BEYOND COVERAGE: SINGLE-VIEW SIGNALS FOR POST-HOC UNCERTAINTY ESTIMATION IN GAUSSIAN SPLATTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-hoc uncertainty estimation for Gaussian Splatting aims to predict rendering errors without modifying a trained scene representation. Existing approaches largely rely on multi-view coverage statistics, requiring training-view rendering, ground-truth comparison, and back-projection onto Gaussian primitives. We revisit where predictive uncertainty signals originate and show that much of the signal is already present in a single rendering at the novel pose. Four local appearance statistics extracted from the colour buffer are as predictive as the entire coverage feature family, while a simple gradient-magnitude map remains competitive without any learned regressor. The two sources are largely complementary: a representation carrying both, together with scale-invariant statistics from the rendered depth buffer, improves Pearson correlation by 23%, Spearman correlation by 17%, and reduces AUSE by 14% over the strongest post-hoc baseline across 23 scenes from four datasets. This scale-invariant representation further enables a single estimator trained across scenes to transfer to unseen scenes without target-scene fitting. We also study monotone view-dependent feature drift as a possible source of evaluation bias, and show that an evidential head on within-view ranks, invariant to such reparameterizations on a fixed pixel set, provides predictive intervals with two-level coverage measurements and similar ranking accuracy. Reliable post-hoc uncertainty in Gaussian Splatting therefore depends not only on how the scene was observed but also on signals intrinsic to the rendered view.

1 INTRODUCTION

Volume-rendering-based scene representations (Mildenhall et al., 2021; Kerbl et al., 2023; Zwicker et al., 2001) have enabled photorealistic novel-view synthesis, with Gaussian Splatting (GS) (Kerbl et al., 2023) used in SLAM (Matsuki et al., 2024) and active view planning (Jiang et al., 2023). In such systems, rendered views are often consumed by downstream algorithms, making it important to identify where predictions can be trusted. Retraining a scene representation solely for uncertainty estimation is costly, motivating *post-hoc* uncertainty estimation (UE), where the trained splatting is frozen and rendering error is predicted from quantities already available from the representation (Jiang et al., 2023; Gottwald et al., 2026). In our comparisons, all estimators use the same frozen splatting, which we verify by re-rendering (Appendix B); performance differences therefore arise from the uncertainty estimator rather than the underlying reconstruction.

Existing methods draw on information gathered from multiple views: stochastic-rendering variance (Shen et al., 2022; Li & Cheung, 2024; Lyu et al., 2024; Aira et al., 2025), Fisher information induced by training views (Jiang et al., 2023), or training-view errors and coverage statistics projected onto Gaussian primitives and regressed onto per-pixel rendering error (Gottwald et al., 2026). These methods exploit the intuitive relation between sparse observations and reconstruction uncertainty. We ask a complementary question: *how much of the predictive signal attributed to cross-view coverage is already present in the novel rendering itself?*

Local contrast and texture are established image-quality cues (Wang et al., 2004; Mittal et al., 2012; Zhang et al., 2018). We test whether they predict rendering error from a single novel view relative to the full coverage pipeline. Under the same view-disjoint protocol, four local appearance

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.

statistics computed from a single rendered image achieve nearly the same predictive correlation as the complete coverage family, while their predictions remain only weakly correlated. Combining both sources improves further. Our goal is therefore *signal accounting*: identifying where post-hoc uncertainty information originates and measuring the contribution of each source.

This accounting depends critically on evaluation design. Holding the estimator and feature set fixed, changing only the split from pixels sampled within the same views to disjoint evaluation views reduces Pearson correlation from 0.508 to 0.138 for a weighted-sum estimator and from 0.637 to 0.378 for a threshold-based estimator. Same-view evaluation can therefore substantially overestimate predictive performance and distort conclusions about feature utility. We analyze one contributing mechanism, *view-dependent monotone feature drift*. Under an explicit latent-law assumption, within-view ranks remove coordinate-wise monotone reparameterizations at the population level, whereas raw weighted sums may incur cross-view excess risk. This model does not capture all sources of cross-view degradation, such as changes in visibility, occlusion, or scene composition. We further justify pooled scene–view training: under representative sampling, pooled fitting approaches the predictor minimizing expected risk over the underlying scene–view population. Whether such statistical sharing compensates for the absence of target-scene fitting at finite sample sizes is evaluated empirically in §4.3.

These observations motivate a distributional estimator built on rank-normalized features. Rank normalization provides invariance to the modeled monotone drift, while a lightweight linear Normal–Inverse-Gamma head produces predictive intervals in addition to scalar uncertainty scores. Empirical interval coverage is evaluated at two nominal levels; ranking performance is statistically similar to that of the gradient-boosting baseline. A precision-adaptive fusion variant also degrades more gracefully when selected feature groups are corrupted at test time, although this robustness is demonstrated only under the source-failure settings considered here.

We make three contributions.

- We quantify the relative predictive value of multi-view coverage and single-view appearance under a unified view-disjoint protocol. Four appearance statistics from a single rendering approach the predictive power of the full back-projection pipeline, yet remain complementary enough that their combination improves both correlation and sparsification performance. A single gradient-magnitude map already captures most of the appearance signal without a learned regressor.
- We develop a unified representation comprising coverage, four appearance statistics, and four scale-invariant geometry statistics. Across 23 scenes, it improves Pearson correlation, Spearman correlation, and AUSE over the strongest post-hoc baseline using the same frozen splatting. The scale-invariant statistics are especially useful for cross-scene generalization, where a shared estimator trained on other scenes empirically matches the per-scene estimator at the reported operating point without target-scene fitting.
- We formalize view-dependent monotone feature drift at three levels. Within-view rank is a maximal invariant to coordinate-wise monotone reparameterization; under a view-invariant latent law, rank normalization removes the induced population-level cross-view bias of linear predictors; and under an exchangeable scene–view population, pooled fitting converges to the population-optimal shared predictor. These results motivate the rank-based evidential estimator while clarifying the distribution shifts it does not address.

2 Related Work

Uncertainty estimation for NeRFs and Gaussian Splatting. For NeRFs, uncertainty is obtained by making the radiance field stochastic — ensembles, conditional flows, provenance fields — or by post-processing it (Sünderhauf et al., 2022; Shen et al., 2022; Nakayama et al., 2024; Goli et al., 2024). For GS, most methods instead estimate *parameter* uncertainty over the primitives and render it: var3DGS (Li & Cheung, 2024) and manifold sampling (Lyu et al., 2024) learn distributions over primitive parameters and read pixel uncertainty off the sampling variance, Aira et al. (2025) and Wilson et al. (2025b) model primitive variance directly, and FisherRF (Jiang et al., 2023) and POp-GS (Wilson et al., 2025a) use the Fisher information of the training views. The stochastic variants alter reconstruction quality because they change GS training and rendering; the Fisher-information

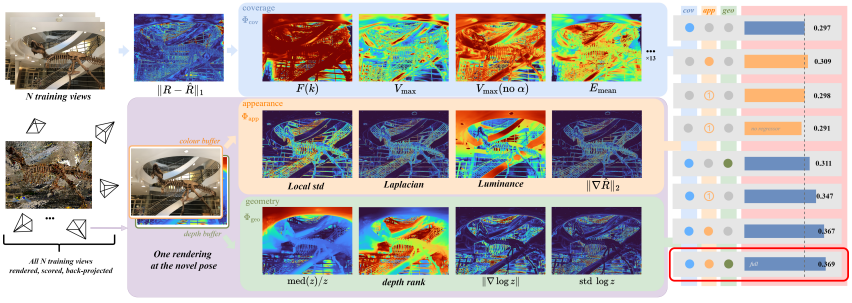

Figure 1: The three feature families. *Top:* the coverage family Φ_{cov} of Gottwald et al. (2026) needs all N training views rendered, scored against ground truth and back-projected onto the primitives before any of its thirteen maps exists at the novel pose (four shown). *Bottom:* Φ_{app} and Φ_{geo}, four maps each, come from the colour and depth buffers of that one rendering. *Right:* Pearson per combination of families, $[O]$, 23 scenes (Table 2); a circled 1 is the gradient map alone, with and without a regressor.

variants are post-hoc but describe how strongly a primitive was constrained rather than how large the rendering error will be. The same maps drive calibration (Amini-Naieni et al., 2024) and active view selection (Pan et al., 2022; Kopanas & Drettakis, 2023). More broadly, uncertainty-aware learning has been studied for incomplete multi-view and multimodal data through divergence-based probabilistic modelling and generative recovery (Xue et al., 2025; Zhang et al., 2025; Qiu et al., 2026). What unites all of these, and what matters here, is the source of signal: each quantity is a function of how the scene was observed across the training views.

Post-hoc regression, local appearance statistics, and evidential regression. The setting we adopt — estimating *predictive* rather than parameter uncertainty by regressing the rendering error from maps rendered out of a frozen splatting — was introduced by Gottwald et al. (2026), whose coverage family we reuse unmodified, both as one of three families and as the comparison point, and whose two premises we revisit: coverage-only features, and a per-scene regressor. That local contrast and high-frequency structure predict where reconstruction error will be visible is long-standing in image quality assessment (Wang et al., 2004; Mittal et al., 2012; Zhang et al., 2018); we use it as one arm of a measurement, the novelty being the comparison against the coverage pipeline under a single protocol. For the distributional part of our method we use evidential regression (Amini et al., 2020; Sensoy et al., 2018; Ulmer et al., 2021), whose identifiability pathologies are documented (Meinert et al., 2023; Bengs et al., 2022); this is why we report empirical interval coverage rather than likelihood as our primary calibration number (Guo et al., 2017; Kuleshov et al., 2018; Gneiting et al., 2007).

3 Method

The main estimator keeps the coverage-only baseline's gradient-boosting regressor and adds eight maps computed from the novel rendering to its thirteen coverage maps. Keeping the regressor fixed makes feature-family comparisons interpretable. We separately study a rank-input evidential head on the baseline's thirteen coverage maps; Algorithm 1 describes this distributional extension, rather than the 21-map regressor that produces the main result.

3.1 Setup and representation

Gaussian Splatting. A splatting represents a scene by primitives k with a mean, a covariance, an opacity and a view-dependent colour $c_k(d)$ in spherical harmonics (Kerbl et al., 2023), composited front to back as $C(\bar{x}) = \sum_k c_k(d)\,\alpha_k(\bar{x})\,T_k(\bar{x})$ with $T_k = \prod_{j<k}(1-\alpha_j)$. Replacing c_k by a per-primitive scalar and reusing the weights $\alpha_k T_k$ renders that scalar as a map at any pose, which is how primitive-level statistics become feature maps, and substituting primitive depth yields the depth buffer z; both are byproducts of one rendering pass.

Problem setup. A scene has views $\mathcal{V}$, view v has pixels $\mathcal{P}_v$, a frozen splatting renders $\hat{R}(\bar{x})$ against ground truth $R(\bar{x})$, and the target is the per-pixel ℓ_1 error

$$e_{\bar{x}} \;=\; \big\|R(\bar{x}) - \hat{R}(\bar{x})\big\|_1, \tag{1}$$

the same target as Gottwald et al. (2026), so that numbers remain directly comparable. A feature map $\phi : p \mapsto \phi(p) \in \mathbb{R}^d$ is read off a novel pose and a post-hoc estimator f_θ predicts $\hat{u}(p) = f_\theta(\phi(p))$; f_θ never observes R.

Three feature families. Our representation is $\Phi = \Phi_{\mathrm{cov}} \cup \Phi_{\mathrm{app}} \cup \Phi_{\mathrm{geo}}$, and the distinction that organises the paper is one of cost. Φ_{cov}, the *coverage* family, is the primitive-based representation of Gottwald et al. (2026): a field-of-view counter plus coverage and back-projected error maps under three aggregations, with and without the opacity term and in direction-independent and direction-dependent variants, thirteen maps in total. Producing any one of them requires every training view to be rendered, its error computed and projected onto the primitives. We use the family unmodified, so §4.2 is like-for-like.

Φ_{app}, the *appearance* family, consists of four maps computed from the rendered image alone and never from ground truth: the Sobel gradient magnitude $\|\nabla \hat{R}\|_2$, the local standard deviation of intensity, the Laplacian response and the local mean luminance. Local texture separates reducible from irreducible error — a wrongly rendered flat wall could be fixed with more coverage, whereas foliage carries error that coverage does not remove — and bright and shadowed surfaces carry different aleatoric structure. Φ_{geo}, the *geometry* family, consists of four maps from the depth buffer z, each invariant to a per-view rescaling of depth: the median-normalised inverse depth $\mathrm{median}_{\mathcal{P}_v}(z)/z$, the within-view depth percentile rank, $\|\nabla \log z\|_2$ and the local standard deviation of $\log z$. The first is motivated by triangulation, where the information a stereo pair carries about a point at depth z with baseline B scales as $(B/z^2)^2$, in a ratio form that removes the scene's absolute scale; the last two mark occlusion boundaries and geometric complexity. This family is the least predictive of the three and adds little within a scene, so we justify it where its invariance pays off, in §4.3.

3.2 A DRIFT MODEL AND WHAT IT IMPLIES

Feature distributions can change with viewpoint. We isolate one possible effect by modelling a reparameterization of fixed latent content.

Assumption 1 (Monotone per-view drift)**.** *On a fixed indexed pixel set, let $\phi(p)$ be a latent feature field. In view v, feature j is observed as $\phi_v^{(j)}(p) = g_v^{(j)}(\phi^{(j)}(p))$, where each unknown $g_v^{(j)}$ is strictly increasing.*

For feature j and pixel p in view v, define the within-view empirical rank

$$r_v^{(j)}(p) = \frac{1}{|\mathcal{P}_v|}\Big|\big\{q \in \mathcal{P}_v : \phi_v^{(j)}(q) \le \phi_v^{(j)}(p)\big\}\Big|. \tag{2}$$

On a fixed pixel set, this rank cancels the modeled monotone transformation and is a maximal invariant when feature values are distinct (Proposition 1). This motivates rank inputs for the linear evidential head; the tree baseline retains raw features, since ranks discard potentially predictive scale (Pearson 0.311 with raw features versus 0.285 with ranks under $[O]$). The model does not cover changing visibility or scene composition. Cross-view statements additionally require a shared latent feature–error law (Proposition 2); the pooled-fitting result concerns population linear risk rather than the finite-sample gradient-boosting comparison of §4.3 (Proposition 3). Their assumptions and proofs remain in Appendix A.

3.3 EVIDENTIAL HEAD AND ACCOUNTING

Evidential head on rank features. For a controlled distributional comparison we apply (2) to the same thirteen coverage maps used by the baseline, and feed the ranks to a linear Normal–Inverse-Gamma head (Amini et al., 2020) with a ranking term and residual weighting. Thus $y \mid \mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$, $\mu \sim \mathcal{N}(\gamma, \sigma^2/\lambda)$ and $\sigma^2 \sim \Gamma^{-1}(a, b)$. The head returns a predictive distribution rather

Algorithm 1: Post-hoc estimation with the evidential head

Input: frozen splatting; training views $\mathcal{V}_{\mathrm{tr}}$, held-out views $\mathcal{V}_{\mathrm{ho}}$, evaluation view v; coverage-map groups $\{\mathcal{F}_k\}_{k=1}^{K}$
Per view: render the thirteen coverage maps $\phi \leftarrow \Phi_{\mathrm{cov}}$, rank every map within the view by (2)
Fit: the all-group head H_0 and the group heads H_k on the ranks and errors of $\mathcal{V}_{\mathrm{tr}}$; then $\omega \geq 0$, $\sum_k \omega_k \leq 1$ on $\mathcal{V}_{\mathrm{ho}}$, and freeze ω
At v**:** $\theta_0 \leftarrow H_0(r)$, $\theta_k \leftarrow H_k(r^{(k)})$
if *precision-weighted* **then**
  $\hat{\theta} \leftarrow \sum_k \pi_k \theta_k$, $\pi_k = \lambda_k / \sum_j \lambda_j$
else
  $\hat{\theta} \leftarrow \theta_0 + \sum_k \omega_k (\theta_k - \theta_0)$
end
$u \leftarrow \hat{b}(1+\hat{\lambda})/[\hat{\lambda}(\hat{a}-1)]$ by (3), aleatoric and epistemic parts available separately

than a score and decomposes the uncertainty as

$$\underbrace{\mathbb{E}[\sigma^2] = \frac{b}{a-1}}_{\text{aleatoric}}, \qquad \underbrace{\mathrm{Var}[\mu] = \frac{b}{\lambda(a-1)}}_{\text{epistemic}}, \qquad u_{\mathrm{tot}} = \frac{b(1+\lambda)}{\lambda(a-1)}. \tag{3}$$

Marginalising μ and σ^2 gives the exact predictive distribution $y \mid \theta \sim \mathrm{StudentT}_{2a}(\gamma, \sqrt{b(1+\lambda)/(a\lambda)})$, from which the interval coverage of §4.4 is computed; the likelihood reported there is the Gaussian-equivalent NLL of the same predictive mean and variance, and is labelled as such. We attach one head per semantic group of input maps plus a head on all groups, and combine them by residual fusion. With $\theta = (\gamma, \lambda, a, b)$, θ_0 for the all-group head and θ_k for group $k \in \{1, \ldots, K\}$, the fused parameters are

$$\hat{\theta} \;=\; \theta_0 + \sum_{k=1}^{K} \omega_k \, (\theta_k - \theta_0), \qquad \omega_k \geq 0, \;\; \textstyle\sum_k \omega_k \leq 1, \tag{4}$$

with ω fitted on held-out views, projected onto the constraint set above, and frozen. Rearranged, (4) is a convex combination of θ_0 and the group parameters, so the fused parameter stays in $\{\lambda > 0, a > 1, b > 0\}$ and $\omega = 0$ recovers the all-group head. This is an admissible parameter interpolation, not an exact Bayesian combination of NIG distributions. Because ω is frozen at test time, the source-failure experiment uses a separate precision-weighted variant that averages group parameters with weights proportional to each group's NIG λ_k (Appendix D). Its corruption curve is not the residual-fusion result reported for clean inputs.

Accounting quantities. Every claim in §4.2 is a value of one of four quantities defined in Appendix B, with ρ Pearson throughout: the predictive power $\mathcal{P}(A) = \mathbb{E}_s[\rho(f_A, e)]$ of a feature subset averaged over scenes, the marginal contribution $\Delta(A \mid B)$, the complementarity $\mathcal{C}(A, B)$, and the correlation $r_\times(A, B)$ between two subsets' predictions, which says whether two families predict the same thing rather than whether they measure the same thing.

4 Experiments

Datasets and protocols. We use 23 scenes: the nine MipNeRF360 scenes (Barron et al., 2022), the eight LLFF scenes (Mildenhall et al., 2019), truck and train from Tanks and Temples (Knapitsch et al., 2017) and the four Light Field scenes (Yücer et al., 2016). All splattings are vanilla GS (Kerbl et al., 2023) trained for 30k iterations (100k for Light Field) at spherical-harmonic degree 4, at resolutions indoor /2, outdoor /4, LLFF 1600px and Light Field full. Two protocols, labelled in every table and never mixed, run on the same 23 scenes. $[P]$, the pipeline protocol, trains the regressor on one hold-out view per scene and evaluates all remaining views with the baseline's hyperparameters unchanged; it carries the main and the evidential comparison. $[O]$, an offline protocol (150k/150k subsampled pixels, four training views, two on Light Field; disjoint views), carries the ablations and the cross-scene experiment. The statistical unit is the scene: its four configurations are averaged before a paired t-test, with a Wilcoxon signed-rank test as a check.

Table 1: Rendering-error UE on the three datasets for which published numbers exist, protocol $[P]$, every row run by us. Reconstruction is listed per method because the stochastic methods alter the representation to obtain their estimate. Bold marks the best UE value in a column; per-scene values are in Appendix C.

		LF					MipNeRF360					LLFF				
method	venue	AUSE ↓	Pears. ↑	PSNR ↑	SSIM ↑	LPIPS ↓	AUSE ↓	Pears. ↑	PSNR ↑	SSIM ↑	LPIPS ↓	AUSE ↓	Pears. ↑	PSNR ↑	SSIM ↑	LPIPS ↓
estimate requires the method's own representation																
manifold (Lyu et al., 2024)	SA'24	0.412	0.015	23.72	0.815	0.228	0.446	0.042	27.41	0.812	0.221	0.377	0.042	24.86	0.801	0.220
var3DGS (Li & Cheung, 2024)	NeurIPS'24	0.363	0.351	23.10	0.826	0.233	0.493	0.210	25.83	0.744	0.324	0.402	0.259	25.84	0.868	0.110
post-hoc, one shared frozen splatting																
FisherRF (Jiang et al., 2023)	ECCV'24	0.906	−0.144	28.86	0.918	0.063	0.738	−0.054	27.76	0.826	0.125	0.988	−0.088	24.70	0.792	0.137
PRIMU (Gottwald et al., 2026)	CVPR'26	0.233	0.436	28.86	0.918	0.063	0.374	0.236	27.76	0.826	0.125	0.280	0.369	24.70	0.792	0.137
full representation (ours)	—	**0.193**	**0.513**	28.86	0.918	0.063	**0.324**	**0.299**	27.76	0.826	0.125	**0.233**	**0.458**	24.70	0.792	0.137

Baselines. Our primary baseline is the coverage-only post-hoc regressor of Gottwald et al. (2026) in its direction-dependent variant, the stronger of its two variants and the strongest published post-hoc estimator of rendering error on these datasets. We reproduce it on our own splattings and use the reproduction everywhere (agreement with the published numbers in Appendix C). We also run FisherRF (Jiang et al., 2023), var3DGS (Li & Cheung, 2024) and manifold sampling (Lyu et al., 2024), and add two reference points no prior work reports: a *regressor-free* baseline using a single view-normalised gradient-magnitude map directly as the uncertainty, and the same map fed to the regressor.

Metrics. Pearson and Spearman correlation between the predicted uncertainty and the true error, and the Area Under the Sparsification Error curve (AUSE) (Ilg et al., 2018), with the total error per view normalised to one so that views are comparable. For the distributional variant we add empirical coverage of the nominal 50% and 90% intervals, our primary calibration number (Kuleshov et al., 2018; Meinert et al., 2023), and the Gaussian-equivalent predictive negative log-likelihood.

Why disjoint views. The training and evaluation split matters for all three metrics. Holding model and features fixed, a weighted-sum estimator reports Pearson 0.508 on same-view pixels against 0.138 on disjoint views, and histogram gradient boosting 0.637 against 0.378. Figure 3 shows these four means over the 18 scenes on which all four were run; the five later scenes reproduce the direction of the effect (Appendix B). The drops are 0.370 and 0.259, respectively, a difference of 0.111 in sensitivity to the split, not a measured decomposition of the causes. Spatial dependence, visibility and feature drift may all contribute. We therefore use disjoint views for every comparison below.

Implementation. The appearance and geometry maps need no new supervision, no change to the rasteriser and no GPU kernel; filter sizes, normalisation, hardware, hyperparameters and wall-clock are in Appendix B.

4.1 Comparison with baselines

Table 1 places the full representation next to the published GS uncertainty methods on the three datasets for which published numbers exist, under $[P]$, every row run by us. The paired comparison that carries the claim is against the coverage-only baseline on the same splatting: over all 23 scenes Pearson rises by 0.073 ($0.315 \rightarrow 0.388$, $+23\%$), Spearman by 0.063 ($+17\%$) and AUSE falls by 0.046 (-14%), all $p < 10^{-10}$; the claim rests on the size and the consistency of the per-scene effect rather than on that p-value, and all three metrics improve on 23 of 23 scenes (Appendix C). By dataset the gain runs from $+0.089$ Pearson on LLFF to $+0.047$ on Tanks and Temples, the smallest single scene being truck ($+0.028$). The reconstruction columns make the asymmetry explicit: the stochastic methods buy their estimate by changing the representation and pay for it in reconstruction quality (var3DGS is 1.9 dB below the shared splatting on MipNeRF360), whereas every post-hoc row shares one frozen splatting.

Table 2: Family-wise accounting, offline protocol $[O]$, all 23 scenes. "Cost" is what must be computed before the estimator can be evaluated at a novel pose. Rows are mutually comparable but not comparable to Table 1, which uses $[P]$. The two Tanks and Temples scenes appear only in the *all* columns. The regressor-free row uses the view-normalised gradient map directly as the uncertainty.

input	cost	MipNeRF360 (9) AUSE ↓	MipNeRF360 (9) Pears. ↑	LLFF (8) AUSE ↓	LLFF (8) Pears. ↑	Light Field (4) AUSE ↓	Light Field (4) Pears. ↑	all 23 scenes AUSE ↓	all 23 scenes Pears. ↑
coverage Φ_{cov} (13 maps)	full cross-view back-projection pipeline	0.424	0.246	0.320	0.356	0.276	0.428	0.366	0.311
appearance Φ_{app} (4 maps)	one rendering	0.398	0.248	0.318	0.378	0.286	0.380	0.355	0.309
gradient map only (1 map)	one rendering, one Sobel	0.438	0.249	0.361	0.357	0.315	0.360	0.391	0.304
gradient map only, *no regressor*	one rendering, one Sobel, no training	0.453	0.246	0.352	0.342	0.311	0.330	0.392	0.292
geometry Φ_{geo} (4 maps)	one rendering (depth buffer)	0.519	0.125	0.393	0.245	0.390	0.271	0.456	0.190
$\Phi_{\text{cov}} \cup \Phi_{\text{geo}}$ (17 maps)	pipeline + depth buffer	0.415	0.261	0.304	0.370	0.269	0.438	0.355	0.325
$\Phi_{\text{cov}} \cup \Phi_{\text{app}}$ (17 maps)	pipeline + one rendering	0.365	0.309	0.271	0.441	0.240	0.493	0.317	0.379
$\Phi_{\text{cov}} \cup \{\text{gradient}\}$ (14 maps)	pipeline + one Sobel	0.393	0.291	0.280	0.419	0.244	0.481	0.334	0.361
full Φ (21 maps)	pipeline + one rendering	**0.360**	**0.310**	**0.268**	**0.444**	**0.236**	**0.497**	**0.313**	**0.381**

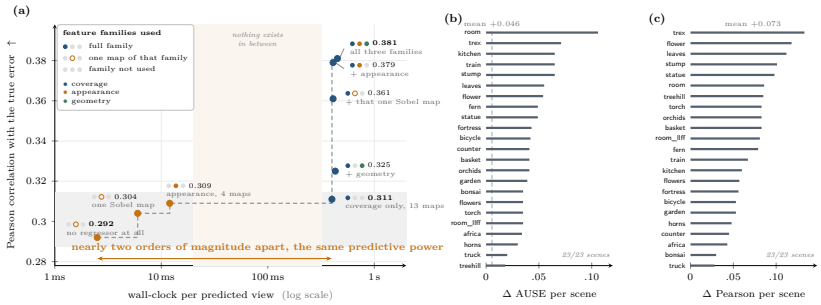

Figure 2: **(a)** The subsets of Table 2 by feature computation cost per scene (horizontal) against uncertainty-estimation performance, protocol $[O]$. **(b, c)** Per-scene AUSE and Pearson improvement of the full representation over the coverage-only baseline under the protocol of Table 1.

4.2 Ablation study: signal accounting

Table 2 is the central measurement of the paper: the upper block gives the predictive power of each family in isolation, the lower block the effect of adding families to the coverage baseline, all rows under $[O]$. The appearance family alone, four maps from one rendering, matches the coverage family in correlation (Pearson 0.309 against 0.311) and is better on the sparsification metric (AUSE 0.355 against 0.366), ahead on MipNeRF360 and LLFF and behind on the dense Light Field captures where coverage is strongest. Most of that value sits in one map: the gradient magnitude alone reaches 0.304, and used directly as the uncertainty with no fitting still reaches 0.292, within 0.02 of the pipeline, though on AUSE it stays behind (0.391 regressed, 0.392 raw, against 0.366; the two differ little because AUSE depends only on the ordering). The geometry family alone is the weakest (0.190 / 0.456). The two main families are only weakly related — $\bar{r}_{\text{feat}} = 0.152$ between an appearance map and the coverage maps, $r_{\times} = 0.290$ between their predictions — so the appearance signal is not a derivative of the coverage signal. Their union gains $\Delta(\text{app} \mid \text{cov}) = +0.068$ Pearson and -0.049 AUSE over coverage, and still $\mathcal{C} = +0.070 / -0.038$ over appearance. The conclusion is not that rendering error is "really" a local appearance phenomenon — were coverage redundant, the full representation could not beat the appearance family by that margin — but that two comparable and largely complementary sources of signal differ in cost by orders of magnitude (Figure 2).

Table 3: Global against per-scene fitting under $[O]$, mean Pearson over 23 leave-one-scene-out folds. The global estimator is trained on the other 22 scenes; the target scene supplies feature maps but no labelled hold-out view for regressor fitting. A non-significant paired test does not establish equivalence.

input	global (22 scenes)	per-scene fit	Δ	paired test
coverage Φ_{cov} (13 maps)	0.292	0.311	-0.020	$p = 0.25$
$\Phi_{\text{cov}} \cup \Phi_{\text{app}}$ (17 maps)	0.367	0.379	-0.012	$p = 0.50$
full Φ (21 maps)	**0.383**	0.381	$+0.002$	$p = 0.93$ (13/23)

The appearance family is partly redundant — the gradient map alone recovers three quarters of its Pearson gain — and helps on all 23 scenes, the geometry family on 19, within ± 0.04 Pearson (Appendix C).

4.3 Cross-scene generalisation

We now ask whether a regressor must be fitted on the target scene at all. Under $[O]$, the per-scene fit uses four labelled target-scene views (two on Light Field), while the global fit uses the other 22 scenes. This tests the practical value of target-scene fitting at their stated training budgets. Proposition 3 concerns population risk for linear squared-loss predictors and does not predict the finite-sample outcome of this gradient-boosting comparison.

The global estimator is the same gradient-boosting model as the per-scene one, fitted on the raw (not rank-normalised) maps of the other 22 scenes. "Zero-shot" means that no estimator is fitted on the target scene and no hold-out view of it is labelled; the coverage family is still computed there, which needs its training views and their ground truth, whereas appearance and geometry need only the rendering being evaluated. In the 23 leave-one-scene-out folds of Table 3, the full global estimator obtains 0.383 Pearson against 0.381 for the per-scene fit (paired $p = 0.93$). Their means are close at this operating point; the non-significant test is not an equivalence guarantee. Coverage alone transfers at 0.292, below its per-scene value of 0.311. Adding appearance raises the global score to 0.367; adding the four scale-invariant geometry maps raises it further to 0.383 ($+0.016$, $p = 0.002$, Wilcoxon $p = 0.001$). Within-scene addition of geometry yields only $+0.002$. The transfer setting therefore reveals a benefit that is small in the per-scene ablation.

One estimator trained on other scenes can thus be used without labelled hold-out views or fitting a regressor on the target scene at the evaluated budgets. What transfers is the predictor, not the coverage-map computation: those maps still require the target scene's training views and ground truth. Because the global and per-scene fits use different pixel budgets, this result does not isolate scene diversity from data quantity. Gottwald et al. (2026) report transfer of a coverage-only regressor but do not report this global-against-per-scene comparison.

4.4 Calibration and robustness to source failure

On the same thirteen coverage maps, the residual-fusion evidential head has no statistically detected difference from gradient boosting on Pearson (-0.009, $p = 0.33$), Spearman ($+0.002$, $p = 0.81$) or AUSE (-0.004, $p = 0.49$), paired over 23 scenes under $[P]$. It also supplies predictive intervals: empirical coverage is 0.506 at nominal 50% ($p = 0.87$) and 0.878 at nominal 90% ($p = 0.064$). This supports approximate calibration at these two levels only; the aleatoric, epistemic and total channels are defined in (3). Under the specific test-time corruption of back-projected error maps, the separate precision-weighted fusion variant degrades more slowly than gradient boosting: it starts with higher AUSE at $\eta = 0$ and ends with lower AUSE at $\eta = 1$. The evidence λ of corrupted groups contracts while the reported intact groups remain stable. Appendix D reports the measured curve in Figure 5, along with the values and capture-specific breakdown.

4.5 Scope and limitations

All claims concern the RGB rendering error of whole views: we evaluate depth uncertainty only on Light Field, where the evidential head transfers poorly (Appendix D). The sign of the geometry family's within-scene contribution is scene-dependent and three depth-shape statistics fail to predict it; with 23 paired scenes we lean on the consistency of direction rather than on the mean difference.

5 Conclusion

The signal available for post-hoc rendering-error prediction comes from both multi-view coverage and local appearance. Measuring them with the same regressor and view-disjoint protocol shows that a single rendering supplies much of the predictive value of the coverage pipeline, while the combined representation performs best. Scale-invariant geometry adds value to cross-scene transfer, allowing one shared regressor to attain a similar mean score to a per-scene fit at the reported budgets. The measured same-view inflation makes the evaluation split part of the result, not a reporting detail. The rank-based evidential head is a separate, controlled extension that supplies predictive intervals on the coverage baseline's inputs.

Ethics statement

This work studies uncertainty estimation for scene reconstruction on public, previously released datasets containing no human subjects and no personally identifying information. Its intended effect is to make failures of novel-view synthesis easier to detect, which we regard as risk-reducing for downstream robotics and medical applications. The residual risk is misplaced confidence: our results cover RGB rendering error on the scene types evaluated here, and should not be assumed to hold for depth uncertainty or for deployment domains unlike these.

Reproducibility statement

§4 states the datasets, resolutions, hyperparameters, baselines and hold-out protocol, and Appendix C reports the reproduction of the published baseline that serves as the comparison point for every result. Appendix A proves every formal statement of §3.2. Appendix B reports the measurement establishing that all compared estimators share one frozen splatting. The two evaluation protocols are labelled $[P]$ or $[O]$ in every table and numbers from the two are never compared. The same appendix gives the definitions of the appearance and geometry maps, including filter sizes and normalisation, and lists the hardware, every hyperparameter of both heads, the per-stage wall-clock, and the random seeds, all of which are fixed. Code is provided as supplementary material and will be released publicly upon publication.

References

Luca Savant Aira, Diego Valsesia, and Enrico Magli. Modeling Uncertainty for Gaussian Splatting. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):11657–11663, 2025.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep Evidential Regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020.

Niki Amini-Naieni, Tomas Jakab, Andrea Vedaldi, and Ronald Clark. Instant uncertainty calibration of NeRFs using a meta-calibrator. In *European Conference on Computer Vision*, pp. 309–324. Springer, 2024.

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.

Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Pitfalls of Epistemic Uncertainty Quantification through Loss Minimisation. *Advances in Neural Information Processing Systems*, 35: 29205–29216, 2022.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic Forecasts, Calibration and Sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2): 243–268, 2007.

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.

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.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 652–667, 2018.

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.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

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.

Georgios Kopanas and George Drettakis. Improving NeRF Quality by Progressive Camera Placement for Free-Viewpoint Navigation. In *Vision, Modeling, and Visualization*, pp. 11–20, 2023.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate Uncertainties for Deep Learning Using Calibrated Regression. In *International Conference on Machine Learning*, pp. 2796–2804. PMLR, 2018.

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.

Linjie Lyu, Ayush Tewari, Marc Habermann, Shunsuke Saito, Michael Zollhöfer, Thomas Leimkühler, and Christian Theobalt. Manifold Sampling for Differentiable Uncertainty in Radiance Fields. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Hidenobu Matsuki, Riku Murai, Paul H. J. Kelly, and Andrew J. Davison. Gaussian Splatting SLAM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18039–18048. IEEE, 2024.

Nis Meinert, Jakob Gawlikowski, and Alexander Lavin. The Unreasonable Effectiveness of Deep Evidential Regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9134–9142, 2023.

Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (ToG)*, 38(4):1–14, 2019.

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.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

Kiyohiro Nakayama, Mikaela Angelina Uy, Yang You, Ke Li, and Leonidas J. Guibas. ProvNeRF: Modeling Per Point Provenance in NeRFs as a Stochastic Field. *Advances in Neural Information Processing Systems*, 37:99145–99160, 2024.

Xuran Pan, Zihang Lai, Shiji Song, and Gao Huang. ActiveNeRF: Learning Where to See with Uncertainty Estimation. In *European Conference on Computer Vision*, pp. 230–246. Springer, 2022.

Xihang Qiu, Yuhao Fang, Qing Zhou, Bin Zhai, Jialong Hong, Wanpeng Zhang, Yao Lu, Ye Zhang, and Chun Li. Beyond missing modalities: Hypergraph conditioned diffusion for uncertainty-aware multimodal emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22953–22963, June 2026.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential Deep Learning to Quantify Classification Uncertainty. 31, 2018.

Jianxiong Shen, Antonio Agudo, Francesc Moreno-Noguer, and Adria Ruiz. Conditional-Flow NeRF: Accurate 3D Modelling with Reliable Uncertainty Quantification. In *European Conference on Computer Vision*, pp. 540–557. Springer, 2022.

Niko Sünderhauf, Jad Abou-Chakra, and Dimity Miller. Density-aware NeRF Ensembles: Quantifying Predictive Uncertainty in Neural Radiance Fields. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9370–9376, 2022.

Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. Prior and Posterior Networks: A Survey on Evidential Deep Learning Methods for Uncertainty Estimation. *Transactions on Machine Learning Research (TMLR)*, 2023, 2021.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE transactions on image processing*, 13(4): 600–612, 2004.

Joey Wilson, Marcelino Almeida, Sachit Mahajan, Martin Labrie, Maani Ghaffari, Omid Ghasemalizadeh, Min Sun, Cheng-Hao Kuo, and Arnab Sen. POp-GS: Next Best View in 3D-Gaussian Splatting with P-Optimality. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3646–3655. IEEE, 2025a.

Joey Wilson, Marcelino Almeida, Min Sun, Sachit Mahajan, Maani Ghaffari, Parker Ewen, Omid Ghasemalizadeh, Cheng-Hao Kuo, and Arnie Sen. Modeling Uncertainty in 3D Gaussian Splatting through Continuous Semantic Splatting. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3284–3290. IEEE, 2025b.

Zhipeng Xue, Yan Zhang, Ming Li, Chun Li, Yue Liu, and Fei Yu. Uncertainty quantification for incomplete multi-view data using divergence measures. *IEEE Transactions on Image Processing*, 34:4328–4342, 2025. doi: 10.1109/TIP.2025.3579987.

Kaan Yücer, Alexander Sorkine-Hornung, Oliver Wang, and Olga Sorkine-Hornung. Efficient 3D Object Segmentation from Densely Sampled Light Fields with Applications to 3D Reconstruction. *ACM Transactions on Graphics (TOG)*, 35(3):1–15, 2016.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595. IEEE, 2018.

Yan Zhang, Ming Li, Chun Li, Zhaoxia Liu, Ye Zhang, and Fei Richard Yu. Uncertainty quantification via hölder divergence for multi-view representation learning. *IEEE Transactions on Multimedia*, 27:8263–8275, 2025. doi: 10.1109/TMM.2025.3604966.

Matthias Zwicker, Hanspeter Pfister, Jeroen Van Baar, and Markus Gross. Ewa volume splatting. In *Proceedings Visualization, 2001. VIS'01.*, pp. 29–538. IEEE, 2001.

A THEORY

Consequences of the drift assumption. Assumption 1 is a modelling choice rather than an empirical guarantee: occlusion, changes in visible content and feature ties can violate its conditions. Three properties follow (Lemma 1): an estimator reading per-feature thresholds is unaffected by a monotone transformation applied alike to training and evaluation values if thresholds are transformed too; the within-view empirical rank in Eq. (2) cancels g_v exactly on a fixed pixel set; weighted sums $f(\phi) = \psi(\sum_j w_j \phi^{(j)})$, including linear heads, do not generally have this invariance. The first consequence characterizes a drift-invariant representation.

Invariance lemma and proofs.

Lemma 1 (Invariance)**.** *Under Assumption 1: **(i)** if f depends on the features only through per-feature threshold comparisons $\mathbb{1}[\phi^{(j)} \le \tau]$, as decision trees and histogram-based gradient boosting do, then a strictly increasing map g applied alike to training and evaluation features changes no prediction once the thresholds are mapped as $\tau \mapsto g(\tau)$, and a drift that differs between training and evaluation views changes a prediction only where it moves a feature value across a split threshold; **(ii)** on a fixed view's pixel set, the within-view empirical rank (2) is unchanged when $g_v^{(j)}$ is applied to every value of coordinate j; **(iii)** estimators of the form $f(\phi) = \psi(\sum_j w_j \phi^{(j)})$, including linear heads and the first layer of an MLP, do not in general have this invariance.*

Proof of Lemma 1. A strictly increasing map preserves order and a rank depends only on order, which gives (ii). (i) follows from $\mathbb{1}[g(\phi^{(j)}) \le g(\tau)] = \mathbb{1}[\phi^{(j)} \le \tau]$, and a prediction of f can change only if one of its indicators does. For (iii), $\sum_j w_j g_v^{(j)}(\phi^{(j)})$ is in general not a function of $\sum_j w_j \phi^{(j)}$; two features with $g_v^{(1)} = \mathrm{id}$ and $g_v^{(2)} = 2\,\mathrm{id}$ form a counterexample. □

Proposition 1 (The within-view rank is the maximal invariant)**.** *Let G act on a view's feature field by $\phi^{(j)} \mapsto g^{(j)} \circ \phi^{(j)}$ with every $g^{(j)}$ strictly increasing, and let each feature take distinct values within the view. A statistic of that field is G-invariant if and only if it is a function of the within-view ranks (2).*

Proof. Strictly increasing transformations preserve each coordinate's ordering, hence its ranks. Conversely, for two finite feature fields with the same ranks, map each coordinate's ordered observed values in the first field to the corresponding values in the second. Strictly increasing piecewise-linear interpolation, extended beyond the observed range, supplies a transformation in G. The two fields therefore lie in the same orbit. An invariant statistic is constant on each orbit and is consequently a function of ranks. □

View-invariant latent law. Assumption 1 characterizes within-view feature distortion, but monotone transformations alone do not guarantee shared latent content across views. We therefore impose the following distributional condition.

Assumption 2 (View-invariant latent law under monotone drift)**.** *There exists a latent pair (ϕ, e), with $\phi \in \mathbb{R}^d$, whose joint distribution is shared across all views. View v observes $\phi_v = g_v(\phi) = \big(g_v^{(1)}(\phi^{(1)}), \dots, g_v^{(d)}(\phi^{(d)})\big)$, where each $g_v^{(j)}$ is strictly increasing. The rendering error e depends on the latent feature rather than its view-specific monotone parameterization, and all required second moments are finite.*

This assumption isolates *pure feature drift*; changes due to occlusion, visible content, or scene composition are not represented by g_v and are therefore excluded from the result below.

Proposition 2 (Cross-view bias of raw features and invariance of population ranks)**.** *Under Assumptions 1–2, define the squared risk in view v as $R_v(w) = \mathbb{E}\big[(e - w^\top \phi_v)^2\big]$, with $\Sigma_v = \mathbb{E}[\phi_v \phi_v^\top] \succ 0$, $w_v = \arg\min_w R_v(w)$. Then, for any views v and v',*

$$R_{v'}(w_v) - R_{v'}(w_{v'}) = (w_v - w_{v'})^\top \Sigma_{v'} (w_v - w_{v'}) \ge 0, \tag{5}$$

with equality if and only if $w_v = w_{v'}$. Let $F_{v,j}(x) = \Pr[\phi_v^{(j)} \le x]$ be the view-specific marginal CDF and define the population within-view rank representation

$$r_v^{(j)} = F_{v,j}\Big(\phi_v^{(j)}\Big). \tag{6}$$

Strict monotonicity implies

$$r_v^{(j)} = F_j\Big(\phi^{(j)}\Big), \tag{7}$$

where F_j is the marginal CDF of $\phi^{(j)}$. Hence the joint distribution of (r_v, e) is identical across views. If $\Sigma_r = \mathbb{E}[r_v r_v^\top] \succ 0$, the population-optimal rank-based predictor

$$w_v^r = \arg\min_w \mathbb{E}\big[(e - w^\top r_v)^2\big] \tag{8}$$

is therefore view-invariant:

$$w_v^r = w_{v'}^r \quad \textit{for all } v, v'. \tag{9}$$

Thus, rank normalization removes cross-view bias caused solely by coordinate-wise monotone drift, but not shifts in the underlying joint distribution of scene content and rendering error.

Proof. Let $b_v = \mathbb{E}[e\phi_v]$. Expanding the risk yields

$$R_v(w) = \mathbb{E}[e^2] - 2w^\top b_v + w^\top \Sigma_v w. \tag{10}$$

Since $\Sigma_v \succ 0$, the unique minimizer satisfies $\Sigma_v w_v = b_v$. Hence,

$$\begin{aligned} R_{v'}(w) - R_{v'}(w_{v'}) &= w^\top \Sigma_{v'} w - 2w^\top \Sigma_{v'} w_{v'} + w_{v'}^\top \Sigma_{v'} w_{v'} \\ &= (w - w_{v'})^\top \Sigma_{v'} (w - w_{v'}). \end{aligned} \tag{11}$$

Setting $w = w_v$ proves Eq. (5).

For the rank representation, since $\phi_v^{(j)} = g_v^{(j)}(\phi^{(j)})$ and $g_v^{(j)}$ is strictly increasing,

$$\begin{aligned} F_{v,j}\Big(g_v^{(j)}(t)\Big) &= \Pr\Big[g_v^{(j)}(\phi^{(j)}) \le g_v^{(j)}(t)\Big] \\ &= \Pr[\phi^{(j)} \le t] = F_j(t). \end{aligned} \tag{12}$$

Therefore,

$$r_v^{(j)} = F_j(\phi^{(j)}), \tag{13}$$

which is independent of v. Since the joint law of (ϕ, e) is shared across views, so is that of (r_v, e). Their second and cross moments are therefore identical, yielding the same unique population minimizer for every view. □

Remark. Equation (9) is a population result. The empirical within-view rank in Eq. (2) is a finite-sample estimate of Eq. (6). Although it is exactly invariant to strictly monotone reparameterizations within a fixed sample, different views need not share the same empirical rank–error distribution. Occlusion, view-dependent scene composition, or changes in $p(e \mid r, v)$ may therefore still produce cross-view prediction error.

The next result concerns average population risk, not a guaranteed ordering of finite-sample estimators.

Assumption 3 (Exchangeable scene–view population)**.** *Let S denote a random scene and G a random view-dependent drift. An unseen test pair (S', G') is drawn independently from the same population as the training scene–view pairs. For each (S, G), define*

$$R_{S,G}(w) = \mathbb{E}\left[\big(e - w^\top G(\phi)\big)^2 \,\middle|\, S, G\right], \tag{14}$$

where the distribution of (ϕ, e) may depend on S. Assuming finite second moments, define the population risk

$$\overline{R}(w) = \mathbb{E}_{S,G}\big[R_{S,G}(w)\big]. \tag{15}$$

We assume that $\overline{R}$ admits a unique minimizer

$$w^\star = \arg\min_w \overline{R}(w). \tag{16}$$

Proposition 3 (Population justification for pooled cross-scene fitting)**.** *Under Assumption 3, let $\widehat{w}$ be any estimator independent of an unseen test pair (S', G'). Then*

$$\begin{aligned}\mathbb{E}\big[R_{S',G'}(\widehat{w})\big] &= \mathbb{E}\big[\overline{R}(\widehat{w})\big] \\ &\geq \overline{R}(w^\star) = \mathbb{E}\big[R_{S',G'}(w^\star)\big].\end{aligned} \tag{17}$$

If $\overline{R}$ is strictly convex, equality holds only if $\widehat{w} = w^\star$ almost surely.

Let $\widehat{w}_n$ further denote the empirical risk minimizer over n independent scene–view samples. Under standard consistency conditions for squared-loss regression—including finite second moments, convergence of the empirical second-moment and feature–error cross-moment matrices, and a positive-definite population second-moment matrix—we have

$$\widehat{w}_n \xrightarrow{p} w^\star, \tag{18}$$

and consequently

$$\overline{R}(\widehat{w}_n) \xrightarrow{p} \overline{R}(w^\star). \tag{19}$$

Hence, pooling observations across independent scenes and views is justified at the population level: with sufficiently representative data, the fitted predictor converges to the population-optimal predictor for a new scene–view draw. This does not *imply that a finite-sample pooled estimator must outperform one fitted using labelled views from the specific target scene; that comparison also depends on sample size, scene-specific shift, and model misspecification, and is evaluated empirically in §4.3.*

Proof. Since (S', G') is independent of $\widehat{w}$ and follows the population in Eq. (15),

$$\mathbb{E}\left[R_{S',G'}(\widehat{w}) \mid \widehat{w}\right] = \mathbb{E}_{S,G}\left[R_{S,G}(\widehat{w})\right] = \overline{R}(\widehat{w}). \tag{20}$$

Taking expectation over $\widehat{w}$ gives

$$\mathbb{E}\big[R_{S',G'}(\widehat{w})\big] = \mathbb{E}\big[\overline{R}(\widehat{w})\big]. \tag{21}$$

Because $w^\star$ minimizes $\overline{R}$, $\overline{R}(\widehat{w}) \geq \overline{R}(w^\star)$ for every realization of $\widehat{w}$, which yields Eq. (17). Strict convexity further implies that equality requires $\widehat{w} = w^\star$ almost surely.

For the consistency statement, the empirical normal equations depend on the empirical second-moment matrix and the empirical feature–error cross moment. Under the stated assumptions, both converge in probability to their population counterparts. Positive definiteness ensures a unique population solution, so

$$\widehat{w}_n \xrightarrow{p} w^\star. \tag{22}$$

Continuity of $\overline{R}$ then yields Eq. (19). □

Remark. Proposition 3 allows the latent distribution $p_S(\phi, e)$ to vary across scenes. Even if a common deterministic relation $e = h(\phi)$ holds, scene-wise risks need not coincide because the distribution of ϕ may differ across scenes. The result therefore concerns risk averaged over the scene–view population rather than equality of individual scene-wise risk functions.

B Implementation and protocol

Feature definitions. Every map below is a deterministic function of the colour buffer $\hat{R}$ or the depth buffer z of the single rendering at the evaluated pose; none of them reads ground truth, and none has a learned parameter. The colour buffer is linear RGB clamped to $[0, 1]$, with no tone mapping anywhere in the pipeline. Gradients use a 3×3 Sobel kernel, the Laplacian a 3×3 kernel and every local statistic a 7×7 box window, all with nearest-neighbour boundary handling. No per-view normalisation is applied to the maps before the regressor: the geometry maps are invariant to a rescaling of depth by construction, and the within-view rank normalisation of the evidential head (§3) is part of that model rather than a preprocessing step.

Appearance family Φ_{app}. (i) The Sobel gradient magnitude of each colour channel, averaged over channels; (ii) the local standard deviation of each channel, averaged over channels; (iii) the absolute Laplacian response of the luminance $Y = 0.299R + 0.587G + 0.114B$; (iv) the local mean of Y. The regressor-free baseline of Table 2 uses (i) alone, normalised within the view to $[0, 1]$, directly as the uncertainty estimate.

Geometry family Φ_{geo}. Non-finite and non-positive depths are first replaced by the view's median depth. (i) $\text{median}_{\mathcal{P}_v}(z)/z$, the inverse depth normalised by the within-view median, which removes the scene's absolute scale; (ii) the within-view percentile rank of z in $[0, 1]$, invariant to any monotone rescaling of depth by Lemma 1(ii); (iii) the Sobel gradient magnitude $\|\nabla \log z\|_2$; (iv) the local standard deviation of $\log z$, $\left(\overline{(\log z)^2} - \overline{\log z}^{\,2}\right)^{1/2}$ with the bar denoting the 7×7 box mean. All four are unchanged by a rescaling $z \mapsto cz$: (i) and (ii) by construction, (iii) and (iv) because it shifts $\log z$ by a constant.

Hyperparameters and computational cost. All timings are measured on a single L40S GPU with a four-thread CPU allocation, per scene unless stated otherwise. The asymmetry that motivates the accounting of §4.2 is in the third and fourth rows: the coverage family needs every training view re-rendered and back-projected, a third to half a second per rendered view, while the eight appearance and geometry maps are pure NumPy functions of buffers the renderer already emits — milliseconds per view, no GPU, and no storage, since recomputing them is cheaper than reading them back.

Table 4: Cost and configuration. Gradient-boosting settings are those of PRIMU, left unchanged; the eight new maps have no hyperparameters beyond the filter sizes given above, being deterministic functions of the rendered colour and depth buffers. All seeds are fixed.

cost	
GS training (30k iterations)	8–10 min, L40S
GS training, LF (100k iterations)	∼30 min, L40S
coverage family, 13 maps (test views)	60–120 s per scene, L40S; 0.33–0.56 s per view
appearance + geometry, 8 maps	∼4 s per scene, CPU; one Sobel map is 5.6 ms per view
gradient boosting, fit / inference	1–2 min per configuration / $<$ 1 s per view, CPU
evidential head, fit	2–4 min per configuration, CPU
global estimator, fit / new scene	∼10 min once for 22 scenes / $<$ 1 s per view with no fitting, CPU
feature storage	1.6× the coverage family
configuration	
gradient boosting	`max_iter`=150, `learning_rate`=0.3, `validation_fraction`=0.2, `random_state`=42
evidential head	linear, `view_norm`=rank, ranking weight $\beta = 1.0$, ℓ_2 0.01, 60 epochs, lr 3×10^{-3}, batch 4096
global estimator	gradient-boosting settings as above; training set = the other 22 scenes, 5M subsampled pixels
protocol $[P]$	one hold-out training view, all remaining views evaluated; scene as the statistical unit
protocol $[O]$	150k/150k subsampled pixels, four training views (two on Light Field), disjoint views
seeds	gradient boosting 42, subsampling 0, selection probe 0

Accounting quantities. For a feature subset A with estimator f_A fitted on it, and with ρ the Pearson correlation between a prediction and the error (1),

$$\mathcal{P}(A) = \mathbb{E}_s\big[\rho(f_A, e)\big], \qquad \Delta(A \mid B) = \mathcal{P}(A \cup B) - \mathcal{P}(B),$$

$$\mathcal{C}(A, B) = \mathcal{P}(A \cup B) - \max\{\mathcal{P}(A), \mathcal{P}(B)\}, \qquad r_\times(A, B) = \mathbb{E}_s\big[\rho(f_A, f_B)\big],$$

all expectations being means over the scenes s of the stated scene set. The input-level quantity $\bar{r}_{\text{feat}}$ is the mean of $|\rho|$ between one map and each map of another family, over pixels and then over scenes. Each quantity is also reported with AUSE in place of ρ where the corresponding AUSE is available.

These quantities also let the joint predictive power of the two main families be split between them. With two sources the Shapley value of A against B is $\varphi_A = \frac{1}{2}\mathcal{P}(A) + \frac{1}{2}[\mathcal{P}(A \cup B) - \mathcal{P}(B)]$, taking $\mathcal{P}(\emptyset) = 0$ since an uninformative estimate has zero correlation with the error. Substituting the measured values of Table 2 ($\mathcal{P}(\Phi_{\text{app}}) = 0.309$, $\mathcal{P}(\Phi_{\text{cov}}) = 0.311$, $\mathcal{P}(\Phi_{\text{cov}} \cup \Phi_{\text{app}}) = 0.379$) gives $\varphi_{\text{app}} = 0.189$ and $\varphi_{\text{cov}} = 0.191$, a split of 49.7% against 50.3%; the two marginal contributions are likewise close, $\Delta(\text{app} \mid \text{cov}) = +0.068$ against $\Delta(\text{cov} \mid \text{app}) = +0.070$. This is the sense in which one of the two sources of signal, the one obtainable from a single rendering, is half of it.

Protocol diagnostics. Figure 3 shows the protocol effect of §4 as the 2×2 measurement it is: the same model and the same features, with only the split changed. The four cells are means over the 18 scenes for which all of them were run; the pattern reproduces on the five scenes added afterwards, where the tree falls from 0.822 to 0.139 on train and from 0.693–0.801 to 0.473–0.503 on Light Field, and the weighted sum from 0.244 to 0.019 on train and from 0.193–0.475 to 0.179–0.426 on Light Field.

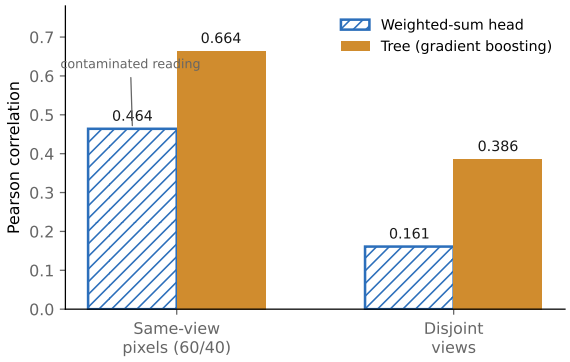

Figure 3: Same model, same features; only the pixel split changes. Both estimator families are inflated when training and evaluation pixels are drawn from the same views, and the weighted-sum head is inflated more, ending far below the tree once the views are separated.

Every number in the paper uses all inputs of the stated feature set, with no feature selection of any kind, so no ground-truth information enters at any stage other than as the regression target on the training views.

The splatting is unchanged by all estimators. Post-hoc UE is defined not to change the 3D model. Rather than assert this, we measured it. After all regressor experiments had completed we re-rendered test views from the same checkpoints and compared them pixel by pixel against the predicted RGB stored in the result archives, which were written before any regressor experiment ran. Rendering is deterministic, so an unchanged model must produce a bit-identical output; the maximum absolute per-pixel difference is $0.000\text{e}{+00}$ for bicycle (6 views), garden (6), bonsai (6) and fern (3). Bit-identity implies that the checkpoint never changed during the UE experiments, and hence that all estimators compared here consume the same splatting; this is what licenses reporting the reconstruction columns of Table 5 once rather than per method.

C Additional results

Per-scene results. Table 5 collects everything we report per scene: the reconstruction quality of the shared splatting on the left, and on the right the UE metrics of the three estimators compared in this paper — the coverage-only baseline, the full representation, and the evidential head on the baseline's own thirteen maps. All UE columns use the pipeline protocol $[P]$. Pearson improves and AUSE decreases on all 23 scenes when the appearance and geometry families are added.

The baseline columns are our reproduction of the direction-dependent coverage-only regressor of PRIMU, the stronger of its two published variants (MipNeRF360 Pearson 0.236 against 0.203). Its AUSE/Pearson means are $0.374/0.236$ on MipNeRF360 against the published $0.378/0.236$

and $0.280/0.369$ on LLFF against $0.281/0.368$; on Light Field they are $0.233/0.436$ against $0.286/0.351$, since there both our splatting and our reproduction differ from the published ones (see the Light Field paragraph below).

The reconstruction columns are a property of the splatting rather than of any UE method (Appendix B), which is why one set of them serves all three estimators. Our MipNeRF360 means are 27.76 / 0.826 / 0.125 against the published 27.51 / 0.817 / 0.215; the LPIPS difference arises because our evaluation images are not clamped before the perceptual metric, a known measurement difference unrelated to any claim in this paper. Our LLFF means are 23.70 / 0.792 / 0.137 against 23.81 / 0.801 / 0.226, from a different preprocessing version of the data. On Light Field our means at full resolution and 100k iterations are 28.86 / 0.918 / 0.063 against the published 34.18 / 0.969 / 0.043; the difference is discussed below.

Table 5: Per-scene results. Left block: reconstruction quality of the shared splatting. Right block: UE metrics under protocol $[P]$ for the coverage-only baseline (13 maps), the full representation (21 maps) and the evidential head (13 maps), with the change of the full representation over the baseline. Δ columns are computed before rounding.

	reconstruction			13 maps (baseline)		21 maps (ours)		change		evidential (13)	
scene	PSNR ↑	SSIM ↑	LPIPS ↓	AUSE ↓	Pears. ↑	AUSE ↓	Pears. ↑	ΔAUSE	ΔPears.	AUSE ↓	Pears. ↑
MipNeRF360											
bicycle	25.71	0.780	0.128	0.436	0.158	0.394	0.211	−0.042	+0.053	0.389	0.248
flowers	21.85	0.621	0.274	0.334	0.278	0.299	0.335	−0.035	+0.057	0.377	0.198
garden	27.73	0.875	0.057	0.309	0.346	0.269	0.398	−0.039	+0.053	0.320	0.256
stump	26.91	0.783	0.111	0.430	0.151	0.364	0.252	−0.065	+0.101	0.442	0.162
treehill	22.73	0.650	0.243	0.350	0.249	0.331	0.333	−0.018	+0.085	0.316	0.270
bonsai	32.30	0.948	0.062	0.341	0.286	0.306	0.316	−0.035	+0.030	0.320	0.287
counter	29.19	0.916	0.096	0.348	0.243	0.307	0.288	−0.041	+0.045	0.320	0.279
kitchen	31.75	0.934	0.062	0.403	0.213	0.338	0.272	−0.065	+0.060	0.395	0.207
room	31.69	0.927	0.089	0.414	0.203	0.309	0.289	−0.106	+0.086	0.401	0.225
LLFF											
fern	23.19	0.787	0.148	0.234	0.450	0.186	0.529	−0.049	+0.079	0.235	0.427
flower	26.76	0.834	0.118	0.265	0.388	0.211	0.506	−0.054	+0.118	0.277	0.351
fortress	29.59	0.873	0.080	0.304	0.384	0.261	0.440	−0.043	+0.056	0.331	0.359
horns	24.85	0.847	0.133	0.273	0.382	0.243	0.429	−0.030	+0.048	0.276	0.384
leaves	17.68	0.596	0.191	0.239	0.406	0.184	0.518	−0.055	+0.112	0.242	0.393
orchids	18.76	0.624	0.152	0.203	0.476	0.162	0.559	−0.041	+0.083	0.214	0.441
room	25.49	0.912	0.163	0.332	0.244	0.297	0.325	−0.035	+0.081	0.338	0.196
trex	23.26	0.862	0.110	0.394	0.224	0.323	0.357	−0.071	+0.133	0.358	0.255
Tanks and Temples											
truck	27.63	0.946	0.020	0.388	0.285	0.368	0.313	−0.020	+0.028	0.317	0.303
train	22.78	0.888	0.047	0.465	0.135	0.399	0.202	−0.065	+0.067	0.414	0.169
Light Field											
africa	24.99	0.873	0.092	0.251	0.478	0.217	0.521	−0.034	+0.043	0.240	0.486
basket	32.34	0.949	0.037	0.211	0.433	0.170	0.516	−0.041	+0.083	0.237	0.357
statue	33.03	0.963	0.032	0.270	0.384	0.221	0.482	−0.049	+0.098	0.307	0.341
torch	25.07	0.888	0.090	0.200	0.449	0.165	0.532	−0.035	+0.083	0.225	0.435
mean (23)	26.32	0.838	0.110	0.321	0.315	0.275	0.388	−0.046	+0.073	0.317	0.306

Light Field. All Light Field numbers are at full resolution and 100k training iterations. With the full representation the four scenes follow the other datasets: Pearson $0.436 \rightarrow 0.513$ and AUSE $0.233 \rightarrow 0.193$ against our reproduction of the baseline, with all four scenes improving on both (Table 5). The gain does not depend on the training setting: it is $+0.085$ / -0.042 at half resolution and 30k iterations, and $+0.074$ / -0.040 at full resolution and 30k iterations. Our splatting reaches 28.86 / 0.918 / 0.063 against the published 34.18 / 0.969 / 0.043. We have not identified the source of this difference: matching the resolution, extending training from 30k to 100k iterations ($+1.3$ dB) and masking the black image borders do not close it, and evaluated on training views our splatting stays near 29 dB. Because the splatting differs, Light Field is compared only against our own reproduction on the same splatting.

Scene-level behaviour of the added families. Under $[O]$, single-map permutation importance over the full representation ranks three appearance maps first — luminance ($+0.047$ AUSE), gra-

dient magnitude ($+0.040$) and local standard deviation ($+0.034$) — while deleting the family as a whole costs 0.049, so the four maps are partly but not wholly redundant.

The appearance family contributes positively on all 23 scenes, so its value is not carried by a subset of captures. The geometry family contributes positively on 19 of 23 and its contribution is small either way, spanning -0.008 to $+0.039$; the four scenes on which it costs anything are bounded indoor captures, and three depth-shape summary statistics fail to predict which (§4.5). The scenes on which appearance most clearly overtakes coverage (bicycle, flower, leaves, stump) share texture density — grass, foliage, flower beds — rather than unboundedness, since flower and leaves are forward-facing. We make no mechanistic claim from either pattern.

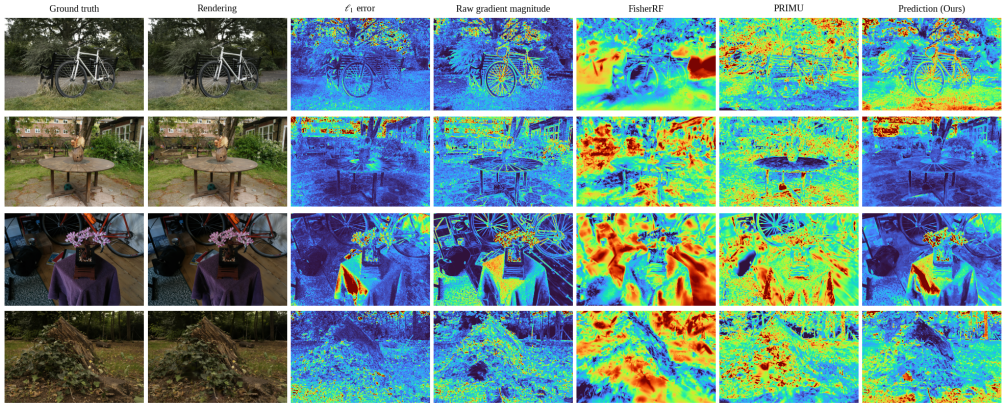

Figure 4: Qualitative comparison on four MipNeRF360 scenes. From left: ground truth, rendering, ℓ_1 rendering error, raw gradient magnitude of the rendering with no regression, FisherRF (Jiang et al., 2023), the coverage-only estimator of PRIMU, and our estimator on the full representation (21 maps). Blue is low and red is high. Much of the structure of the error map is already present in the regression-free gradient column. FisherRF, which measures how strongly the primitives were constrained rather than the size of the rendering error, produces spatially smooth maps that follow the error only loosely; the coverage-only estimator and ours both follow it more closely.

Qualitative comparison.

D EVIDENTIAL HEAD: FURTHER ANALYSIS

Source failure. Table 6 collects the two dimensions of §4.4 that the ranking metrics do not see, and Figure 5 shows the two halves of the source-failure experiment: the evidence λ of each semantic group at $\eta = 0$ and $\eta = 1$, and the AUSE difference between the two heads as a function of η. Across the eleven scenes with saved group-evidence measurements, the evidence of the error group falls from $\lambda = 2.385$ to 0.505 and that of the opacity-free error group from 3.378 to 0.398, while the coverage and field-of-view groups stay at 4.269 and 0.423 to the printed digit. The error-group contraction is also visible in the newly added scenes (for instance train $5.87 \to 1.00$ and statue $2.50 \to 0.82$); this observation requires neither retraining nor an explicit failure detector. The AUSE difference between the two heads is $+0.039$, $+0.021$, $+0.006$, $+0.008$, -0.041 at $\eta = 0, 0.25, 0.5, 0.75, 1$. Note that the $\eta = 0$ point uses the precision-weighted fusion variant and therefore does not coincide with the residual-fusion numbers in the first two columns of Table 6: the two are different fusion variants of the same head, and the comparison within the curve is between the precision-weighted variant and gradient boosting at every η.

Breakdown by capture geometry. The aggregate tie of §4.4 is a mean over captures that behave differently. What the evidential head is, concretely, matters for reading this: it is not a larger model but a *linear* Normal-Inverse-Gamma head whose inputs are the within-view ranks (2) of the same thirteen maps the baseline consumes. It is invariant to the per-view drift of Assumption 1 by Lemma 1(ii), so nothing it learns is specific to the statistics of the single training view; and being linear in the ranks it cannot carve out the sharp, high-order interactions a depth-limited tree ensemble

Table 6: Dimensions the ranking metrics do not see; protocol $[P]$, all 23 scenes, identical inputs for both heads. Nominal coverage is 0.5 and 0.9; NLL is the Gaussian-equivalent predictive negative log-likelihood. Gradient boosting, fitted with a squared-error loss as in PRIMU, outputs a point estimate, so coverage and NLL are not defined for it.

head	Pears. ↑	AUSE ↓	cov@0.5	cov@0.9	NLL ↓
gradient boosting	0.315	0.321	n/a	n/a	n/a
evidential (ours)	0.306	**0.317**	0.506	0.878	−1.835

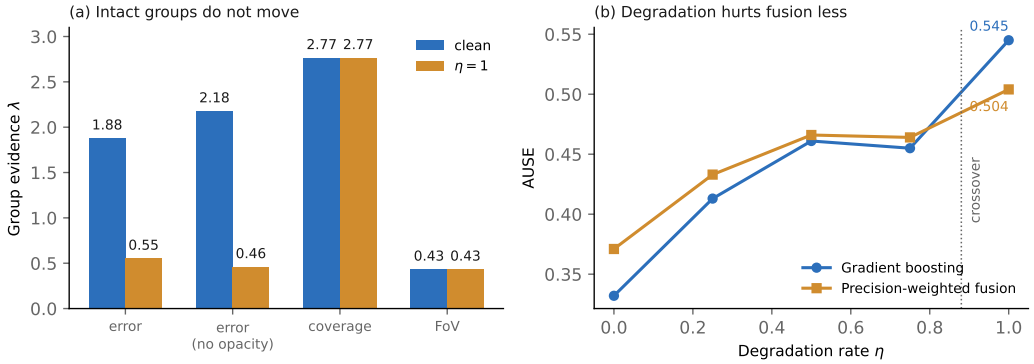

Figure 5: Source failure with precision-weighted fusion, distinct from the residual-fusion head used for clean interval results. **(a)** Group evidence λ at $\eta = 0$ and 1 on the eleven scenes with saved group measurements. **(b)** Aggregate AUSE as the corruption probability η increases; the fused head starts behind gradient boosting and ends ahead.

can. Against a strong signal the second property should cost more than the first gains, and against a weak one the trade may reverse.

Table 7 separates the 360° inward-facing captures (MipNeRF360 and Tanks and Temples) from the forward-facing ones (LLFF) and the four dense Light Field captures. On the eleven inward-facing scenes the evidential head matches gradient boosting in Pearson ($+0.005$, $p = 0.74$) and lowers AUSE by 0.019 ($p = 0.089$, Wilcoxon $p = 0.083$), ahead on $8/11$ scenes on each metric; on the eight LLFF scenes the direction reverses in Pearson (-0.018, $p = 0.076$) with AUSE essentially unchanged ($+0.003$), and the four Light Field captures follow LLFF rather than the inward-facing group (-0.031 Pearson on four scenes, which is too small a sample to test). Within the inward-facing captures the pattern has a gradient: the weaker a scene's baseline signal, the larger the evidential head's gain in Pearson (Spearman $\rho = -0.61$, $p = 0.047$, $n = 11$; within MipNeRF360 alone $\rho = -0.63$, $p = 0.07$, $n = 9$). The three weakest-signal scenes in the sample all improve (train $0.135 \to 0.169$, stump $0.151 \to 0.162$, bicycle $0.158 \to 0.248$, the last being the largest gain of any scene), while the two strongest-signal MipNeRF360 scenes lose (garden $0.346 \to 0.256$, flowers $0.278 \to 0.198$) and inside LLFF there is no such gradient ($\rho = -0.24$).

We report this as a consistent directional pattern and not as an established effect. The partition was made after inspecting the per-scene numbers, neither contrast reaches 0.05, and we therefore claim no mechanism for it; the case for the head in the main text rests on calibration and graceful degradation instead. The per-scene numbers behind the table are the last two columns of Table 5.

Direction-independent coverage maps. Both heads were re-run with the direction-independent (sh0) variant of the coverage family on all 23 scenes. The two heads tie there as they do at sh4: gradient boosting reads Pearson/AUSE $0.298/0.330$ at sh0 against $0.315/0.321$ at sh4, and the evidential head $0.288/0.324$ against $0.306/0.317$. The benefit of direction dependence is therefore nearly the same for the two heads ($+0.017$ and $+0.018$ Pearson), reproducing the baseline's "small but consistent" finding, and every claim of §4.4 holds for both variants of the coverage family.

Table 7: Evidential head against gradient boosting on the same thirteen coverage maps, pipeline protocol $[P]$, mean over the four configurations of each scene, split by capture geometry (all 23 scenes; the four Light Field captures form their own row, as they are dense inward captures that belong to neither group). Δ is evidential minus gradient boosting; p from a paired t-test and a Wilcoxon signed-rank test over scenes; "wins" counts scenes on which the evidential head is better. Paired tests are not reported for the two-scene subgroup. Δ columns are computed before rounding. The split was made post hoc, and the Light Field row has only four scenes, so its p values carry little weight.

		Pearson ↑						AUSE ↓					
group	n	GB	evid.	Δ	p_t	p_W	wins	GB	evid.	Δ	p_t	p_W	wins
inward-facing	11	0.232	0.237	$+0.005$	0.74	0.32	8/11	0.384	0.365	-0.019	0.089	0.083	8/11
MipNeRF360	9	0.236	0.237	$+0.001$	0.97	0.57	6/9	0.374	0.364	-0.009	0.35	0.30	6/9
Tanks and Temples	2	0.210	0.236	$+0.026$	n/a	n/a	2/2	0.427	0.366	-0.061	n/a	n/a	2/2
forward-facing (LLFF)	8	0.369	0.351	-0.018	0.076	0.11	2/8	0.280	0.284	$+0.003$	0.60	0.20	1/8
Light Field	4	0.436	0.405	-0.031	0.35	0.38	1/4	0.233	0.252	$+0.019$	0.30	0.63	1/4
all 23	23	0.315	0.306	-0.009	0.33	0.45	11/23	0.321	0.317	-0.004	0.49	0.64	10/23

Depth uncertainty. On the Light Field scenes we also ran the depth target, where our reproduction of the coverage-only baseline reads Pearson/AUSE $0.681/0.150$ against the published $0.728/0.118$. The evidential head does not transfer to this target: paired over the four scenes its Pearson is 0.177 lower ($p < 0.001$) and its AUSE 0.035 higher ($p = 0.017$). The head was designed around the heavy-tailed distribution of RGB rendering error, and we restrict every claim about it to that target. We did not evaluate the appearance and geometry families on the depth target, so the accounting of §4.2 is likewise stated for RGB only.