CONFIDENCE IS NOT COVERAGE: BUDGETED CONFIDENCE RANKING CAN FAIL UNDER FIXED RECONSTRUCTION BUDGETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Feed-forward multi-view reconstruction models produce dense 3D predictions with per-point confidence, but downstream representations typically retain only a fixed number of points or primitives. A simple strategy is therefore to rank predictions by confidence before sampling. We show that this seemingly natural use of confidence can be mismatched to fixed-budget reconstruction: because confidence scores points independently, ranking can remove all candidates from parts of the scene before a spatial sampler is ever applied. We isolate the effect of confidence ranking with a matched pool test that compares confidence-ranked and uniformly sampled candidate pools of identical size under the same sampler, budget, and initialization. We also derive a coverage ceiling that upper-bounds the reconstruction quality attainable by any sampler that selects from a given pool without moving its points, allowing irreversible support loss to be detected from reference geometry before sampling. Across multiple reconstruction models and benchmarks, confidence-ranked pools can provide substantially less scene support than equal-sized random pools. Mixing ranked and random candidates restores most of the lost performance. Our results show that confidence remains useful for rejecting unreliable predictions, but is not by itself a suitable allocation rule for a tight scene-level budget.

1 INTRODUCTION

Feed-forward multi-view models now produce dense 3D predictions in a single forward pass. The Visual Geometry Grounded Transformer (VGGT) (Wang et al., 2025a) takes a set of images and returns a 3D point and a confidence for each pixel of each view. Most of these predictions are thrown away at once, because downstream representations and applications impose a fixed primitive budget, which we call the output budget. EcoSplat, a feed-forward splatting model, keeps the Gaussians with the highest predicted opacity in each view to meet its budget (Park et al., 2026). The predictions that survive determine which parts of the scene the downstream representation can hold.

What should the budget be spent on? The assumed answer is confidence. The network emits one for each point, and it ranks pointwise error well (Deng et al., 2026). Ranking has precedent: Fast3R keeps only its 15% most confident points for pose estimation (Yang et al., 2025). A threshold on confidence removes unreliable points, and a fixed count can then be drawn uniformly from the points that pass. A budget can instead raise that threshold until only the budget's worth of points survive, and confidence then stops filtering points and starts allocating the budget. A per-point score judges each point alone and says nothing about where the other points lie.

We show that at tight budgets confidence ranking can leave whole parts of the scene with no candidate. On DTU, with a pool four times the budget, ranking keeps a fifth of the recall and gains no precision over ALL (Table 3). Figure 1 shows one scan (Jensen et al., 2014). A pool of the most confident points, the candidates a sampler may draw from, reaches a fraction of the reference surface, in dense stripes where the network is surest. A random pool of the same size reaches nearly all of it. Once a region has no candidate, no sampler that keeps the predicted coordinates can put a point there.

To isolate the loss, we vary only pool membership. Eligible predictions have finite coordinates and a confidence above a fixed threshold. From them we build two pools of the same size M: the top

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.

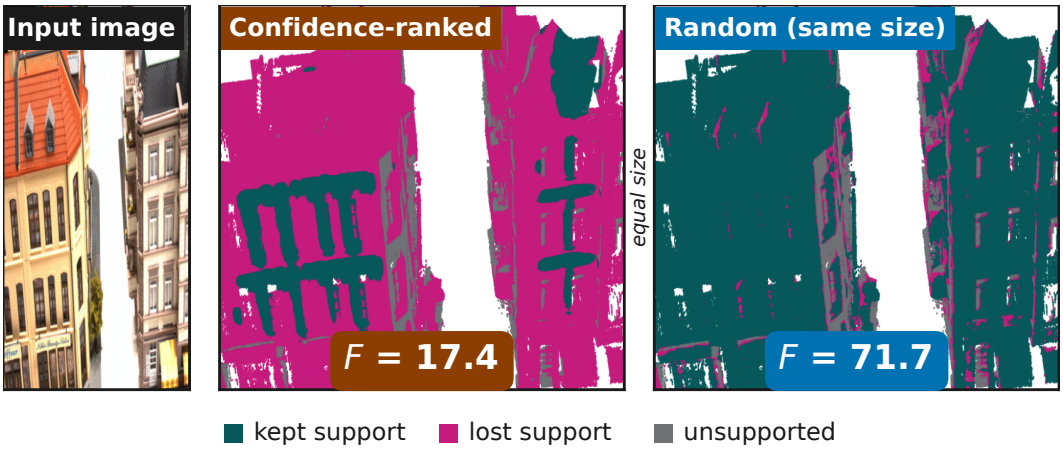


Figure 1: **A random pool keeps surface that confidence ranking discards.** DTU scan 21 (Jensen et al., 2014), view 25, the scan whose ceiling gap sits at the median of the 21. Each panel is a mask over the reference surface. Colors mark three cases: reference surface the pool still supports, surface lost relative to all eligible predictions, and surface no prediction reaches. F is this scan's F-score in percent after sampling. Both pools hold 96,428 candidates and keep 24,107 points under the same sampler and start. Table 3 gives the 21-scene means.

M by confidence, and M drawn uniformly at random. Each passes through the same farthest-point sampler (FPS) (Qi et al., 2017) to the same output budget and start. We call this the matched test. On 21 DTU scans with a pool four times the budget, which keeps 2.18% of eligible points on average, the ranked pool reaches a mean F-score of 23.52, the random pool 77.62, and the full eligible set 78.44 (Table 3). The gap shrinks as the pool widens, to 24.67 points at 25% retention (Table 1). We developed the protocol on DTU. The gap also appears on BlendedMVS (Yao et al., 2020), a disclosed VGGT training source, in sign on held-out ETH3D scenes (Schöps et al., 2017), and as lost support on the VGGT-free Depth Anything 3 (Section 4.4).

To tell a poor pool from a poor sampler, we say a part of the reference surface is supported by a pool if some candidate lies within tolerance of it. The supported fraction bounds the recall of any subset of the pool, and with it the subset's F-score. We call this F-score bound the coverage ceiling (Section 3.2). When the ranked pool's ceiling falls below the F-score already achieved from the full eligible set, no subset of the ranked pool can match that score.

Confidence still works as a filter. On raw network outputs, invalid predictions included, VGGT's default confidence filter beats both a random pool and the unfiltered set (Section 4.3). The loss comes one step later, when the same confidence ranks points that have passed a threshold.

Contributions.

(1) **Confidence ranking can discard scene support.** Among points that pass a confidence threshold, a ranked pool can reach much less of the scene than a random pool of the same size. A matched test isolates this by changing only pool membership (Section 3.1); drawing the pool in proportion to confidence, as Mini-Splatting and Taming-3DGS draw Gaussian primitives, recovers 97% of the ranked-to-random gap on DTU (Appendix E).

(2) **A coverage ceiling that bounds any selection from a pool**, computed from reference geometry before sampling. When it falls below an achieved same-budget score, no coordinate-preserving selection from the pool can match that score. For narrow ranked pools it sits near the achieved score, so the loss is in the pool (Table 3). On DTU it rules out recovery from the ranked pool in 84 of 84 VGGT conditions, 78 of 84 for π^3, and 80 of 84 for Depth Anything 3 (Table 5).

(3) **How far the effect reaches.** Random pools gain recall and give up some precision, so their lead is largest at the narrowest pools and closes as the pool widens (Section 4.2). The effect also holds on a backbone that does not initialize from VGGT weights (Section 4.4) and on an alignment task that does not use the F-score. There, random pools agree with the full-set alignment in 92.5% of disturbed DTU cases against 39.0% for ranked ones (Section 4.6). Where budgets are tight, a half-random pool recovers 99% of what ranking costs (Section 4.5).

2 Related Work

Feed-forward dense geometry. Classical multi-view stereo matches calibrated views and filters the result geometrically (Schönberger et al., 2016; Yao et al., 2018). Feed-forward predictors instead produce dense point observations. DUSt3R regresses pairwise pointmaps (Wang et al., 2024), MASt3R adds dense matching (Leroy et al., 2024), and Spann3R, Fast3R, CUT3R, and AMB3R extend prediction to many views (Wang & Agapito, 2025; Yang et al., 2025; Wang et al., 2025b; Wang & Agapito, 2026). MapAnything, MV-DUSt3R+, VGGT with its variant StreamVGGT, and Depth Anything 3 likewise produce geometry for flexible inputs (Keetha et al., 2026; Tang et al., 2025; Wang et al., 2025a; Zhuo et al., 2026; Lin et al., 2026), and VGGT-X adapts VGGT to dense view synthesis (Liu et al., 2025). Explicit-primitive renderers (Kerbl et al., 2023) run under fixed primitive budgets in practice, and the methods that meet those budgets differ in how they select. LightGaussian and PUP 3D-GS prune by a global score (Fan et al., 2024; Hanson et al., 2025). Mini-Splatting and Taming-3DGS instead draw primitives with probability proportional to a score (Fang & Wang, 2024; Mallick et al., 2024), and Fang & Wang (2024) give the reason: deterministic selection disrupts local geometry, since neighboring primitives carry similar scores. Those scores measure a primitive's contribution to the rendered image, while a feed-forward predictor's confidence estimates correctness.

Confidence and completeness. Prior work treats confidence as a threshold or as pointwise uncertainty. Hillemann et al. (2026) sweep VGGT confidence thresholds on DTU and measure accuracy–completeness trade-offs. Trust3R evaluates evidential pointmap uncertainty with pointwise risk–coverage measures (Zhu et al., 2026). Deng et al. (2026) audit the confidence of seven feed-forward backbones on thirteen datasets and find that it ranks error well while its magnitude is miscalibrated. Their count falls as the threshold rises; we fix the count and add an equal-size random pool. To our knowledge, no prior work compares confidence-ranked and random equal-size subsets of one eligible population while fixing coordinates, sampler, output budget, and start.

Confidence as a selection signal. Confidence can derive from predictive variance, ensembles, or evidential distributions (Kendall & Gal, 2017; Lakshminarayanan et al., 2017; Amini et al., 2020), and is judged by selective risk, sparsification, and calibration (Geifman & El-Yaniv, 2017; Ilg et al., 2018; Poggi et al., 2020; Guo et al., 2017). A score can meet those criteria and still remove the only candidates covering part of a scene.

Point selection. Prior point-selection work changes the sampler. We keep the sampler and change its input pool. Keypoint detectors meet a fixed count by raising a suppression radius until that count remains (Brown et al., 2005; Bailo et al., 2018) or by filling grid cells (Mur-Artal et al., 2015), so the score competes only within a neighborhood, and Poisson-disk sampling keeps a minimum spacing (Bridson, 2007). FPS spreads samples spatially (Qi et al., 2017), and learned samplers optimize a task objective (Dovrat et al., 2019; Lang et al., 2020), compress localization maps (Yang et al., 2022), weight points by semantic scores (Chen et al., 2022), or reweight key points for robustness (Li et al., 2025). RecGen3D samples uniformly and then runs FPS, using confidence only to weight an alignment (Huang et al., 2026b). Data selection sees a related cost in a trained model. Batches chosen only by uncertainty can be redundant (Sener & Savarese, 2018; Ash et al., 2020), score-ranked coresets fall below random ones at high pruning rates (Zheng et al., 2023), and the better end of the ranking depends on how much data remains (Sorscher et al., 2022). Edge-focused point sampling trades away global uniformity (Wu et al., 2025), and attention-ranked token pruning loses spatial layout (Huang et al., 2026a). For a model trained on the pruned set, Ayed & Hayou (2023) prove asymptotically that score-based pruning can fall below random pruning when little data is kept, and that mixing in random samples restores consistency. In 3D the cost is spatial, and the ceiling measures it per scene before sampling. The confidence that works as a threshold (Section 4.3) removes support under a tight budget.

3 Method

The matched test (Section 3.1) changes only how the candidate pool is chosen, and neither it nor the coverage ceiling (Section 3.2) uses anything specific to VGGT. The ceiling bounds what any sampler that keeps the pool's coordinates can reach. The mixed pool (Section 3.3) tests whether random members restore the score.

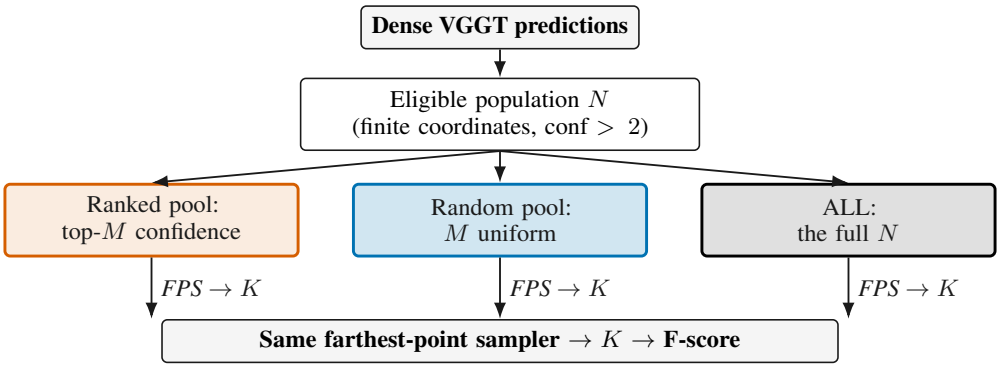


Figure 2: **The candidate pool limits what the sampler can choose.** Three variants start from the same eligible predictions: the ranked pool (top-M by confidence), the random pool (M drawn uniformly), and ALL (every eligible point). Each passes through the same farthest-point sampler to the same output budget K from the same start. Schematic only.

3.1 The Matched Test: Only Pool Membership Differs

Pools. The ranked and random variants differ only in which eligible points enter the candidate pool. A prediction is eligible when its coordinates and direct-head confidence are finite and its confidence exceeds 2.0, the threshold Hillemann et al. (2026) report as a good choice for filtering VGGT's raw output on DTU. We call the whole eligible set ALL. The pool width $w \in \{4, 8, 16, 32\}$ is the number of candidates per output point. With N eligible points and output budget K, each pool has $M = \min(wK, N)$ original points, selected by top confidence or uniformly at random. FPS then returns exactly K of them. Predictions at the same location count as separate candidates. The variants share coordinates, output budget, and sampler (Figure 2).

Reference geometry. On DTU we use 21 of 22 scans, each at its two fixed output budgets. We align the variants of a scan with one shared camera-derived similarity. We score precision against the scanner's observation mask and recall against scanner points above the table plane, both at a 5 mm tolerance. On BlendedMVS we use seven validation scenes with 32 calibrated views and budgets of 25,000 and 50,000 points. We take reference points from the rendered depths, and the tolerance is 1% of the median valid depth.

Sampling and statistics. Our main contrast is the matched $w = 4$ random-minus-ranked difference with two shared FPS starts per scene. Each 95% interval pairs the two pools within scene and resamples scenes (Table 21). Five pool seeds per scan test the narrow pool at each output budget (Table 2, Appendix H).

3.2 The Coverage Ceiling: When a Pool Has Already Lost Surface

A sampler can only choose among the candidates it is given. If a part of the reference surface has no candidate within tolerance in the pool, no subset of the pool reaches it. So the fraction of the reference the pool supports caps the recall of any subset, and with it the F-score.

We use calligraphic capitals for point sets. $\mathcal{C}$ is a pool; $\mathcal{S} \subseteq \mathcal{C}$ is a sampled subset. Let $r_{\mathcal{C}}$ be the fraction of reference observations within tolerance τ of a candidate in $\mathcal{C}$. Here $P(\mathcal{S})$, $R(\mathcal{S})$, and $F(\mathcal{S})$ denote precision, recall, and their harmonic mean. Any subset keeps the same coordinates and observation identities, so $R(\mathcal{S}) \le r_{\mathcal{C}}$. Because precision is at most one,

$$F(\mathcal{S}) = \frac{2P(\mathcal{S})R(\mathcal{S})}{P(\mathcal{S}) + R(\mathcal{S})} \le \frac{2R(\mathcal{S})}{1 + R(\mathcal{S})} \le \frac{2r_{\mathcal{C}}}{1 + r_{\mathcal{C}}} =: U(\mathcal{C}). \tag{1}$$

If a reconstruction A obtained at the same budget has $F(A) > U(\mathcal{C})$, no subset of $\mathcal{C}$, and therefore no sampler applied to $\mathcal{C}$, can match it. The bound holds at any precision and budget and need not be attainable, so it can only rule a pool out. Unlike the core-set bound (Sener & Savarese, 2018), which also scores coverage, the ceiling needs no smoothness assumption and is read before any sampler runs. FPS is the greedy k-center rule (Gonzalez, 1985; Eldar et al., 1997).

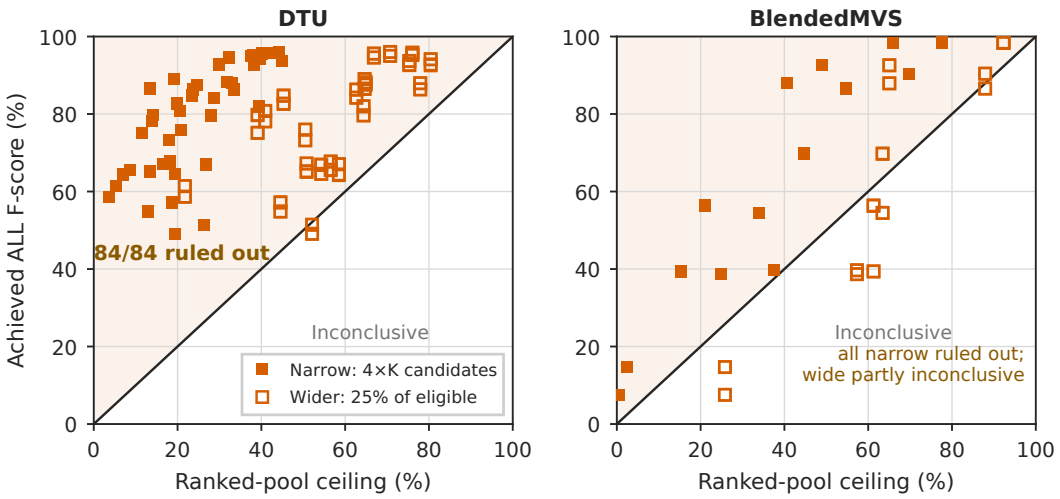


Figure 3: **Every narrow ranked pool falls below the ALL baseline.** Each marker pairs the F-score achieved by sampling from ALL eligible points (vertical axis) with the coverage ceiling $U(\mathcal{C})$ of the ranked pool (horizontal axis) at the same budget and start. A point above the diagonal means no subset of that pool can reach the ALL score ("ruled out"), and a point below is inconclusive. Every narrow condition is ruled out, and only some wider pools are inconclusive.

Table 1: **The random pool scores higher at the narrow width on both datasets and at 25% retention on DTU.** Random minus ranked in selected F-score and in coverage ceiling (computed without sampling), in F-score points (↑). The narrow $4 \times K$ pool keeps 2.18% of the eligible DTU points on average over scans and budgets, and the wider rows reuse the same scenes.

Candidates	Dataset	ΔF-score (pts) ↑	Δceiling (pts) ↑
$4 \times K$	DTU	**+54.10**	**+60.94**
$4 \times K$	BlendedMVS	+32.17	+43.66
25%	DTU	+24.67	+30.79
25%	BlendedMVS	+2.87	+23.64

The ceiling holds while coordinates, alignment, reference weights, masks, and tolerance are shared, as in our ceiling comparisons. Repeated views and nonnegative weights are allowed. We derive it in Appendix F, state the evaluation assumptions in Appendix H, and test realigned selections in Section 4.6.

3.3 Mixed Pools: The Constructive Check

If pool membership explains the loss, we should recover it by replacing part of the ranked pool with uniformly sampled candidates at fixed pool size. The check follows the baseline-substitution diagnostics of Tatarchenko et al. (2019). We keep the scans, eligibility rule, sampler, starts, and metrics of the preceding sections. The variants are ranked-only, random-only, and mixed pools at ranked fractions $f \in \{0.25, 0.50, 0.75\}$, with 50/50 as the main configuration. The mixed-pool study uses the smaller DTU output budget, where the ranked and random variants score 20.54 and 76.63 in mean F-score. On BlendedMVS it uses both budgets, where they score 38.22 and 70.36 (Appendix H explains the gap to Table 3). The DTU values of 23.52 and 77.62 in the Introduction average both budgets.

4 Experiments

We score reconstructions with precision, recall, and F-score at a fixed distance tolerance, following Tanks and Temples (Knapitsch et al., 2017). Sections 4.1 and 4.2 use VGGT predictions with confidence above two, and Section 4.3 applies VGGT's default validity masks to its raw predictions.

Table 2: **Independent random-pool seeds keep the narrow-pool contrast at both fixed output budgets.** Random minus ranked F-score, in percentage points, for $4 \times K$ pools. DTU entries carry a sample standard deviation over the five pool seeds, and BlendedMVS entries use one pool seed. "Small" and "large" are each dataset's two fixed output budgets.

Dataset	Output budget	ΔF-score (pts) ↑	SD over seeds
DTU	small	**+56.11**	0.037
DTU	large	+52.09	0.024
BlendedMVS	small (25,000)	+35.01	—
BlendedMVS	large (50,000)	+29.34	—

Table 3: **At the narrow width the ranked pool's recall falls while its precision holds.** Selected precision (P), recall (R), F-score (F), and coverage ceiling (U) in percent for $4 \times K$ pools and ALL; Wins counts conditions in which the selection beats ALL. Shaded rows are the ranked pools, with their recall in bold. On DTU their recall is a fifth of the ALL recall. Means cover 84 DTU and 28 BlendedMVS conditions (Tables 8 and 9 give every width). The ranked pool's ceiling sits close to its F-score, so sampling loses little beyond what the pool already lost.

Dataset	Pool	P (%) ↑	R (%) ↑	F (%) ↑	U (%) ↑	Wins ↑
DTU	ALL	80.71	76.90	78.44	89.67	0/84
DTU	Ranked $4 \times K$	80.42	**14.17**	23.52	24.23	0/84
DTU	Random $4 \times K$	83.09	73.38	77.62	85.18	37/84
BlendedMVS	ALL	63.69	63.92	62.49	90.00	0/28
BlendedMVS	Ranked $4 \times K$	86.47	**26.30**	38.22	38.42	0/28
BlendedMVS	Random $4 \times K$	74.00	69.18	70.39	82.08	28/28

4.1 Ranked and Random Pools at the Narrow Width

On both datasets the advantage comes from recall (Table 3, Figure 4). On DTU the three variants have similar precision (ranked 80.42, random 83.09, ALL 80.71) but ranked recall is only 14.17. On BlendedMVS ranked precision is higher than random (86.47 against 74.00) while ranked recall is far lower (26.30 against 69.18).

Each ranked DTU pool from $4 \times K$ to $32 \times K$ has a ceiling below the ALL F-score at the matched budget (Table 8), so no coordinate-preserving sampler could have matched that reconstruction from it. Figure 3 shows each condition for the $4 \times K$ and 25% pools. Wider ranked pools improve on average. On BlendedMVS some become wide enough that the ceiling no longer settles the comparison, and some of those then beat ALL.

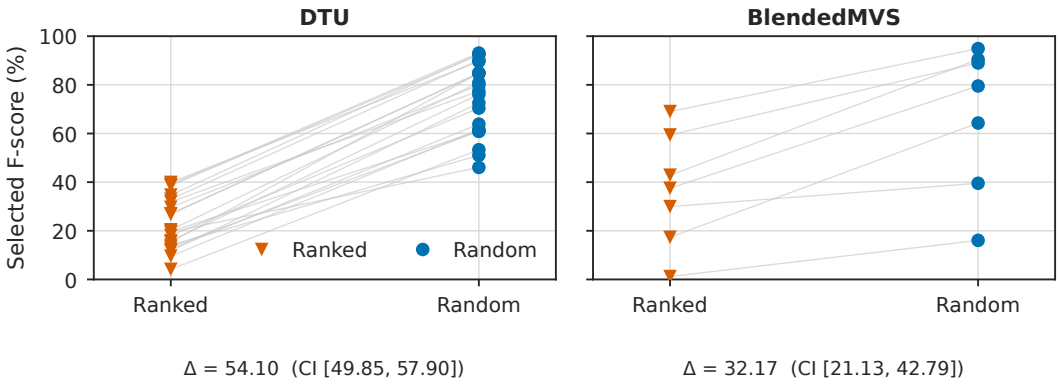


Figure 4: **At four candidates per output point, the random pool beats the ranked pool on both datasets.** Lines join each scene's ranked and random F-scores (%). The annotated Δ is the mean random-minus-ranked contrast with its interval (Table 1). The lines rise for nearly every scene.

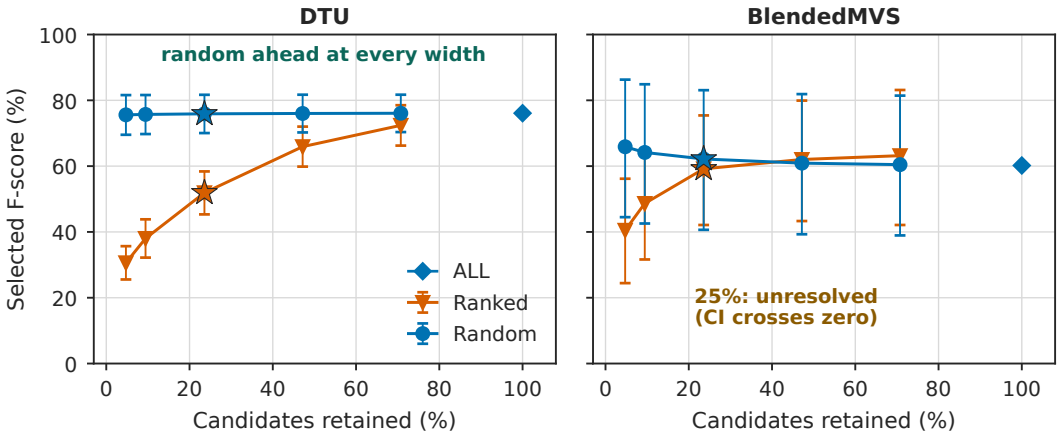


Figure 5: **The random advantage is largest at the tightest pools.** Mean selected F-score (%) against the fraction of eligible points retained, with 95% scene-level intervals. Stars mark the 25% setting, and ALL is the right-hand endpoint. The random pool stays ahead at every DTU fraction. On BlendedMVS the interval of the marked 25% contrast crosses zero.

Table 4: **On raw predictions, VGGT's confidence filter beats both of its controls.** VGGT mask + FPS minus each same-head control (an equal-size random pool, and ALL raw predictions), in F-score points (↑).

Dataset	Head	Δvs random (pts) ↑	Δvs ALL (pts) ↑
DTU	Depth	**+50.60**	**+52.12**
DTU	Point	+41.66	+44.32
BlendedMVS	Depth	+46.97	+47.82
BlendedMVS	Point	+42.83	+43.98

4.2 The Loss Is Largest at Tight Pools

At the 25% setting, random membership improves DTU F-score by 24.67 points, with an interval clear of zero (Table 1, Figure 5). We sweep retention from 5% to 75% of eligible points, with ALL as a separate endpoint. On BlendedMVS the random pool keeps more support at each fraction, and its F-score lead is separated from zero up to 10% retention. At 25% the 2.87-point mean is not separated from zero, and at 75% the ranked pool leads, with an interval that excludes zero (Table 11).

4.3 VGGT's Filter Helps on Raw Outputs

On raw predictions, where invalid points remain, applying VGGT's mask before budgeted sampling helps. For each output head, we rerun VGGT and compare the masked and unfiltered predictions. VGGT's own filter retains points at or above the 50th confidence percentile for its depth-derived and direct-point outputs (Appendix B). We reproduce that mask, ties at the threshold included. For each head, FPS draws exactly K points from the masked set, an equal-size random pool, and ALL raw predictions from one shared start.

VGGT's filter scores higher on both datasets and both heads (Table 4; Appendix B). The random pools retain higher coverage ceilings yet reconstruct worse, and ALL never exceeds a filtered pool's ceiling (Table 14). Support is lost when confidence ranks further within an already-filtered population. The two protocols differ in inference, eligibility, and alignment, so we do not compare their effect sizes.

4.4 The Loss Repeats on Other Backbones and New Scenes

Other backbones. Our camera-alignment criterion requires a backbone's own cameras to align to the DTU reference within a residual bound (Appendix C). Among backbones that do not initialize from VGGT weights, Depth Anything 3 (Lin et al., 2026) meets it on all 21 scans, and it shows the same loss. Its ceiling gap is 35.54 points, with recovery ruled out in 80 of 84 conditions, and its ALL

Table 5: **The ranked pool's ceiling falls below the achieved ALL score in every VGGT condition and most π^3 and DA3 conditions.** Mean ceiling gap is the ALL F-score minus the coverage ceiling $U(\mathcal{C})$ of the ranked $4 \times K$ pool, in F-score points (↑); "ruled out", in bold, counts conditions with a positive gap, where no coordinate-preserving sampler can recover the ALL score. The shaded row is the primary study. Table 21 gives the scene-level 95% intervals, DA3's under Appendix C. The DUSt3R and MapAnything diagnostic rows are in Appendix C.

Model / dataset	Scenes	Mean ceiling gap (pts) ↑	Ruled out
VGGT / DTU	21	54.21	**84/84**
VGGT / BlendedMVS	7	24.07	**28/28**
π^3 / DTU	21	19.95	**78/84**
DA3 / DTU	21	35.54	**80/84**

Table 6: **Random exceeds ranked in every ETH3D budget and pool condition.** Means over five paired pool seeds on the ten usable scenes, with the increment random minus ranked. Table 21 gives per-seed standard deviations and paired 95% scene-bootstrap intervals, and all four exclude zero.

Budget	Matched pool	Ranked F (%)	Random F (%)	Increment (pts) ↑
small	$4 \times K$	8.47	23.47	**15.00**
small	25% eligible	15.48	23.34	**7.86**
large	$4 \times K$	12.89	25.72	**12.83**
large	25% eligible	15.96	25.52	**9.56**

F-score is 85.18. π^3 and AMB3R share pretrained components with VGGT and repeat the effect. On the DTU scans, π^3 (Wang et al., 2026b) keeps the achieved ALL F-score above the ranked pool's ceiling across the resampling range at both budgets. Recovery is ruled out in 78 of 84 conditions (Table 5, Figure 8). AMB3R refines a frozen VGGT front-end with a volumetric backend (Wang & Agapito, 2026) and shows the same trend under a tighter-budget protocol. There, a random $2 \times K$ pool beats the ranked one by 4.0 F-score points at a 2.5% budget, against 5.3 for VGGT, both with intervals clear of zero (Appendix C, AMB3R geometry).

DUSt3R and MapAnything fall outside the criterion, so we report their DTU results as diagnostics. DUSt3R, architecturally independent of VGGT, shows the same loss: 19.83 points with its estimated cameras and 39.21 with the DTU calibration supplied, where the cameras come from the benchmark rather than the model (Appendix C). MapAnything, whose ALL F-score is 13.97 under its estimated cameras, shows no separated ceiling gap, since at that error level it recovers little surface to lose.

New scenes. The matched membership contrast also holds on ETH3D (Schöps et al., 2017), which lies outside both VGGT's disclosed training sources and our protocol development. On the ten of 13 scenes with usable predictions, random exceeds ranked at both output budgets and both matched-pool definitions (Table 6). Absolute ETH3D scores are low, 8.5 to 26.5 F-score points across the pools we report, and under the per-scene alignment of Appendix C seven of ten median camera residuals exceed 5 cm. We therefore use ETH3D only to test the sign of the effect. That appendix gives the residuals, the coordinate convention, and the three excluded scenes.

Low-reach regime. Ranking can remove only surface the predictions reach. On VGGT-Omega (Wang et al., 2026a), a retrained model with 10.8% of points within tolerance, the contrast shrinks: each variant scores below 13 F-score points, ranked leads in the mean, and random leads on 6 of the 21 scans (Appendix E).

4.5 Mixed Pools Close the Gap

Replacing part of the ranked pool with random candidates recovers most of the lost F-score (Figure 6). With the 50/50 pool, DTU reaches 76.19, which recovers 99% of the ranked-to-random gap, a paired difference of $+55.65$ points over ranked-only. BlendedMVS reaches 71.56, more than the full gap, with $+33.34$ points. Table 12 lists every mixed pool, and Table 21 gives the intervals.

On BlendedMVS the ranked pool covers far less of the reference within tolerance than the random pool (Table 3, column U), so a mixed pool keeps support that the pure ranked variant discards.

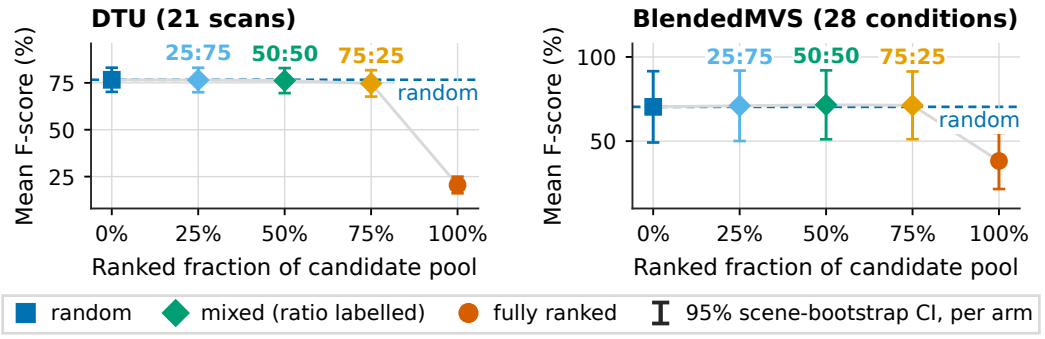


Figure 6: **A 50/50 mixed pool closes the ranked–random gap.** Mean F-score against the ranked fraction of the candidate pool (0% = random, 100% = ranked) at the smaller DTU output budget and both BlendedMVS budgets. Error bars are per-arm 95% scene-bootstrap intervals, and the paired differences in Section 4.5 are tighter.

On DTU the mixed pool sits within half a point of random selection (-0.44), and on BlendedMVS its $+1.20$ difference is not separated from zero. Most of the loss therefore comes from how the candidate pool is built. A confidence-proportional pool recovers most of what ranking loses but stays below random on all 21 DTU scans at $4 \times K$, and two selectors that change more than membership do no better (Appendix E); mixing and proportional draw are two separate remedies that each recover at least 97% of the gap on DTU with VGGT.

4.6 Pool Membership Affects Alignment Recovery

Membership also changes the outcome of an alignment test that does not use the F-score. We align ALL, ranked and random selections at 25% retention to the reference by iterative closest point (ICP) under twelve shared disturbances. Random selections recover the common alignment in 92.5% of disturbed DTU cases against 39.0% from ranked ones, a difference of 53.47 percentage points. On BlendedMVS the rates are 81.5% and 45.8% (Table 17). We fit the reference transform to ALL, so success measures agreement with the ALL alignment and cannot say which policy is more accurate on its own. The disturbed rates include each policy's unperturbed agreement: 42.9% of DTU cases for ranked against 100% for random (Appendix D).

5 Conclusion

Confidence tells us how good a point is and nothing about where the other points lie. It works as a filter, but under a tight output budget ranking on it alone can keep accurate points and still remove support from entire regions. The random pool gains recall at some cost in precision, so which rule leads depends on how much is kept. Once the pool is wide enough that recall stops being the bottleneck, ranking's higher precision wins, as on BlendedMVS at wide retention (Table 11).

5.1 In practice

For the VGGT settings we studied, the better rule depends on pool width. Under a tight budget, draw the pool at random after thresholding, mix in random candidates, or draw it in proportion to confidence as Mini-Splatting and Taming-3DGS draw Gaussian primitives; each recovers nearly all of what ranking costs (Section 4.5). Where reference geometry exists, compute the ceiling before sampling (Appendix F). A method that fills a budget from a confidence-ranked pool can also report an equal-size random pool through the same sampler.

Reproducibility Statement

Appendices G and H state what every reported number was computed from: the scene lists, the reference geometry with its alignment and distance tolerance, the eligibility rule, the pool seeds and the shared sampler starts, the evaluation settings, the resampling construction behind every reported

interval, and the fixed per-scan output budgets (Table 19). Selection is deterministic given a scene, budget, start, and pool seed, so these settings fix every contrast we report. Appendix F derives the coverage ceiling in full. Licensed datasets remain subject to their original terms.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs (a draft of the coverage-ceiling derivation in Appendix F, from a statement and proof strategy we fixed), propose or refine hypotheses and design or provide feedback on research methodology or experiments (proposing candidate analyses, controls, and pre-registered predictions for us to approve), implement methods (the selection and evaluation scripts), and interpret results (as a second reader of our analyses). We have not used generative AI tools to formulate mathematical claims or provide critical ingredients for proving them, and generating synthetic data sets, translation, cleaning and reformatting datasets, and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to draft parts of this paper from an author-written evidence brief, edit it to improve readability, summarize existing literature and identify relevant literature, source and search for information, create or modify scientific figures, create or edit software code, and brainstorm. We have reviewed all AI-assisted work. We set every research question, experimental design, and decision rule before each study, ran and inspected the experiments, checked every number and figure against the result files, checked that every reference exists and matches its bibliographic record, and checked every step of the derivation. No language model generated, altered, imputed, or selected any experimental result, dataset, or reference. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This paper studies how candidate selection under a fixed output budget affects feed-forward multi-view reconstruction quality. It involves no human subjects, no human-derived data, and no personally identifiable information. All experiments use standard public benchmark datasets under their original terms, none of which are redistributed: DTU (freely released for multi-view stereo evaluation), BlendedMVS (Creative Commons Attribution 4.0), ETH3D (Creative Commons Attribution-NonCommercial-ShareAlike 4.0), and ARKitScenes (Baruch et al., 2021) (Apple's dataset license).

REFERENCES

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *Advances in Neural Information Processing Systems*, 2020.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2020.

Fadhel Ayed and Soufiane Hayou. Data pruning and neural scaling laws: Fundamental limitations of score-based algorithms. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=iRTL4pDavo.

Oleksandr Bailo, Francois Rameau, Kyungdon Joo, Jinsun Park, Oleksandr Bogdan, and In So Kweon. Efficient adaptive non-maximal suppression algorithms for homogeneous spatial keypoint distribution. *Pattern Recognition Letters*, 106:53–60, 2018. doi: 10.1016/j.patrec.2018.02.020.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitScenes: A diverse real-world dataset for 3D indoor scene understanding using mobile RGB-D data. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/66f041e16a60928b05a7e228a89c3799-Paper-round1.pdf.

Robert Bridson. Fast Poisson disk sampling in arbitrary dimensions. In *ACM SIGGRAPH 2007 Sketches*, pp. 22, 2007. doi: 10.1145/1278780.1278807.

Matthew Brown, Richard Szeliski, and Simon Winder. Multi-image matching using multi-scale oriented patches. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pp. 510–517, 2005. doi: 10.1109/CVPR.2005.235.

Chen Chen, Zhe Chen, Jing Zhang, and Dacheng Tao. SASA: Semantics-augmented set abstraction for point-based 3D object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 221–229, 2022. doi: 10.1609/aaai.v36i1.19897.

Nanxing Nick Deng, Qing Cheng, Niclas Zeller, and Daniel Cremers. A calibration audit of confidence in feed-forward 3D reconstruction models. *arXiv preprint arXiv:2608.29705*, 2026. URL https://arxiv.org/abs/2608.29705.

Oren Dovrat, Itai Lang, and Shai Avidan. Learning to sample. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2760–2769. IEEE, June 2019. doi: 10.1109/CVPR.2019.00287.

DUSt3R contributors. DUSt3R released demo filtering code. https://raw.githubusercontent.com/naver/dust3r/3cc8c88c413bb9e34c41db0e0eef99c2ee010b12/dust3r/demo.py, 2026. Accessed 7 September 2026.

Y. Eldar, M. Lindenbaum, M. Porat, and Y. Y. Zeevi. The farthest point strategy for progressive image sampling. *IEEE Transactions on Image Processing*, 6(9):1305–1315, 1997. doi: 10.1109/83.623193.

ETH3D. ETH3D dataset documentation, 2026. URL https://www.eth3d.net/documentation. Accessed 7 September 2026; laser scan coordinate transforms.

Zhiwen Fan, Kevin Wang, Kairun Wen, Zehao Zhu, Dejia Xu, and Zhangyang Wang. LightGaussian: Unbounded 3D Gaussian compression with 15x reduction and 200+ FPS. In *Advances in Neural Information Processing Systems*, 2024.

Guangchi Fang and Bing Wang. Mini-Splatting: Representing scenes with a constrained number of Gaussians. In *European Conference on Computer Vision*, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985. doi: 10.1016/0304-3975(85)90224-5.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

Alex Hanson, Allen Tu, Vasu Singla, Mayuka Jayawardhana, Matthias Zwicker, and Tom Goldstein. PUP 3D-GS: Principled uncertainty pruning for 3D Gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5949–5958, 2025.

Markus Hillemann, Robert Langendörfer, Steven Landgraf, and Markus Ulrich. Uncertainty quality of VGGT: An analysis on the DTU benchmark dataset. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XI-2-2026:665–672, 2026. doi: 10.5194/isprs-annals-XI-2-2026-665-2026.

Yihong Huang, Fei Ma, Yihua Shao, Jingcai Guo, Zitong Yu, Laizhong Cui, and Qi Tian. Nüwa: Mending the spatial integrity torn by VLM token pruning. In *International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=C9yclwdquU.

Zhisheng Huang, Jiahao Chen, Cheng Lin, Chenyu Hu, Hanzhuo Huang, Zhengming Yu, Mengfei Li, Yuheng Liu, Zekai Gu, Zibo Zhao, Yuan Liu, Xin Li, and Wenping Wang. RecGen3D: Reconstruction-guided 3D generation in a shared canonical space. In *SIGGRAPH Asia 2026 Conference Papers*, 2026b. arXiv:2604.01479.

Eddy Ilg, Ozgun Cicek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *European Conference on Computer Vision*, 2018.

Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

Nikhil Keetha, Norman Müller, Johannes 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 Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. In *International Conference on 3D Vision (3DV)*, pp. 499–509, 2026. doi: 10.1109/3DV69130.2026.00054.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, 2017.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and Temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4):1–13, July 2017. doi: 10.1145/3072959.3073599.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Itai Lang, Asaf Manor, and Shai Avidan. SampleNet: Differentiable point cloud sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7578–7588, 2020.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *European Conference on Computer Vision*, 2024.

Chongshou Li, Pin Tang, Tianrui Li, Yuheng Liu, and Xinke Li. Enhancing sampling protocol for point cloud classification against corruptions. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 421–429, 2025. doi: 10.24963/ijcai.2025/48. URL https://www.ijcai.org/proceedings/2025/48.

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*, pp. 141261–141285, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/e4cd50120b6d7e8daff1749d6bbaa889-Paper-Conference.pdf.

Yang Liu, Chuanchen Luo, Zimo Tang, Junran Peng, and Zhaoxiang Zhang. VGGT-X: When VGGT meets dense novel view synthesis. *arXiv preprint arXiv:2509.25191*, 2025. URL https://arxiv.org/abs/2509.25191.

Saswat Subhajyoti Mallick, Rahul Goel, Bernhard Kerbl, Markus Steinberger, Francisco Vicente Carrasco, and Fernando De la Torre. Taming 3DGS: High-quality radiance fields with limited resources. In *SIGGRAPH Asia Conference Papers*, 2024.

Raúl Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. doi: 10.1109/TRO.2015.2463671.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14:265–294, 1978. doi: 10.1007/BF01588971.

Open3D. Open3D 0.19.0 documentation: ICPConvergenceCriteria, n.d. URL https://www.open3d.org/docs/0.19.0/python_api/open3d.pipelines.registration.ICPConvergenceCriteria.html. Version-specific software documentation. Accessed September 7, 2026.

Jongmin Park, Minh-Quan Viet Bui, Juan Luis Gonzalez Bello, Jaeho Moon, Jihyong Oh, and Munchurl Kim. EcoSplat: Efficiency-controllable feed-forward 3D Gaussian splatting from multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26010–26020, 2026.

Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, 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*, pp. 2538–2547, 2017. doi: 10.1109/CVPR.2017.272.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: Beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems*, 2022.

Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander 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*, pp. 5283–5293, 2025. doi: 10.1109/CVPR52734.2025.00498.

Maxim Tatarchenko, Stephan R Richter, Rene 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)*, 2019.

VGGT contributors. VGGT released visualization and demo filtering code. https://github.com/facebookresearch/vggt/tree/a288dd0f14786c93483e45524328726ab7b1b4ce, 2026. visual_util.py, demo_gradio.py and demo_colmap.py; accessed 7 September 2026.

Hengyi Wang and Lourdes Agapito. 3D reconstruction with spatial memory. In *International Conference on 3D Vision*, 2025.

Hengyi Wang and Lourdes Agapito. AMB3R: Accurate feed-forward metric-scale 3D reconstruction with backend. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14612–14625, 2026.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5294–5306, 2025a. URL https://arxiv.org/abs/2503.11651v1.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω, 2026a. URL https://arxiv.org/abs/2605.15195. CVPR 2026.

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*, 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*, pp. 20697–20709, 2024.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/11a09e0aaa74867c6b0719c639fc09f8-Paper-Conference.pdf.

Chengzhi Wu, Yuxin Wan, Hao Fu, Julius Pfrommer, Zeyun Zhong, Junwei Zheng, Jiaming Zhang, and Jürgen Beyerer. SAMBLE: Shape-specific point cloud sampling for an optimal trade-off between local detail and global uniformity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1342–1352, 2025.

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*, 2025.

Luwei Yang, Rakesh Shrestha, Wenbo Li, Shuaicheng Liu, Guofeng Zhang, Zhaopeng Cui, and Ping Tan. SceneSqueezer: Learning to compress scene for camera relocalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8259–8268, 2022.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *European Conference on Computer Vision*, 2018.

Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. BlendedMVS: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1790–1799, 2020.

Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. Coverage-centric coreset selection for high pruning rates. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=QwKvL6wC8Yi.

Zihao Zhu, Wenyuan Zhao, Nuo Chen, Chao Tian, and Zhiwen Fan. Trust it or not: Evidential uncertainty for feed-forward 3D reconstruction with Trust3R. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.19539v1. Author preprint arXiv:2605.19539v1.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming visual geometry transformer. In *International Conference on Learning Representations*, pp. 88055–88072, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/8da04a60948be713dc766f0c7e3a5b1f-Paper-Conference.pdf. StreamVGGT.

APPENDIX

A THE GAP SHRINKS AS THE POOL WIDENS

Ranking does its damage when the budget is tight and the pool has little room left for coverage. On both datasets the random pool's advantage is largest at the narrowest width and shrinks as the pool widens. Tables 8 and 9 and Figure 7 show this by width, and Tables 10 and 11 by retained fraction, the sweep of Section 4.2. To see where the gap closes, the fixed-fraction study keeps $M = \lfloor qN \rfloor$ for $q \in \{0.05, 0.10, 0.25, 0.50, 0.75\}$ of the same eligible points, at the same budgets and starts. Figure 1 uses the smaller budget and the first repetition, and its panels show the reference observations each pool supports, without the sampled output.

What the random pool trades. The random pool's $4 \times K$ advantage is a recall gain (Table 3), and its precision is close to the ranked pool's on DTU and lower on BlendedMVS. Table 7 follows both terms across the fractions of Tables 10 and 11. On DTU the recall gain stays clear of zero at each fraction and the precision cost stays within 3.61 points, so the random pool leads throughout. On BlendedMVS the precision cost is larger, reaching 20.38 points at 25% retention. By 50% retention the interval on the recall gain covers zero while the interval on the precision cost excludes it, and the F-score difference changes sign there. The crossover is measured on seven scenes. Across the BlendedMVS rows, the pattern is consistent with a recall gain that runs out while the precision cost remains.

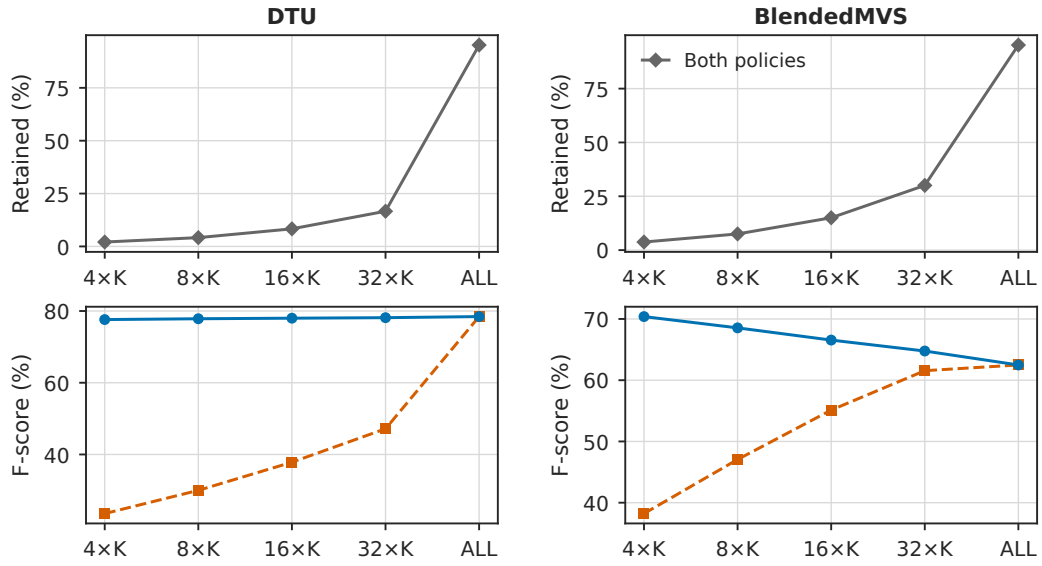


Figure 7: **The ranked pool's F-score deficit is largest at the narrowest width and shrinks as the pool widens.** Retained fraction and selected F-score against candidate width on DTU and BlendedMVS, for ranked pools (squares, dashed) and random pools (circles, solid). ALL is a separate endpoint.

B THE RELEASED FILTER ON RAW OUTPUTS

On raw predictions the released mask beats a random pool of equal size, even though the random pool keeps more coverage, so confidence works as a filter. Table 13 gives the released outputs before any fixed-budget adaptation, and Table 14 the ceilings behind Table 4.

VGGT's own pipeline keeps points at or above the selected confidence percentile, floored at 10^{-5}. The released COLMAP export thresholds on confidence and then fills its hard 100,000-point budget with a uniform draw without replacement from the survivors (VGGT contributors, 2026). The released pipelines we examined cut at a confidence threshold or a confidence percentile. The budgeted rule we test raises that cut until a fixed number of points remain, and that step is what Section 3.1 isolates.

Table 7: **The F-score sign changes where the recall gain runs out.** Random minus ranked recall and precision, in percentage points, with 95% paired scene-bootstrap intervals, over the conditions of Tables 10 and 11.

Population	Retained	Recall difference	Precision difference
DTU	5%	+54.11 [+48.26, +59.51]	+0.35 [−4.78, +6.10]
	10%	+48.20 [+42.40, +53.69]	−1.26 [−5.86, +4.01]
	25%	+35.26 [+30.31, +40.15]	−3.61 [−6.61, −0.39]
	50%	+18.71 [+15.61, +22.03]	−3.61 [−4.66, −2.60]
	75%	+8.04 [+6.34, +9.96]	−2.24 [−2.69, −1.74]
BlendedMVS	5%	+38.87 [+25.88, +49.92]	−15.19 [−35.01, +2.60]
	10%	+29.72 [+18.42, +40.39]	−18.44 [−36.90, −2.40]
	25%	+15.80 [+6.67, +26.69]	−20.38 [−35.41, −7.12]
	50%	+6.35 [−1.76, +15.21]	−10.87 [−16.16, −6.06]
	75%	+0.14 [−4.93, +4.52]	−6.58 [−9.26, −4.13]

Table 8: **Every DTU ranked width has a ceiling below the achieved ALL score.** Means in percent. P, R, F are selected precision, recall, and F-score and U is the pool ceiling. E counts ceilings that do not settle the comparison and W counts selected wins against ALL, each out of 84 conditions.

Policy	Width	P	R	F	U	E	W
ALL	ALL	80.71	76.90	78.44	89.67	84	0
Ranked	4	80.42	14.17	23.52	24.23	0	0
Random	4	83.09	73.38	77.62	85.18	84	37
Ranked	8	81.61	19.01	29.99	31.35	0	0
Random	8	82.76	74.09	77.84	86.49	84	36
Ranked	16	83.10	25.58	37.83	40.07	0	0
Random	16	82.37	74.70	78.00	87.52	84	37
Ranked	32	84.26	34.45	47.16	50.55	0	0
Random	32	81.97	75.29	78.14	88.32	84	35

Table 9: **Some wider BlendedMVS ranked pools exceed the ALL baseline.** Means in percent for P, R, F, and U, with E and W counted as in Table 8, each out of 28 conditions.

Policy	Width	P	R	F	U	E	W
ALL	ALL	63.69	63.92	62.49	90.00	28	0
Ranked	4	86.47	26.30	38.22	38.42	0	0
Random	4	74.00	69.18	70.39	82.08	28	28
Ranked	8	87.21	34.74	47.06	47.55	2	2
Random	8	71.16	68.38	68.55	84.74	28	28
Ranked	16	87.73	43.10	55.12	56.21	8	4
Random	16	68.54	67.07	66.55	86.76	28	28
Ranked	32	84.34	51.00	61.55	64.93	16	16
Random	32	66.33	65.78	64.76	88.23	28	28

Table 10: **DTU keeps a random-pool advantage at every retained fraction.** Random minus ranked selected F-score and ceiling, in percentage points, with 95% scene-bootstrap intervals. 25% is the primary fixed-fraction setting.

Retained	Selected F-score difference	Ceiling difference
5%	+46.35 [+41.58, +50.92]	+53.71 [+49.25, +58.16]
10%	+38.87 [+33.91, +43.75]	+46.01 [+40.94, +51.13]
25%	+24.67 [+20.46, +29.25]	+30.79 [+26.18, +35.84]
50%	+10.48 [+8.18, +12.93]	+14.32 [+11.89, +16.79]
75%	+3.81 [+2.62, +5.18]	+5.71 [+4.33, +7.18]

The export's uniform draw from the survivors is the membership choice that Section 5.1 recommends under a tight budget. We rerun inference for both heads and apply exact-K FPS to the released mask, to an equal-size random pool, and to all same-head observations, with starts matched within head. Scores use the surviving observations without evaluator downsampling.

Table 11: **On BlendedMVS the wider random pools keep more support, and at 75% the ranked pool selects better.** Random minus ranked selected F-score and ceiling, in percentage points, with 95% scene-bootstrap intervals. At 25% the F-score interval covers both signs, and the mean changes sign between the two budgets.

Retained	Selected F-score difference	Ceiling difference
5%	+26.75 [+16.52, +35.96]	+41.71 [+32.65, +52.11]
10%	+16.40 [+6.42, +25.56]	+34.94 [+24.20, +46.50]
25%	+2.87 [−5.82, +12.31]	+23.64 [+14.09, +32.91]
50%	−1.33 [−6.14, +4.26]	+14.04 [+7.94, +19.84]
75%	−3.39 [−6.56, −0.59]	+5.24 [+3.11, +7.42]

Table 12: **Mixed pools stay near the random pool until the pool is fully ranked.** Mean F-score (%) by ranked fraction of the candidate pool at the smaller DTU output budget and both BlendedMVS budgets, as plotted in Figure 6. Section 4.5 gives the paired differences.

Dataset	0% (random)	25%	50%	75%	100% (ranked)
DTU	76.63	76.52	76.19	74.71	20.54
BlendedMVS	70.36	71.00	71.56	71.26	38.22

Table 13: **The released masks return variable counts.** Returned count, F-score, and retained fraction as scene means at the 50th-percentile setting, without FPS. BlendedMVS keeps every observation in one scene because of ties.

Population / head	Returned count	F-score (%)	Retained (%)
DTU / depth	4,974,873	85.07	50.00
DTU / point	4,974,872	76.89	50.00
BlendedMVS / depth	3,713,024	73.57	57.14
BlendedMVS / point	3,713,024	72.19	57.14

Table 14: **On raw outputs the released mask beats random pools despite their higher ceilings.** Achieved ALL F-score minus the released-pool ceiling, and random-pool minus released-pool ceiling at equal candidate count, in percentage points with 95% scene-bootstrap intervals.

Population / head	ALL F − released ceiling	Ceiling difference
DTU / depth	−60.90 [−63.84, −57.98]	+6.25 [+4.73, +7.94]
DTU / point	−52.67 [−56.59, −48.71]	+9.31 [+7.01, +11.70]
BlendedMVS / depth	−75.58 [−80.83, −69.04]	+7.56 [+3.58, +12.14]
BlendedMVS / point	−70.33 [−77.60, −62.35]	+8.99 [+4.33, +14.61]

C THE EFFECT ON OTHER BACKBONES AND SCENES

A different metric from each source paper. The percentages above are F-scores under our own protocol: precision and recall at a 5 mm tolerance, at a per-scan output budget K, on the 21 camera-aligned DTU scans (Appendix G). They are not the metric each source paper reports for DTU. VGGT, DUSt3R, π^3, and Depth Anything 3 report DTU as Chamfer-distance accuracy/completeness/overall in millimeters (Depth Anything 3's own number is 1.85 mm, on its own 22-scan protocol). AMB3R reports relative-depth, accuracy, and completeness error. MapAnything does not evaluate DTU. No claim in this paper depends on these numbers agreeing with a source paper's own table. The loss carries over to a backbone whose predictions reach the surface and whose cameras meet the camera-alignment criterion. Depth Anything 3 meets the criterion without any VGGT initialization and shows the loss, and so do π^3 and AMB3R, which start from VGGT components (Tables 15 and 16). Where reconstruction error dominates, as for MapAnything, the comparison says little about ranking. DUSt3R's own cameras fall outside the criterion, so both DUSt3R runs are diagnostics. On ETH3D the loss appears in the sign of every condition (Table 6), conditional on a per-scene alignment. It weakens on the retrained checkpoint of Appendix E. Table 21 gives the intervals for the gaps below.

Table 15: **Ceiling results for every backbone and dataset.** F-score and ceiling in percent, averaged over scenes with both budgets feasible. R4 and R8 are the ranked $4 \times K$ and $8 \times K$ pools. Depth Anything 3 (DA3) meets the camera-alignment criterion on all 21 scans without initializing from VGGT weights. The estimated-camera DUSt3R and MapAnything rows fall outside that criterion; the known-pose DUSt3R row meets it by construction, since the DTU calibration is supplied. All DUSt3R and MapAnything rows are diagnostics.

Population	Pool	Scenes	F	Ceiling	Ruled out
DTU π^3	ALL	21/21	65.32	94.83	0/84
DTU π^3	R4	21/21	42.51	45.37	78/84
DTU π^3	R8	21/21	49.80	54.07	64/84
DTU DA3	ALL	21/21	85.18	97.93	0/84
DTU DA3	R4	21/21	47.18	49.64	80/84
DTU DA3	R8	21/21	56.61	60.13	76/84
DTU DUSt3R	ALL	21/21	29.17	58.47	0/84
DTU DUSt3R	R4	21/21	8.76	9.33	84/84
DTU DUSt3R	R8	21/21	12.00	13.36	74/84
DTU DUSt3R, known poses	ALL	21/21	71.05	98.15	0/84
DTU DUSt3R, known poses	R4	21/21	31.16	31.84	82/84
DTU DUSt3R, known poses	R8	21/21	41.84	43.19	80/84
DTU MapAnything	ALL	21/21	13.97	58.82	0/84
DTU MapAnything	R4	21/21	10.87	13.26	53/84
DTU MapAnything	R8	21/21	13.98	18.03	33/84
ETH3D	ALL	8/10	26.49	39.43	0/32
ETH3D	R4	8/10	9.86	10.39	28/32
ETH3D	R8	8/10	14.73	16.20	24/32

π^3 on DTU. On π^3 the ALL-minus-ceiling gap is 19.95 points, ruling out recovery in 78 of 84 conditions (Figure 8). π^3's encoder and attention modules initialize from VGGT (Wang et al., 2026b, Appendix A.2), so the independence test rests on Depth Anything 3.

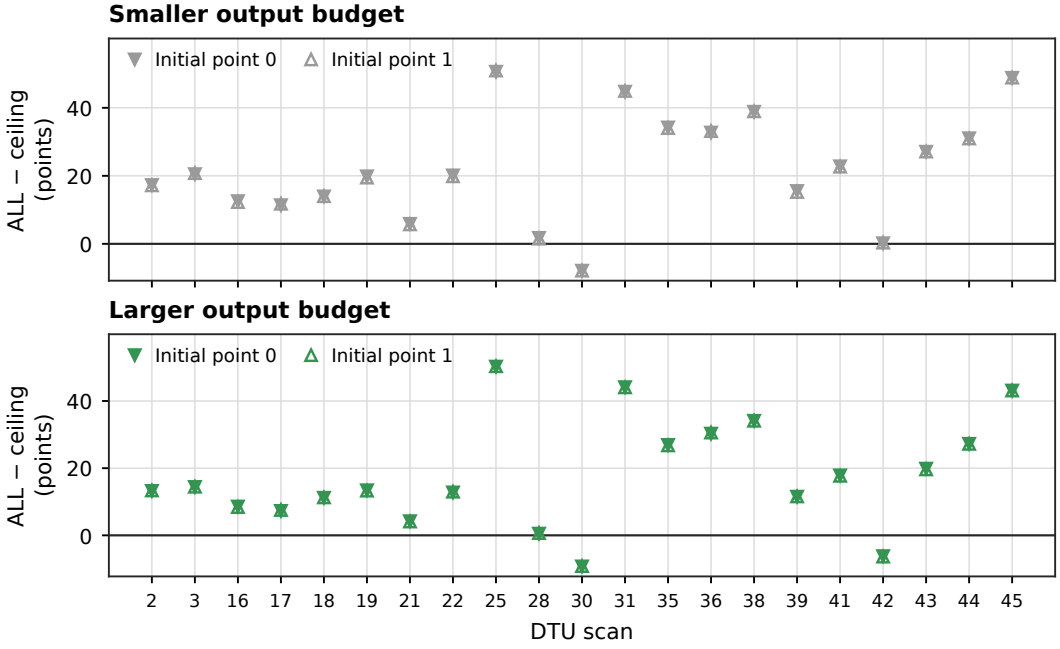


Figure 8: **The π^3 ceiling rules out recovery in 78 of 84 DTU conditions.** ALL F-score minus the ranked-$4 \times K$ ceiling, in F-score points, for every DTU scan and FPS start, one panel per output budget. A positive gap rules out recovery from that pool, and a nonpositive gap leaves the comparison open (Appendix F).

Depth Anything 3 on DTU. Depth Anything 3 (Lin et al., 2026) trains from its own initialization on a DINOv2-family encoder. Its ceiling gap is 35.54 points, ruling out recovery in 80 of 84 conditions. Its camera-center residuals, from 5.21 to 30.12 mm, stay below the 32.7 mm bound on every scan, so DA3 meets the camera-alignment criterion. The bound is π^3's own largest per-scan camera-center residual on these 21 scans, measured in DTU millimeters. It is separate from the per-scene check of Table 20. DTU lies outside its disclosed training corpus, which includes BlendedMVS. On ETH3D its camera residuals exceed 5 cm on seven of 13 scenes. Its eligibility cut is set on DA3's own confidence.

DUSt3R on DTU. Both DUSt3R runs go the same way as VGGT, and we report them as diagnostics. With its own estimated cameras the residuals, 69–490 mm, exceed the same 32.7 mm bound on every scan and fall outside the camera-alignment criterion. The ceiling gap is 19.83 points. Presetting the DTU calibration zeroes the residuals by construction and raises the gap to 39.21 points.

MapAnything on DTU. MapAnything (Keetha et al., 2026) is trained from scratch on a DINOv2 encoder. Its camera-center residuals of 77–410 mm exceed the same bound on every scan, outside the camera-alignment criterion, and at that error its reconstruction error dominates. The ceiling gap is 0.71 points at $4 \times K$ and -4.07 at $8 \times K$, ruled out in 53 and 33 of 84 conditions.

ETH3D with VGGT. On ETH3D the direct-point ceiling gap is 16.10 points on the eight of ten scenes the ceiling run covers. Of 13 ETH3D scenes, ten are usable. Meadow, playground, and terrains fail on too few eligible points or a zero F-score. We use one VGGT pass with scanner poses in the documented convention (ETH3D, 2026) and align each scene by a camera-center similarity. Seven of the ten median residuals are larger than a strict 5 cm tolerance, so the ETH3D conclusions are conditional on this alignment. Table 6 averages five paired repetitions per scene.

AMB3R geometry. AMB3R (Wang & Agapito, 2026) refines a frozen VGGT front-end through a volumetric backend with a metric-scale head, which gives it different coordinates and confidences. We restrict candidates to the top 0.9 confidence quantile. The gap is largest at the tightest budget and shrinks to about one point at 5% on both backbones (Table 16), since the candidates are already high-confidence.

Table 16: **A random pool beats the confidence-ranked pool on AMB3R and VGGT geometry at a tight budget.** Mean DTU F-score (%) over 21 scans. The difference is random minus ranked at $2 \times K$ with a 95% scene-bootstrap interval from 10,000 draws, and the last column counts scans the random pool wins.

Backbone	Budget	Ranked $2 \times K$	Random $2 \times K$	Random $4 \times K$	Difference	Scans won
AMB3R	2.5%	83.3	87.3	87.5	+4.0 [2.9, 5.3]	21/21
AMB3R	5%	87.5	88.3	88.3	+0.8 [0.3, 1.2]	15/21
VGGT	2.5%	68.5	73.9	74.2	+5.3 [3.8, 7.1]	21/21
VGGT	5%	74.4	75.4	75.5	+1.0 [0.5, 1.6]	16/21

D REGISTRATION AS A DOWNSTREAM ALIGNMENT TASK

Random pools recover a common alignment more often than ranked ones (Table 17), so the support loss also reaches a task that does not use the F-score.

Protocol. We align cached ALL, ranked, and random selections at 25% retention with a separate point-to-point ICP evaluator, whose solver settings are in Table 20. The target is a subsample of at most 100,000 reference points with centroid and 95th-percentile radius ρ, and an unperturbed ALL fit defines the reference transform T_A per scene, budget, and repetition. Twelve shared disturbances rotate each signed axis by 5° and 15° with translations of 0.02ρ and 0.05ρ along that axis. We count a success when the recovery error of $\widehat{T}DT_A^{-1}$, for disturbance D and estimate $\widehat{T}$, has rotation at most 2° and translation at most 0.01ρ.

Table 17: **Random selections recover the common alignment far more often than ranked ones.** Success rates in percent and random-minus-ranked differences in points with 95% scene-bootstrap intervals. The reference transform is fit to ALL, so ALL always succeeds unperturbed.

Population / input	Ranked	Random	ALL	Random − ranked
DTU / unperturbed	42.9	100.0	100.0	n/a
DTU / disturbed	39.0	92.5	92.3	+53.47 [+33.43, +73.12]
BlendedMVS / unperturbed	71.4	100.0	100.0	n/a
BlendedMVS / disturbed	45.8	81.5	83.3	+35.71 [+10.12, +58.33]

E BOUNDARY OF THE EFFECT AND ALTERNATIVE SELECTORS

The effect ends where predictions stop reaching the surface, as a retrained checkpoint whose points mostly miss it shows, and where recall stops being the bottleneck, as in the BlendedMVS reversal at wide retention (Appendix A). Three alternative selectors do no better than the random pool.

VGGT-Omega checkpoint. On the VGGT-Omega-1B 416-pixel reproduction checkpoint (Wang et al., 2026a), whose points mostly miss the surface, ranking leads in the mean, by a random-minus-ranked difference of -3.13 points (Table 21). We ran the five-variant mixed-pool protocol of Section 4.5 on it at its training resolution, with the scans, eligibility, $4 \times K$ pool, starts, and evaluator used for VGGT. Mean F-score is 11.08 ranked and 7.95 random, and the 25/50/75% mixed pools reach 8.48, 9.53, and 12.10, so the 75% mixed pool is best. The random variant has the higher precision and the lower recall, the reverse of the usual pattern, at 29.8 against 9.8 precision and 4.6 against 13.3 recall. The reason is where this checkpoint's points sit. Their median distance to the DTU surface is 31.0 mm against 2.1 mm for the released checkpoint, and only 10.8% lie within the 5 mm tolerance against 74%. With so little of the population near the surface, ranking has little reached surface to remove. The contrast shrinks to a few points and reverses in the mean, with random ahead on 6 of the 21 scans.

Alternative selectors. Three alternative selectors stay at or below the random pool. A coverage-aware selector that adds a coverage term to the ranking rule scores below both ranked and random at the smaller budget. A learned reliability estimator with a monotone submodular coverage objective (Nemhauser et al., 1978), trained on VGGT and Trust3R geometric-support features and evaluated on twelve ARKitScenes visits (Baruch et al., 2021) at 5 cm, did no better. Both change coordinates, counts, or initialization along with membership, so both sit outside the matched test. Drawing the pool with probability proportional to confidence, the rule Mini-Splatting and Taming-3DGS apply to rendering scores (Fang & Wang, 2024; Mallick et al., 2024), recovers most of what ranking loses. At $4 \times K$ on the 21 DTU scans with VGGT, it reaches an F-score of 75.83, between 23.52 for ranked and 77.63 for random, or 97% of the ranked-to-random gap (Table 18). It beats the ranked pool and falls below the random pool on all 21 scans, and its ceiling is 82.28 against 85.18 for random. We drew it without replacement from the same eligible candidates, with the FPS start shared by all three pools included as one member, so only membership changes. The arm and its decision rule were fixed before the run. We tested one width, one dataset, and one backbone.

Table 18: **A pool drawn in proportion to confidence lands close to the random pool and below it on every scan.** DTU F-score and coverage ceiling (U) in percent for $4 \times K$ VGGT pools over 21 scans at the two budgets of Table 19. Each scan averages five repetitions, each with its own pool seed and FPS start, so the random F-score differs slightly from Table 3, which averages two. Differences are proportional minus the named pool, paired by scan; the 95% scene-bootstrap interval (10,000 draws) and scans won refer to the budget average.

Pool	F, small	F, large	F	U
Ranked $4 \times K$	20.54	26.49	23.52	24.23
Random $4 \times K$	76.66	78.59	77.63	85.18
Proportional $4 \times K$	74.75	76.91	75.83	82.28
Difference	small	large	Average [95% interval]	Scans won
Proportional − ranked	+54.20	+50.41	+52.31 [+48.08, +56.17]	21/21
Proportional − random	-1.91	-1.69	-1.80 [-2.28, -1.34]	0/21

F COVERAGE-CEILING DERIVATION

The ceiling of Section 3.2 holds for any pipeline and coordinate-preserving selector, and is computed from reference geometry before a sampler runs.

What the bound rules out. The bound holds for every precision and for every number and arrangement of selected points. For an achieved same-budget score $F(A)$, it decides the comparison when $F(A) > U(C)$. When $F(A) \le U(C)$, equality included, the comparison stays open and no matching subset is implied. It applies while coordinates, one shared alignment, reference weights, masks, and tolerance stay the same. For narrow ranked pools the ceiling sits close to the achieved score (Table 3), so the loss relative to ALL is already fixed by membership when the sampler starts. A policy realigned on its own, as by the ICP evaluator of Appendix D, falls outside it. Spatial restrictions may differ between precision and recall provided every selected point used for recall remains in the corresponding restricted pool.

Running the test on another pipeline. Neither the matched test nor the ceiling uses anything specific to VGGT. With a per-item score and a budget K, fix the eligible set, the selector, K, and its start. Form the top-M pool and an equal-size uniform pool from the eligible set, run the selector on both and on the full set, and compare the downstream metric. If a reference defines what must be covered and the metric is recall-limited at a fixed tolerance, compute $U(\mathcal{C})$ for the ranked pool. A pool whose ceiling falls below an achieved same-budget score cannot recover it under any coordinate-preserving selector.

Derivation. For fixed reference observations G, define $H(C) = \{g \in G : \min_{x \in C} \|g - x\| < \tau\}$. A coordinate-preserving selection $S \subseteq C$ satisfies $H(S) \subseteq H(C)$, so $R(S) \le r_C = |H(C)|/|G|$. Since precision is at most one,

$$F(S) \le \frac{2R(S)}{1 + R(S)} \le \frac{2r_C}{1 + r_C} = U(C).$$

The first step uses $P \le 1$ and that F is nondecreasing in P ($\partial F/\partial P = 2R^2/(P+R)^2 \ge 0$). The second uses that $2r/(1+r)$ is increasing ($2/(1+r)^2 > 0$). We set $F = 0$ when $P + R = 0$. Empty pools have zero support and zero score, and the reference is nonempty. Nonnegative reference weights with positive total preserve the argument. In counts, $U = 2h/(n+h)$ for support count h and reference count n. Here P, R, and r are fractions, and tables report them in percent.

Decomposing recall loss. For ranked selection S, paired ALL selection A, and full ranked pool C, the signed losses obey

$$\underbrace{100[R(A) - R(S)]}_{L} = \underbrace{100[R(A) - R(C)]}_{D_R} + 100[R(C) - R(S)].$$

D_R is the loss fixed by membership before sampling. The final term need not be attainable at budget K. A negative D_R means the pool could in principle exceed the achieved ALL recall.

G SCENES, REFERENCES, AND POOLS

Both pools of the matched test share the scenes, references, and eligibility fixed here.

Scenes. On DTU (Jensen et al., 2014) we use the 21 scans of Table 19, each with its two fixed output budgets. These are the scans in the DTU release we used that have complete rectified images and scanner references, less one scan we had inspected during development. On BlendedMVS we use the seven official validation scenes present in the released archive. On ETH3D we use the ten usable scenes of the 13 available (Appendix C, ETH3D with VGGT).

References, alignment, and tolerance. On DTU every variant shares one camera-derived similarity per scan. We score precision against the scanner's observation mask and recall against scanner points above the table plane, at a strict 5 mm tolerance with in-volume predictions only (a 60 mm patch). On BlendedMVS we render reference depth from the reconstructed textured meshes (Yao et al., 2020). The tolerance is 1% of the median valid depth.

Eligibility and pools. A VGGT prediction is eligible when its coordinates and direct-head confidence are finite and its confidence exceeds 2.0. Every pool of a scene shares its FPS starts, and the random pools come from fixed seeds, so a repeat draws the same ones, so ranked and random pools differ only in membership (Table 20). The released-output study (Section 4.3) instead starts from raw same-head predictions without the confidence cut, using one camera-derived similarity per scene shared across heads and policies.

Table 19: **Per-scan DTU output budgets.** Every policy and both starts for a scan share its small and large budget.

Scan	small	large	Scan	small	large	Scan	small	large
2	31,777	63,555	3	25,498	50,996	16	26,001	52,003
17	20,515	41,029	18	26,797	53,594	19	29,460	58,920
21	24,107	48,213	22	26,298	52,596	25	22,416	44,831
28	25,806	51,612	30	39,689	79,378	31	26,429	52,857
35	24,861	49,722	36	21,662	43,325	38	20,868	41,736
39	23,762	47,523	41	29,737	59,473	42	37,092	74,184
43	22,516	45,032	44	24,216	48,433	45	24,882	49,764

H EVALUATION PROTOCOL

Table 20: **Replication specification.** Settings behind the matched test, the backbone studies, and the registration evaluator. ρ and τ are defined in Appendices D and F.

Item	Setting
Pool seeds	Seeded permutations indexed by study, population, scene, and repetition, nested across widths and budgets.
FPS starts and ties	Two shared starts per scene from the intersection of the smallest pools. Ties broken by point index, as are tied FPS distances once distinct coordinates are exhausted. Exactly K original point IDs per selection, no extra anchor point inserted.
Scene inputs	BlendedMVS: 32 uniformly spaced unmasked views per scene, output budgets of 25,000 and 50,000 points, reference depth back-projected at every fourth pixel with finite positive depth, each weighted equally. ETH3D: up to 32 uniformly selected calibrated images per scene.
Primary resampling	Starts and budgets averaged within scene, then 20,000 scene resamples for scene-level percentile intervals.
Pool-seed control	Per budget, a 21×5 matrix of paired per-scan random-minus-ranked differences. 10,000 replicates, each drawing 21 scenes, then five pool seeds per drawn scene, both with replacement. Interval: 2.5th and 97.5th percentile of replicate means.
Fixed fraction, mixed pool	Two repetitions averaged within scene, with both output budgets where both were run, before 20,000 scene resamples (10,000 for Table 7).
Camera alignment	Camera centers normalized to RMS radius 1000. Discarded if the trajectory is degenerate, the scale is nonpositive, the median residual is above 100 normalized units, or fewer than 1,000 reference observations are valid.
π^3	Finite points with sigmoid confidence above 0.1, demo edge filter omitted, the same 21 scans, budgets, and starts.
Depth Anything 3	Released DA3-LARGE checkpoint, 49 views per scan, 504-pixel preprocessing. Eligibility cut per scan at $\min(\max(1.05, c_{40}), c_{90})$, where c_p is the per-scan pth confidence percentile.
DUSt3R	Official checkpoint (DUSt3R contributors, 2026), 512-pixel preprocessing, MST (minimum spanning tree) global alignment, the 21 scans.
MapAnything	Released checkpoint, 49 views per scan, 518-pixel preprocessing, eligibility at the confidence percentile of 10 per scan.
AMB3R	$2 \times K$ pools, robust similarity camera fit, budgets of 2.5% and 5% of the candidate count, five random pools per scan.
ICP evaluator	Open3D 0.19.0 point-to-point ICP. Correspondence limits of 0.20ρ, 0.10ρ, 0.05ρ, and 2τ, each run for up to 20 iterations with default relative fitness and RMSE thresholds of 10^{-6} (Open3D, n.d.).
Compute	Eight CPU cores per dataset job, no GPU, jobs run concurrently.

Development and data. We developed the protocol on DTU and tuned no model, threshold, or baseline hyperparameter on any reported score. BlendedMVS is on VGGT's disclosed training list (Wang et al., 2025a, Sec. 3.4), so the BlendedMVS results are measured on scenes VGGT may have seen in training. ETH3D is the only one of the three outside both VGGT's training sources and our protocol development. Its ten scenes are public benchmark scenes used in earlier work.

Alignment and evaluator. We score the π^3 selected outputs through the evaluator's own downsampling, while the coverage ceiling uses each in-volume candidate, so we report the two counts separately. Within π^3 we apply it to every variant, on the achieved-F side only. The other backbones' selected outputs are scored without downsampling. It keeps a subset of the selected points, so achieved recall never exceeds the pool support and the inequality holds for the reported scores. Both sides of the ceiling inequality use the same reference set, tolerance, and weights.

Statistics. Intervals for the primary $4 \times K$ contrast resample scenes. To check that pool seeds do not drive the contrast, the pool-seed control (Table 2) also resamples pool seeds within each drawn scene. For DTU we also report the paired-seed mean and its sample SD. The fixed-fraction and mixed-pool analyses average repetitions within scene before resampling scenes. The DTU mixed-pool study runs at the smaller budget only. Because these analyses average repetitions where the primary contrast averages sampler starts, the BlendedMVS random-pool mean they report differs slightly from Table 3. The ranked pool is deterministic and identical in both.

Reproducibility. We recompute each reported table from the per-scan predictions and the settings above.

I INTERVALS

Of the main-text contrasts in Table 21, two have a 95% interval that covers zero: the BlendedMVS F-score difference at 25% retention, which Table 11 also flags, and the BlendedMVS mixed-minus-random difference of Section 4.5. Tables 1, 2, and 4–6 print point estimates, and their intervals and dispersions are below, followed by the gaps quoted in Appendices C and E. The intervals are pointwise 95% percentile bootstrap intervals, and the resampling unit is named per block. Standard deviations are sample SDs over the five pool seeds and measure seed-to-seed spread.

Multiple comparisons. Table 21 lists 38 interval contrasts, drawn from a small set of underlying comparisons (random vs. ranked vs. ALL, replicated across datasets, budgets, and backbones), so they are not 38 independent tests, and no correction for multiple comparisons is applied. The four intervals that include zero are exactly the contrasts already reported as null: BlendedMVS at 25% retention, the BlendedMVS mixed-minus-random contrast, and MapAnything's ceiling gap at both pool widths.

Table 21: **The uncertainty behind each reported contrast.** Resampled over scenes unless stated. The final column marks intervals that exclude zero.

Table	Row	Quantity	Estimate	95% interval	Excl. 0
Table 1: random minus ranked, resampled over scenes					
1	$4 \times K$ / DTU	ΔF-score	+54.10	[+49.85, +57.90]	yes
1	$4 \times K$ / DTU	Δceiling	+60.94	[+57.27, +64.48]	yes
1	$4 \times K$ / BlendedMVS	ΔF-score	+32.17	[+21.13, +42.79]	yes
1	$4 \times K$ / BlendedMVS	Δceiling	+43.66	[+34.38, +54.19]	yes
1	25% / DTU	ΔF-score	+24.67	[+20.46, +29.25]	yes
1	25% / DTU	Δceiling	+30.79	[+26.18, +35.84]	yes
1	25% / BlendedMVS	ΔF-score	+2.87	[−5.82, +12.31]	no
1	25% / BlendedMVS	Δceiling	+23.64	[+14.09, +32.91]	yes
Table 2: nested over scenes and pool seeds (BlendedMVS: scenes only); SD over five pool seeds					
2	DTU / small	ΔF-score	+56.11	[+51.94, +60.05]	yes
2	DTU / small	SD over seeds	0.037	—	—
2	DTU / large	ΔF-score	+52.09	[+47.86, +55.98]	yes
2	DTU / large	SD over seeds	0.024	—	—
2	BlendedMVS / small	ΔF-score	+35.01	[+23.74, +45.34]	yes
2	BlendedMVS / large	ΔF-score	+29.34	[+17.59, +40.99]	yes
Table 4: released mask minus each control, resampled over scenes					
4	DTU / Depth	Δvs random	+50.60	[+47.07, +54.41]	yes
4	DTU / Depth	Δvs ALL	+52.12	[+48.65, +55.84]	yes
4	DTU / Point	Δvs random	+41.66	[+36.80, +46.74]	yes
4	DTU / Point	Δvs ALL	+44.32	[+39.47, +49.30]	yes
4	BlendedMVS / Depth	Δvs random	+46.97	[+25.88, +65.78]	yes
4	BlendedMVS / Depth	Δvs ALL	+47.82	[+26.27, +66.97]	yes
4	BlendedMVS / Point	Δvs random	+42.83	[+22.06, +62.19]	yes
4	BlendedMVS / Point	Δvs ALL	+43.98	[+22.65, +63.77]	yes
Table 12: 50/50 mixed pool, paired over scenes					
12	mixed / DTU	ΔF vs ranked	$+55.65$	[51.27, 59.77]	yes
12	mixed / BlendedMVS	ΔF vs ranked	$+33.34$	[22.11, 44.00]	yes
12	mixed / DTU	ΔF vs random	-0.44	$[-0.75, -0.16]$	yes
12	mixed / BlendedMVS	ΔF vs random	$+1.20$	$[-0.07, 2.86]$	no
Table 5: ALL F-score minus ranked ceiling, resampled over scenes					
5	VGGT / DTU	Mean ceiling gap	54.21	[50.07, 57.84]	yes
5	VGGT / BlendedMVS	Mean ceiling gap	24.07	[15.78, 32.85]	yes
5	π^3 / DTU	Mean ceiling gap	19.95	[13.37, 26.65]	yes
Table 6: ETH3D, paired over scenes; SD over five pool seeds (ranked: only the FPS start varies)					
6	small / $4 \times K$	Increment	15.00 ± 0.04	[6.08, 25.45]	yes
6	small / 25% eligible	Increment	7.86 ± 0.02	[3.60, 12.82]	yes
6	large / $4 \times K$	Increment	12.83 ± 0.03	[4.80, 22.82]	yes
6	large / 25% eligible	Increment	9.56 ± 0.03	[4.98, 14.44]	yes
6	small / $4 \times K$	SD, ranked / random	0.002, 0.036	—	—
6	small / 25% eligible	SD, ranked / random	0.004, 0.015	—	—
6	large / $4 \times K$	SD, ranked / random	0.003, 0.030	—	—
6	large / 25% eligible	SD, ranked / random	0.004, 0.034	—	—
Appendices C and E: ALL F-score minus ranked ceiling unless stated, resampled over scenes					
15	DA3 / DTU	Mean ceiling gap	35.54	[27.71, 43.24]	yes
15	DUSt3R / DTU	Mean ceiling gap	19.83	[16.53, 23.05]	yes
15	DUSt3R, known poses	Mean ceiling gap	39.21	[33.01, 44.13]	yes
15	MapAnything / $4 \times K$	Mean ceiling gap	0.71	$[-4.54, 5.44]$	no
15	MapAnything / $8 \times K$	Mean ceiling gap	-4.07	$[-9.86, 1.21]$	no
15	ETH3D, 8 of 10	Mean ceiling gap	16.10	[6.63, 27.78]	yes
—	VGGT-Omega / DTU	ΔF-score	-3.13	$[-6.26, -0.79]$	yes