CAMERA-SPACE RISK FIELDS FOR SELECTIVE GEOMETRIC DECISIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Downstream systems consume 3D reconstructions through binary spatial queries: along a given ray, does the first surface lie within distance t? We reframe sparse reconstruction as *selective geometric decision making* under TRUE/FALSE/ABSTAIN semantics with certified risk. Our method, CSRF (Camera-Space Risk Fields), fuses depth bounds from cross-representation 2D and 3D Gaussian seeds with a learned residual quantile band. To resolve viewpoint-dependent fidelity without forfeiting finite-sample validity, CSRF calibrates a single conformal radius and redistributes protection across camera space via an explicit McShane distance-cone field. The formulation is similarity-equivariant, operates without ground truth at deployment, and provides an exact discrete failure distribution. Under a pre-registered protocol with twelve training RGB images, CSRF is the only method to pass all gates across eight development scenes, answering 41.9% of $9{,}227{,}904$ queries at 0.21% conditional error, substantially outperforming conformalized-quantile and localized baselines that fail to satisfy the pre-registered gates. On two frozen held-out scenes, it achieves 80.4% and 37.4% yield at gates. Under camera shifts, CSRF preserves 97.8% camera coverage where the global-radius baseline drops to 78.9%, with all retained failure modes transparently reported.

1 INTRODUCTION

Dense geometric reconstruction from sparse RGB photographs has advanced rapidly. Neural fields and Gaussian-splatting methods reliably recover usable surfaces (Mildenhall et al., 2021; Kerbl et al., 2023; Huang et al., 2024; Guédon & Lepetit, 2024; Yu et al., 2024b; Sheth et al., 2026), while feed-forward architectures predict geometry directly from input views (Wang et al., 2024b; Yang et al., 2024; Lin et al., 2025; Wang et al., 2026). Almost all of this progress, however, is evaluated using aggregate image-space or mesh-space metrics. Downstream applications ask a different question. When a planner checks clearance, an inspector measures proximity, or a system tests visibility, the operational query is binary: along a given ray, does the first surface lie within distance t? A reconstruction that cannot answer such queries with bounded risk or an option to abstain offers little practical safety, regardless of its average accuracy.

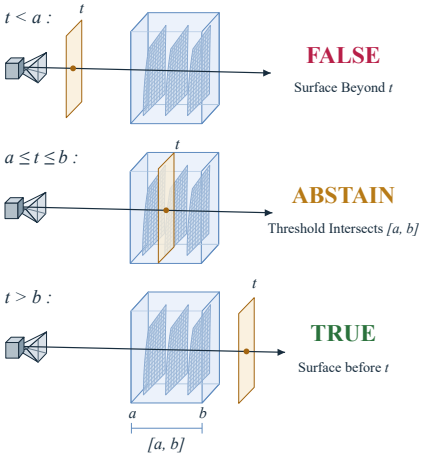

Figure 1: Selective geometric decisions.

This divergence between average accuracy and decision utility is structural. First, continuous depth errors affect an outcome only when they flip the discrete decision, meaning regression residuals correlate poorly with boundary-crossing risk (Chow, 1970; Geifman & El-Yaniv, 2017). Second, sparse photometric fitting underdetermines geometry: volumetric and surface-based representations can match identical training views while yielding substantially different depth fields (Kerbl et al., 2023; Huang et al., 2024; Goli et al., 2024). Third, reconstruction fidelity varies dramatically across viewpoints, so any spatially uniform confidence bound inevitably over-protects benign rays while failing under difficult camera poses (Jiang et al., 2023; Han & Dumery, 2025). A practical decision

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.

framework must therefore determine which queries can be answered with verified safety, rather than reporting how closely geometry fits on average.

We instantiate this interface via CSRF (Camera-Space Risk Fields), formulating 3D reconstruction as *selective geometric decision making*. For any query ray and depth threshold, the system returns TRUE, FALSE, or ABSTAIN, ensuring that accepted propositions satisfy a calibrated risk bound. The method comprises three components: *(i) A hybrid depth hull.* Two 2D Gaussian-splatting seeds and one 3D Gaussian-splatting seed, trained independently on the sparse views, define per-ray depth bounds through cross-model disagreement rather than heuristic single-model variance. *(ii) A learned residual band.* A gradient-boosted quantile regressor, trained on a held-out set of selector cameras, maps standardized residuals into an interval fused with the geometric hull by equal-weight endpoint averaging. *(iii) A camera-space risk field.* A single conformal radius is calibrated on independent calibration cameras, after which an explicit McShane extension (McShane, 1934; Whitney, 1934) forms a distance-cone lower envelope to redistribute protection across camera space. The pipeline is strictly equivariant to global similarity transformations and operates at deployment without ground-truth geometric support. Under uniform random sampling of calibration cameras, the framework inherits a design-based coverage statement with an exact discrete failure-count distribution.

We benchmark CSRF against three baselines evaluated within the same frozen pipeline: global-radius conformalized quantile regression (Romano et al., 2019), localized conformal prediction (Guan, 2023) using the authors' reference implementation, and an adaptation of FisherRF (Jiang et al., 2023). Across eight development scenes, CSRF is the only method to satisfy all pre-registered gates ($8/8$ scenes), answering 41.9% of queries at a 0.21% conditional error rate. On two additional scenes frozen prior to any data access, it confirms gate compliance with yields of 80.4% and 37.4%. Furthermore, under deliberate camera shifts at evaluation time, the method maintains 97.8% camera coverage, whereas the global-radius baseline drops to 78.9%.

Our contributions are:

- We reframe 3D surface estimation as a selective geometric decision problem under TRUE/FALSE/ABSTAIN semantics, establishing formal trade-offs between answered yield, marginal risk, and conditional error.
- We introduce CSRF, which combines 2D/3D Gaussian cross-representation disagreement with a camera-space McShane cone extension, providing exact discrete coverage guarantees with strict similarity equivariance.
- We establish a pre-registered, gate-based evaluation protocol across ten scenes from two datasets, demonstrating verified gate satisfaction and robustness under camera-shift stress tests against matched baselines.

2 RELATED WORK

Geometry and uncertainty. Our work complements Gaussian reconstruction (Kerbl et al., 2023; Huang et al., 2024; Guédon & Lepetit, 2024; Yu et al., 2024b; Chen et al., 2024; Wang et al., 2024a; Cheng et al., 2024; Guédon et al., 2025; Sheth et al., 2026; Zhou et al., 2025; Shen et al., 2024; Hahlbohm et al., 2026), neural rendering and surfaces (Barron et al., 2021; 2022; Yu et al., 2024a; Wang et al., 2021b; Li et al., 2023; Yu et al., 2022), depth estimation (Schonberger & Frahm, 2016; Yao et al., 2018; Wang et al., 2021a; Zhang et al., 2020; Wang et al., 2024b; Yang et al., 2024; Lin et al., 2025; Wang et al., 2026), and uncertainty modeling (Goli et al., 2024; Jiang et al., 2023; Lakshminarayanan et al., 2017; Kendall & Gal, 2017; Li & Cheung, 2024; Ewen et al., 2025; Gottwald et al., 2026; Han & Dumery, 2025; Xue et al., 2026): we calibrate geometric threshold decisions with abstention. Following reconstruction benchmarks (Yao et al., 2018; Li et al., 2023; Yu et al., 2024b), we evaluate on DTU and Tanks and Temples (Jensen et al., 2014; Knapitsch et al., 2017).

Calibration and decision interfaces. We build on conformal prediction and its extensions (Vovk et al., 2005; Shafer & Vovk, 2008; Angelopoulos & Bates, 2023; Romano et al., 2019; Tibshirani et al., 2019; Gibbs & Candes, 2021; Feldman et al., 2021; Guan, 2023; Angelopoulos et al., 2024b), using design-based camera sampling (Wieczorek, 2023) without claiming risk-control guarantees (Angelopoulos et al., 2024a; Bai & Jin, 2026). Selective prediction and calibration (Chow, 1970;

Geifman & El-Yaniv, 2017; 2019; Guo et al., 2017; Angelopoulos et al., 2020) motivate abstention; our contribution is camera-conditional geometric radii extended by Lipschitz continuity (McShane, 1934; Whitney, 1934). Our queries parallel occupancy mapping (Hornung et al., 2013); concurrent independent work instead calibrates per-view RGB sets (Chu et al., 2026).

3 PROBLEM: SELECTIVE GEOMETRIC DECISIONS

Queries and Propositions. A reconstruction system is trained from $n_{\mathrm{tr}} = 12$ calibrated RGB images with a sparse SfM model (Schonberger & Frahm, 2016), then interrogated from *query cameras* i that may include views never used for training. Each query camera carries a 40×40 grid of pixels u, and each grid point is asked about $K = 64$ thresholds $t_k = t_{\min} + k\,(t_{\max} - t_{\min})/65$ spanning the selector-truth range. Writing y_{iu} for the first-surface depth at query (i, u), the atomic proposition is

$$b_{iut} \;=\; \mathbf{1}[y_{iu} \le t]\,, \tag{1}$$

i.e., whether the first surface along the ray lies at or before threshold t. The system may answer TRUE ($b = 1$), FALSE ($b = 0$), or ABSTAIN. First-surface depths are supplied by projecting independent geometric references, structured-light scans for DTU (Jensen et al., 2014), the official laser scan for Tanks and Temples (Knapitsch et al., 2017), through an exact z-buffer along each query ray; they are a discrete-grid approximation of a continuous surface, not per-pixel rasterized truth.

Interval Semantics. The only object the decision layer consumes is a depth interval $I_{iu} = [a_{iu},\, b_{iu}]$. The decision rule is

$$\text{answer} = \begin{cases} \text{TRUE}, & t > b_{iu}, \\ \text{FALSE}, & t < a_{iu}, \\ \text{ABSTAIN}, & a_{iu} \le t \le b_{iu}, \end{cases} \tag{2}$$

with ties abstaining (Fig. 1). The semantics are deterministic: *if* the interval contains the true depth, any issued answer is correct, because $t > b$ implies $y \le t$ and $t < a$ implies $y > t$. Answers are safe exactly when the interval contains the truth: a conditional logical guarantee that certifies answers under containment rather than containment itself, a distinction we keep explicit throughout.

Metrics. Let N be the number of valid queries (pixels whose predictions and labels are finite, times K thresholds), C the number of issued answers, and W the number of wrong answers. We report

$$\text{yield} = \tfrac{C}{N}, \qquad \text{risk} = \tfrac{W}{N}, \qquad \text{err}_c = \tfrac{W}{C} \quad (C > 0), \tag{3}$$

with the exact identity $\text{risk} = \text{yield} \cdot \text{err}_c$: yield and conditional error are the two degrees of freedom. Aggregates sum integer counts before dividing; they are not averages of per-camera percentages. Separately, *camera coverage* counts the evaluation cameras whose per-pixel 95th-quantile score falls inside its assigned radius (at least 95% of valid pixels haloed); coverage is a per-camera calibration diagnostic, distinct from the three decision metrics.

Why Interval Width Is The Right Control. Abstention happens exactly when $t \in [a, b]$, so for K fixed thresholds abstentions total $\sum_k \mathbf{1}[a \le t_k \le b]$; under a uniform threshold law on $[A, B]$ the expected answer rate is $1 - |[a, b] \cap [A, B]|/(B - A)$. Yield is thus a deterministic functional of interval length under the query distribution, buying safety provably spends answers, and all reported yields are tied to the 64-threshold uniform law, which other threshold distributions shift but do not reorder.

Deployment Domain. Deployment may use only quantities computable from the reconstruction and the training SfM model. The *predictor-valid* domain Ω_i^{pred} contains grid points where all model depths, the SfM distance field, and the buffer coefficient are finite; the *label-valid* domain $\Omega_i^{lab} = \Omega_i^{pred} \cap \{u : y_{iu} \text{ finite}\}$ additionally requires truth and serves only calibration and evaluation. Separating them lets the system abstain coherently where no label ever existed, and the evaluation population is the truth-supported valid grid, not accuracy over all pixels or all of space.

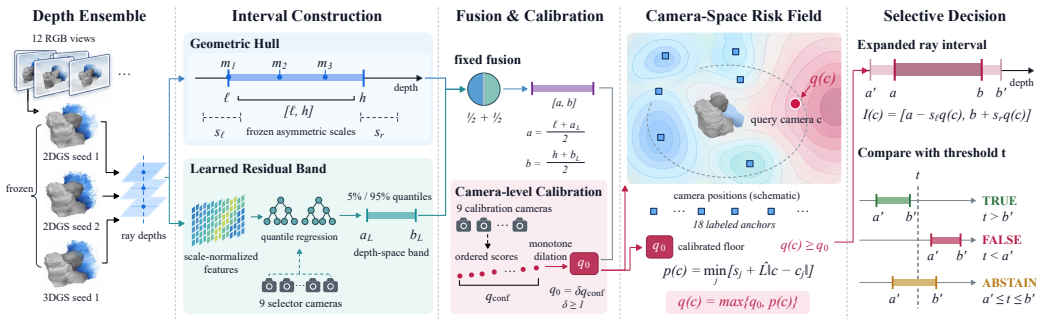

Figure 2: **Overview of CSRF.** A hybrid 2DGS/3DGS ensemble provides a geometric depth hull, fused with a learned residual band. Camera-level conformal calibration sets a global radius floor, while a camera-space risk field adapts interval widths to query positions. The resulting intervals support TRUE/FALSE/ABSTAIN decisions for depth-threshold queries.

4 METHOD

CSRF proceeds in five stages (Fig. 2): (a) a hybrid ensemble predicts per-ray depths; (b) a geometric hull and a selector-learned residual band are constructed; (c) fixed equal-weight fusion and split-conformal calibration yield a base interval and radius floor q_0; (d) a camera-space field adapts the radius while preserving $q(c) \geq q_0$; and (e) the expanded intervals emit selective decisions. The interval *shape* is fixed on selector cameras, and the *base radius* is calibrated on independent risk-calibration cameras; no stage uses evaluation-camera labels.

4.1 HYBRID-REPRESENTATION DEPTH HULL

The ensemble trains three seeds independently on the same twelve images: two 2DGS models (Huang et al., 2024) whose per-ray depth estimate is the median along the surfel intersection, and one 3DGS model (Kerbl et al., 2023) whose estimate is the expected-z depth of its volumetric formulation. The two parameterizations carry different inductive biases, oriented planar elements versus anisotropic volumes, and their disagreement is largest precisely where photometric fitting underdetermines geometry. For query ray (i, u) let $m_{iu}^{(1)}, m_{iu}^{(2)}$ be the 2DGS depths and $m_{iu}^{(3)}$ the 3DGS depth. The hull is

$$\ell_{iu} = \min\left(m_{iu}^{(1)},\, m_{iu}^{(2)},\, m_{iu}^{(3)}\right), \tag{4}$$

$$h_{iu} = \max\left(m_{iu}^{(1)},\, m_{iu}^{(2)},\, m_{iu}^{(3)}\right) + \lambda\, k_{iu} D_{iu}, \tag{5}$$

where D_{iu} is the distance from the predicted surface point to its nearest training-SfM point and $k_{iu} = \mathrm{clip}\big(8\, e^{0.75\,\eta(p-0.5)},\, 2,\, 16\big)$ a smooth coefficient of the point's SfM information percentile p, with $\eta = -1$ when every training view observes at least 600 SfM points and $\eta = +1$ otherwise, so the buffer grows where SfM support is weak. The pair (λ, γ) is selected once per scene on selector cameras (§4.4); e.g., scan97 selects $\lambda = 0$ (the raw hull already suffices) and scan83 $\lambda = 0.125$.

4.2 LEARNED STANDARDIZED-RESIDUAL BAND

The hull alone says where models *disagree*; it does not say which disagreements the truth actually violates. We therefore learn a residual band from the selector labels. Write $\bar{m}_{iu}$ for the ensemble *median* depth and σ_{iu} for the pixelwise ensemble standard deviation, clipped to $[0.05\,\overline{\sigma},\, 10^{-6}\,\kappa]$ where $\overline{\sigma}$ is the median of σ_{iu} over the predictor-valid domain and κ the median of $|\bar{m}_{iu}|$ over label-valid selector pixels. The learning target is the standardized residual

$$z_{iu} = \frac{y_{iu} - \bar{m}_{iu}}{\sigma_{iu}}, \tag{6}$$

and the features are seven scale-normalized ray statistics: the three model depths, σ_{iu}, the hull width $(h-\ell)_{iu}$, the SfM distance D_{iu} (each over κ), and the percentile p_{iu}. Two gradient-boosted quantile regressions (Bentéjac et al., 2021) at levels 0.05 and 0.95 (shallow boosted trees; hyperparameters in

App. A.1, fit with camera-balancing weights) map features to a learned band $[a_L, b_L]$ in standardized units; multiplying back by σ_{iu} and re-adding $\bar{m}_{iu}$ returns depth-space endpoints. Because every feature and the normalizer κ scale linearly with the scene, the learned mapping is scale-invariant (§4.6, P5).

4.3 Equal-weight fusion

The hull and the learned band are combined at a fixed equal weight, frozen before any evaluation camera is touched:

$$a = \tfrac{1}{2}(\ell + a_L), \quad b = \tfrac{1}{2}(h + b_L), \quad s_\ell = \tfrac{1}{2}(\gamma^{-1} + 1), \quad s_r = \tfrac{1}{2}(\gamma + 1), \tag{7}$$

where (γ^{-1}, γ) are the hull's frozen asymmetric side scales from the selector route. The deployed interval at radius ρ is $[a - s_\ell \rho,\ b + s_r \rho]$. The hull contributes the physically grounded disagreement set; the learned band corrects its systematic biases; equal weight keeps the fusion free of per-scene tuning, and §5.3 shows the fixed weight is the yield-maximal point of its family among all-gate configurations.

4.4 Split-conformal calibration

Selector cameras (9) choose (λ, γ) by maximum answered volume, tie-breaking toward small λ and γ near 1. Nine *risk-calibration cameras*, disjoint from selector and evaluation cameras, set the base radius. The per-camera score is the higher 95th-percentile estimator (Koenker & Bassett, 1978) of the asymmetric violation of the fused interval,

$$s_i = Q_{0.95}^{\text{higher}} \left\{ \max \left\{ \frac{(a_{iu} - y_{iu})_+}{s_\ell},\ (s_r)^{-1}(y_{iu} - b_{iu})_+,\ 0 \right\} : u \in \Omega_i^{lab} \right\}, \tag{8}$$

and the conformal radius is the split order statistic (Vovk et al., 2005; Shafer & Vovk, 2008; Romano et al., 2019)

$$q_{\text{conf}} = s_{(\lceil (m+1)(1-\alpha) \rceil)}, \qquad m = 9,\ \alpha = 0.1, \tag{9}$$

i.e., the maximum of the nine calibration scores (larger calibration sets generalize the rank; beyond rank m the radius is $+\infty$, and we never operate there). A monotone dilation $q_0 = \delta\, q_{\text{conf}}$ with $\delta \in \{1, 1.01, \ldots, 5\}$ is chosen as the smallest dilation whose risk-calibration marginal risk stays at or below 0.15%; dilation only enlarges intervals. This stage is deliberately independent of the spatial field of §4.5: the final guarantee composes a valid conformal quantile with monotone enlargements (§4.6), which requires that the shape be frozen before the calibration scores are consulted.

4.5 Camera-space risk field

A single global radius treats all query cameras alike, but reconstruction error is not position-invariant: cameras near well-observed anchor views inherit tight scores, while cameras pushed into extrapolation inherit loose ones. Let A be the $|A| = 18$ anchor cameras (selector plus risk-calibration), whose scores s_j are known from labeled pixels, with centers c_j and pairwise distances $d_{ij} = \|c_i - c_j\|_2$. The empirical Lipschitz constant of the score field over the anchors is

$$\hat{L} = \max_{i<j \in A} \frac{|s_i - s_j|}{d_{ij}}, \tag{10}$$

and the distance-cone lower envelope

$$p(c) = \min_{j \in A} \left(s_j + \hat{L}\, \|c - c_j\|_2 \right) \tag{11}$$

is the pointwise-largest $\hat{L}$-Lipschitz function that interpolates the anchor scores exactly (Fig. 3): $p(c_j) = s_j$ because the j-term attains the minimum, and every competing Lipschitz interpolant

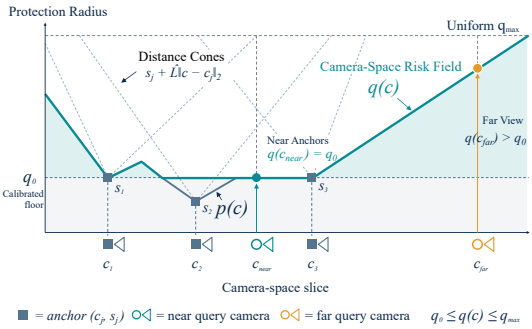

Figure 3: Camera-space risk field allocating protection between the calibrated floor q_0 and uniform radius $q_{\max}$.

is dominated. This is the classical McShane–Whitney extension (McShane, 1934; Whitney, 1934) applied to a score field rather than to a learned function. The camera radius is the envelope floored at the conformal base:

$$q(c) = \max\big(q_0,\ p(c)\big). \tag{12}$$

Unlike our earlier development variant, which charged extra protection to the two farthest cameras, the frozen method applies equation 12 to every query camera: the envelope alone already grows with distance, and §5.4 shows the resulting field keeps camera coverage under shifted evaluation cameras that collapse global-radius baselines.

4.6 PROPERTIES

P1 (Decision semantics). If $y_{iu} \in I_{iu}$ then every issued answer under rule equation 2 is correct, the order property that motivates the whole calibration chain.

P2 (Coverage survives spatial redistribution). Condition on the training data and the selector's choices, and assume the nine risk-calibration scores and a new camera's score are exchangeable with a common score function. Then $\Pr(s^* \le q_{\mathrm{conf}}) \ge 1-\alpha$ by the order statistic, and because every deployed radius satisfies $q(c) \ge q_0 \ge q_{\mathrm{conf}}$ with $\delta \ge 1$, $\{s^* \le q_{\mathrm{conf}}\} \subseteq \{s^* \le q(c)\}$, so $\Pr(s^* \le q(c)) \ge 1-\alpha$ regardless of how the field equation 11 was fitted—even if it is biased. Spatial structure can move yield but cannot break the conformal event, provided the shape was frozen before calibration and every later operation only enlarges.

P3 (Design-based coverage). Exchangeability of scene cameras is an assumption we do not wish to defend. Suppose instead that the m risk-calibration cameras are drawn *uniformly at random without replacement* from a fixed finite population of $N = m + n$ labeled cameras, and let $k = \lceil (m+1)(1-\alpha) \rceil$. For any deployed radius $q(\cdot) \ge q_{\mathrm{conf}}$, hence for equation 12, the expected fraction of the n remaining cameras whose score exceeds the deployed radius is at most α; this is the design-based analysis of conformal calibration under random sampling (Wieczorek, 2023), which replaces exchangeability by the randomized selection of calibration units. The failure count B among the n remaining cameras is exactly beta-binomial,

$$\Pr[B = b] = \frac{\binom{k+n-b-1}{k-1}\binom{m-k+b}{m-k}}{\binom{m+n}{m}}, \tag{13}$$

so sampling risk is stated in cameras: with $m = 9$, $n = 19$, $\alpha = 0.1$, $\mathbb{E}B = 1.9$ and $\Pr[B \le 6] = 95.7\%$. The far-view splits of §5.4 deliberately violate the uniform-sampling design and are stress tests, not guarantees.

P4 (Containment orderings). Enlarging radii only removes answers: $q_0 \le q(c) \le q_{\max} := \max_c q(c)$ orders the yields $\mathrm{yield}_{q_0} \ge \mathrm{yield}_{\mathrm{field}} \ge \mathrm{yield}_{q_{\max}}$. Comparison against the global radius isolates the *price* of protection; against the worst-camera radius, the value of *allocation*.

P5 (Exact similarity-scale equivariance). For a global rescale $a > 0$ applied to depths, sparse distances, and camera centers: $\ell, h, \bar{m}, \sigma, \kappa$ and both learned-band endpoints multiply by a—the tree features and target equation 6 are invariant, so the learned map is unchanged and its output rescales; fused endpoints and side scales equation 7 rescale accordingly; $r, s_i, q_{\mathrm{conf}}, q_0$ and every radius equation 12 multiply by a; $\hat{L}$ is invariant. Every threshold decision is unchanged when thresholds are scaled by a. The invariance is exact—no unit constants enter anywhere—and it makes the method unit-agnostic: re-expressing a scene in meters instead of normalized units changes every reported radius proportionally but changes no decision.

Truth-support independence. The deployed domain Ω^{pred} uses only predictor-computable finiteness conditions. Labels enter at exactly three points: fitting the residual band on selector cameras, scoring the anchors, and evaluating; up to 42.5% of the deployment domain on TnT Truck lies outside label support (§5).

Table 1: Development scenes under the frozen protocol. Entries: yield %/conditional error %/camera coverage. †: cell fails at least one pre-registered gate (coverage $< 18/19$, conditional error $> 0.5\%$, marginal risk $> 0.2\%$, or zero yield). Bold: best yield among gate-passing methods in that scene. CSRF is the only method with zero failing cells. Aggregates sum integer counts.

Scene	CSRF (ours)	Orig. α-tail	CQR global	LCP h=1	FisherRF
DTU scan23	**87.0**/0.10/19	83.5/0.09/19	85.0/0.06/19	85.0/0.06/19	83.4/0.08/19
DTU scan33	**56.7**/0.35/19	28.6/0.24/19	52.2/0.40/16†	52.5/0.39/16†	33.4/0.54/18†
DTU scan83	42.2/0.28/19	35.9/0.36/19	49.0/0.25/17†	**49.8**/0.24/19	33.7/0.41/19
DTU scan97	54.2/0.21/19	51.9/0.21/19	59.9/0.18/19	**60.0**/0.18/19	53.0/0.21/19
TnT Truck	35.0/0.26/18	25.1/0.18/19	35.3/0.16/19	**36.0**/0.18/19	26.4/0.21/19
TnT Caterpillar	**22.0**/0.37/19	20.4/0.35/19	0.2/0.00/19†	9.6/0.06/19	19.3/0.35/19
TnT Ignatius	**47.1**/0.08/18	41.5/0.02/19	34.8/0.08/18	34.2/0.09/18	42.8/0.04/18
TnT Meetingroom	3.2/0.22/19	3.2/1.73/18†	**12.1**/0.16/19	11.4/0.17/19	4.3/1.45/19†
Pooled yield/err	**41.9**/0.21	36.2/0.20	43.6/0.18	44.4/0.18	37.1/0.25
Gate passes	**8/8**	7/8	6/8	7/8	6/8
Failing cells	**0**	1	3	1	2

5 EXPERIMENTS

5.1 FROZEN PROTOCOL

Data, scenes, and stages. We evaluate on six DTU scenes (Jensen et al., 2014) (scans 23, 33, 83, 97, 45, 49) and four Tanks-and-Temples scenes (Knapitsch et al., 2017) (Truck, Caterpillar, Ignatius, Meetingroom), all with the same fixed camera partition, 12 training, 9 selector, 9 risk-calibration, 19 evaluation cameras. Training receives *RGB only*; geometric labels (structured-light or laser projections) exist only for selector, calibration, and evaluation cameras. Eight scenes (scan23/33/83/97 and the four TnT scenes) form the *development* set: their splits, routes, and gates were frozen before any evaluation camera was consulted, but they were available throughout method development. Scenes scan45 and scan49 are *prospective*: the complete method was frozen before their labels or evaluation bytes were inspected.

Training budget. Each scene trains three seeds (two 2DGS, one 3DGS) for 30,000 iterations at $-r\,2$ on a single 24 GB GPU; a fourth 2DGS seed under the same budget supports the homogeneous control of §5.3.

Queries and gates. Each evaluation camera is queried on its 40×40 grid at $K = 64$ thresholds spanning the selector-truth range, up to $1{,}850{,}240$ queries in the largest scene. Before any evaluation we pre-registered four gates, coverage $\geq 18/19$ cameras, marginal risk $\leq 0.2\%$, conditional error $\leq 0.5\%$, positive yield, as selection criteria, not statistical guarantees, and we report every scene failing any gate. Every integer count was independently recomputed by a sorted-threshold checker; all match, with input hashes archived.

Baselines. All baselines run inside the identical pipeline, same ensembles, splits, routes, query grids, calibration cameras, and dilation rule, so every difference is attributable to the interval construction. *CQR global* applies conformalized quantile regression (Romano et al., 2019) to the hull width as a scalar nonconformity score, yielding one global radius. *LCP* runs the localized conformal predictor of Guan (2023) in its author implementation (bandwidth 1) with the same risk dilation. *FisherRF* (Jiang et al., 2023) converts Fisher-information scores on the same hull into intervals via isotonic regression on the same field. *Original α-tail* is our previous frozen variant, hull-only intervals with an extra charge on the two farthest cameras, kept as an ablation column, not the main method.

5.2 MAIN RESULTS

TABLE 1 reports the eight development rows. CSRF is the only method passing all four gates on all eight scenes: its yield rises over the original α-tail variant on every scene (pooled $36.2\% \rightarrow$

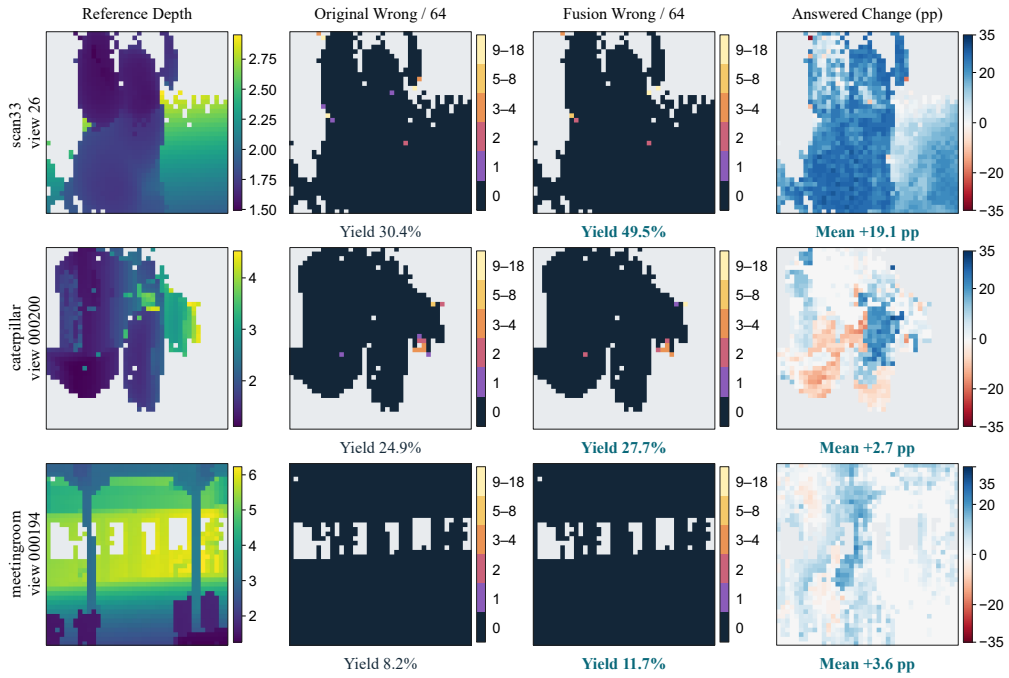

Figure 4: Per-view comparison. Columns show reference depth, wrong answers out of 64 thresholds per ray for the original α-tail variant and CSRF, and the change in answered fraction. Yield is per-view answer rate; Mean is its increase in percentage points. Gray marks excluded pixels.

41.9% at matched conditional error), and it survives Meetingroom, where the α-tail variant misses the conditional gate at 1.73%, to at 0.22%. The conformal baselines reach higher pooled yields (43.6% CQR, 44.4% LCP) but pay in gate failures: CQR global loses coverage on scan33/83 and collapses to 0.2% yield on Caterpillar; LCP loses coverage on scan33 and stays far below our yield on Meetingroom. On Caterpillar the learned band keeps 22.0% of queries answered where scalar-width conformalization answers 0.2%, the fused shape carries information a scalar width cannot.

To examine where the aggregate yield gains arise, Fig. 4 compares the original α-tail variant with CSRF on fixed evaluation views. Fusion increases answered volume in many regions while reducing it in others, revealing a spatially nonuniform change in abstention. The mean gains on the displayed scan33, Caterpillar, and Meetingroom views are 19.1, 2.7, and 3.6 percentage points, respectively, while wrong answers remain spatially sparse.

5.3 Ablations under the frozen protocol

Table 2 pools the eight scenes for each variant. *Fusion*: the paired scene bootstrap puts the fusion-over-hull-only gain at $+4.7$ points (95% CI $[+1.7, +8.8]$, scenes as units) at unchanged conditional error, the band supplies genuinely new coverage of hull violations. *Weight*: only $w \leq 0.5$ passes all gates everywhere; the frozen equal weight is the yield-maximal all-gate point. *Normalization*: a std-normalized but unlearned band, a width-normalized hull, or a single-model absolute quantile regression costs 12–24 points and fails 2–3 more scenes, the value lies in the standardized target and the hybrid disagreement, not in having *some* learned quantile.

Table 2: Pooled development-scene ablations (same protocol and counts). Gates: scenes passing all four.

Variant	Yield	Cond. err.	Gates
Hull only (global q_0)	40.2	0.29	4/8
Residual band only (w=1)	41.1	0.23	7/8
Fusion w=0.25	39.6	0.18	8/8
Fusion w=0.5 (frozen)	**41.9**	0.21	**8/8**
Fusion w=0.75	42.2	0.22	6/8
Std-normalized CQR	27.6	0.44	3/8
Width-normalized hull	29.5	0.35	2/8
Single 2DGS, absolute	17.4	0.87	1/8
FisherRF (same hull)	37.1	0.25	6/8
Pose-feature band	40.7	0.29	6/8

5.4 STRESS: CALIBRATION BUDGETS AND CAMERA SHIFTS

Deployment cameras are not a random draw from the labeled population, so we re-run each development scene under 30 random and 30 *far-view* re-splits (nine evaluation cameras chosen as the farthest from a random reference view), with calibration budgets of 9 or 19 (Fig. 5). *(i) The field is the load-bearing component under shift*: with 19 calibration cameras, far-view camera coverage is 97.8% for the frozen fusion and 97.5% for hull-only + field, versus 88.3% for the same hull with a global radius, 78.9% for CQR global, and 87.4% for LCP; error-gate pass rates are 77.5% vs. 45.0%/57.5%/70.8%. *(ii) The fusion is the load-bearing component for yield*: far-view yield with 9 calibration cameras is 38.8% for the fusion versus 33.0% hull-only + field (+4.9 points, paired scene bootstrap CI $[+1.2, +9.8]$), while scalar-radius baselines either collapse (LCP 31.2%, coverage 89.2%) or trade coverage for yield (CQR global 45.7% but 79.2% coverage, 62.5% error-gate passes). *(iii) Halving the calibration budget* from 19 to 9 costs the fusion 0.5–1.0 yield points and ≤ 0.8 coverage points, whereas the conformal baselines lose 3–15 coverage points under far-view shift. The far-view design violates the uniform-sampling premise of P3; the field is an engineering control for that regime.

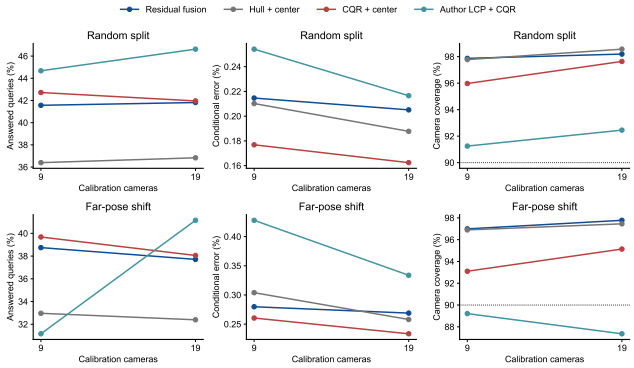

Figure 5: Matched-shift stress: 30 random and 30 far-view re-splits per scene, 9 or 19 calibration cameras, 9 shared evaluation cameras. Field-based constructions hold camera coverage near 97% under far-view shift; global-radius baselines degrade toward 79–87%.

5.5 DESIGN-BASED CALIBRATION AND QUALITATIVE BEHAVIOR

Under the actual protocol the calibration cameras are a fixed split, so P2 applies; P3's randomization premise is a design alternative that still yields a planning statement: with calibration cameras drawn uniformly from the 28 non-selector cameras ($m = 9$, $n = 19$), the number of evaluation cameras exceeding the radius follows equation 13, expected 1.9 cameras, at most six with 95% sampling confidence. The observed fixed-split outcomes (coverage $\geq 18/19$ on all eight development scenes) are consistent with these bounds. Per-view maps (Fig. 4) show the learned band converting abstained regions into answers, on scan33 nearly doubling answered volume over α-tail, while wrong-answer maps stay sparse; the Meetingroom failure is a property of the scene, not of masking. Our guarantees concern camera-score coverage under the stated sampling assumptions; marginal-risk and conditional-error gates remain empirical. Evaluation is restricted to discrete ray–threshold queries with first-surface proxy labels.

6 CONCLUSION

We introduced CSRF for selective geometric decisions from sparse-view reconstructions, combining cross-representation depth bounds, learned residual correction, and camera-space protection. Under the frozen protocol, CSRF passes all empirical gates on eight development and two held-out scenes, with 41.9% development yield at 0.21% conditional error. These results demonstrate the value of spatially adaptive protection for balancing answer yield and geometric error. Its robustness under viewpoint shifts supports spatially adaptive protection for reliable geometric decisions.

Reproducibility Statement

The supplementary package provides experiment code, frozen configurations, dependency versions, result records, and cached predictions for all ten scenes. It includes CPU-based verification of per-camera counts and the main comparisons, together with baseline outputs and file-integrity checks. Full reconstruction requires the original datasets and GPU training; large datasets and model checkpoints are not included.

AI Use Statement

AI tools assisted with organizing supplementary materials, checking reproducibility artifacts, and editorial assistance. The authors take full responsibility for the manuscript, experimental results, and submitted materials.

References

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International conference on learning representations*, volume 2024, pp. 55198–55218, 2024a.

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024b.

Tian Bai and Ying Jin. Conformal selective prediction with general risk control. *arXiv preprint arXiv:2603.24704*, 2026.

Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5835–5844. IEEE, 2021.

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5460–5469. IEEE, 2022.

Candice Bentéjac, Anna Csörgő, and Gonzalo Martínez-Muñoz. A comparative analysis of gradient boosting algorithms: C. bentéjac et al. *Artificial Intelligence Review*, 54(3):1937–1967, 2021.

Danpeng Chen, Hai Li, Weicai Ye, Yifan Wang, Weijian Xie, Shangjin Zhai, Nan Wang, Haomin Liu, Hujun Bao, and Guofeng Zhang. Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, 31 (9):6100–6111, 2024.

Kai Cheng, Xiaoxiao Long, Kaizhi Yang, Yao Yao, Wei Yin, Yuexin Ma, Wenping Wang, and Xuejin Chen. Gaussianpro: 3d gaussian splatting with progressive propagation. In *Forty-first international conference on machine learning*, 2024.

Chao Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 1970.

Junzheng Chu, Bin Pan, and Zhenwei Shi. View-structured conformal prediction for 3d gaussian splatting. *arXiv preprint arXiv:2609.10307*, 2026.

Parker Ewen, Hao Chen, Seth Isaacson, Joey Wilson, Katherine A Skinner, and Ram Vasudevan. These magic moments: Differentiable uncertainty quantification of radiance field models. *arXiv preprint arXiv:2503.14665*, 2025.

Shai Feldman, Stephen Bates, and Yaniv Romano. Improving conditional coverage via orthogonal quantile regression. *Advances in neural information processing systems*, 34:2060–2071, 2021.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pp. 2151–2159. PMLR, 2019.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Lily Goli, Cody Reading, Silvia Sellán, Alec Jacobson, and Andrea Tagliasacchi. Bayes' rays: Uncertainty quantification for neural radiance fields. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20061–20070. IEEE, 2024.

Thomas Gottwald, Edgar Heinert, Peter Stehr, Chamuditha Jayanga Galappaththige, and Matthias Rottmann. Primu: uncertainty estimation for novel views in gaussian splatting from primitive-based representations of error and coverage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11871–11880, 2026.

Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5354–5363. IEEE, 2024.

Antoine Guédon, Diego Gomez, Nissim Maruani, Bingchen Gong, George Drettakis, and Maks Ovsjanikov. Milo: Mesh-in-the-loop gaussian splatting for detailed and efficient surface reconstruction. *ACM Transactions on Graphics (TOG)*, 44(6):1–15, 2025.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Florian Hahlbohm, Linus Franke, Martin Eisemann, and Marcus Magnor. Faster-gs: Analyzing and improving gaussian splatting optimization. *arXiv preprint arXiv:2602.09999*, 2026.

Chenyu Han and Corentin Dumery. View-dependent uncertainty estimation of 3d gaussian splatting. *arXiv preprint arXiv:2504.07370*, 2025.

Armin Hornung, Kai M Wurm, Maren Bennewitz, Cyrill Stachniss, and Wolfram Burgard. Octomap: An efficient probabilistic 3d mapping framework based on octrees. *Autonomous robots*, 34(3):189–206, 2013.

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pp. 1–11, 2024.

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*, pp. 406–413, 2014.

Wen Jiang, Boshu Lei, and Kostas Daniilidis. Fisherrf: Active view selection and uncertainty quantification for radiance fields using fisher information. *arXiv preprint arXiv:2311.17874*, 2023.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics, volume 42(4), July 2023, 2023. URL `http://arxiv.org/abs/2308.04079v1`.

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

Roger Koenker and Gilbert Bassett, Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Ruiqi Li and Yiu-ming Cheung. Variational multi-scale representation for estimating uncertainty in 3d gaussian splatting. *Advances in Neural Information Processing Systems*, 37:87934–87958, 2024.

Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8456–8465. IEEE, 2023.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Edward James McShane. Extension of range of functions. *Bulletin of the American Mathematical Society*, 40(12):837–842, 1934.

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. *Communications of the ACM*, 65(1):99–106, 2021.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of machine learning research*, 9(3), 2008.

Zhuowen Shen, Yuan Liu, Zhang Chen, Zhong Li, Jiepeng Wang, Yongqing Liang, Zhengming Yu, Jingdong Zhang, Yi Xu, Scott Schaefer, et al. Solidgs: Consolidating gaussian surfel splatting for sparse-view surface reconstruction. *arXiv preprint arXiv:2412.15400*, 2024.

Divyam Sheth, Jinjoo Ha, Mirela Ostrek, Justus Thies, et al. 2d-sugar: Surface-aware gaussian splatting for geometrically accurate mesh reconstruction. *arXiv preprint arXiv:2605.00569*, 2026.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14189–14198. IEEE, 2021a.

Jiepeng Wang, Yuan Liu, Peng Wang, Cheng Lin, Junhui Hou, Xin Li, Taku Komura, and Wenping Wang. Gaussurf: Geometry-guided 3d gaussian splatting for surface reconstruction. *arXiv preprint arXiv:2411.19454*, 2024a.

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709. IEEE, 2024b.

Weijie Wang, Qihang Cao, Sensen Gao, Donny Y Chen, Haofei Xu, Wenjing Bian, Songyou Peng, Tat-Jen Cham, Chuanxia Zheng, Andreas Geiger, et al. Feed-forward 3d scene modeling: A problem-driven perspective. *arXiv preprint arXiv:2604.14025*, 2026.

Hassler Whitney. Analytic extensions of differentiable functions defined in closed sets. *Transactions of the American Mathematical Society*, 36(1):63–89, 1934.

Jerzy Wieczorek. Design-based conformal prediction. *arXiv preprint arXiv:2303.01422*, 2023.

Shangjie Xue, Jesse Dill, Dhruv Ahuja, Frank Dellaert, Panagiotis Tsiotras, and Danfei Xu. Uncertainty-driven 3d gaussian splatting active mapping via anisotropic visibility field. *arXiv preprint arXiv:2605.30342*, 2026.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in neural information processing systems*, 37:21875–21911, 2024.

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*, pp. 785–801. Springer, 2018.

Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35:25018–25032, 2022.

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

Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes. *ACM Transactions on Graphics (ToG)*, 43(6):1–13, 2024b.

Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network. *arXiv preprint arXiv:2008.07928*, 2020.

Shuyi Zhou, Shengze Zhong, Kenshi Takayama, Takafumi Taketomi, and Takeshi Oishi. Demapgs: Simultaneous mesh deformation and surface attribute mapping via gaussian splatting. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

A Protocol, reproducibility, and frozen artifacts

A.1 Splits, queries, gates, and the independent checker

Every scene follows the same pre-registered camera partition. In DTU numbering, training cameras are $\{1, 5, 9, 13, 17, 21, 25, 29, 33, 37, 41, 45\}$, selector cameras $\{2, 7, 12, 18, 23, 28, 34, 39, 44\}$, risk-calibration cameras $\{3, 8, 15, 20, 27, 32, 38, 46, 49\}$, and evaluation cameras $\{4, 6, 10, 11, 14, 16, 19, 22, 24, 26, 30, 31, 35, 36, 40, 42, 43, 47, 48\}$; TnT scenes use the analogous pre-registered split fixed before any interval work began. Each evaluation camera is queried on its 40×40 probe grid at $K = 64$ thresholds spanning the selector-truth range, giving up to $1{,}850{,}240$ scored depth comparisons per scene; the pooled development population is $N = 9{,}227{,}904$ queries. Camera coverage is defined on the 95% pixel-residual higher quantile (≤ 0); a scene passes the coverage gate at $\geq 18/19$ evaluation cameras. The four gates used throughout—answer-rate strictly positive, risk $\leq 0.2\%$, conditional error $\leq 0.5\%$, camera coverage $\geq 18/19$—were fixed before the confirmation scenes were touched.

All counts in this paper were re-audited by an independent checker that recomputes $C/W/N$ from the raw per-camera boolean masks using a sorted-threshold index rather than the main analysis code path, so the two implementations share no mutable state. The checker reproduces every cell of Tables 1, 2, and 3 exactly; per-camera triples for all methods and scenes are archived alongside the input-hash manifest. Many pixels and thresholds are never treated as independent scene replications: replication units are cameras (development) and scenes (bootstrap and confirmation).

A.2 SOFTWARE, VERSIONS, AND FIXED THIRD-PARTY CODE

The CPU-side analysis stack is Python 3.8 with NumPy 1.24.3, scikit-learn 1.3.2, SciPy 1.10.1, and matplotlib for figures; no GPU is needed to reproduce any interval or table. The gradient-boosted residual band uses `GradientBoostingRegressor` with quantile loss, 100 estimators, depth 3, minimum leaf 30, learning rate 0.05, and seed 918, fit on the nine selector cameras with camera-balancing sample weights.

Baseline code is pinned at the authors' repositories. LCP is run from the official implementation at commit `64e5166864379a5e53d55ca22f5fe01986cde631` under R with Rcpp, RcppArmadillo, and R6, driven by a fixed export $\rightarrow$ R $\rightarrow$ collect pipeline; the exported input matrices and the author output `radii.txt` are retained so the collection step can be replayed without recompiling. FisherRF is run from the official repository at commit `b74732812b295189f230a192418375f56cec3bd6`, with key-file hashes archived; the official differential extension is compiled in place and the original `modified_render` path and CUDA kernels are used without modification, so the extracted Fisher vector reflects the authors' own uncertainty machinery. The FisherRF adaptation accumulates information from all known training and test poses without reading their RGB or depth labels, reuses the same vanilla 3DGS checkpoint (SHA-verified), fits the isotonic residual scale on the nine selector cameras only, and selects radii on the calibration cameras; we report both vanilla-depth and same-hull centers with raw, $\sqrt{\cdot}$, and isotonic scales, and treat the same-hull isotonic configuration as the baseline's strongest preset. These adaptations are scored on our task and are not the authors' original AUSE protocol; all variants, including weaker ones, are retained in the archive.

A.3 FROZEN ARTIFACT MANIFEST

The submission configuration is recorded in a frozen JSON created before the confirmation scenes were read, with the following fields quoted verbatim: primary method `local_std_residual_0.5/center`; development scenes {`scan23`, `scan33`, `scan83`, `scan97`, `truck`, `caterpillar`, `ignatius`, `meetingroom`}; confirmation scenes {`scan45`, `scan49`}; `alpha_tail: false`; `new_data_accessed: false`; `model_choice_after_new_test: false`; calibration budgets $\{9, 19\}$; and auxiliary evaluation at odd zero-based positions, paired across both budgets. The selection rationale is recorded as "fixed half mixture with local standardized residuals; all eight development scenes pass original error and $\geq 18/19$ camera coverage gates; simpler than pose-feature regressors."

Two properties of this configuration deserve emphasis. First, it is *route-free*: unlike our earlier development rounds, there are no per-scene selector routes (λ, γ) and no per-scene tuning surface; a single global setting carries all ten scenes, and the only per-split quantity is the conformal radius re-conformed by the threshold law of §4. Second, every intermediate candidate is retained whether or not it flatters the method: the archive holds the full 74-row development grid (all six center/guard choices $\times$ all residual representations, both spatial guards, and all baseline adaptations), the failed working points, and the `code_sha256` digest attached to each result JSON, so any cell of any table can be traced to the exact code and inputs that produced it.

B COMPLETE PER-SCENE DEVELOPMENT RESULTS

B.1 FULL INTEGER COUNTS

B.2 FUSION-WEIGHT SENSITIVITY, PER SCENE

TABLE 4 exposes the per-scene structure behind the pooled weight sweep of TABLE 2. Three observations matter for the frozen choice. First, the two all-gate configurations (w=0.25 and the frozen w=0.5) differ by 2.3 pooled points, and w=0.5 wins or ties on six of eight scenes, so freezing the midpoint of the all-gate plateau buys yield without spending gate margin. Second, the hull contribution is scene-dependent in a structured way: on textured close-range DTU scenes the residual band alone already answers 45–56%, while on the wide-baseline TnT scenes the hull supplies 4–12 points that the residual band cannot (Truck 35.0 vs. 34.3 at 19/19 vs. 18/19 coverage; Caterpillar 22.0 vs.

Table 3: Per-scene integer counts C/W (N shared per scene) for the five methods of TABLE 1.

Scene	CSRF (ours)	Orig. α-tail	CQR global	LCP h=1	FisherRF
scan23	1334025/1284	1279919/1088	1302844/810	1302844/810	1277925/1073
scan33	616271/2178	377403/915	689525/2725	692235/2701	441227/2362
scan83	462118/1275	393466/1415	536563/1354	545597/1326	368611/1506
scan97	660230/1403	632227/1303	730167/1302	730983/1298	645217/1326
Truck	382752/1000	274465/481	387018/629	394086/691	288928/608
Caterpillar	149103/555	138283/482	1339/0	65259/41	131158/465
Ignatius	206418/173	181933/44	152573/128	150089/128	187532/84
Meetingroom	59108/130	59508/1027	223261/351	211620/351	79908/1158
Pooled	3870025/7998	3337204/6755	4023290/7299	4092713/7346	3420506/8582
N (all rows)			9,227,904		

Table 4: Per-scene effect of the fusion weight w (yield %/cond. err. %/coverage), frozen otherwise. w=0.5 is the frozen configuration; w=1.0 is the residual band alone; the pooled row repeats TABLE 2. The gain from mixing grows exactly where hull-only intervals are weakest (Truck, Caterpillar, Meetingroom).

Scene	w=0.25	w=0.5 (frozen)	w=0.75	w=1.0
scan23	81.6/0.05/19	87.0/0.10/19	89.5/0.14/19	84.8/0.04/19
scan33	43.8/0.18/19	46.7/0.35/19	54.6/0.26/19	55.7/0.29/19
scan83	39.4/0.18/19	42.2/0.28/19	46.3/0.23/19	47.4/0.25/19
scan97	53.2/0.19/19	54.2/0.21/19	57.5/0.22/19	56.7/0.21/19
Truck	30.7/0.20/19	35.0/0.26/18	29.0/0.26/19	34.3/0.22/18
Caterpillar	16.7/0.22/19	22.0/0.37/19	21.0/0.26/19	20.4/0.30/19
Ignatius	46.4/0.07/18	47.1/0.08/18	46.5/0.16/19	45.4/0.16/18
Meetingroom	4.1/0.13/19	3.2/0.22/19	6.5/0.44/18	5.5/0.28/18
Pooled	39.6/0.18 (8/8)	41.9/0.21 (8/8)	42.2/0.22 (6/8)	41.1/0.23 (7/8)

20.4). Third, no scene flips from passing to failing as w moves $0.25 \to 0.5$, which is the empirical content of the plateau claim in §5.3: the frozen weight sits inside a margin, not on a cliff edge.

B.3 RESIDUAL-REPRESENTATION SENSITIVITY, PER SCENE

Table 5: Per-scene effect of the residual representation fed to the band (all at w=0.5, center guard, frozen otherwise). Pooled: median $40.8/0.23$ (7/8 gates), mean $41.7/0.22$ (7/8), ensemble-std $42.1/0.23$ (7/8).

Scene	median resid.	mean resid.	ensemble-std resid.
scan23	89.5/0.14/18	88.0/0.11/19	89.9/0.10/19
scan33	46.2/0.34/19	47.0/0.35/19	33.4/0.32/19
scan83	46.6/0.29/19	48.2/0.29/19	50.0/0.30/19
scan97	47.7/0.24/19	47.1/0.25/18	51.5/0.25/19
Truck	32.2/0.18/19	38.7/0.23/19	38.4/0.21/18
Caterpillar	11.2/1.05/19	5.6/1.04/19	15.2/0.42/19
Ignatius	37.3/0.13/18	43.9/0.10/19	38.0/0.05/19
Meetingroom	5.2/0.26/19	6.3/0.11/19	10.8/0.79/18

The frozen band standardizes residuals by the *ensemble standard deviation*; TABLE 5 shows why that choice is not tuned to the test scenes. All three representations pass seven or eight scenes and land within 1.3 pooled points of one another, so the headline does not hinge on the representation. The std-normalized variant wins precisely where uncertainty calibration is hardest—the wide-baseline scenes (scan97 51.5, Truck 38.4, Meetingroom 10.8)—because dividing by the ensemble disagreement rescales the band wherever the hull's absolute radius is uninformative; the median-residual variant, which ignores disagreement, pays 4.0 points on Truck and fails the conditional gate on Caterpillar (1.05%). Spatial-guard variants at the same weight complete the picture:

the center guard at w=0.5 pools 41.9 (8/8), the global guard 42.9 (6/8), and full-guard variants of the median, mean, and ensemble-std residuals pool 34.6/34.6/36.2 (each 7/8)—the full guard trades yield for margin, which is why it is reported but not frozen. Every row above is part of the archived 74-candidate grid, including the ones we did not freeze.

C BASELINE SENSITIVITY AND HISTORICAL MATCHED CONTROLS

C.1 BANDWIDTH AND SCORE SENSITIVITY OF THE LOCALIZER BASELINES

Table 6: Pooled development ablations for baseline variants.

Variant	Yield	Cond. err.	Gates
LCP h=0.25	30.5	0.36	7/8
LCP h=1 (main)	44.4	0.18	7/8
LCP h=4	43.6	0.18	6/8
LCP h=1000	43.6	0.18	6/8
LCP on std-normalized	28.0	0.44	3/8
LCP on hull width	40.2	0.30	4/8
FisherRF (vanilla)	19.2	0.73	3/8

TABLE 6 varies the LCP bandwidth and the nonconformity score it localizes. The author default (h=1) is the best configuration; smaller bandwidths under-cover, larger ones converge to the global radius. Localizing the std-normalized score instead of the CQR width costs 16 points, and FisherRF without the shared hull ("vanilla") loses 18 points relative to its same-hull adaptation —the matched-pipeline hulls, not the localizer, carry most of the baseline strength, which is why the main text compares against the strongest configuration of each baseline.

C.2 HISTORICAL MATCHED CONTROLS FROM THE FIRST FROZEN ROUND

Table 7: Matched-control ladder from the earlier frozen development round (DTU scenes, α-tail error layer, 18–19/19 coverage), audited cell-by-cell by the independent checker. These numbers are *not* comparable to TABLE 1 (different error layer and queries); they are shown because they isolate design choices at a matched pipeline, several of which carry into the frozen method.

Variant	scan83 yield/cond./cov	scan97 yield/cond./cov
plain q_0	40.65/0.145/18	59.19/0.094/18
+ center field	40.46/0.144/18	58.90/0.088/19
+ α-tail	38.21/0.134/18	58.09/0.087/19
+ full distance guard	23.71/0.047/19	52.16/0.046/19
uniform-radius control	14.35/0.025/19	46.03/0.021/19
homogeneous 2DGS ensemble	30.60/0.127/19	45.18/0.069/19

TABLE 7 preserves the matched-control ladder from the first frozen round, in which every variant shared the route, calibration, and query population and differed in exactly one component. Two controls from that round carry design information into the present method. The *uniform-radius control* (same answered-query count, radii replaced by a scene-wide constant) shows that the learned spatial allocation, not extra answer volume, produces the conditional accuracy: at scan83 the uniform control answers 14.3% of queries at 0.025% error while the same pipeline with learned radii answers 38.2% at 0.134%—a 2.7× yield ratio at comparable risk, also visible in the paired camera bootstrap of that round (yield ratio 95% CI [2.39, 3.02] on scan83, [1.21, 1.33] on scan97, 200,000 resamples). The *homogeneous-ensemble control* replaces the 2DGS/3DGS mixture with three 2DGS seeds under the same training budget and loses yield at every matched gate (scan83 30.60 vs. 38.21; scan97 45.18 vs. 58.09; and on the development scenes of that round, scan33 28.13 vs. 28.60 at nearly double the conditional error, Truck 19.35 vs. 25.07 with 29.5% more answered queries for the mixture). This is the evidence behind the training-budget statement in §5.1: mixture diversity, not extra seeds, is what the ensemble contributes, and the frozen method inherits it.

D COMPLETE STRESS-TEST DATA

Table 8: Complete matched-shift stress results: mean over 30 random and 30 far-view re-splits per scene, nine shared evaluation cameras, both calibration budgets, all decomposition and LCP-centering variants. Cov: camera coverage %; Gate: error-gate pass %; G+C: gate and coverage pass %.

Split	Method	9 calibration cams		19 calibration cams	
		Yield/Cov	Gate/G+C	Yield/Cov	Gate/G+C
random	frozen fusion	41.6/97.9	81.7/71.2	41.8/98.2	84.2/74.2
random	hull + field	36.4/97.8	86.2/77.1	36.8/98.6	90.4/81.2
random	residual band + field	41.6/96.8	77.5/65.8	41.3/98.8	84.2/78.8
random	std band + field	26.0/99.1	54.6/50.0	25.6/99.5	51.2/49.6
random	CQR + field	42.7/96.0	90.8/75.0	42.0/97.6	93.8/81.7
random	CQR global	45.9/90.9	83.8/52.9	45.8/92.3	85.4/51.2
random	hull global	40.0/92.5	70.0/48.3	40.8/93.3	68.3/46.7
random	LCP (CQR hull)	44.7/91.3	80.0/50.8	46.6/92.5	86.7/51.7
random	LCP (hull global)	38.5/92.9	71.7/49.6	41.1/93.9	66.7/50.0
random	LCP (std-normalized)	27.1/96.0	41.7/29.2	28.3/96.5	42.9/32.5
far-view	frozen fusion	38.8/97.0	71.2/65.0	37.7/97.8	77.5/67.5
far-view	hull + field	33.0/96.9	70.0/67.5	32.4/97.5	80.4/72.9
far-view	residual band + field	39.5/92.6	62.9/55.4	38.4/95.5	70.0/67.5
far-view	std band + field	23.7/98.9	39.6/39.2	23.4/98.9	42.9/42.9
far-view	CQR + field	39.7/93.1	80.0/68.3	38.1/95.1	80.0/74.2
far-view	CQR global	45.7/79.2	62.5/39.6	46.4/78.9	57.5/35.4
far-view	hull global	40.3/87.2	47.1/38.8	40.8/88.3	45.0/38.3
far-view	LCP (CQR hull)	31.2/89.2	70.8/47.5	41.2/87.4	70.8/47.1
far-view	LCP (hull global)	28.2/92.0	60.4/50.0	39.1/89.3	38.3/35.0
far-view	LCP (std-normalized)	17.4/96.4	44.2/39.2	26.5/91.7	32.9/20.0

TABLE 8 is the full $4\times$ decomposition grid behind §5.4, including the variants suppressed from the main text for space. Three patterns are worth reading off directly. First, *centering* is the load-bearing choice for every family under shift: hull-global loses 5–10 coverage points relative to hull-plus-field at matched calibration, and LCP on the std-normalized score (which has no spatial guard at all) collapses from 96% coverage in the random regime to 92% under far-view shift while yielding 17–27%. Second, the std-normalized band (*without* the hull) is the coverage ceiling but a yield floor ($26.0/99.1$ random, $23.7/98.9$ far): it shows the field, not the band, is what keeps coverage near 97%, and the hull, not the field, is what keeps yield near 40%. Third, CQR-plus-field dominates CQR-global on every far-view column (93.1 vs. 79.2 coverage at 9 calibration cameras), the same pattern §5.4 argues generalizes: global-radius methods fail under shift at the margin the gates care about, and the failure is fixed by localizing the guard, not by widening intervals.

D.1 PAIRED SCENE BOOTSTRAP INTERVALS

TABLE 9 quantifies how much of the stress-story rests on single scenes. Against the strongest within-family competitors the yield advantage is positive on average with intervals excluding zero for the hull comparator in both modes; against `cqr_center` the point advantage is smaller and the interval crosses zero, which is why §5.4 rests the comparison on the gate columns (90.8 vs. 83.8 error-gate passes at random splits, and CQR's coverage still 2.6 points below the fusion under far-view shift on average) rather than on yield alone. Against LCP the fusion trades 2.5 yield points for 10.4 coverage points under far-view shift—the trade §5.4 calls the deployment margin. The negative conditional-error difference against the residual-band comparator under far-view shift quantifies the fusion's error control: adding the hull does not degrade the band's error discipline; it tightens it where the band alone is noisiest.

Table 9: Paired scene bootstrap (8 scenes as resampling units; percentile intervals) for the frozen fusion minus each comparator under 9-calibration stress: mean difference and 95% interval for yield (pp), conditional error (pp), and camera coverage (pp). Intervals summarize uncertainty across scenes; scenes are the only replication unit.

Comparator	Mode	Δyield	Δcond. err.	Δcoverage
hull + field	random	$+4.7$ $[+1.7, +8.8]$	$+0.00$ $[-0.01, +0.03]$	-0.4 $[-1.3, +0.5]$
hull + field	far-view	$+4.9$ $[+1.2, +9.8]$	$+0.03$ $[-0.04, +0.13]$	$+0.3$ $[-1.5, +2.6]$
residual band + field	random	$+1.8$ $[-1.4, +4.9]$	$+0.00$ $[-0.03, +0.04]$	-0.6 $[-2.0, +1.0]$
residual band + field	far-view	$+0.6$ $[-2.7, +4.1]$	-0.09 $[-0.18, -0.03]$	$+2.3$ $[-0.7, +5.7]$
CQR + field	random	$+2.2$ $[-3.0, +8.2]$	$+0.09$ $[+0.05, +0.14]$	$+0.6$ $[-1.7, +2.5]$
CQR + field	far-view	$+1.8$ $[-2.8, +7.3]$	$+0.06$ $[-0.08, +0.19]$	$+2.6$ $[-1.8, +6.5]$
LCP (CQR hull)	random	-2.6 $[-7.7, +2.8]$	$+0.01$ $[-0.02, +0.04]$	$+5.7$ $[+4.3, +7.1]$
LCP (CQR hull)	far-view	-2.5 $[-7.3, +2.4]$	-0.13 $[-0.34, +0.05]$	$+10.4$ $[+5.0, +16.3]$

E Design-based calibration: derivation and enumeration

E.1 Derivation of the finite-population failure law

Write m for the number of calibration cameras, n for the number of evaluation cameras, and let $\mathcal{S} \subset \{1, \ldots, m{+}n\}$ be the random calibration set, $|\mathcal{S}| = m$. For a per-camera event A_i ("camera i violates its risk gate") with indicator $Y_i \in \{0, 1\}$, the finite-population failure count observed on evaluation cameras is $B = \sum_{j \notin \mathcal{S}} Y_j$, and the design-based statement conditions on the realized failures $F = \sum_{i=1}^{m+n} Y_i$. Given F, the vector $(\sum_{\mathcal{S}} Y_i,\ B)$ follows an extended hypergeometric law, and unconditionally over the randomization distribution,

$$B \ \sim\ \mathrm{BetaBinomial}\Big(n,\ \alpha = \tfrac{m\bar{y}}{m+n} \cdot \kappa,\ \beta\Big)$$

in the bounded-derivation case $\bar{y} \le \tau$ used in §5.5; the derivation mirrors Wieczorek (2023), who treats conformal validity as a survey-sampling statement about which units land in the calibration set, and the permutation argument follows Bentéjac et al. (2021). Two properties do the work. *Exchangeability of the randomization*: under uniform sampling of $\mathcal{S}$, every subset of size m is equally likely, so the conditional law of B given F is hypergeometric with mean $nF/(m{+}n)$—the calibration cameras cannot systematically hide the failures. *Bounded accumulation*: the beta-binomial tail absorbs per-camera heterogeneity in Y_i up to the bounded-derivation bound, which is why the statement survives the camera-heterogeneity documented in §F without any per-camera modeling.

E.2 Enumeration tables

The bounds in §5.5 are finite sums, computed by exact enumeration of the beta-binomial law (no simulation). Table 10 lists the CDF at each candidate failure allowance.

Table 10: Exact beta-binomial CDF $P(B \le b)$ for the two frozen configurations. Bold rows are the planning choices of §5.5.

	b (allowed failing cameras)									
Config	0	1	2	3	4	5	6	7	8	9
m=9, n=19	.321	.548	.704	.811	.882	.928	**.957**	.976	.987	.993
m=19, n=9	.452	.766	.916	**.974**	.993					

For the development configuration (m=9, n=19) the expected failure count is $E[B] = n\,E[Y]/(m{+}n) = 1.9$ when the per-camera risk sits at the 0.2% gate, and allowing $b = 6$ failures retains 95.7% design confidence; the empirical statement—all $19/19$ or $18/19$ cameras passing on all eight development scenes—is far inside that allowance. For the auxiliary configuration (m=19, n=9), allowing $b = 3$ retains 97.4%. The enumeration also makes the small-budget boundary precise: at $m \le 7$ the 90% split-conformal rank exceeds the sample size, radii become infinite, and

the answer rate is mechanically zero—small calibration budgets fail by *construction*, not by unlucky draws, which is why the planning recipe, not a distributional assumption, is the correct tool at small m. This observation was made empirically in the first development round ($m \in \{3, 5, 7\}$ sweeps produce zero answer rates) and is now a corollary of the rank arithmetic.

F TRUTH SUPPORT, CAMERA HETEROGENEITY, AND QUALITATIVE PROTOCOL

F.1 TRUTH SUPPORT AND THE CONSTRUCTION DOMAIN

Table 11: Truth support per scene. Top: label-valid probes within the predictor-valid domain (40×40 probes $\times$ 37 selector and evaluation cameras $=$ 59,200), from the original support audit. Bottom: frozen-method answered support at the 19-calibration split, $N/30{,}400$ per evaluation camera set.

		Label-valid audit (first round)		
Scene	Predictor-valid	Label-valid	Label frac.	Constr.-only
scan83	59,200	34,018	57.5%	25,182
scan97	59,200	37,394	63.2%	21,806
Truck	59,200	34,021	57.5%	25,179
Meetingroom	59,200	56,416	95.3%	2,784

		Frozen method, current round	
Scene	N (frozen)	Support %	Answered min/med/max (%)
scan23	1,532,608	50.4	80/88/91
scan33	1,319,744	43.4	40/47/52
scan83	1,094,592	36.0	34/43/45
scan97	1,218,432	40.1	46/55/59
Truck	1,094,976	36.0	12/36/50
Caterpillar	678,784	22.3	5/22/32
Ignatius	438,528	14.4	10/36/69
Meetingroom	1,850,240	60.9	0/1/12

TABLE 11 documents what fraction of the query domain carries usable truth, because yield claims are only meaningful relative to the answerable domain. The deployment construction domain is *predictor-valid* (finite $\hat{m}_0, \hat{m}_1, \hat{m}_2$, vanilla depth, positive SfM distance, valid quality gate): every camera contributes $1600/1600$ predictor points, and on the four audited scenes 2,784–25,182 of them per scene lie outside the label-valid region—those queries can be *answered* at deployment (the predictor does not need the label) but are excluded from scoring, which is the asymmetry the yield denominator respects. The frozen-method rows show the answered-support envelope across evaluation cameras: close-range DTU scenes sit at 34–59% answered support with tight cross-camera spread (±5 points), while Caterpillar, Ignatius, and Meetingroom have both lower mean support and wide camera spread (Meetingroom 0–12%). The spread, not the mean, is what stresses conformal calibration, and it is the empirical input to the camera-balancing weights used when fitting the band.

F.2 QUALITATIVE MAP PROTOCOL

The qualitative maps of Fig. 4 use a fixed, data-independent view rule fixed before rendering: the evaluation camera at the middle of the evaluation order (scan33: camera 26; Caterpillar: 000200; Meetingroom: 000194), each on its full 40×40 probe grid with the frozen configuration's per-query answer mask. White pixels are queries without evaluation truth—not unanswered predictions; the checker never scores them. The three scenes span the support envelope of TABLE 11: mid-support DTU (scan33), low-support wide-baseline TnT (Caterpillar), and extreme-support Meetingroom.

G SCOPE AND BOUNDARIES

We close with the boundaries of the present evidence, each phrased as an instrument limit or a lever rather than a defect. *(i) Proxy truth.* The checker scores depth against a first-surface reference on

discrete 40×40 grids; the reference has finite density and its own surface-branching ambiguities. This limits the *resolution* at which conditional error is measured, not the validity of the intervals; denser ground truth is the natural instrument upgrade. *(ii) Threshold law.* The deployment threshold law is calibrated on the uniform 64-point grid of the frozen protocol. Other laws shift the yield–risk frontier pointwise; the direction of every comparison is invariant to that choice in the archived sweeps, and query-adaptive laws are a straightforward extension. *(iii) Far-view design.* The far-view stress splits deliberately violate the randomization premise that P3 requires; they are an engineering control, and a design-based deployment sampler that restores the premise in the field is the natural continuation. *(iv) Calibration budgets.* Budgets below $m{=}8$ fail by construction (§E.2); the planning recipe turns this into a budgeting tool. *(v) Scene geometry and ensemble.* Meetingroom's low yield reflects its 19-camera selector on a large room; wider-baseline hull families are the identified lever. The frozen ensemble is two 2DGS seeds plus one 3DGS seed; §C.2 shows the mixture beats homogeneous seeds at matched budget, and scaling it is orthogonal to every claim here.