was scored for every zero-shot model's aligned cloud on the 52 audit windows that carry zero-shot rows, 647 model cells, for the wrong scene of each of their 71 audit cells and for the flattened target (Table M.4). The rule, fixed before the grid was run, asks for both substitutes rejected on 70% of every low-relief domain's cells and 90% of every relief domain's, and a four-outcome agreement within 0.10 of Chamfer's on every domain. With raw agreement no candidate passes; with agreement beyond completeness 93 do, but a selection repeated with each domain held out picks a candidate that passes the held-out domain on 2 of the 5 folds, so the grid yields no formula that transfers and none is proposed from it. What it does show is where Section 5's formula, fixed before any of this, stands: at 0.2 m it agrees with navigation beyond completeness above Chamfer on every domain, and at 0.1 and 0.05 m it rejects both substitutes on every AMEDS, Erfoud and TartanGround cell of the grid, 10 of 11 LuSNAR cells and 7 of 8 ETH3D cells. The grid's rasters are rebuilt from the aligned clouds, so its counts can differ by a cell or two from Table 3's.

Alternatives a reader may propose. Four variations on Equation 3 are cells of the grid, and on the grid's own cells, where the published formula rejects both substitutes on 11 of LuSNAR's 11, 10 of AMEDS's 10, 9 of Erfoud's 10, 6 of ETH3D's 8 and 21 of TartanGround's 22, none rejects both on more of the 31 low-relief cells than its 30. Its rank version, Spearman's ρ on the same rasters, rejects both on 8, 10, 9, 7 and 21. Lin's concordance also charges a vertical offset and a vertical scale mismatch, and so answers the objection that relief correlation credits an exaggerated relief; it rejects both on 10, 10, 9, 5 and 21. Both agree with navigation beyond completeness within 0.09 of relief correlation on every domain. The amplitude ratio alone, a pure vertical-scale term, rejects both on 6, 10, 10, 7 and 19. A distance between the two detrended rasters, the raster form of a Chamfer on what is left after each cloud's own plane is removed, fails where Chamfer fails: its root mean square rejects the flattened target on only 4 of LuSNAR's 11 cells, because the plane's residual is zero and the models' residual error exceeds the target's relief. Computed on the residual point clouds themselves rather than on rasters, that Chamfer rejects both substitutes on 4 of LuSNAR's 11 cells, 2 of AMEDS's 10 and 2 of Erfoud's 10, failing on the flattened target each time. On ETH3D, whose walls a height raster represents poorly, a slope field does what the height does not: slope correlation rejects both substitutes on all 8 ETH3D cells under every trend removal, and a Gaussian high-pass

of 0.5 or 1 m recovers one of the two cells relief correlation loses there. Neither transfers: slope correlation keeps 9 of LuSNAR's 11 and agrees with navigation beyond completeness by $+0.18$ on AMEDS against relief correlation's $+0.41$, and the two high-passes keep 9 and 10 of LuSNAR's 11 and 9 of AMEDS's 10.

How much relief correlation depends on its construction. With the correlation made rank-based, winsorised, Huber-weighted or trimmed, and with the prediction thinned to a tenth of its points, it rejects both substitutes on 27 to 30 of the 31 low-relief cells of the grid, and every raster cell from 0.05 to 0.8 m keeps 30; the trend removal matters more than the cell.

Orientation metrics. Normal consistency and normal concordance rank the models well but cannot see flattening: on LuSNAR they reject the flattened target on at most 2 of 11 cells.

Weights, dominance and the sign of the score. The equal weighting of relief correlation with completeness sits on a plateau, weights of 0.5 and 0.6 on relief correlation each rejecting both substitutes on 9 of 11 LuSNAR cells, and Pareto dominance is too weak a rule for this pair.

The control per-view depth cannot face. Per-view depth is still not a replacement for the surface metric, and the depth audit says why. The plane control records no depth of its own, but the target's least-squares plane has a depth map in every target camera, met along each pixel's ray in closed form, and scored that way through the evaluator's own per-view depth path, with a ray that misses the plane charged the depth cap, it beats every model's median-scaled depth error on 4 of Erfoud's 5 windows, where the cameras look down at nearly planar sand, and on 1 of LuSNAR's 11, counts that do not change when a missed ray is left out rather than charged; on AMEDS and ETH3D it loses on every window, its error a median 4.2 and 417 times the best model's. Depth is therefore not immune to the plane control, only harder for it to pass, and on Erfoud the plane beats it, which is why Section 7 runs the plane control on depth, ray-cast, as on every metric reported.

A joint depth-and-surface score. A rank composite of per-view depth and Chamfer buys control-passing at little ranking cost on four of the five domains and gives the control back on Erfoud, so depth is better reported beside the surface metric than blended into it.

Slope, the competitor from planetary mapping. Slope and roughness error, the forms planetary mapping reports, fail the plane, since on nearly planar terrain a plane has nearly the right slope field; slope correlation passes the control and beats Chamfer on LuSNAR but loses to it on coverage on every other domain, so it is a place check rather than a replacement.

Does it rank the models differently? Relief correlation and Chamfer were computed for every model on every zero-shot frame window, not only the control cells, and each domain's models ranked within each scene under both. Where Chamfer passes the control the two orderings are close: they agree at $\rho = +0.92$ over TartanGround's 12 models and 20 scenes, with the same winner and the same top pair, and at $\rho = +0.66$ on ETH3D, again with the same winner. Where Chamfer fails the control they are not: on LuSNAR's 14 models the two orderings agree at $\rho = +0.12$ and name different second-place models, and on AMEDS and Erfoud they agree at $+0.56$ and $+0.52$ and name different winners. The choice between the two metrics is therefore consequential exactly where the control says the surface metric has stopped measuring the scene, which is the regime this paper is about, and of no consequence where it says the metric works.

F Further checks of the control and its mechanism

This appendix reports checks run on the 106 audit cells of Section 4's plane control, the populations being those of Appendix D: the target's ground truth, the placed wrong scene and the best zero-shot model's aligned reconstruction of the same frames. The model's saved cloud reproduces its logged Chamfer on every cell. Checks run before Erfoud's clouds were saved cover the other 96 and say so.

Why a flat surface wins: a sufficient condition. Let the target's ground truth be a height field h_t over a plane P, measured along the plane's normal, and let a submission be a height field h_s

over the same plane. Four hypotheses are used, and they are stated because the clouds the evaluator scores meet them only in part. (i) Each surface is single-valued over P, which a wall or an overhang violates; NordicHarbor's harbour walls below are an instance. (ii) Both are sampled on a common raster over the same footprint, so that every submitted point has a target point directly above or below it and every target point a submitted one. (iii) Accuracy averages over submitted points and completeness over target points, so $\mathbb{E}_x$ below stands for the measure of whichever direction it bounds, and the bound holds for each direction separately. (iv) Chamfer is the evaluator's, the *mean* of the two directed mean distances, $\frac{1}{2}(\mathrm{acc} + \mathrm{comp})$, so that a bound on each direction is a bound on the whole; under the unweighted sum every constant below doubles and the threshold moves from $\mathrm{dev}(h_t)$ to $2\,\mathrm{dev}(h_t)$. Under (i) and (ii) every submitted point $(x, h_s(x))$ lies within $|h_s(x) - h_t(x)|$ of the target point directly above or below it, and every target point equally close to a submitted one, so each nearest-neighbour distance in either direction is at most the vertical gap at its location, and

$$\mathrm{CD}(h_s, h_t) \;\le\; \mathbb{E}_x\,|h_s(x) - h_t(x)| \;\le\; \mathrm{dev}(h_s) + \mathrm{dev}(h_t), \tag{4}$$

where $\mathrm{dev}(h)$ is the mean absolute deviation of a surface from P. Without (ii) the nearest target sample to a submitted point at x lies at some x' in the plane, and for an L-Lipschitz h_t each nearest-neighbour distance gains a term of at most $(1 + L)\,|x - x'|$, which does not shrink as the terrain flattens. Measured between the best model and the ground truth on the audit cells, the mean in-plane offset is a median of 0.38 times $\mathrm{dev}(h_t)$ on LuSNAR, 0.71 on AMEDS and 2.33 on Erfoud, against 0.08 on ETH3D and 0.06 on TartanGround, and it is set by the two clouds' footprints rather than their point spacing, which on LuSNAR is under a centimetre. Equation 4 therefore holds for co-rasterised surfaces over a common footprint and is only indicative for the scored clouds, least reliable on exactly the flat ground this paper concerns. That is why its wrong-scene form below is tested cell by cell rather than assumed, and why the rule the paper applies rests on Equation 2 alone. For the flattened target the bound is tight on one side by construction. The submission is the orthogonal projection of the ground truth onto P, so it shares the target's footprint and raster and hypotheses (ii) and (iii) hold by construction. Every one of its points lies in P, so a target point is at least its own perpendicular distance from all of them, and its own foot is among them: completeness is therefore *exactly* $\mathrm{dev}(h_t)$. Each projected point has its preimage in the target, so accuracy is at most $\mathrm{dev}(h_t)$. With Chamfer the mean of the two,

$$\tfrac{1}{2}\,\mathrm{dev}(h_t) \;\le\; \mathrm{CD}(h_{\mathrm{plane}}, h_t) \;=\; \tfrac{1}{2}\big(\mathrm{acc} + \mathrm{dev}(h_t)\big) \;\le\; \mathrm{dev}(h_t),$$

which is Equation 2 of Section 4 and gives a characterisation in both directions and with no free constant: a model whose Chamfer exceeds $\mathrm{dev}(h_t)$ must lose to the plane, a model whose Chamfer falls below $\frac{1}{2}\mathrm{dev}(h_t)$ must beat it, and between the two the verdict is decided by the plane's accuracy, which is a statistic of the ground truth alone. That the flattened target's Chamfer is at most $\mathrm{dev}(h_t)$ is thus an identity and not a prediction, and this paper does not treat it as evidence. Measured, the plane's accuracy term, $2\,\mathrm{CD} - \mathrm{dev}(h_t)$, is a median 0.77 of $\mathrm{dev}(h_t)$ over the 123 frame windows (10th and 90th percentiles 0.38 and 0.97) and below a quarter of it on 1, so the plane usually sits near the top of its band. Over the 84 audit windows the ratio falls on rougher targets (Spearman -0.54 with $\mathrm{dev}(h_t)$) and weakly with point density (-0.24), and does not follow the footprint's elongation or its holes ($+0.08$ and -0.07). The same argument bounds any error score that aggregates nearest-neighbour distances monotonically, a median or HyperCD's hyperbolic distance, with dev replaced by the same aggregate of $|h_t|$. A thresholded similarity runs the other way and the bound has to be stated the other way with it. Every point of the flattened target is within $|h_t(x)|$ of the target, so at tolerance τ the plane's F-score is at least $\mathbb{P}_x(|h_t(x)| \le \tau)$, a floor rather than a ceiling, and a model loses to the plane when its F-score falls below that empirical cumulative probability. Since flattening moves each ground-truth point along the normal, the flattening itself is a one-to-one matching of cost $\mathrm{dev}(h_t)$, which bounds the earth mover's distance between the full, equal-mass clouds too; the earth mover's distances of Table E.1 are between independently drawn subsamples, for which that matching is not a feasible coupling, so the corollary does not bound those numbers. A different place laid on the same plane wins whenever the model errs by more than its vertical gap to the target, which for two low-relief surfaces is at most the sum of their deviations. The condition is sufficient, not necessary: it says when a surface metric must fail, and a model erring by less than the terrain varies can still lose to a donor that happens to resemble it. Deviation from the fitted plane is not the paper's relief, which is measured in 1 m columns and so excludes any slope or undulation wider than a column; for the flat scenes the two are of the same order, and the bound uses the former because it is the quantity the argument needs. Planetary stereo mapping makes the same comparison for terrain models: relief only a few times a model's vertical precision is at the limit of what it

resolves (Kirk et al., 2008), and whether a featureless area is smooth or merely unresolved depends on that precision (Kirk et al., 2021).

The bound on the control cells. Measured, the flattened ground truth's Chamfer is 0.58 to 1.00 times $\mathrm{dev}(h_t)$ over the 123 plane-control windows, a median of 0.88, which is inside Equation 2's band on all of them as the identity requires and says only that the identity was computed correctly. What is not automatic is that *no* cell's best model errs by an amount between the plane's own Chamfer and $\mathrm{dev}(h_t)$, the interval in which the one-sided rule would be wrong: that interval is empty here, and it is why the sufficient condition decides every cell rather than merely most. The condition itself, that the best model's Chamfer exceeds $\mathrm{dev}(h_t)$, holds on exactly the 38 cells the plane control wins and on no other: all 31 low-relief cells, where the model errs by a median of 6.2 and 3.5 times $\mathrm{dev}(h_t)$ on LuSNAR and AMEDS and 4.5 on Erfoud; none of ETH3D's 8 or electro's 10 cells, where it errs by 0.34 and 0.26 of it; and 7 of the 57 TartanGround cells, the same 7 the plane wins. Section 4's plane control is therefore decided, cell by cell and on every domain it covers, by one inequality between a model's error and the target's deviation from its own plane. For the wrong scene the vertical gap is measured directly, as the mean absolute difference between the two ground truths' median heights in shared 0.25 m columns of the target's plane frame. The bound itself, the wrong scene's Chamfer below that gap, holds on 9 of the 11 LuSNAR cells, 4 of the 10 AMEDS cells, 6 of the 10 Erfoud cells, 7 of the 8 ETH3D cells, all 10 electro cells and 56 of the 57 TartanGround cells; where it fails, the placed donor does not share the target's plane or footprint closely enough for the height-field comparison. Used as a prediction, a model erring by more than the gap, it marks 5 of LuSNAR's 10 wins, all 7 of AMEDS's and all 7 of Erfoud's, no terrestrial cell, and one cell on which the wrong scene loses, an Erfoud cell where the bound itself fails.

The plane control on the whole benchmark. Both quantities in that inequality are free of any model: the plane is built from the ground truth and the models' Chamfers are already recorded, so the plane control can be run on every cell of the benchmark without a donor scene and without inference. Over the 123 frame windows of Table 1, StereoLunar's counted once per illumination, rebuilding each cell's ground truth by the evaluator's own steps, the rule decides all 123: the plane beats the best model on 19 of LuSNAR's 20 windows, 5 of AMEDS's 5, 5 of Erfoud's 5, 39 of StereoLunar's 42, 4 of TartanGround's 40 and none of ETH3D's 11, and on every window the plane wins exactly when the best model's Chamfer exceeds $\mathrm{dev}(h_t)$. Equation 2 forces that verdict on the 117 windows outside its band; the 6 inside it, where the rule is a finding rather than an identity, all went to the model. The one further window inside the band that this paper reports, Moon_1 at seventeen views (below), went to the plane, so off the frozen protocol the rule can fail inside the band, which is why the plane's own Chamfer is the safer screen. The median model error is 6.1 times $\mathrm{dev}(h_t)$ on LuSNAR, 4.5 on Erfoud, 3.5 on AMEDS and 3.0 on StereoLunar against 0.29 on ETH3D and 0.21 on TartanGround. On the 52 windows that also carry an audit cell this reconstruction reproduces the plane Chamfer of Section 4 to a relative 7×10^{-8}, so the two measurements are of the same quantity. The audit cells also lie on 32 TartanGround windows outside the 123, all in the 16 unseen environments of Table 2, and the rule decides every TartanGround audit cell as well: the plane wins on the 7 of the 57 where the best model's Chamfer exceeds $\mathrm{dev}(h_t)$ and on no other. The descent camera is the addition here: StereoLunar carries no wrong-scene verdict that is free of its altitude confound (Appendix D), but its ground truth is flat enough at the scale of those models' error that a plane beats them on 39 of 42 windows.

Why degrading the input can improve Chamfer. Corrupting the input images improves Chamfer on LuSNAR in 17 of 21 conditions (Section 6.2), while the models carry about five times the ground truth's relief in 1 m columns (below). Equation 4 suggests a reading, a smoother output moving toward a plane and so toward a lower Chamfer, but flattening the best model's own output beats the model on only 3 of LuSNAR's 11 cells, and Appendix H finds the improvement is not mainly flattening. Either way a surface metric cannot be left unchecked on this terrain: degrading the input makes the number better.

How flat is too flat? Relief is the obvious candidate for where the risk begins, and it is the wrong quantity. Over the 81 of the 123 windows that carry a local relief, a logistic regression of whether the plane beats the best model on $\log_{10}$ of that relief puts the crossing at 0.032 m, but with a scene bootstrap interval from 0.0050 to 0.083 m and no separation behind it: the plane still wins on a

ConstructionSite window whose relief is 0.4724 m and already loses on a NordicHarbor window whose relief is 0.0015 m, with 53 of the 81 windows lying between the two. StereoLunar's 42 windows are excluded because relief is the spread of height inside a 1 m column and a descent camera's ground truth is not sampled that finely. Those two windows say why relief cannot carry the threshold: relief is measured inside a column and so cannot see that ConstructionSite deviates from its own plane by 0.96 m while its best model errs by 10.05 m, or that NordicHarbor's flat columns sit on harbour walls that put its deviation at 1.35 m against an error of 0.33 m. The quantity that does carry a threshold is the one the bound names, and on it the data separate completely: over the same 81 windows the plane loses wherever the best model's error is at most 0.845 times $\mathrm{dev}(h_t)$ and wins wherever it is at least 1.085 times it, so every threshold in between fits equally well, a logistic fit has no maximum-likelihood estimate, and the data locate the boundary no more precisely than that interval. On error in units of relief there is no separation, and the same logistic fit crosses at 11 with a scene bootstrap interval from 5 to 60. The separating interval should not be read as confirming a constant of 1: Equation 2 puts the true per-cell threshold at the plane's own Chamfer, between a half and one times $\mathrm{dev}(h_t)$, and no window falls between that threshold and $\mathrm{dev}(h_t)$, which is why the gap sits where it does. To judge a new domain, fit a plane to its ground truth and ask whether the expected error exceeds that plane's own Chamfer, the per-window threshold of Equation 2, which needs no model, donor or inference; the mean deviation only brackets it, and the seventeen-view Moon_1 window of this appendix falls between the two.

A designed defect at the plane's own height error. A defect placed on the ground truth at exactly the plane's own height error is preferred to the plane on every one of the 63 audit windows with relief, but on low-relief ground the plane is the better submission on 9 of the 21 windows for additive noise and 11 of the 21 for band-limited microrelief: the thesis with a designed defect in place of a donor scene.

Non-local matching. Neither HyperCD nor a geodesic Chamfer restores the low-relief verdict, the flattened scene winning all 31 low-relief cells under both, and the earth mover's distance discriminates less than Chamfer on every domain but LuSNAR.

Criteria read from the images. A reference-free point-map consistency, a reprojection through the target's camera matrices, and LPIPS and DINOv2 on renders of the geometry pass the control where the ground has relief and fail it on low relief, and the reference-free one is provably blind to the plane (Appendix K); Table E.1 gives their counts.

Beyond the plane: curved, low-variance substitutes. Equation 2 is stated for a flat submission, but its argument uses only that each substitute point sits above or below one target point: any surface S that replaces the target's heights while keeping one point per ground-truth point has accuracy and completeness at most $\mathrm{dev}_S = \mathrm{mean}\,|h - S|$, so its Chamfer is at most dev_S and it beats every model that errs by more. The other side of the band needs completeness to equal dev_S, which holds for the plane but only approximately for a curved surface, whose points a ground-truth point may find nearer than its own. Seven such surfaces were scored on the 84 audit windows whose ground-truth cloud is stored (Table F.1): the plane; the least-squares quadric; the target's heights low-passed with Gaussians of 0.5 to 4 m; and a bowl, the plane plus a paraboloid whose mean deviation from the plane equals $\mathrm{dev}(h_t)$, a curved surface of the right amplitude and the wrong shape. The guaranteed side decides all 209 window and surface pairs it covers. The other side holds on 293 of its 296; the 3 exceptions are the 0.5 m low-pass on TartanGround, whose completeness falls to $0.53\,\mathrm{dev}_S$. The bowl beats every model on 9 of LuSNAR's 11 windows, 3 of AMEDS's 5 and 5 of Erfoud's 5, against 4 of TartanGround's 54: the failure needs a low-variance surface, not a plane. Against the model best on it on each window, relief correlation rejects the bowl on 20 of the 21 low-relief windows and the 1 m low-pass on 0, which is right: that surface is the target with its finest relief removed, in the right place, and it beats every model on 21 of the 21 low-relief windows under Chamfer too. The smoother surfaces are credited in the same way: the quadric, the target's own broad shape, is rejected on only 9 of the 21 and the 4 m low-pass on 6. Relief correlation scores relief shared with the target at whatever scale it lies, so it separates a surface of the wrong shape from the target, not a smoothed copy of the target from a reconstruction.

Table F.1: Curved substitutes: windows on which each beats the best zero-shot model's Chamfer, and, on the three low-relief domains, the windows on which relief correlation ranks it below the model best on that score. Each keeps one point per ground-truth point, so its Chamfer is at most its own dev_S, the guaranteed side of Equation 2's band.

Substitute	LuSNAR	AMEDS	Erfoud	ETH3D	electro	TartanGround	Relief corr.
Plane	11 / 11	5 / 5	5 / 5	0 / 8	0 / 1	6 / 54	21 / 21
Quadric	11 / 11	5 / 5	5 / 5	2 / 8	0 / 1	13 / 54	9 / 21
Low-pass, $\sigma = 0.5$ m	11 / 11	5 / 5	5 / 5	4 / 8	1 / 1	31 / 54	0 / 21
Low-pass, $\sigma = 1$ m	11 / 11	5 / 5	5 / 5	3 / 8	1 / 1	23 / 54	0 / 21
Low-pass, $\sigma = 2$ m	11 / 11	5 / 5	5 / 5	2 / 8	0 / 1	14 / 54	1 / 21
Low-pass, $\sigma = 4$ m	11 / 11	5 / 5	5 / 5	1 / 8	0 / 1	8 / 54	6 / 21
Bowl of matched amplitude	9 / 11	3 / 5	5 / 5	0 / 8	0 / 1	4 / 54	20 / 21

How the plane is fitted. Robust, piecewise and quadratic fits to the ground truth each beat the best model on every low-relief cell, the richer the surface the wider the margin, and flattening the best model's own reconstruction helps it on no cell with relief.

What the best model's error is made of. In the ground truth's plane frame, each reconstructed point's height is compared with the ground truth's median height in its 0.25 m column, and the mean squared difference is split into a constant offset, a planar tilt, a quadratic bend and the remainder. On LuSNAR the tilt carries a median of 74% of it, the offset 14%, the remainder 13% and the bend under 1%; on AMEDS the offset carries 31% and the tilt 23%. Each share is the median of that component's share over cells, taken independently of the others, so the four need not sum to one. A fit to the geometry would remove the tilt, but the alignment here is fitted to camera poses precisely so that it cannot absorb geometric error (Appendix B). The remainder alone has a median RMS 18 times the ground truth's relief on LuSNAR and 1.4 times on AMEDS, both measured in the same frame, and in the same 1 m columns the LuSNAR reconstructions carry 5.2 times the ground truth's relief: the models add relief the terrain lacks rather than smoothing it away. On terrestrial cells the split is the opposite: the remainder carries a median of 86% on ETH3D, 70% on electro and 84% on TartanGround, and the tilt 1.3%, 4.2% and 6.1%, so there the error is local detail rather than a single pose-like term. Neither lunar pattern carries information about which scene was reconstructed, which is why the variant these errors suggest removes each cloud's own plane before comparing, as relief correlation does (Appendix E).

Adaptation does not restore discrimination. The controls are scored against the best zero-shot model, so the failure could be one of zero-shot evaluation rather than of the metric. The adaptation axis answers this without new inference: its models are fine-tuned on six LuSNAR scenes, Moon_2, Moon_3, Moon_4, Moon_5, Moon_7 and Moon_9, with Moon_6 for validation, and scored only on Moon_1 and Moon_8, which no schedule saw. Three of the six are among the five evaluated scenes, which leaves the zero-shot results untouched but is why the adaptation arm is scored on the other two alone. It is scored on the same windows as the control cells there, eight same-domain cells and four cross-domain ones. Fine-tuning works, cutting the best model's Chamfer by a median factor of 3.6, and it moves the verdict without reversing it. The wrong scene still beats the best fine-tuned model on 4 of the 8 same-domain cells, against 7 zero-shot, and is rejected on the 4 cross-domain ones either way; the flattened ground truth still beats it on 6 of the 8, against 8 zero-shot, at a median margin of 1.40. A model taught the domain therefore still loses to a surface with none of the domain's structure, which is what the inequality above predicts: fine-tuning has to bring the error below the terrain's deviation from its own plane, and a factor of 3.6 does not.

More views, and a better scale. The view-count sweep of Appendix H, run at 2 to 65 views on Moon_1 and Moon_8 (113 on Moon_1), was extended to the other three lunar scenes, each run at 3, 5, 9 and 17 views spread over the same 112-frame span, with four batch models (VGGT, MapAnything, Pi3 and MASt3R) and four streaming ones (CUT3R, MUSt3R, Spann3R and StreamVGGT); two-view runs record no Chamfer and are left out. The unit this appendix argues for is $\mathrm{dev}(h_t)$, not relief, and the plane control needs no new inference, so it was run on the seventeen-view footprints themselves: the flattened ground truth beats the best seventeen-view

model on all five scenes. In units of that footprint's $\mathrm{dev}(h_t)$ the best model errs by 0.84, 2.16, 16.4, 7.01 and 23.1 on Moon_1, Moon_3, Moon_5, Moon_7 and Moon_8; Moon_1 is therefore inside Equation 2's undecided band, where the plane's own accuracy decides, and its Chamfer of 0.25 m beats the model's 0.28 m. In the relief units used before, the same errors are 2.4 to 117 times the terrain's relief, and on two scenes the error has stopped falling. From 3 to 17 views it falls by a factor of 1.6 to 3.1, but between 9 and 17 it changes by under 3% on Moon_5 and Moon_8, so more views do not bring the error under the deviation that decides the plane control there. A better scale does not close the gap either: the best similarity fit to the geometry, which removes any error a metric-scale model or a better calibration could remove, still leaves lunar error at 21.8 times the relief over all models at each scene's window 120 (Table D.5).

G FULL RESULTS

Per-domain metric audit. Table G.1 gives every stored metric the control cells carry, relief correlation and its composites apart (Table E.1), counted as Table 3's rejections are: the cells on which the best model on that metric beats the wrong scene. The surface metrics reject the control on a minority of low-relief cells, 1 to 3 of LuSNAR's 11 and, accuracy apart, at most 4 of AMEDS's and Erfoud's 10, and on 87.5% to 100% of the terrestrial cells, 7 or 8 of ETH3D's 8 and 50 to 56 of TartanGround's 57. Accuracy is the exception on AMEDS and Erfoud, rejecting it on 9 of AMEDS's cells and all of Erfoud's, so there the wrong scene outscores the models on completeness. Under the primary depth protocol, a median scale per view, every depth statistic rejects the control on every cell of the four other domains carrying depth and on 9 or 10 of Erfoud's 10. Without that scale the absolute relative error still rejects it on 10 of LuSNAR's 11 cells while the threshold accuracy $\delta < 1.25$ rejects it on none, so a recommendation naming the depth family without the statistic and its scaling could direct readers to the combination that discriminates least. The navigation metrics are mixed on LuSNAR: coverage rejects the control on all 11 cells, traversability and roughness error on 8, slope error on 6 and height error on 4.

Table G.1: The control audit in full. Cells of Table 2 (same-domain and unseen donors) on which the best model on the row's metric beats the wrong scene, in that metric's registered direction. Depth rows use a median scale per view unless marked unscaled; AMEDS's and Erfoud's depth is their LiDAR projected into each view.

Family	Metric	LuSNAR	AMEDS	Erfoud	ETH3D	TartanGround
Surface	Chamfer	1 / 11	3 / 10	3 / 10	8 / 8	56 / 57
Surface	Chamfer, median	1 / 11	3 / 10	0 / 10	8 / 8	51 / 57
Surface	Accuracy	3 / 11	9 / 10	10 / 10	8 / 8	54 / 57
Surface	Completeness	2 / 11	4 / 10	2 / 10	7 / 8	54 / 57
Surface	F at 0.05 m	1 / 11	3 / 10	0 / 10	7 / 8	50 / 57
Surface	F at 0.1 m	1 / 11	3 / 10	1 / 10	8 / 8	51 / 57
Surface	F at 0.2 m	2 / 11	3 / 10	2 / 10	8 / 8	52 / 57
Depth	AbsRel	11 / 11	10 / 10	9 / 10	8 / 8	57 / 57
Depth	SqRel	11 / 11	10 / 10	10 / 10	8 / 8	57 / 57
Depth	log10	11 / 11	10 / 10	9 / 10	8 / 8	57 / 57
Depth	RMSE log	11 / 11	10 / 10	9 / 10	8 / 8	57 / 57
Depth	MAE	11 / 11	10 / 10	9 / 10	8 / 8	57 / 57
Depth	RMSE	11 / 11	10 / 10	9 / 10	8 / 8	57 / 57
Depth	$\delta < 1.25$	11 / 11	10 / 10	9 / 10	8 / 8	57 / 57
Depth	AbsRel, unscaled	10 / 11	2 / 10	2 / 10	8 / 8	56 / 57
Depth	$\delta < 1.25$, unscaled	0 / 11	2 / 10	2 / 10	6 / 8	43 / 57
Navigation	Coverage	11 / 11	3 / 10	6 / 10	7 / 8	49 / 57
Navigation	Height error	4 / 11	9 / 10	7 / 10	7 / 8	47 / 57
Navigation	Slope error	6 / 11	10 / 10	10 / 10	8 / 8	56 / 57
Navigation	Roughness error	8 / 11	10 / 10	10 / 10	7 / 8	54 / 57
Navigation	Traversability F1	8 / 11	10 / 10	10 / 10	6 / 8	53 / 57

Track B. Nerfstudio's two per-scene optimisation baselines, given the ground-truth camera poses, are 4.2 to 9.6 times worse than the best feed-forward model on TartanGround and 7.0 to 60 times on LuSNAR under their dense fit (Appendix M.3).

How many independent navigation outcomes are there? About two: traversability, reachability and safe paths agree at a median of $+0.86$ to $+0.99$, and coverage is the outcome that does not follow them.

Why SLAM-based reconstruction is not a fourth track. Track B, the closest family scored, is close to a hundred times the plane's Chamfer on LuSNAR even with ground-truth poses, the bound decides the outcome before the method is chosen, and eight-frame windows remove what distinguishes SLAM.

Seed variation. Five cells were re-run with everything held fixed but the inference seed, giving 54 zero-shot model-by-cell groups: LuSNAR's Moon_1, TartanGround's Prison, ETH3D's courtyard and StereoLunar's nadir_t01 for every model the domain carries, and Moon_8 for the two models withheld from TartanGround, at three seeds, and at eight for three models on Moon_1 and Prison. Zero-shot inference is deterministic on this hardware for eleven of the fourteen models: 47 of the 54 groups return a Chamfer whose spread across seeds is exactly zero. The three that do not are named rather than averaged away. Fast3R's range reaches 9.7% of its mean, Spann3R's 0.2% and DUSt3R's 0.001%, the last on one StereoLunar cell. Each group's seed range was set against every gap between adjacent models on the zero-shot leaderboard of its own cell, the same scene and frames, 11 or 13 gaps a cell, which for the seven zero-shot groups that vary makes 85 group and gap pairs; the 20 gaps narrower than a model's own seed range are all Fast3R's. Fine-tuned checkpoints behave differently. Thirteen combinations of model and adaptation recipe (Appendix H) were each fine-tuned at three seeds and scored on Moon_1 and Moon_8 at frames 840 to 868, giving 26 groups. Every one of the 26 fine-tuned groups moves with the seed, by a median of 19% of its mean and by as much as 91%, and 231 of the 338 group and gap pairs there, 13 gaps a cell against the zero-shot leaderboard of the same frames, fall inside that range. The fine-tuned groups and the six eight-seed groups ran on two or three nodes, so their spread may include node effects; Fast3R also varies on cells run on one node, and MV-DUSt3R+ does not vary across two. Adaptation is therefore reported in this paper as an effect and never as a ranking: the factor of 3.6 by which fine-tuning cuts the error is an order above that spread, while the order of two fine-tuned checkpoints on one cell is not resolvable at three seeds.

Table G.2: The resolution analysis applied to sixteen rows transcribed from five tables of E3D-Bench (Cong et al., 2025), version 3, a benchmark this study did not produce; the first column names the source table. Cols are the datasets or scenes the published table breaks results down by. W is Kendall's concordance of the columns on the model ordering, Winner first counts the columns in which the table's winner ranks first, and Cols needed is the fewest columns at which any pair could separate under an exact Wilcoxon test with Holm correction at $\alpha = 0.05$.

E3D-Bench table	Metric	Models	Cols	W	Winner first	Cols needed
Table 2	Absolute Relative Error	8	5	0.70	4 / 5	11
Table 2	delta < 1.03 accuracy	8	5	0.79	4 / 5	11
Table 3	Absolute Relative Error	16	6	0.57	4 / 6	13
Table 3	delta < 1.25 accuracy	16	6	0.61	2 / 6	13
Table 4	ATE	11	5	0.42	3 / 5	12
Table 4	RPE rotation	9	6	0.24	4 / 6	11
Table 4	RPE translation	9	6	0.29	2 / 6	11
Table 5, dense	Accuracy	7	5	0.86	4 / 5	10
Table 5, dense	Completeness	7	5	0.81	3 / 5	10
Table 5, dense	Normal Consistency	7	5	0.86	3 / 5	10
Table 5, sparse	Accuracy	8	5	0.68	3 / 5	11
Table 5, sparse	Completeness	8	5	0.69	3 / 5	11
Table 5, sparse	Normal Consistency	8	5	0.79	4 / 5	11
Table S2	ATE	11	8	0.28	4 / 8	12
Table S2	RPE_rot in degrees	11	8	0.32	4 / 8	12
Table S2	RPE_trans	11	8	0.32	3 / 8	12

Which single number, per domain. On TartanGround completeness predicts navigation better than Chamfer ($+0.85$ against $+0.80$), so a single surface number there should be completeness. The gap is one rank correlation against another on one domain and no test separates the two, so this is a tie-break and not a finding.

A planned comparison against an all-pairs one. The floor $m \times 2/2^n$ is a statement about testing every pair. A study with one pre-registered comparison, a proposed method against a named incumbent, has $m = 1$ and a floor of $2/2^n$, which is 0.031 at six columns and 0.0078 at eight. On that basis seven of the sixteen published rows are wide enough to admit a planned comparison, and in each of those seven at least one pair does reach an uncorrected $p < 0.05$. The other nine have five columns, where even $2/2^5 = 0.063$ exceeds α and nothing can be claimed at any m. What survives in every one of the sixteen is the narrower statement: no table separates its own first-ranked model from its runner-up, even as a family of one. The same scope is applied here to this paper's own designs rather than only to other people's: of the twelve designs of Table G.4, seven admit an all-pairs test and nine a planned one, and among those nine the leading pair separates at $m = 1$ in three, TartanGround's per-view depth at an uncorrected $p = 0.008$ against a Holm-adjusted 0.100, and StereoLunar's Chamfer under both scale conventions. Only one of the three is ground-level and it is a depth metric: of the four ground-level Chamfer and coverage designs wide enough for the test, none separates its leading pair even as a family of one, the closest being ETH3D Chamfer at $p = 0.067$. What changes on that one ground-level design is that its head is unseparated by the all-pairs procedure rather than by arithmetic. The barrier this paper reports is therefore a barrier to the all-pairs procedure over dataset-summary columns, not a claim that the underlying datasets cannot separate models; a study that tests across scenes rather than across aggregated columns has as many blocks as it has scenes.

What the navigation correlations can and cannot support. Every claim this paper makes about navigation is a comparison between two correlations computed on overlapping data, so each was re-examined with bootstrap intervals over scenes and over models, a test for dependent correlations, a split-half reliability of the navigation outcome, and a leave-one-model-out jackknife (the recomputation reproduces all 110 published correlations exactly first).

The navigation outcome is not itself noise. Its split-half reliability over frame windows is 0.93 on LuSNAR and 0.81 on TartanGround, and over scenes 0.98 and 0.97, so it is a stable ranking and can support the correlations asked of it.

Uncorrected, four comparisons reach $p < 0.05$ under resampling of the model pool: Chamfer agrees with navigation better on TartanGround than on LuSNAR ($p = 0.010$), Chamfer beats per-view depth on TartanGround ($p = 0.006$ on coverage and 0.036 on traversability), and depth beats Chamfer on LuSNAR coverage ($p = 0.032$). Eleven comparisons were made, and Holm's correction over them keeps none, the smallest adjusted p being 0.062, so these comparisons are exploratory and are read as such. The others do not reach even the uncorrected level. The contrast between Chamfer on ETH3D and on LuSNAR is not resolved ($p = 0.255$), nor is depth's lunar advantage over either terrestrial domain ($p = 0.246$ and 0.063), nor depth over Chamfer on lunar traversability or safe paths ($p = 0.076$ and 0.177), nor Chamfer over depth on the ETH3D hazard map ($p = 0.104$), nor Chamfer's higher lunar agreement with traversability than with coverage ($p = 0.124$). The claim that metric rankings and navigation diverge in both directions therefore rests on the coverage outcome as a point estimate; after correction neither direction is supported, and it is stated only as a point estimate rather than as a general divergence.

Two correlations rest on one model each. TartanGround's depth-coverage figure of $+0.24$ falls to 0.00 without MV-DUSt3R+, and LuSNAR's Chamfer-coverage $+0.20$ falls to 0.05 without Fast3R. Its model-resampling interval is $[-0.30, +0.60]$, and those of LuSNAR's other three Chamfer correlations are $[-0.09, +0.77]$ on traversability, $[-0.12, +0.73]$ on reachability and $[-0.10, +0.75]$ on safe paths, every one of the four containing zero. Finally the two figures most often compared, LuSNAR's Chamfer at $+0.20$ and depth at $+0.66$, are not like for like: over the eleven models that carry both, Chamfer is $+0.16$.

Coverage is one outcome of four. Table 3's agreement column is against navigation coverage, the registered primary outcome, and coverage is a footprint measure: that a point-cloud completeness score agrees with it is partly structural rather than evidence about obstacle clearance. Against the

functional outcomes Chamfer reads higher on LuSNAR than its coverage figure of $+0.20$, at $+0.44$ on traversability, $+0.41$ on safe paths and $+0.39$ on reachability, though that within-domain difference is not resolved under model resampling. LuSNAR is the weakest domain against AMEDS and TartanGround on coverage, reachability and traversability, and is not separated from Erfoud on any outcome; these domain contrasts are uncorrected, since they lie outside the eleven comparisons Holm's correction covers. A reader should not take the single column as a statement about navigation safety, nor these four as four independent readings.

Predictive validity in full. Table G.3 gives every stored geometry metric against every navigation outcome, of which Table 3 shows the coverage column; the coverage agreement of relief correlation and its composite is in Table 3, and their four-outcome agreement on the audit cells in Table M.5.

Table G.3: Predictive validity. Mean within-scene Spearman correlation between the ranking induced by a geometry metric and the ranking induced by a navigation outcome, over all four registered outcomes. Higher is better in every column. A scene contributes only where the outcome varies across models, since a constant outcome induces no ranking, and a cell reads n/a where fewer than three scenes survive that rule, as for safe paths on AMEDS and ETH3D: on ETH3D every model reaches the same safe-path rate on four of the six scenes that carry enough models, which leaves two. AMEDS's and Erfoud's depth is their LiDAR projected into each view.

Domain	Geometry metric	vs coverage	vs traversability	vs reachability	vs safe path	Scenes
LuSNAR (lunar)	Chamfer	$+0.204$	$+0.439$	$+0.387$	$+0.407$	5
	Completeness	$+0.746$	$+0.703$	$+0.617$	$+0.650$	5
	F-score at 0.1 m	$+0.391$	$+0.314$	$+0.279$	$+0.267$	5
	Depth AbsRel	$+0.664$	$+0.831$	$+0.723$	$+0.732$	5
AMEDS (real, lunar analogue)	Chamfer	$+0.837$	$+0.736$	$+0.753$	n/a	4
	Completeness	$+0.850$	$+0.763$	$+0.768$	n/a	4
	F-score at 0.1 m	-0.071	-0.202	-0.097	n/a	4
	Depth AbsRel	$+0.766$	$+0.658$	$+0.550$	n/a	4
Erfoud (real, desert sand)	Chamfer	$+0.520$	$+0.488$	$+0.523$	$+0.493$	5
	Completeness	$+0.663$	$+0.531$	$+0.588$	$+0.548$	5
	F-score at 0.1 m	-0.133	-0.136	-0.097	-0.137	5
	Depth AbsRel	$+0.647$	$+0.740$	$+0.779$	$+0.750$	5
ETH3D (real, laser)	Chamfer	$+0.514$	$+0.605$	$+0.522$	n/a	11
	Completeness	$+0.513$	$+0.628$	$+0.485$	n/a	11
	F-score at 0.1 m	$+0.304$	$+0.427$	$+0.292$	n/a	11
	Depth AbsRel	$+0.108$	$+0.176$	$+0.257$	n/a	11
TartanGround (simulation)	Chamfer	$+0.802$	$+0.745$	$+0.594$	$+0.541$	20
	Completeness	$+0.852$	$+0.765$	$+0.599$	$+0.558$	20
	F-score at 0.1 m	$+0.602$	$+0.529$	$+0.475$	$+0.440$	20
	Depth AbsRel	$+0.240$	$+0.206$	$+0.081$	$+0.153$	20

Resolution in detail: StereoLunar. Under the frozen scale DVLT is distinguishable from StreamVGGT at a corrected $p = 0.011$. That scale varies by a median factor of 18 between calibration subsets there, and under a per-scene fit 49 pairs separate and DVLT keeps first place, the runner-up becoming Pi3 ($p = 0.038$), so the per-scene rows of Tables G.4 and G.5 are the more reliable reading of this domain.

Resolution in detail: the floored designs. Table G.4 prints each design's floor beside its count of separated adjacent pairs, so a count of zero on a design whose floor exceeds α reads as the arithmetic it is rather than as evidence about the models. ETH3D's depth metric, with nine models, is the one small design the floor leaves open, and there the test separates 20 of 36 pairs, none of them adjacent. The pairwise procedure corrects within one domain and one metric, not across the twelve combinations, so the per-combination results are read as exploratory, and nothing here turns on one borderline p-value. The floor is also a property of the procedure, exact rank tests on scene-level scores: LuSNAR's five scenes carry 20 windows, and a hierarchical model that keeps the windows nested in their scenes (Corani et al., 2017) could resolve what the scene averages cannot, so the resolution column of Table 1 describes the procedure used here rather than the data.

Resolution in detail: equivalence. The region of practical equivalence is 5% relative, applied to log values, since a threshold in metres would be large on an easy scene and negligible on a hard one.

Of the 573 ground-level pairs, 504 are undetermined. Only 67 are decided and 2 are established as practically equivalent, and the largest posterior any first-ranked model attains over the second is 0.83. The verdict does not hinge on the width of the region: re-running the identical test at half and twice it leaves 498 and 513 of the 573 pairs undetermined, with 74 and 50 decided, and 1 and 10 practically equivalent, at 2.5% and 10% respectively. Widening the region from 5% to 10% takes 17 pairs out of the decided count, 8 of them to equivalence and 9 to undetermined, so the wider region resolves no more of the benchmark. That distinction, between models shown alike and models about which nothing is resolved, is invisible to a p-value, which is the reason for reporting a second framework at all. StereoLunar decides 149 of its 237 pairs, and its winner's posterior over the second is 0.96 under the frozen scale and 0.93 under the per-scene fit.

Resolution in detail: resampling. Resampling converts the pairwise null into a design quantity, tabulated in Table G.5: scenes are drawn with replacement at the observed count and at twice and four times it, and the whole pairwise procedure is re-run each time. Separating the best model from the worst needs few scenes, 3 to 8 by Demšar's closed form, within what every domain here provides. Separating the top two does not follow from more of the same data: on TartanGround's Chamfer and coverage the top two separate in at most 0.03 of resamples at its 20 scenes and at most 0.16 at 80, and on ETH3D's Chamfer in 0.00 at 11 and 0.74 at 44. Depth behaves differently: TartanGround's top two separate in 0.42 of resamples at 20 scenes and 0.99 at 80. On StereoLunar the frozen-scale top pair separates in 0.97 already, and the per-scene fit's in 0.56 at 14 scenes and in every resample at twice that. The resampling assumes the observed scenes represent the population a larger benchmark would draw from, which is optimistic, since a gap measured on few scenes is selected for being large; it is not reported for LuSNAR's five scenes, which are too few to resample.

The published tables in detail. The sixteen rows of Section 6.1 were transcribed with nothing rescaled or imputed; a model missing any column of a row is left out of that row, which is why the ATE and RPE rows of E3D-Bench's Table 4 carry 11 and 9 of its 12 models. Each is one metric's ordering of the models inside one published table, and the five tables they come from are named in Table G.2: a paper's Table 4 contributing three metrics contributes three rows, not three tables. Spanning 7 to 16 models and 5 to 8 columns, they agree on the model ordering at Kendall's W of 0.24 to 0.86, and the reported winner ranks first in only 2 to 4 of its columns, so the tables are far from random and still cannot separate a pair.

Provenance and qualifications for the published tables. Each table was transcribed with its source, table number and access date recorded so that any cell can be rechecked, and two were verified cell by cell against the rendered source after transcription. Three qualifications are owed on the re-analysis in Section 6.1. Columns are whole datasets rather than scenes, so each is one observation and the counts are not directly comparable to the per-scene analysis. The accuracy, completeness and normal-consistency columns come from the same runs, so the sixteen rows are not sixteen independent tests. And the re-analysis disputes none of the transcribed values, only what can be concluded from them. Dependence between the columns of one table, where two datasets share scenes or capture conditions, cannot weaken the conclusion: $2/2^n$ is the smallest value the exact test returns over n columns whatever their correlation, correlated columns carry fewer than n independent observations and so resolve less rather than more, and Holm's correction controls the family-wise error under any dependence between the tests (Holm, 1979).

H PROTOCOL AXES IN DETAIL

Table H.1 lists every axis. The ten scoring axes re-score stored predictions and, with an alternative camera-centre protocol, form Section 6.2's comparisons, one per metric and level other than the axis's reference; their per-domain entries sum to its 266 and 268. That protocol's choices are four, applied alone and then stacked two, three and four at once: a similarity fit on camera centres, F-score tolerances at 1% and 2% of the diagonal of the ground-truth camera centres' bounding box, each model's own confidence rule, and ground truth kept to 1000 m rather than the frozen range. A model's own confidence rule is not a common standard: across seven backbones, confidence ranks error well but understates its size by a median factor of at least 2.4, and more the more confident the model (Deng et al., 2026), so one rule removes different error populations from different models. The four grid parameters (cell size, minimum points per cell, height quantile and plane window)

Table G.4: Resolution over k models and n scenes. W is Kendall's concordance of the scenes on the model ordering, and Friedman p the Friedman test's p-value in the form of Iman and Davenport. CD is Nemenyi's critical difference in mean ranks at $\alpha = 0.05$. Min. p is the smallest Holm-corrected p an exact Wilcoxon test can return with this many scenes and pairs, $\min(1, m{\cdot}2/2^n)$; where it exceeds 0.05 no pair can separate whatever the data. Adjacent counts the adjacent model pairs that Wilcoxon with Holm correction separates at $\alpha = 0.05$. Best/worst evaluates $N > q_\alpha^2 k(k+1)/(6\delta^2)$ at the observed mean-rank gap δ between the best and worst models. StereoLunar's Chamfer is reported under its frozen scale and under a Sim(3) fit per scene.

Domain	Metric	k	n	W	Friedman p	CD	Min. p	Adjacent	Best/worst
LuSNAR	Chamfer	14	5	0.63	<0.001	8.87	1.00	0 / 13	4
	Coverage	14	5	0.93	<0.001	8.87	1.00	0 / 13	3
	Depth AbsRel	11	5	0.96	<0.001	6.75	1.00	0 / 10	3
ETH3D	Chamfer	12	11	0.28	<0.001	5.02	0.064	0 / 11	8
	Coverage	12	11	0.48	<0.001	5.02	0.064	0 / 11	5
	Depth AbsRel	9	11	0.74	<0.001	3.62	0.035	0 / 8	4
TartanGround	Chamfer	12	20	0.55	<0.001	3.73	<0.001	1 / 11	5
	Coverage	12	20	0.68	<0.001	3.73	<0.001	1 / 11	4
	Depth AbsRel	9	20	0.62	<0.001	2.69	<0.001	1 / 8	4
StereoLunar	Chamfer	14	14	0.98	<0.001	5.30	0.011	11 / 13	3
	Chamfer, Sim(3)	14	14	0.70	<0.001	5.30	0.011	2 / 13	4
	Depth AbsRel	11	14	0.78	<0.001	4.04	0.007	2 / 10	4

Table G.5: Resampling. Share of 200 smoothed bootstrap resamples in which the observed first- and second-ranked models separate under Wilcoxon with Holm correction, at the observed number of scenes n and at twice and four times it. Designs with fewer than ten scenes, LuSNAR's three, are too small to resample and are not included.

Domain	Metric	n	P(top two separate)		
			$1 \times n$	$2 \times n$	$4 \times n$
ETH3D	Chamfer	11	0.00	0.25	0.74
	Coverage	11	0.00	0.00	0.07
	Depth AbsRel	11	0.01	0.03	0.03
TartanGround	Chamfer	20	0.03	0.04	0.16
	Coverage	20	0.03	0.04	0.14
	Depth AbsRel	20	0.42	0.85	0.99
StereoLunar	Chamfer	14	0.97	1.00	1.00
	Chamfer, Sim(3)	14	0.56	1.00	1.00
	Depth AbsRel	14	0.01	0.23	0.64

move rankings least, with 6 of 66 and 4 of 64 comparisons robustly below $\rho = 0.90$ on LuSNAR and TartanGround. Seven of the eight remaining axes change what a model sees, and the eighth, ground-truth extent, what its output is scored against; each is reported below with its own denominator.

Corruption design. Seven corruptions are used, each justified by a deployment failure rather than adopted wholesale from an image-classification corruption suite: blur from platform motion, defocus, additive Gaussian noise and shot noise from low-light sensing, exposure shift from high-contrast illumination, a combined low-light condition, and compression from a constrained downlink. Corruptions that presuppose an atmosphere are excluded, since they have no physical meaning on one of the domains. Blur, exposure and shot noise are applied in linear light, and additive Gaussian noise after encoding and compression to the encoded image, matching where each is modelled to occur; low light adds its photon and read noise in linear light, before the gain that restores brightness. Blur radii are specified as fractions of image width so that one severity denotes the same angular blur across domains with different sensor resolutions.

Table H.1: The protocol axes. For a scoring axis, levels counts the levels compared with the axis's reference, which is the frozen protocol's setting except that the scale track is compared with unscaled predictions and depth scaling with a scale-and-shift fit in depth, so that the frozen scale and the median depth scaling are among the levels there. Each domain's entry is comparisons, those below $\rho = 0.90$ and those that stay below under a scene bootstrap; a level that reproduces the reference and a comparison a level cannot move by construction are excluded (Appendix I). An input axis changes the images or the training and is not in those counts.

Scoring axis	Levels	LuSNAR	TartanGround
Alignment to camera poses	2	11 / 11 / 10	14 / 12 / 10
Scale track	2	14 / 14 / 14	16 / 15 / 13
Injected scale	7	56 / 25 / 11	56 / 20 / 9
Confidence filter	5	38 / 26 / 18	38 / 25 / 15
Depth scaling	4	8 / 8 / 4	8 / 4 / 3
Grid cell size	3	18 / 11 / 3	16 / 7 / 1
Minimum points per grid cell	3	18 / 8 / 2	18 / 9 / 2
Height quantile	4	24 / 8 / 1	24 / 8 / 1
Plane window	1	6 / 1 / 0	6 / 2 / 0
Region radius	3	17 / 6 / 2	16 / 6 / 3
Camera-centre protocol, singly and combined	7	56 / 32 / 20	56 / 40 / 29
Input or ground-truth axis	**Levels**	**Where it was run**	
Corruption	7×3 severities	LuSNAR and TartanGround	
Input resolution	4 long-side sizes	Moon_1, Moon_8, NordicHarbor, Prison	
Input crop	full frame, 4 models	20 lunar and 40 terrestrial windows	
Camera intrinsics	6 perturbations	Moon_1, Moon_8, NordicHarbor, Prison	
View count	fixed stride and span	LuSNAR's 5 scenes and 2 of TartanGround's	
View selection	jittered, diverse window	Moon_1, Moon_8, NordicHarbor, Prison	
Ground-truth extent	scan culled or uncut	ETH3D's 11 scenes	
Adaptation schedule	4 recipes, 1 more on CUT3R	LuSNAR's Moon_1 and Moon_8	

The corruption harness and its scale. An IDENTITY level runs the entire corruption path without changing a pixel and reproduces the clean row to 5.6×10^{-9}, so the harness contributes nothing. Where the two scales are expressible in one unit, this benchmark's hardest defocus is milder than ImageNet-C's severity 1, its hardest Gaussian noise falls between their severities 1 and 2, its hardest shot noise between their 3 and 4, and its hardest JPEG quality equals their severity 5 (Table M.10). On LuSNAR, 17 of the 21 corruption and severity conditions improve Chamfer; the four that do not are the two stronger defocus levels, the mildest brightness and the hardest motion blur (Table H.4).

Ground-truth extent. ETH3D admits two defensible ground-truth choices, its laser scan culled to the union of the scored frusta or left uncut. The choice moves Chamfer by 0.8% on average and preserves the ranking almost exactly ($\rho = +0.993$), a null against this paper's own thesis that protocol choices dominate. It is not a uniform null: 9 of the 12 models score *better* uncut and only 3 worse, because the extra ground truth helps accuracy match more of it while making completeness harder.

Does a degraded input flatten the prediction? Over all 420 LuSNAR corruption cells, the seven corruptions at three severities on Moon_1 and Moon_8 for ten models, Chamfer improves in 233; there the point count is unchanged (median ratio 1.00), and accuracy and completeness improve together in 171, so the prediction does not win by thinning. Chamfer improving under corruption on LuSNAR admits two readings: the model returns a flatter surface, which Equation 4 says has a lower Chamfer ceiling on low relief, or the alignment and scoring interact with a sparser prediction. The corruption sweep's cells at severity 3, and the identity level beside them, were rerun with their aligned predictions saved, and each prediction's own deviation from its least-squares plane, $\mathrm{dev}(h_s)$, measured as the mean distance of its points from that plane. The predictions are not flat to begin with: at the identity level their $\mathrm{dev}(h_s)$ is a median 12.8 times the target's. A corruption changes that by a median factor of 1.00 over the 140 cells, flatter on 70, with the point count unchanged (median ratio 1.00). Of the 75 cells whose Chamfer improves, 44 are flatter, by a median of 4%, and across all cells the change in relief and the change in Chamfer are barely related (Spearman $+0.09$); motion blur flattens most, by a median 12%. Where a degraded input improves Chamfer it trims some of the prediction's spurious relief more often than not, which lowers its Chamfer ceiling a little, but it does not make the surface flat, and the improvement is not mainly flattening.

View count. Eight models, four batch and four streaming, were run at a fixed stride of four frames, from 2 to 16 views, and at a fixed span of 112 frames, from 2 to 65 views (113 on Moon_1): on LuSNAR at the fixed span on Moon_1 and Moon_8 and, at 3, 5, 9 and 17 views, on the other three scenes (Appendix F), and at the fixed stride on Moon_1 and Moon_8, and on TartanGround at both on NordicHarbor and Prison, eleven scene-and-family combinations in all. Adding views at a fixed stride widens the terrain span and tends to make models worse; adding them at a fixed span tends to make them better. The tendency is not a law: every model moves the family's way on 7 of the 11 combinations for Chamfer and on 2 for the F-score. What holds more widely is the comparison between model classes: streaming models degrade less and improve more than batch models on Chamfer in all 11 combinations, on both domains and in both families, and on the F-score in 9, though on slope error in only 5.

View selection. With the count fixed at eight and the span at 112 frames, which frames a model sees was varied: uniform spacing at a stride of 16; uniform spacing with a random offset per slot, over three seeds; and farthest-point sampling on the ground-truth camera centres. Four models were run on Moon_1 and Moon_8 and on NordicHarbor and Prison, 72 cells a domain. The choice moves the ordering below $\rho = 0.90$ in 8 of 22 comparisons on LuSNAR and 10 of 17 on TartanGround, and of the 11 pairs of a metric and a selection rule that both domains score, the two disagree on whether it breaks the ranking for 8. The frozen protocol fixes the uniform rule for every model, so this axis measures what an unstated selection rule could move, not a source of variation in the paper's own numbers.

Input crop. Five models read frames through an input pipeline that centre-crops a square image to 4:3, so on LuSNAR and TartanGround they never see the top and bottom eighths of the frame, while the other models receive it whole. Four of the five (CUT3R, DUSt3R, MASt3R and Fast3R) can be given the full frame instead; MUSt3R cannot, since none of its trained resolutions is square. The axis re-runs those four on the same 20 lunar and 40 terrestrial eight-frame windows at two levels: the default crop, re-run through the axis as its own control, and the full frame. The control reproduces the frozen zero-shot row exactly in 237 of 240 windows, and the remaining three, two DUSt3R windows and one Fast3R window, differ by at most 0.3%. The crop is the models' own preprocessing and stays in the frozen protocol, since forcing a square input puts these checkpoints outside the aspect ratios they were trained on. Table H.2 gives the outcome. Under the full frame CUT3R's lunar F-score rises in 19 of 20 windows, while its Chamfer does not follow. The crop does not favour the controls: with MASt3R's full-frame Chamfer in place of its cropped one, the best model changes on 4 of LuSNAR's 11 audit cells and the wrong scene beats every model on 11 of them rather than 10, the plane still on all 11.

Table H.2: Input crop. Median F-score at 0.1 m and Chamfer under each model's default 4:3 crop and with the full square frame, on the same eight-frame windows, with the windows where the full frame is better and worse in parentheses, and the model's mean rank among all zero-shot models on those windows at each level. The two counts need not sum to the window total: at a fine tolerance the F-score is exactly zero under both crops on the weaker models, and those exact ties are in neither count. Windows share scenes, so the counts are descriptive.

Domain	Model	F-score at 0.1 m		Chamfer	
		median	mean rank	median	mean rank
LuSNAR (lunar)	CUT3R	0.007 → 0.051 (19/1)	10.10 → 8.05	0.639 → 0.672 (9/11)	6.90 → 7.80
	DUSt3R	0.067 → 0.090 (14/6)	7.30 → 7.00	0.690 → 0.779 (9/11)	8.50 → 8.95
	Fast3R	0.008 → 0.005 (6/8)	10.30 → 10.50	1.003 → 0.965 (6/14)	11.90 → 11.95
	MASt3R	0.259 → 0.032 (5/15)	4.50 → 7.58	0.288 → 0.695 (1/19)	1.80 → 7.25
TartanGround (simulation)	CUT3R	0.043 → 0.066 (27/13)	4.16 → 3.95	0.711 → 0.707 (25/15)	3.90 → 3.95
	DUSt3R	0.024 → 0.026 (23/16)	5.50 → 5.53	1.340 → 1.320 (28/12)	6.58 → 6.50
	Fast3R	0.014 → 0.016 (22/12)	7.55 → 7.38	1.386 → 1.370 (30/10)	8.25 → 8.15
	MASt3R	0.039 → 0.113 (27/13)	3.92 → 3.38	0.611 → 0.451 (24/16)	2.85 → 2.98

The remaining axes. Supplied intrinsics reproduce the clean row for Depth Anything 3, whose adapter discards them without extrinsics; the resolution levels are in Table M.9, and ETH3D's culled against uncut scan is a near-null. On Chamfer, streaming models degrade less and improve more with view count than batch models on all 11 scenes and families swept on both domains.

Table H.3: Input corruption. Median ratio of the corrupted score to the clean score, matched on model, scene and frame window. Values above 1 mean corruption made the metric worse, which is the expected direction.

Domain	Metric	Severity 1	Severity 2	Severity 3
LuSNAR (lunar)	Chamfer	0.985	0.965	0.979
LuSNAR (lunar)	Depth AbsRel	1.069	1.140	1.292
TartanGround (simulation)	Chamfer	1.008	1.023	1.083
TartanGround (simulation)	Depth AbsRel	1.000	1.008	1.024

Table H.4: Which corruption drives the effect. Median ratio of corrupted to clean Chamfer, per corruption and severity, matched on model, scene and frame window. Values below 1 mean the corrupted input scored better. On the lunar domain 17 of 21 cells fall below 1, and the four that do not are the two stronger defocus levels, the mildest brightness and the hardest motion blur. Noise, low light and compression improve the lunar score at every severity while the two stronger defocus levels worsen it. Compression and the milder blurs also remove high-frequency texture and still improve it, so the rows do not isolate a texture mechanism, and the improvement is not mainly flattening (Spearman $+0.09$ between the changes in $\mathrm{dev}(h_s)$ and in Chamfer, above).

Corruption	LuSNAR (lunar)			TartanGround (simulation)		
	1	2	3	1	2	3
Brightness	1.038	0.932	0.926	1.005	0.990	1.006
Defocus	0.984	1.020	1.044	1.018	1.150	1.210
Gaussian noise	0.963	0.989	0.975	1.006	1.021	1.016
JPEG	0.974	0.954	0.949	1.072	1.164	1.414
Low light	0.986	0.966	0.965	0.996	1.005	1.032
Motion blur	0.999	0.996	1.012	1.057	1.199	1.177
Shot noise	0.936	0.962	0.983	1.004	1.005	1.069

Adaptation. Four fine-tuning schedules applied to MapAnything, Pi3 and CUT3R do not agree on what helps: each model's own recipe gives its lowest median ratio, and the outcome depends on the pairing of model and recipe (Table M.11, Appendix M.4); at three seeds these orderings are not resolved (Appendix G).

Does the scale protocol treat the method families alike? Neither family is favoured on both domains: MapAnything's metric claim is right to within 2.3% on TartanGround and off by a factor of 1.74 on LuSNAR, where the frozen calibration of the other models transfers better (Appendix M.4).

What a protocol change does to the head of the table. Spearman agreement is a statement about a whole ordering, and a high value is compatible with the one change a reader of a leaderboard would notice: a different model in first place. The same comparisons Section 6.2 counts were therefore re-examined for that event alone. Two things are counted separately, because conflating them overstates the instability: the winner changing, and the first two trading places exactly.

The head is less stable than the aggregate suggests, and the part that matters is the right-hand block of Table H.6. Restricted to the comparisons whose agreement is at or above 0.90, which a reader would have taken as the ranking holding, the top-ranked model still changes in 20 of 116 on LuSNAR and 19 of 120 on TartanGround, about one in six either way, and most of those are an exact swap of the leading pair. Over all comparisons the winner changes in 127 of 266 and 96 of 268. The rate among the comparisons that look stable was one in ten before the comparisons that cannot move by construction left the denominator, sixty per domain of LuSNAR's 326 and TartanGround's 328, all at $\rho = 1.00$, so every one of them had been counted as agreement. This does not contradict Section 6.2's reading that the claim is degraded agreement rather than reversal: ρ's bootstrap interval lies wholly below zero in only 13 of the 266 comparisons on LuSNAR and 2 of the 268 on TartanGround. It does mean that a table reporting ρ alone has not shown its winner to be stable, and

Table H.5: Does the clean model ranking survive corruption? Spearman correlation between the clean and corrupted ordering, min to max over seven corruptions at that severity. Higher is better; a value near zero means the ranking is scrambled. LuSNAR's two severity-1 ranges coincide although each is computed from its own metric's scores; with few models, Spearman's ρ takes few values.

Domain	Metric	Severity 1	Severity 3
LuSNAR (lunar)	Chamfer	$+0.72$ to $+0.96$	$+0.19$ to $+0.73$
LuSNAR (lunar)	Depth AbsRel	$+0.72$ to $+0.96$	$+0.50$ to $+0.88$
TartanGround (simulation)	Chamfer	$+0.28$ to $+0.95$	$+0.17$ to $+0.92$
TartanGround (simulation)	Depth AbsRel	$+0.40$ to $+0.99$	$+0.50$ to $+0.95$

that a benchmark changing any scoring choice should report what happened to first place rather than to the ordering as a whole.

Table H.6: The head of the ranking under a protocol change. Winner changes counts the comparisons whose first-ranked model differs from the frozen protocol's; top two swap counts the stricter event where the same two models appear first in the opposite order, which is a subset of it. The right-hand block restricts to the comparisons whose Spearman agreement with the frozen ordering is at or above 0.90.

Domain	All comparisons			Those at $\rho \geq 0.90$		
	Total	Winner changes	Top two swap	Total	Winner changes	Top two swap
LuSNAR (lunar)	266	127	21	116	20	16
TartanGround (simulation)	268	96	21	120	19	15

Defaults these axes would recommend. The sensitivity study is worth little if it ends in a list of effects, so this is what the measurements here argue a shared benchmark should fix, and in two places what they argue it should *not* fix. Report the input crop and score both, rather than enforcing a full frame: the crop moves the leading lunar model, MASt3R, from a mean Chamfer rank of 1.80 to 7.25 and changes CUT3R's lunar F-score on every window (Table H.2), but forcing a square input puts these checkpoints outside the aspect ratios they were trained on, and no level here pads the frame to a trained ratio that would separate the two causes, so the defensible default is to publish which crop produced which row. Do not treat supplied intrinsics as a protocol variable until the adapter is shown to consume them: one model here accepts them and discards them unless extrinsics come too, and the evidence for that is a bit-identical score, not a reading of the code. State the resolution as an absolute long-side pixel count and check each level against the adapter's own internal size, since two combinations here were no-ops, MASt3R at 512 px exactly and Spann3R at 224 px exactly on three of its four cells and to five significant figures on the fourth, and would otherwise have been counted as robustness. Fit the alignment to camera poses rather than to the geometry, so that the transform cannot absorb the error the metric is meant to report. Finally, the ground-truth extent is the one axis that can be left to the curator: culling the laser scan to the scored frusta or leaving it uncut moves Chamfer by under a percent and preserves the ranking almost exactly, so it needs to be recorded but not standardised.

I Verification and reproducibility guarantees

Every number computed in this study, that is every number other than the values quoted from cited work, among them the E3D-Bench rows of Table G.2 and the corruption levels of Table M.10, is produced under a frozen protocol, inside containers pinned by digest, and read through code that asserts its invariants rather than assuming them. Each run records the git commit, the container image, the GPU model and the protocol it was produced under; the frozen-protocol filter reads that recorded protocol rather than the run identifier, and datasets are read-only images mounted into the containers, so a reproduction runs against the same bytes. The tables are emitted by one generator that reads the evaluation log through one log reader, a shared control filter and a shared protocol filter, and a checker recomputes every number computed in this study that the prose states, appendices included, and fails the build when one disagrees with the evaluation outputs. The seven

checks below are those whose failure modes are not specific to this implementation. Each is stated with what it guards against and with the size of the effect, measured on this benchmark during its development, that it keeps out of the tables; a benchmark with its own data readers and metrics can adopt them without adopting anything else here. Superseded rows are dropped by reading only the newest row of each run, and stale summaries by the checker, which reads the summaries the tables are built from.

Every score carries its scale. Each row records the scale it was scored at and where that scale came from, so a unit scale is visible in the log rather than inferred from a ranking. The check guards against an alignment mode that fits a unit scale to the raw prediction instead of applying the scale calibrated on validation data: of MapAnything, MASt3R, Pi3, VGGT and CUT3R, the four not used at their own metric scale are then scored at between 2.6 and 20 times too small a scale, and the model ranking changes.

Controls never enter a model pool. One shared predicate removes control rows before any model pool is formed, keyed on how a row was constructed rather than on its name. A control borrows the evaluated model's identifier, carries the same task field and passes every protocol filter, so a list of control names does not suffice: a list naming the first control family matches 4 of the 2017 control rows the log holds, which are spread over sixteen run-identifier families that name the batch producing a row rather than its construction. Without the predicate a control can be selected as a cell's best model and report a deficit of $74.2\times$ that does not exist, or inflate one model's mean until a protocol change appears to reverse a ranking that in fact moves by one place.

Comparisons are matched on scene and frame window. A comparison is keyed on scene and frame window, and a pair without a matching window is dropped rather than approximated. A model's score varies more between frame windows than between models, one model's Chamfer on a single lunar scene being 0.166 on one window and 0.524 on another, so a comparison across windows can report a control failure on two domains that is confined to one.

Results are reported per domain. Control and validity results are computed and reported per domain and never pooled. Pooled, Chamfer rejects the wrong scene on 71 of the 96 same-domain and unseen cells of Table 2, 74%, and on 1 of 11 on LuSNAR alone (Table 3), from identical data, because a pooled rate is a weighted average of a failure and a success whose weights are how many cells of each domain were run.

Resampling draws scenes. A within-domain correlation over scenes counts each scene once, and every resample draws scenes rather than rows. Counting control rows instead counts each scene twice where each target has two donors, and turns $p = 0.136$ into $p = 0.028$.

A protocol level must be able to change the score. A level whose every scored cell equals the frozen level's is detected and excluded, and so is a comparison that cannot move by construction because the metric does not read what the level changes; both exclusions are printed with the analysis rather than applied silently, and the protocol and top-of-table analyses share one implementation, so a level that starts to differ counts again by itself. The check guards against placebo levels. The alignment axis carries one: a request for a similarity alignment that the dispatcher honours only when the rotation is taken from camera centres, so on the orientation path the level is a byte-identical copy of the frozen rigid alignment on all 117 LuSNAR and 240 TartanGround cells, while a level that does differ matches on none of them. Counted, its eight comparisons per domain, each at $\rho = 1.00$ by construction, would dilute Section 6.2's share below 0.90, 150 of 266 on LuSNAR and 148 of 268 on TartanGround, to 150 of 274 and 148 of 276. Those denominators already leave out the sixty comparisons per domain that cannot move by construction (Appendix H); counted as well, they would give 150 of 326 and 148 of 328, and 150 of 334 and 148 of 336 with the placebo level. Rows that record an ICP refinement that never ran are the same failure mode, and the log reader drops them. The check works on whole levels of the scoring axes, so a level that leaves one model's score unchanged while moving the others, as MASt3R at 512 px and, on three of its four cells, Spann3R at 224 px do on the resolution axis, is reported in Appendix H instead, where the scale track also answers whether a similarity alignment changes the conclusions.

Ground truth is registered to the images. An ETH3D row enters an aggregate only with its registration recorded. ETH3D ships each scene's laser scans in their scanners' own frames, with a MeshLab project beside them holding the rigid transform that places each scan in the frame the camera poses and the rendered depth share, and a reader that ignores the project places the scans 0.15 m off on one scene and 3.9 m and 8.3 m off on the two scans of another, one of them also rotated by 4.8 degrees. The dataset's own oracle check cannot see this, because it scores the scan against itself; a per-frame check that compares each scored frame's depth with the scan does, finding offsets of 0.11 to 1.56 m without the transform and agreement on 94 to 100% of the scan points projected into the images with it. Every ETH3D surface and navigation number, both ground-truth extents, the controls and the relief measurement depend on it, and the size of the effect is itself a result. Re-scoring all 132 cells from stored predictions with and without the registration moves the Chamfer ranking to a Spearman of $+0.91$ against itself: 9 of 12 models change rank, the largest shift is three places, and the winner changes, while Chamfer's agreement with navigation rises from $+0.41$ $[+0.00, +0.75]$ unregistered to $+0.51$ $[+0.18, +0.79]$ registered. A choice inside the evaluation pipeline moves the leaderboard further than the models differ from each other, which is the finding of Section 6.2 in its most concrete form. Rows scored without the registration carry no registration field, so no such row can re-enter a table.

J ADDITIONAL ANALYSES

This appendix reports the analyses the main text refers to but has no room for, and states what each one leaves open. Their numbers are checked against the analysis outputs like every other number in the paper.

Low-relief terrain that is not lunar. AMEDS is itself terrestrial ground: volcanic terrain on Mount Etna, photographed by a rover and scanned by LiDAR, used because it resembles the Moon. The wrong scene wins 7 of its 10 cells and the plane all 10 (Section 4, Appendix C.1), so the failure is not a property of lunar imagery or of rendering. Erfoud closes the other half of that question: Saharan interdune sand, photographed by a rover at a Mars-analogue site and not for any resemblance to the Moon, where the wrong scene wins 7 of 10 cells and the plane all 10 cells (Section 4) and all 5 windows (Appendix F). No dataset of snow, ice, seafloor or indoor corridors, among other low-relief terrain, with posed images and metric ground truth was available here, so those remain untested; the plane control needs no donor scene and can be run on them directly. A real scene tests the opposite direction and shows that the ratio of error to relief is not sufficient on its own. ETH3D's electro has a local relief of 0.0142 m, lunar-like, and its best model errs by 57 times it, above LuSNAR's 16.7. The control was run there once with each of the other 10 evaluated ETH3D scenes as donor, so no donor was chosen. The wrong scene is rejected against all 10, with a median margin of 0.38 and a largest of 0.53, and completeness, per-view depth and navigation coverage reject it on all 10 as well, the F-score on 9. A high ratio of error to local relief is therefore not sufficient for the failure, and the plane control of Section 4 says why: on all 10 electro cells the plane loses to the best model as well, so electro's ground truth is not one surface at the scale of that error even though its columns are shallow. Scene-wide spread says the same thing in another way. Electro's scene-wide height standard deviation, 7.0 m, is the largest of any ETH3D scene, and measured against that spread its best model errs by 0.12, inside ETH3D's range of 0.10 to 0.62 rather than lunar terrain's 0.37 to 4.27.

Relief within a domain, and two alternatives to it. Relief tracks the margin within a domain, as the mechanism requires: the TartanGround scenes whose ground truth varies least come closest to a wrong-scene win ($\rho = -0.49$, $p = 0.003$, $n = 36$). Two alternatives were tested and rejected: a threshold scaled to each scene's relief leaves the lunar acceptance rate flat across a twenty-fold change in tolerance, and a pairwise terrain-similarity descriptor does not predict the margin ($\rho = -0.22$, $p = 0.094$, $n = 61$, against a smallest detectable $|\rho|$ of 0.25).

Donors that resemble their target. On LuSNAR every donor is regolith and the wrong scene already wins 10 of 11 cells, so a better-matched donor has little left to gain there. The terrestrial donors of the first control run, 15 TartanGround cells and ETH3D's 8, were labelled similar or dissimilar to their targets before scoring; the other 42 TartanGround cells, added when the donor set was extended, carry no label. On TartanGround the 9 similar-donor cells have a median margin

of 0.36 and 1 win, against 0.27 and 0 for the 6 dissimilar ones; on ETH3D the 4 similar ones 0.45 against 0.30 for the 4 dissimilar, with no win either way. Among donors chosen by hand, resemblance moves the margin up modestly and wins one of the fifteen TartanGround cells, a similar donor's, and the pairwise similarity descriptor above does not predict the margin at all.

A donor chosen adversarially is a different matter, and it changes one of the claims above. Each scored window is described by ten ground-truth features, the three quantiles of its within-column height spread, its height range and standard deviation, the logarithm of its footprint and of its column count, and its planarity, linearity and sphericity; all are standardised over the pooled descriptor set and the donor is the nearest eligible window of the same dataset in equal-weight Euclidean distance. The rule reads no Chamfer and no margin, and the donor list was frozen before the first cell was scored. It retrieves a far closer donor than a human chose: on terrestrial targets the median descriptor distance falls from 3.08 to 0.88, and on TartanGround the hand-picked donor sits at a median rank of 37 in the retrieval order.

The study covers 84 target windows: TartanGround's 54, which carry Table 2's 57 cells because three windows have two hand-picked donors; ETH3D's 9, the 8 scenes that carry a hand-picked control cell and electro's one window, whose 10 cells are its ten donors, the other two of its eleven scenes carrying no control cell; LuSNAR's 11; and 5 each for AMEDS and Erfoud, whose 10 cells are two donors per window. Scored on all 84, the rule makes the control win where a hand-picked donor did not. On TartanGround the wrong scene now beats every model on 6 of 54 windows against 1, a median margin of 0.38 against 0.32, and it beats the best hand-picked donor on 37 of the 54; 5 windows change verdict, at margins from 1.01 to 5.39. ETH3D does not move, 0 of 9 either way, and neither does any lunar or lunar-analogue domain: LuSNAR stays at 10 of 11, AMEDS at 4 of 5 and Erfoud at 4 of 5, with no verdict change anywhere. The reading above, that resemblance among hand-picked donors barely moves the verdict, therefore holds for the donors a person picks and not for the donor a retrieval rule picks, and Table 2's terrestrial rows should be read as a statement about hand-picked donors.

What the five flips are is the reason they are not a refutation of the mechanism. They are TartanGround's flattest, most nearly planar environments, two of them windows of BrushifyMoon, which is that dataset's lunar surface, and one of HQWesternSaloon, which is flat desert ground. The relief mechanism of Section 4 reappears inside a terrestrial dataset rather than being contradicted by it. Proximity is in any case a weak predictor of the margin: over the terrestrial cells the rank correlation between descriptor distance and margin is -0.15, and the median margin gain from retrieval is a factor of 1.16, so the effect lives in the tail rather than in the typical cell.

The ten features describe how much height a window has and how it is spread, not the spatial scale at which it varies, so a second arm added the height spectrum's log-log slope and the share of its power above the mid frequency, fixed before any score was read. The spectrum is undefined on a footprint under sixteen grid cells a side or a grid more than three quarters empty, so 42 of the 84 targets and 90 of the 137 pool windows cannot be described and leave that arm. On the 42 it describes, the control wins 12 times against 15 for the ten-feature rule and 13 for the hand-picked donors, and on TartanGround's 24 none against 3 and 1. Two of the three TartanGround wins lost are the spectrum's rather than the smaller pool's: restricted to the same 47 pool windows, the ten-feature rule still retrieves the two donors that win on BrushifyMoon and HQWesternSaloon, which the spectral rule does not; the third, on Downtown, goes with the pool. A descriptor of spatial scale therefore does not strengthen the adversary; the donors that win are the ones matched on relief and planarity, the quantities the mechanism names.

Image-based metrics. Novel-view metrics such as PSNR or SSIM were not compared. The feed-forward models return point maps and cameras rather than a scene representation that renders new views, so they would first require fitting a renderer to each prediction, which changes what is scored. A reprojection through the target's camera matrices and LPIPS and DINOv2 on depth renders of the geometry need no fitted renderer, and were measured: each passes the control where the ground has relief and none rejects both substitutes on more than four of the ten or eleven cells of a low-relief domain. A point-space correspondence consistency needs no renderer, and it was measured: it is a real signal on regolith, it does not predict lunar navigation better than per-view depth, and it cannot distinguish the ground truth from the ground truth flattened onto a plane.

Effect sizes behind the tests. The best model's error in units of relief has a median of 16.690 on LuSNAR with a 95% scene bootstrap interval of 2.68 to 104.84, 1.122 (0.19 to 2.80) on ETH3D and 0.626 (0.27 to 1.54) on TartanGround. The ratio of LuSNAR's median to ETH3D's is 14.9 (2.4 to 147) and to TartanGround's 26.7 (4.3 to 167), neither interval reaching 1, while ETH3D's to TartanGround's is 1.8 (0.2 to 7.0) and does not separate the two terrestrial domains. In units of $\mathrm{dev}(h_t)$, the unit step 6 of the checklist asks for, the medians are 6.05 (5.20 to 10.26) on LuSNAR, 0.29 (0.21 to 0.45) on ETH3D and 0.21 (0.19 to 0.27) on TartanGround, and LuSNAR's is 20.7 times ETH3D's (12.4 to 38.2) and 28.6 times TartanGround's (21.4 to 49.3). With five lunar scenes the lunar interval cannot leave their range and is optimistic. The top pair's effect size on each design is in Table M.8.

K Extended related work and positioning

This appendix places the paper among the work on surface metrics that the main text names only in passing, and states what the prior work established and what is left.

Surface metrics and their critics. Besides accuracy, completeness and the F-score, point-to-plane distance (Chen & Medioni, 1992), normal consistency (Mescheder et al., 2019) and the earth mover's distance (Rubner et al., 2000; Fan et al., 2017) are the standard alternatives to Chamfer, and each is put through the control in Appendices D and F. As a training loss, Chamfer's gradient collapses in a way no local regulariser prevents (Song & Hyde, 2026), a result about optimisation rather than discrimination: density-aware, hyperbolic and geodesic Chamfer change the gradient a network is trained by, while as evaluators they still aggregate nearest-neighbour distances, which Equation 4 bounds for any flat submission. Image-based and geometric metrics disagree on large outdoor captures (Xiong et al., 2024); a large-scale LiDAR-visual benchmark with laser-scanned ground truth finds radiance field methods overfitting to their training poses (Tao et al., 2026); reconstruction error indicates neither pose accuracy nor grasp success (Burde et al., 2025; 2026), and pose-alignment error does not indicate manipulation success (Hietanen et al., 2021); and a street-view benchmark supplies mesh ground truth carrying normals so that orientation can be scored where LiDAR points cannot (Hu et al., 2024), a direction tested here and found insufficient on low-relief ground (Appendix F).

Negative controls are not new to 3D evaluation. They are concurrent in at least three benchmarks. Cui et al. (2026) release a stereo generation benchmark whose evaluator ships calibration controls, among them a right view taken from a different scene under the same aggregation protocol, an unmatched scene substituted into an unmodified evaluator, as here. Shi & Gong (2026) release a Chamfer-scored dynamic-surface benchmark carrying three falsification controls and a ground-truth-side recoverability oracle, and argue in as many words that a protocol-level reality check need not introduce a new method to be a contribution. The closest is SysCON3D (Paul et al., 2026), a controlled benchmark for ground-truth-free multiview-consistency metrics whose levels mix in views of other scenes or noise: MEt3R (Asim et al., 2025) rates a random mixture of scenes as more consistent than a genuine scene, and the same work shows feed-forward models hallucinating geometry and cross-view support for such inputs.

What remains specific to this paper is the regime, and the substitute. That a trivial substitute can outscore learned reconstruction under the field's own metrics is not new: Tatarchenko et al. (2019) find retrieval and classification baselines matching or beating single-view networks. Whether such a control wins is decided by the target: Equation 2 says a plane must beat a model whose Chamfer exceeds the target's deviation from its own plane, so a concurrent benchmark that ran the same plane control on terrain in that regime would see it win too. The contribution is therefore to identify and bound the terrain class on which a flat plane must win, and to measure what the metric then fails to separate. Scoring a submission against a trivial reference is itself old in forecast verification: a mean-square-error skill score measures a forecast against climatology, and its decomposition into a correlation term and two bias terms shows what a correlation alone cannot see (Murphy, 1988). In those terms the flattened target is the terrain's climatological forecast, a model it beats has negative skill, and steps 5 and 6 of Section 7 report that skill. Relief correlation, which removes each cloud's own plane before correlating, keeps the correlation term and discards the bias terms, which is consistent with its passing the control and still ranking navigation below Chamfer (Section 5); combining

the terms as a skill score does, in place of the rank average of Section 5, is not attempted here. The substitute also differs from all three concurrent ones. The stereo controls act on image-space and matching measures and substitute a wrong image, and SysCON3D's feed wrong views to metrics that read no ground truth; here the measure is a nearest-neighbour surface metric and the substitute is a geometrically valid, correctly scaled reconstruction of another place, planted on the target's own ground-truth poses so that its alignment is an identity, and in Appendix J chosen by ten ground-truth descriptors of the target, while the falsification controls of the dynamic benchmark are null conditions its protocol is expected to pass. What the planting contributes is measured in Appendix D: once every submission, the substitutes included, is given the same similarity fit to the geometry, which removes every global misplacement, the wrong scene stops winning on AMEDS and Erfoud and keeps 4 of LuSNAR's 11 cells. On those two domains the metric's failure is therefore to separate a correctly placed wrong surface from a misplaced right one.

The resolution analysis is a transfer, not a discovery. Whether scores measure the intended construct is an established concern elsewhere in machine learning, through dataset bias (Torralba & Efros, 2011), label error (Northcutt et al., 2021), replication on new test sets (Recht et al., 2019), unfair comparisons and tuning on test-set feedback (Musgrave et al., 2020) and the choice of task suite (Dehghani et al., 2021), and test sets too small to support the significance a leaderboard claims are documented in natural language processing (Card et al., 2020), where rankings on standard splits also fail to reproduce under random splits (Gorman & Bedrick, 2019). Every component of Section 6.1 is established statistical practice and is cited as such. The observation that a published leaderboard's displayed ranking can fail a conventional resolution target has also been made independently and earlier in other fields: Kotawala (2026) finds 11 of 40 Open LLM Leaderboard v1 pairwise comparisons and 4 of 9 adjacent MMLU-Pro pairs unresolved at $(\alpha, 1-\beta) = (0.05, 0.8)$, and Kendiukhov (2026) reports pairwise reversal rates of 16.3%, 19.3%, 32.1% and 0.0% across four evaluation-protocol axes in gene-regulatory-network benchmarking, which is the design of Section 6.2 in another discipline. Read together they make the present analysis an instance of a pattern that recurs across leaderboards rather than a finding about 3D reconstruction alone. The attainability floor is not new either: the bound $\min(1, m \cdot 2/2^n)$ on the smallest adjusted two-sided p-value an exact signed-rank test can return over n scenes and m comparisons follows in one step from the exact null distribution, and using the smallest attainable p-value of discrete tests inside a Bonferroni-type family is established (Tarone, 1990). What is added here is its use on published tables, where it decides what a table of a given size could have reported whatever the effect size had been.

One of the tools has been challenged directly. Urbano (2026) argues that the Wilcoxon signed-rank test loses control of its Type I error rate in benchmark comparison and should be abandoned in information retrieval. The direction of that error protects some of this paper's conclusions and not others, and the two should be told apart. A test that rejects too readily can only overstate what a leaderboard separates, so it cannot manufacture the claims Section 6.1 makes about the leading pair on ground-level imagery, which are failures to separate, and the floor argument does not rest on the test's calibration at all, only on the number of distinct sign assignments available. The same direction puts that section's positive separations at risk: the tail pairs TartanGround separates, the 89 of 91 StereoLunar pairs and DVLT's lead over StreamVGGT at a corrected $p = 0.011$ are exactly what an anti-conservative test would inflate, and the per-scene check of that lead in Appendix G tests DVLT against Pi3 rather than against StreamVGGT.

Auditing a metric against constructed cases is not new either. Tomsett et al. (2020) apply psychometric reliability analysis to the metrics used to rank saliency methods and conclude that comparative rankings built on them can be untrustworthy, which is the method-level move this paper makes for a surface metric; the paper already cites the randomisation tests that began that lineage (Adebayo et al., 2018). Closest of all in method, Langerman et al. (2025) enumerate the pitfalls of automated metrics for structured 3D reconstruction through systematic unit tests, compare metric rankings against expert modellers' preferences and distil a learned metric from those judgments. The difference is the standard a metric is held to. There the standard is human preference, which is surveyed; here it is a geometrically valid wrong answer, which a metric can be shown to fail and, on a target whose relief falls below the models' error, must fail. Wu et al. (2025a) reach the neighbouring conclusion from perturbation, finding depth metrics severely under-sensitive to curvature, and Coupry et al. (2025) reach it from reconstruction practice, concluding that ground truth and metric should be chosen in view of the target's frequencies. Neither substitutes a different scene, which is

the perturbation that breaks the link between a score and the identity of what was reconstructed, and neither gives the threshold at which the failure becomes certain.

Criteria that need no ground truth. Every criterion in the search of Section 5 and Appendix E is hand-designed and referenced to ground truth, which sits awkwardly beside this paper's own recommendation that a designer run the control on their own data without additional ground truth. A reference-free family exists and is active: cross-reference metrics score a novel view against other captures of the same scene rather than against a reference (Hermann et al., 2025). One member of that family is measured here. Appendix F reports a criterion that uses no ground truth, no renderer and no intrinsics, scoring a prediction by whether the pixels it claims see one surface point agree in brightness. On the terrain where Chamfer fails worst, under the dense warp that is its loosest rejection radius, it rejects the wrong scene on all 11 LuSNAR cells where Chamfer rejects 1, and rejects both substitutes on 10 of those 11 where Chamfer rejects none, the flattened one only in the sense below. It is still not a replacement: on the same 11 cells it ranks the best model above the target's own ground truth on 10, cell for cell the same ones, so what it rewards is a point map's self-consistency with the images rather than its fidelity to the scene, and the flattened target, which it scores exactly as that ground truth, loses to the best model on those cells for the same reason. It is blind to the plane by construction, since flattening a height field moves every point along one normal, which is injective, leaving the correspondence it scores unchanged. The blindness just described is a property of matching in point space, and an optical reprojection moves a flattened point by roughly $fB\,\Delta h/Z^2$ across a baseline B, so its successor built on camera matrices was run, beside learned metrics fit to human judgment (Langerman et al., 2025) on renders of the geometry (Appendix F). Neither passes the control on low-relief ground, and the first shows why: there the images prefer the flattened target to the true geometry on four of ten AMEDS and Erfoud cells.

Adjacent benchmarks that are not competitors. Three further recent results sit close enough that a reader may name them. Wang et al. (2026d) benchmark four of this pool's models over 51 sites across satellite, aerial and ground altitudes and find that adding satellite imagery often degrades them; the axis is altitude and viewpoint, and no claim is made about whether the scoring metric discriminates. Michalkiewicz et al. (2026) make the same rhetorical move as this paper on a different object, showing that models strong on standard benchmarks struggle to reconstruct even Platonic solids; their subject is what the models understand, and the subject here is what the measurement can establish, so the two are complementary rather than overlapping. Du et al. (2025) replace retrospective geometric accuracy with precision relevant to completing a navigation task, under which passive stereo SLAM performs at a precision comparable to LiDAR SLAM's in typical indoor environments, which is the SLAM-side counterpart of scoring surface metrics against navigation outcomes in Section 5; it is evidence that task-driven re-evaluation of geometry benchmarks is an active parallel line rather than a device invented here.

Which models the pool contains, and why. The fourteen of Table B.1 were fixed on 2 September 2026, the date the last of them was scored. The criterion was a general-purpose model that predicts scene geometry from unposed images, with released weights, not specialised to a single domain and not an efficiency or quantisation variant of a model already present. That criterion admits more models than the pool contains, so the pool is a sample of the family rather than an enumeration of it. The nearest candidates it does not contain are OmniVGGT, AMB3R and Dens3R, all available before the pool closed and absent because the pool is a sample, not because they were judged and set aside; OmniVGGT's released weights could not be confirmed from a primary source. The bound is a statement about the target's geometry and the error a model would have to reach, so it holds whatever is ranked against it.

Method sources for the two secondary tracks. The corruption axis follows common-corruption methodology (Hendrycks & Dietterich, 2019), with corruptions chosen for these deployment settings rather than adopted unchanged. Neural radiance fields (Mildenhall et al., 2020) and Gaussian splatting (Kerbl et al., 2023), run through an open implementation (Tancik et al., 2023), form a second track whose comparability to feed-forward inference is limited by the supply of ground-truth poses and, in the dense fit, by view count (Appendix G).

L THE CHECKLIST WITH ITS EVIDENCE

Section 7 states the checklist; Table L.1 gives each step with the result it rests on and where that result is established. The first two steps read only the ground truth, needing no model, no donor and no inference, so the score every model must beat is known before any model is run.

Table L.1: The nine steps of Section 7, each with the result it rests on.

Step	What it rests on	Where
1. Compute $\mathrm{dev}(h_t)$ for every target.	The plane control turns on it; one-metre relief is neither necessary nor sufficient, the plane winning on a ConstructionSite window with 0.4724 m of relief and losing on a NordicHarbor window with 0.0015 m.	Appendix F
2. Run the plane control on every metric reported, per-view depth through the plane ray-cast into the target's cameras.	Equation 2 decides it outside its band, and the model won all six windows inside it; the ray-cast plane beats every model's depth on 4 of Erfoud's 5 windows and 1 of LuSNAR's 11.	Section 4; Appendices F and E
3. Run the wrong-scene control on the target's poses and on each model's.	On the target's poses it beats every model on 77% of the low-relief cells; on a model's own it wins 214 of 420 pairs and loses the verdict on AMEDS.	Section 4; Appendix D
4. Where a control wins, give every submission the same similarity fit.	After the fit the wrong scene wins no AMEDS or Erfoud cell, against 7 of 10 each under the frozen protocol, and 4 of LuSNAR's 11; the plane keeps 7 of LuSNAR's windows, one of AMEDS's and three of Erfoud's.	Appendix D, Table D.3
5. Where a control beats the best model, withdraw the surface metric on those cells and rank by the criterion of step 7; elsewhere print the control's score as a floor.	A control that beats the best model beats every model, so the metric orders nothing on that cell; a table marks each withdrawn cell and says how many it withdrew.	Section 4
6. Express surface error in units of $\mathrm{dev}(h_t)$.	In those units the plane's verdict separates all 123 windows at one, which relief in metres does not.	Appendix F
7. Report relief correlation beside completeness, or rank by their rank composite.	The composite rejects both substitutes on 28 of the 31 low-relief cells, where Chamfer rejects none, at parity with Chamfer on navigation coverage.	Section 5; Appendix E
8. Score the primary result under one frozen protocol, and report other settings as sensitivity.	Scoring choices alone push rank agreement below 0.90 in 150 of 266 comparisons on LuSNAR and 148 of 268 on TartanGround; a level is counted only once it is shown to move the metric.	Section 6.2; Appendices H and I
9. Compute the smallest p-value the design admits.	Holm's floor $\min(1, m \cdot 2/2^n)$ forbids separating any pair in tables of the size the field publishes; a comparison planned in advance has the floor $2/2^n$, which seven of the sixteen published rows clear.	Section 6.1; Appendix G

M ANALYSES IN FULL

This appendix holds the full versions of analyses that the earlier appendices summarise in a sentence or a table row: COLMAP and the control under other aggregations; the cell-level metric study and the search over pairings; the per-scene baselines, cost and effect sizes; and the protocol axes that Appendix H states only by their outcome.

M.1 COLMAP AND OTHER AGGREGATIONS

A classical pipeline loses to the plane as well. COLMAP (Schönberger & Frahm, 2016) was run to ask whether the collapse belongs to Euclidean surface metrics on near-planar geometry or to the priors of feed-forward networks. It is scored through the unmodified evaluator on the benchmark's own eight-frame windows, with exhaustive matching, SIFT at four times the default feature budget and a lowered peak threshold, affine shape and, after Dong & Soatto (2015), domain-size pooling on, and dense PatchMatch stereo (Schönberger et al., 2016); it is given the dataset's own intrinsics,

which no network receives, and placed on the same calibrated-scale track by a scale frozen on each domain's validation scenes. Every departure from the defaults is in its favour.

It reconstructs well where there is texture. On Erfoud it registers all eight images on all five windows and fuses a median of 1.8 million points, and its Chamfer of 0.26 m beats the best of the fourteen on 2 of the 5 cells. On LuSNAR it registers all eight on all twenty windows. On AMEDS's basalt it does not: one window registers none of its eight images at all, which is itself a result about classical matching on that ground.

The plane beats it anyway. Of the 30 low-relief windows attempted, 20 LuSNAR, 5 AMEDS and 5 Erfoud, 29 produce a surface, and the flattened ground truth beats COLMAP on 28 of those 29 under the most generous alignment the evaluator offers, the lowest Chamfer any of its four alignments gives COLMAP on the cell, at a median margin of 28.0; on the 15 that also clear the protocol's coverage guard it is 15 of 15. A different place beats it on 18 of the 20 of those surfaces whose window carries a hand-picked donor, and on 10 of the 11 on which that different place was also replanted on COLMAP's own estimated poses. On Erfoud, where it reconstructs well and beats at least 11 of the 14 networks on every window, the plane still beats it on 5 of 5.

Two qualifications belong beside that. Learned priors are not exonerated as *reconstruction*: classical multi-view stereo beats the best network on 2 of Erfoud's 5 windows. And COLMAP exposes a limit of the calibrated-scale track rather than passing it: a frozen per-domain scale is coherent for a network that regresses depth in a learned unit and incoherent for structure from motion, whose gauge is fixed independently in every reconstruction, so its fitted scale varies by 92 times across LuSNAR's cells and 14 of the 29 surfaces fall below the evaluator's coverage guard. The verdicts above are therefore read under the most generous alignment, where all 29 surfaces are present.

The high-relief cells are where the separation is only partial. On eight wide-baseline frames COLMAP's error is large there too: of the 16 ETH3D and TartanGround windows it was run on with known intrinsics, the 4 ETH3D windows whose images share one calibration and 12 TartanGround windows, 6 of them in environments the study never ranks models on, it produces 13 surfaces, and the plane beats it on 7 of them, all 3 on ETH3D and 4 of 10 on TartanGround, at a median margin of 1.40 against 28.0 on low relief. One inequality decides both: the plane wins wherever COLMAP's error exceeds the target's deviation from its plane, loses wherever it is below half of it, and wins the two cells between on its own accuracy. A pipeline with no learned prior therefore fails the plane control on low relief, so the failure does not need one; at this view count it also fails on about half of the high-relief cells, so what low relief does to a sparse classical reconstruction is multiply the margin twentyfold rather than decide the verdict. Five further ETH3D cells do not share one calibration across their frames; under self-calibration four register no image and one registers five of eight, with no frozen self-calibrated scale to score it by.

Structural metrics and tolerance ladders. Every lunar and ETH3D control cell of Table 2, and every model on the same cells, was re-scored from stored predictions under four further families: symmetric point-to-plane distance, aggregated by mean and by median, which removes the in-plane component that sampling density contributes; Chamfer after both clouds are quantised to one voxel grid at the 0.10 m tolerance, so density cannot favour either side; and normal consistency, the mean absolute cosine between each point's normal and its nearest neighbour's in the other cloud. The F-score was also re-read at eight absolute tolerances from 0.01 to 2 m and at seven relative ones from 0.05% to 5% of the ground truth's bounding-box diagonal. The re-scoring reproduces the frozen Chamfer of all 272 rows it can be matched to, to a relative difference below 10^{-12}, and a control cell counts only when every model scored on it was re-scored as well. TartanGround, whose packed frames need a staging step, was re-scored separately: 709 rows, 648 of them the 12 roster models on each of the 54 windows its 57 control cells sit on, and 61 control rows, the 57 and four NordicHarbor and Prison pairs at windows no model is scored on. The control rows reproduce the frozen Chamfer to 3×10^{-16} and its model rows to 1.2×10^{-8}, the larger figure being a float32 difference between a prediction as the sweep computed it and the same prediction read back from its stored file, identical on the CPU and on the GPU and about 1600 times smaller than the closest TartanGround margin to a tie. On the lunar cells the wrong scene wins under every distance family and under every absolute tolerance up to 0.5 m, and acceptance falls to about half only at 1 and 2 m, tolerances far above the lunar relief of 0.0151 m, so no tolerance restores discrimination either. Normal consistency is the exception: it accepts the wrong scene on 3 of the 11 lunar cells. TartanGround answers as ETH3D

does. Every distance family rejects the wrong scene on at least 53 of its 57 cells, at median margins of 0.23 to 0.31 beside ETH3D's 0.39 to 0.48 and nowhere near lunar terrain's 1.12 to 7.25, and acceptance falls as the tolerance widens, monotonically on the absolute ladder and on all but one step of the relative one, instead of persisting as it does on the Moon. The tail that remains, 10 of 57 cells at the tightest 0.01 m tolerance, is proportionally no larger than ETH3D's 2 of 8, so the failure is not in how distances are aggregated on the largest domain either.

Table M.1: The control under metric families that aggregate differently. Control wins counts the cells on which the wrong scene scores better than every model on that row's metric, so each metric is allowed its own best model; the margin is the best model's value divided by the control's (the control's divided by the best model's for normal consistency, where higher is better), so values above 1 mean the control won.

Metric	LuSNAR, 11 cells		ETH3D, 8 cells		TartanGround, 57 cells	
	Wins	Margin	Wins	Margin	Wins	Margin
Chamfer (frozen)	10 / 11	3.01	0 / 8	0.43	1 / 57	0.31
Point-to-plane, mean	10 / 11	3.24	0 / 8	0.48	1 / 57	0.28
Point-to-plane, median	10 / 11	7.25	0 / 8	0.45	4 / 57	0.25
Density-equalised	7 / 11	1.12	0 / 8	0.39	1 / 57	0.23
Normal consistency	3 / 11	0.97	0 / 8	0.63	3 / 57	0.85

Table M.2: The control under dense F-score tolerance ladders: cells on which the wrong scene's F-score beats every model's. Relative tolerances are a percentage of the ground truth's bounding-box diagonal.

Tolerance	Value	LuSNAR	ETH3D	TartanGround
Absolute	0.01 m	10 / 11	2 / 8	10 / 57
	0.02 m	10 / 11	1 / 8	9 / 57
	0.05 m	10 / 11	1 / 8	7 / 57
	0.1 m	10 / 11	0 / 8	6 / 57
	0.2 m	9 / 11	0 / 8	5 / 57
	0.5 m	9 / 11	0 / 8	5 / 57
	1 m	6 / 11	0 / 8	5 / 57
	2 m	5 / 11	0 / 8	4 / 57
Relative	0.05%	10 / 11	1 / 8	9 / 57
	0.1%	10 / 11	1 / 8	8 / 57
	0.2%	10 / 11	1 / 8	6 / 57
	0.5%	9 / 11	0 / 8	6 / 57
	1%	9 / 11	0 / 8	5 / 57
	2%	7 / 11	0 / 8	6 / 57
	5%	6 / 11	0 / 8	1 / 57

M.2 The metric study and the search over pairings

Chamfer's own agreement. Table M.3's agreement block is computed on the control cells only, over the models this study scores relief correlation on, so Chamfer's own values there, $+0.82$ on AMEDS, $+0.40$ on ETH3D and $+0.75$ on TartanGround, differ from Table 3's $+0.84$, $+0.51$ and $+0.80$ over every scored window and model, while LuSNAR's and Erfoud's are $+0.20$ and $+0.52$ in both: the difference is the pool, not the measurement.

The pairing the leaderboard points to, and a search to check it. On the two axes of rejecting both substitutes and agreeing with navigation within each cell, each criterion counted against its own best model, slope correlation is the strongest discriminator, best or tied on five of the six domains, all but Erfoud; relief correlation is best or tied on four, Erfoud among them; and Chamfer is the incumbent ranker, so pairings of the three are the ones to try first.

The whole space was then scored as a check rather than as the source of the claim: thirteen base criteria, Chamfer, relief correlation and eleven terrain metrics (slope correlation, Pearson and Spear-

Table M.3: How often the zero-shot model best on Chamfer beats the wrong scene under each candidate metric, on the 96 control cells of Table 2 with same-domain and unseen donors; electro's ten are reported in Appendix J. Table 3 judges relief correlation against the model best on relief correlation instead, as it judges every metric, which raises AMEDS and Erfoud to 10 of 10 and TartanGround to 57 of 57. For correlations and skills higher is better. The last block is each metric's agreement with navigation coverage, the mean within-cell Spearman correlation over every zero-shot model on the same cells, AMEDS's and Erfoud's averaged over their five windows.

Metric	LuSNAR	AMEDS	Erfoud	ETH3D	TartanGround
Chamfer	1 / 11	3 / 10	3 / 10	8 / 8	56 / 57
Relief correlation, 0.2 m raster	9 / 11	8 / 10	8 / 10	6 / 8	56 / 57
Relief correlation, 0.1 m raster	9 / 11	8 / 10	8 / 10	6 / 8	55 / 57
Relief correlation, 0.5 m raster	9 / 11	8 / 10	8 / 10	6 / 8	56 / 57
Correlation, quadratic trend removed	9 / 11	6 / 10	6 / 10	6 / 8	55 / 57
Relief skill (MSE against own-plane residual)	5 / 11	7 / 10	7 / 10	3 / 8	55 / 57
High-pass skill ($\sigma = 1$ m)	5 / 11	7 / 10	8 / 10	5 / 8	45 / 57
Agreement with navigation, mean within-cell Spearman ρ					
Relief correlation	+0.06	+0.51	+0.41	+0.24	+0.41
Chamfer	+0.20	+0.82	+0.52	+0.40	+0.75

man; slope error, mean and RMS; and roughness, normal, height, relief, step, profile and aspect errors), on the 106 audit cells. Each error is the mean absolute difference, or for slope also the root mean square, between the two clouds' fields on the evaluator's 0.10 m navigation grid, over the grid cells both fill: a grid cell's height; its relief, the spread between the 10th and 90th percentiles of its points' heights; its step, the largest height difference to a neighbouring grid cell; slope, normal and roughness from a plane fitted to its 3×3 neighbourhood, roughness being the root mean square residual about that plane and normal error an angle; aspect, the downhill direction; and profile, the height along the camera path. All 91 combinations of one or two are scored, 13 single criteria and 78 pairs. All thirteen bases are referenced to a ground-truth cloud, and the one reference-free criterion measured is treated in Appendix F. The composite is the mean of the two within-cell ranks, each in its registered direction, which needs no normaliser and so cannot be tuned. Relief correlation is scored on 56 of TartanGround's 57 cells here: one model has no calibrated alignment on the last, and the search needs every criterion for every model of a cell. Table 3 keeps that cell, judged against the models that carry the score, and counts 57 of 57 for relief correlation and 56 of 57 for its composite with completeness.

A combination passes when, on every domain, it rejects both substitutes on at least as many cells as Chamfer and its four-outcome agreement is no more than 0.10 below Chamfer's. Six pass (Table M.5): Chamfer with Pearson slope correlation; Chamfer with Spearman slope correlation; relief correlation with Chamfer; relief correlation with Pearson slope correlation; relief correlation with slope error; and relief correlation with normal error. Chamfer with Pearson slope correlation is never worse than Chamfer at rejecting both substitutes on any domain and is better on four, taking LuSNAR from 0 of 11 to 6 of 11, AMEDS from 0 of 10 to 8 of 10, Erfoud from 0 to 4 of 10 and TartanGround from 49 to 55 of 57, while its ranking agreement, the mean over the four navigation outcomes, is higher on four domains and lower by 0.016 on AMEDS and 0.008 on TartanGround. Relief correlation with Chamfer rejects more on the low-relief domains, 8 of 11 on LuSNAR, 10 of 10 on AMEDS and 9 of 10 on Erfoud, and its ranking agreement is higher on four domains and lower by 0.043 on LuSNAR and 0.002 on TartanGround. Slope correlation with relief correlation, one of the three passing pairings without Chamfer, rejects both substitutes on 9 of 11 LuSNAR cells and on every other cell but one of Erfoud's. Under a cell bootstrap within each domain the whole verdict survives 98.1% of 2000 resamples for Chamfer with Pearson slope correlation and 90.0% for relief correlation with Chamfer, whose rejection half survives every resample on every domain and whose ranking half fails only on LuSNAR; it survives 73.5% for the Spearman variant, 70.8% for slope with relief correlation and under 44% for the pairings of relief correlation with slope or normal error, so the six are not interchangeable.

Table M.4: Named rows of the metric grid, each judged against the model best on it. Both rejected counts the grid's audit cells on which the wrong scene and the flattened target each score worse than that model; beyond completeness is the four-outcome agreement partialled on completeness, over every zero-shot model on the grid's windows. The grid is scored on its own 71 audit cells on 52 windows (Appendix E), not on the 106 of Table M.5, which is why TartanGround counts 22 here; electro, a single window reported beside ETH3D, carries no agreement column.

	Both substitutes rejected						Beyond completeness, ρ				
Candidate	LuSNAR	AMEDS	Erfoud	ETH3D	electro	TartanGround	LuSNAR	AMEDS	Erfoud	ETH3D	TartanGround
Chamfer, recomputed	0 / 11	0 / 10	0 / 10	8 / 8	10 / 10	20 / 22	-0.19	-0.20	-0.11	+0.19	+0.11
Completeness, recomputed	2 / 11	2 / 10	0 / 10	7 / 8	10 / 10	21 / 22	n/a	n/a	n/a	n/a	n/a
Relief correlation, 0.2 m (Section 5)	11 / 11	10 / 10	9 / 10	6 / 8	10 / 10	21 / 22	+0.00	+0.41	+0.24	+0.35	+0.26
Relief correlation, 0.1 m	10 / 11	10 / 10	10 / 10	7 / 8	10 / 10	22 / 22	+0.00	+0.38	+0.28	+0.38	+0.31
Relief correlation, 0.05 m	10 / 11	10 / 10	10 / 10	7 / 8	10 / 10	22 / 22	+0.02	+0.40	+0.29	+0.35	+0.32
Slope correlation, 0.2 m	9 / 11	9 / 10	8 / 10	8 / 8	10 / 10	21 / 22	+0.22	+0.18	+0.31	+0.18	+0.18
Relief correlation with completeness	7 / 11	10 / 10	8 / 10	8 / 8	10 / 10	21 / 22	0.00	+0.36	+0.21	+0.29	+0.22
Slope correlation with completeness	4 / 11	6 / 10	6 / 10	8 / 8	10 / 10	21 / 22	+0.23	+0.18	+0.23	+0.16	+0.12

Table M.5: The search for a criterion that could replace Chamfer. Rejection counts the control cells on which the criterion's own best model beats both the wrong scene and the flattened plane, and ρ is the mean within-cell agreement with the four navigation outcomes. Every row uses the same definitions, so the columns are comparable with each other but not with Table M.3, which holds the compared model fixed instead.

Criterion	Rejects both substitutes						ρ with navigation					
	AMEDS	electro	Erfoud	ETH3D	LuSNAR	TartanGround	AMEDS	electro	Erfoud	ETH3D	LuSNAR	TartanGround
Chamfer	0 / 10	10 / 10	0 / 10	8 / 8	0 / 11	49 / 57	+0.64	+0.64	+0.51	+0.41	+0.28	+0.63
Slope correlation	10 / 10	10 / 10	5 / 10	8 / 8	10 / 11	57 / 57	+0.43	+0.64	+0.43	+0.44	+0.47	+0.54
Relief correlation	10 / 10	10 / 10	10 / 10	6 / 8	9 / 11	56 / 56	+0.56	+0.56	+0.41	+0.47	+0.12	+0.48
Slope corr. with Chamfer	8 / 10	10 / 10	4 / 10	8 / 8	6 / 11	55 / 57	+0.62	+0.74	+0.57	+0.50	+0.46	+0.62
Relief corr. with Chamfer	10 / 10	10 / 10	9 / 10	8 / 8	8 / 11	55 / 56	+0.71	+0.65	+0.56	+0.49	+0.24	+0.63
Slope with relief corr.	10 / 10	10 / 10	9 / 10	8 / 8	9 / 11	56 / 56	+0.60	+0.69	+0.50	+0.55	+0.35	+0.58

M.3 Per-scene baselines, cost and effect sizes

Track B. The per-scene optimisation baselines are reported separately because they optimise the test scene and are given its ground-truth camera poses, so the alignment is an identity for them by construction. Each is fitted on every eighth frame of its scene, 137 to 169 frames on LuSNAR and 360 to 398 on TartanGround, and scored on the ground truth of an eight-frame window, as a feed-forward model is (Table M.6). On the windows that also carry feed-forward rows, all eight on LuSNAR and four of eight on TartanGround, they are 4.2 to 9.6 times worse than the best feed-forward model on TartanGround and 7.0 to 60 times on LuSNAR. Fitted instead on the window's own eight frames, Gaussian splatting's error falls to 0.67 to 3.5 times the best feed-forward model's on TartanGround and 0.94 to 6.6 times on LuSNAR, below it on one window of each, and the neural field's to 1.2 to 8.0 and 1.5 to 19 times, so the dense fit is not the configuration most favourable to them: the comparison is one of these methods on this benchmark's windows, not of the method class.

Table M.6: Track B, Nerfstudio's two per-scene optimisation baselines, per domain and fit. Every cell is scored on the ground truth of an eight-frame window, as a feed-forward model is; the dense fit uses every eighth frame of the scene, and the sparse fit the window's own eight frames. Matched cells are those whose window also carries feed-forward rows, and the last column is the baseline's Chamfer over the best feed-forward model's on them.

Domain	Method	Fitted on	Cells	Median Chamfer (m)	Matched cells	Chamfer / best feed-forward
LuSNAR	Splatfacto	every eighth frame, 137 to 169	8	3.33	8	10.77 to 57.17
LuSNAR	Nerfacto	every eighth frame, 137 to 169	8	3.29	8	7.03 to 60.06
LuSNAR	Splatfacto	the window's 8 frames	8	0.40	8	0.94 to 6.61
LuSNAR	Nerfacto	the window's 8 frames	8	1.40	8	1.49 to 18.63
TartanGround	Splatfacto	every eighth frame, 360 to 398	8	3.11	4	4.21 to 7.73
TartanGround	Nerfacto	every eighth frame, 360 to 398	8	3.02	4	5.01 to 9.63
TartanGround	Splatfacto	the window's 8 frames	8	0.56	4	0.67 to 3.50
TartanGround	Nerfacto	the window's 8 frames	8	1.54	4	1.20 to 7.97

Cost as a tie-break. Where accuracy does not separate two models, cost spreads them further: between-model spread on TartanGround is 11.5× on runtime against 3.9× on Chamfer, and 8 of the

12 models are dominated on both axes at once (Table M.7). The same resolution standard applies to cost as to accuracy, and it answers differently. No TartanGround scene carries a cost record for all twelve models, so the twelve-model design admits no pairwise test at all, and an empty block is not a null result. Restricted to the widest complete block at each roster size, the cost axes do separate pairs: at five models on twelve shared scenes, runtime separates 7 of the 10 pairs under the same Holm correction the accuracy designs use. Peak VRAM separates all 10, but only because it is constant within a model across scenes (median scene CV 0.000), so every paired difference of a pair is the same number, and every pair returns the same p, 0.0006 under the normal approximation such ties force and 0.006 after Holm's correction, whatever the size of the difference; its differences are point values, not tested separations, and runtime is the one cost axis the test measures. At six models and above the floor $m \times 2/2^n$ exceeds α and nothing can separate whatever the numbers are. What decides the cost axes is therefore the block structure rather than the axis, which is the same finding this paper reports for accuracy.

Table M.7: Cost and accuracy on TartanGround, restricted to a single GPU type so that hardware cannot confound runtime. Scene CV is the median across models of the coefficient of variation of that model's per-scene means. Best model is the lowest mean over the scenes each model was scored on rather than over a common block, so it need not match a per-scene ranking such as Table M.8's. Peak VRAM is constant within a model, so its separations are point values rather than tested ones (text).

Axis	Between-model spread	Median scene CV	Best model
Runtime (s)	11.5×	0.145	CUT3R
Peak VRAM (GB)	5.7×	0.000	MUSt3R
Chamfer	3.9×	0.869	MASt3R

Table M.8: Effect size of the top pair. Rank 1 and rank 2 are the first two models by mean rank over the scenes on which every model was scored. Per scene, rank 2's value is divided by rank 1's (inverted for coverage, where higher is better), so a ratio above 1 favours rank 1; the table gives the median over scenes with a 95% percentile interval from 10,000 scene resamples. Mean rank is taken over all models, so rank 1 need not beat rank 2 head to head, as on TartanGround's coverage.

Domain	Metric	Rank 1	Rank 2	Scenes	Median ratio [95% interval]	Rank 1 better
LuSNAR (lunar)	Chamfer	MASt3R	Spann3R	5	1.54 [1.34, 2.26]	5 / 5
ETH3D (real, laser)	Chamfer	MapAnything	CUT3R	11	1.26 [0.97, 2.28]	8 / 11
TartanGround (simulation)	Chamfer	MapAnything	MASt3R	20	1.31 [0.91, 1.56]	13 / 20
StereoLunar (lunar descent)	Chamfer, frozen scale	DVLT	StreamVGGT	14	4.34 [3.08, 6.71]	13 / 14
StereoLunar (lunar descent)	Chamfer, per-scene Sim(3)	DVLT	Pi3	14	1.76 [1.31, 2.61]	13 / 14
TartanGround (simulation)	Coverage	MASt3R	MapAnything	20	0.93 [0.84, 1.03]	8 / 20
ETH3D (real, laser)	Depth AbsRel	DVLT	VGGT	11	1.06 [0.83, 1.85]	6 / 11
TartanGround (simulation)	Depth AbsRel	Pi3	VGGT	20	1.19 [1.05, 1.39]	15 / 20

Resolution in detail: effect sizes. The intervals in Table M.8 are per comparison and uncorrected, and they describe how large a difference is rather than test it. Four of them exclude 1: LuSNAR's Chamfer, over five scenes whose resamples cannot leave the range of five values, StereoLunar's under both scale conventions, which the corrected test also separates, and TartanGround's depth, where rank 1 is better on 15 of 20 scenes but the Holm-corrected Wilcoxon test over all 36 pairs of that metric leaves the pair at $p = 0.10$. A point estimate favouring the first-ranked model is therefore common; a difference the design can certify once every pair is tested is not.

M.4 Protocol axes

Seeding and separability. Corruption is seeded from a hash of the seed, corruption, severity, dataset, scene and frame identifier, with no model term, so every model sees identical pixels and a comparison across models is not a comparison of noise draws. Every corrupted row records the corruption, severity, seed, realised parameters and a digest of the corrupted pixels, and the protocol filter drops such rows from the primary analyses by reading that record rather than a string. A corrupted row is the most dangerous kind of ablation to pool, because unlike a control it is an ordinary measurement of a real model that is only worse.

Intrinsics. There are six perturbation levels, five of the focal length and one of the principal point, and four of the focal ones are drawn from measured calibration error magnitudes rather than round numbers: repeat-calibration scatter, thermal drift, and mechanical cycling in both directions; the fifth is a catalogue focal length in place of a calibration. The axis returned a result about the interface, since these models accept unposed and uncalibrated images and estimate their own calibration, and Depth Anything 3, which appears to accept intrinsics, discards them unless extrinsics are supplied, which would mean supplying ground-truth poses. That reading of its adapter code is confirmed by measurement: run at each of the five focal levels on four scenes of LuSNAR and TartanGround, its Chamfer is bit-identical to the exact-calibration row in 20 of 20 comparisons.

Resolution. Levels are absolute long-side pixel counts chosen against each adapter's fixed internal size, and downsampling only. Two combinations of model and level are no-ops, arising where an adapter's internal resize reproduces the requested one exactly: MASt3R at 512 px, whose Chamfer is bit-identical to its own axis control on all four cells, and Spann3R at 224 px, bit-identical on three of the four and agreeing to five significant figures on the fourth. Both are excluded rather than counted as robustness. Table M.9 gives every model's Chamfer at each level over its axis control, the two no-ops printing 1.00: at 112 px the five models' order on LuSNAR no longer agrees with the control's ($\rho = -0.05$) and MASt3R's error grows 5.52-fold, while from 336 px on both domains agree at $\rho \geq 0.90$.

Table M.9: The input-resolution axis. Each entry is a model's Chamfer at a long-side size over its Chamfer at the axis's own control, averaged over the axis's two scenes on each domain; the last row is the Spearman agreement of the five models' order with the control's, averaged over the scenes. MASt3R at 512 px and Spann3R at 224 px are the no-op combinations of the text, which print 1.00 and are not evidence of robustness.

	LuSNAR				TartanGround			
Model	112 px	224 px	336 px	512 px	112 px	224 px	336 px	512 px
MapAnything	0.84	0.93	0.92	0.99	1.62	1.29	1.00	0.97
MASt3R	5.52	1.72	0.96	1.00	1.69	1.72	1.11	1.00
Pi3	1.14	0.74	0.65	0.77	0.84	0.90	0.95	0.98
Spann3R	2.72	1.00	1.00	1.00	1.16	1.00	1.00	1.00
VGGT	1.94	0.95	0.82	1.06	1.63	1.00	1.15	1.01
ρ with the control	-0.05	$+0.75$	$+0.90$	$+0.95$	$+0.65$	$+0.70$	$+0.95$	$+1.00$

Adaptation. Four fine-tuning schedules were applied to three models under a fixed budget, MapAnything, Pi3 and CUT3R. The other eleven were not adapted, and two of them could not be: VGGT-Ω releases no training code, and DVLT's released trainer depends on a data pipeline its authors have not published. The recipes are each model's own published recipe (R0), MASt3R's schedule transplanted unchanged (R1), and that warmup at one tenth of the learning rate and at the geometric midpoint between the two (R2, R3), with a heads-only recipe (R5) for CUT3R (Table M.11); R4, a low-rank adaptation declared in the recipe file, was not implemented, so the numbering skips it. Models are fine-tuned on the six training scenes of the LuMon split (Sekmen et al., 2026) and scored on its two held-out test scenes, Moon_1 and Moon_8, at four windows each. Against each model's own zero-shot row, MapAnything improves under every recipe, to 0.484 of its zero-shot Chamfer under its own and consistently across all 8 windows; CUT3R improves only under its own recipe, to 0.367, and worsens under all four of the others by up to 2.775×; Pi3 improves in the median under all four recipes, but on every window under only three of them. Each model's own recipe gives its lowest median ratio, and MASt3R's schedule and its two variants help MapAnything and Pi3 but hurt CUT3R, so the outcome depends on the pairing of model and recipe; at three seeds these orderings are not resolved (Appendix G).

Does the scale protocol treat the method families alike? Thirteen of the fourteen models are rescaled by a frozen per-model factor fitted on a validation scene; MapAnything claims a metric scale and is therefore rescaled by nothing, so a wrong claim would penalise the one model that made it. The scale track answers this directly, because its ORACLE level fits a per-scene similarity, which is the seven-degree-of-freedom alignment the rigid protocol is sometimes asked to justify.

Table M.10: Corruption severities. Top: every corruption's three levels in its own parameter; low light is a short exposure with photon and read noise followed by the gain that restores brightness, and its parameter is the exposure fraction. Bottom: the four corruptions whose parameter converts to ImageNet-C's, beside ImageNet-C's five levels, transcribed from that benchmark's generation code. Brightness, motion blur and low light have no counterpart in a shared unit: brightness here is a linear gain that clips highlights, the lunar failure mode, rather than a shift in HSV value.

Corruption	Parameter	Levels 1 / 2 / 3
Brightness	gain	2 / 3.5 / 6
Defocus	radius, % width	0.25 / 0.5 / 1
Gaussian noise	σ, DN	6 / 14 / 28
JPEG	quality	30 / 15 / 7
Low light	exposure fraction	0.25 / 0.1 / 0.04
Motion blur	length, % width	0.8 / 1.6 / 3.2
Shot noise	full well, e^-	60 / 20 / 6

Corruption	This benchmark (1 / 2 / 3)	ImageNet-C (1 to 5)
Defocus	0.25 / 0.5 / 1	1.34 / 1.79 / 2.68 / 3.57 / 4.46
Gaussian noise	6 / 14 / 28	20.4 / 30.6 / 45.9 / 66.3 / 96.9
JPEG	30 / 15 / 7	25 / 18 / 15 / 10 / 7
Shot noise	60 / 20 / 6	60 / 25 / 12 / 5 / 3

Table M.11: Adaptation. Median fine-tuned Chamfer divided by the same model's zero-shot Chamfer on the identical scene and frame window, over eight cells per entry. Values below 1 mean fine-tuning helped. R4, a low-rank recipe, was not implemented and is omitted; R5 trains CUT3R's heads only and was not run on the other two models.

Model	R0 own recipe	R1 transplanted	R2 lr/10	R3 midpoint	R5 heads only
MapAnything	0.484	0.551	0.648	0.511	n/a
Pi3	0.132	0.771	0.401	0.303	n/a
CUT3R	0.367	2.775	1.020	2.011	2.218

Comparing the scale each model was actually given with the scale that fit would have chosen, as $|\log(\text{oracle}/\text{published})|$ read from the evaluator's own alignment record, separates the two cases.

The answer depends on the terrain, and it separates the two ways of obtaining a scale. On TartanGround MapAnything's gap is 0.023, the *smallest* of the twelve models scored there against a median of 0.476: its metric claim is right to within 2.3%, and giving it no rescaling is giving it the correct scale. On LuSNAR the same claim is off by a factor of 1.74, a gap of 0.552, the third largest of the eleven models the scale track scores there. The frozen calibration the other models receive moves the other way: its median gap is 0.218 on LuSNAR, a factor of 1.24, against 0.476 on TartanGround, a factor of 1.61, so a scale fitted once on a validation scene transfers *better* across LuSNAR's homogeneous regolith than across TartanGround's varied environments. The advantage therefore changes sign with the terrain, although two rescaled models are further from their oracle scale than MapAnything is even on LuSNAR, and neither family is favoured on both domains. It also decides the head of the table: scored after the per-scene similarity fit, as Table G.4 scores StereoLunar, MapAnything falls from first to third on TartanGround Chamfer, behind MUSt3R and MASt3R, and from first to fifth on ETH3D, where DVLT leads, while LuSNAR's first model, MASt3R, stays first; no top pair separates under the pairwise test of Section 6.1 on either track.