LEARNING WHICH REGIONS TO TRUST: SUPPORT-AWARE RISK RANKING FOR SELECTIVE NOVEL VIEW SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Novel-view synthesis (NVS) produces spatially heterogeneous errors, making it important to identify which regions of a rendered view can be trusted. We study spatially selective NVS, where regions within each view are ranked to minimise error at a prescribed coverage under a declared loss. Cross-view consistency provides strong spatial evidence when source observations are available, but becomes incomplete under occlusion, extrapolation, and limited view overlap. Moreover, the desired reliability ordering depends on the deployment objective: pixel, structural, and perceptual losses induce different regional rankings. We derive a pairwise decomposition of selective ranking error by source-view support, identifying where cross-view evidence is informative and where complementary evidence is needed. Guided by this analysis, we introduce L2R-GS, a target-conditioned learning-to-rank framework that combines external cross-view evidence, internal Gaussian-state evidence, and observability cues according to source support and the declared error objective. Across 11 scenes from the Mip-NeRF 360, Tanks and Temples, and Deep Blending datasets, L2R-GS lowers L1, DSSIM, and LPIPS AURC by 4.9%, 9.8%, and 15.8% relative to the strongest published methods on the same held-out views. On held-out views, pairs involving unsupported regions account for more of the ranking gain as source support weakens. A separate study with human annotations shows that training for perceived artifacts improves the selection of reliable regions.

1 INTRODUCTION

3D Gaussian Splatting (3DGS) produces fast, high-quality novel views [15], yet a plausible render may contain local geometric or perceptual errors. A thin structure may disappear or a disoccluded surface may be hallucinated while the rest of the image looks plausible. This makes it important to identify reliable regions within each rendered view.

We study spatially selective NVS: at a fixed coverage, retain the regions ranked safest under a declared loss (Figure 1A). Coverage is imposed within each view, so the task is to order regions of the same render, not merely to identify difficult cameras. The same regional ordering serves every coverage budget; we assess it with the area under the risk–coverage curve (AURC) [5, 31].

How can this ordering be estimated when the target photograph is unavailable? Comparing a novel render with training views is a natural starting point for assessing its reliability [20, 27]. The training images show the same scene from other cameras: where a source observation can be reprojected into the target view, disagreement with the novel render gives a local cue to rendering error. Model sensitivity, Gaussian state, and visibility offer other reliability cues [6, 13, 28].

Two difficulties matter. First, the cross-view check requires a valid source observation. Occlusion, limited overlap, and extrapolation can remove correspondence entirely; the discrepancy is then undefined rather than merely noisy. Gaussian-state signals can still describe the reconstruction in these regions [6, 8]. Because one view can contain both supported and unsupported regions, a cross-view score may rank the former but cannot by itself place the latter among them. Putting every unsupported region last assumes that each is less reliable than every supported one, which need not hold: a region with no valid nearby source may still be rendered correctly, while a supported thin structure

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.

may be wrong. The ranker must therefore use source support to decide how much to trust each kind of evidence.

Second, the desired ordering depends on the error target. Pixel, structural [26], and perceptual [30] losses can rank the same regions differently: a local color mismatch and a structural distortion need not have the same relative severity under each loss. Reusing the oracle LPIPS ordering raises AURC by 41% under L1 and 41% under DSSIM on our benchmark (Figure 1B); a target-swap control for the learned rankers appears in Appendix C.5. This disagreement persists even without estimation error in the realised losses. Proposition A.4 gives the corresponding conditional-risk result: if two targets reverse a pair's preferred order given the same evidence, any fixed, target-independent instance of our score incurs positive expected AURC regret for at least one.

The two difficulties meet in the ordering decision. AURC penalises pairs of regions placed in the wrong order, especially when their losses differ greatly (Proposition 1). In a partially observed view, a pair may contain two supported regions (SS), two unsupported regions (UU), or one of each (SU). Correspondence-based discrepancies can help with SS pairs; UU pairs lack that measurement. Ordering an SU pair requires deciding whether the unsupported region is safer than the supported one. The relative frequency of these pair types changes with source support, so a pairwise decomposition predicts when combining the evidence can help (Proposition 2). Always placing unsupported regions last incurs a quantifiable SU error (Proposition 3).

L2R-GS uses 10 frozen regional signals grouped by what they measure: *external* evidence from comparisons across views, *internal* evidence from the Gaussian model and its render, and *observability* cues measuring source support and image texture (Figure 1C). It converts their values to within-view ranks and learns target-specific linear weights that vary with source support and evidence availability. The support and availability variables tell the ranker when a cross-view measurement is weak or missing; they need not themselves predict the loss. The reconstruction and evidence producers remain frozen.

Across eleven scenes from Mip-NeRF 360, Tanks and Temples, and Deep Blending, L2R-GS lowers L1, DSSIM, and LPIPS AURC by 4.9%, 9.8%, and 15.8% relative to the strongest published methods on the same held-out views. On disjoint views, a view-normalised pairwise diagnostic tracks how structural and perceptual gains vary with support. We also test target-declared selection with human-perceived artifacts instead of a reconstruction loss (Section 4.5).

Our contributions are:

- **Target-declared spatial selection.** We formulate fixed-coverage selection as ranking regions within each rendered view for a declared loss. This evaluates where errors occur within a view and makes explicit why pixel, structural, and perceptual targets may require different orderings.
- **A pairwise theory of selective reliability under partial observation.** We express AURC regret as loss-weighted regional misorderings, decompose them by source support, characterise the gain from complementary evidence, and derive a tight placement bound across the support boundary.
- **L2R-GS, a target-conditioned ranker with support-aware evidence weights.** It learns a loss-specific combination of frozen cross-view and Gaussian-state cues, with explicit interactions for source support and evidence availability. On held-out reconstructions it improves all three target AURCs over published methods; experiments across capture regimes, target swaps, and support ablations show when the combination helps.

2 Related Work

Reliability evidence in neural rendering. Bayes' Rays and FisherRF estimate uncertainty from model sensitivity [7, 13]; PPU predicts 3DGS errors from primitive-level state [6], and visibility-based methods measure observation support [28, 29]. In contrast, WarpRF uses geometric consistency across views [20], while CrossScore compares a query image with multiple same-scene views using learned attention [27]. PuzzleSim combines references with perceptual similarity [10]. These signals capture different failure modes. Cross-view discrepancies additionally depend on valid source correspondences, which may disappear under occlusion or extrapolation. We analyse

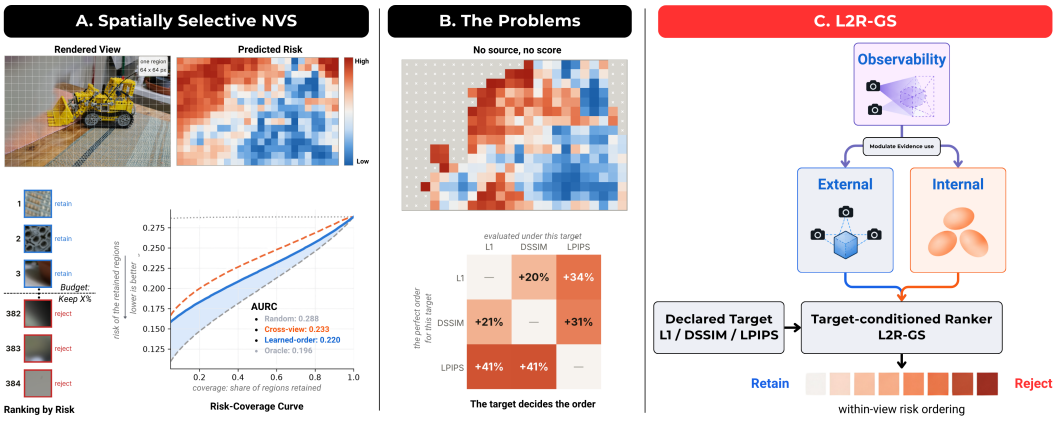


Figure 1: **Overview of spatially selective NVS. (A)** Regional selection within a view. **(B)** Missing cross-view evidence and target-dependent rankings. **(C)** L2R-GS combines evidence according to source observability and the declared target.

how that availability affects the regional ordering and combine frozen external and internal evidence accordingly.

Spatial confidence and region-level decisions. Stochastic Gaussian Splatting estimates pixelwise uncertainty and optimises sparsification [21]; CoMapGS uses covisibility maps to weight sparse-view supervision [11]; and UAR-Scenes directs selective refinement with uncertainty maps [2]. These works make spatial reliability actionable. A confidence map, however, does not determine a region ordering for a declared loss and coverage. We learn that ordering on frozen renders, including unsupported regions for which cross-view evidence is undefined.

Selective ranking. Selective prediction studies the risk–coverage trade-off induced by rejecting uncertain predictions [4, 5]. AURC summarises the trade-off; AUGRC and population analyses clarify its evaluation [23, 31]. Here coverage is imposed separately in each rendered view, so the decision is an ordering of regions within that view. This motivates our pairwise regret analysis and a pairwise learning objective [3]. Missingness can itself carry predictive information [25]; here, missing cross-view evidence affects the placement of unsupported regions relative to supported ones. The desired ordering also changes with the declared error target, rather than being a single task-independent uncertainty score.

3 Method

3.1 Spatially Selective Novel-View Synthesis

As shown in Figure 2, a trained 3D Gaussian Splatting model [15] renders a novel view $\hat{I}_v$, which we partition into a fixed set of equal-area regions $\mathcal{P}_v = \{p_1, \ldots, p_{N_v}\}$. For a declared target t, such as L1, DSSIM, or LPIPS, let $e_t^v(x)$ be its dense loss map computed from the render and ground-truth view. The regional loss is $\ell_t(p) = |p|^{-1} \sum_{x \in p} e_t^v(x)$. Computing the map before pooling preserves the neighbourhood context of DSSIM and LPIPS. The ground-truth view $I_v^\star$ is available only during training and evaluation. Given a coverage $c \in (0, 1]$, a selector retains $K_v(c) = \lceil cN_v \rceil$ regions of every view independently. A score $q_t(p)$, with lower values indicating lower risk, induces a within-view ordering π. The selective risk at coverage c is

$$R_t^v(c) = \frac{1}{K_v(c)} \sum_{i=1}^{K_v(c)} \ell_t(\pi_i), \tag{1}$$

and we evaluate the complete ordering using the discrete area under the risk–coverage curve,

$$\mathrm{AURC}_t(v) = \frac{1}{N_v} \sum_{k=1}^{N_v} R_t^v(k/N_v) = \sum_{i=1}^{N_v} w_i \, \ell_t(\pi_i), \qquad w_i = \frac{1}{N_v} \sum_{k=i}^{N_v} \frac{1}{k}, \tag{2}$$

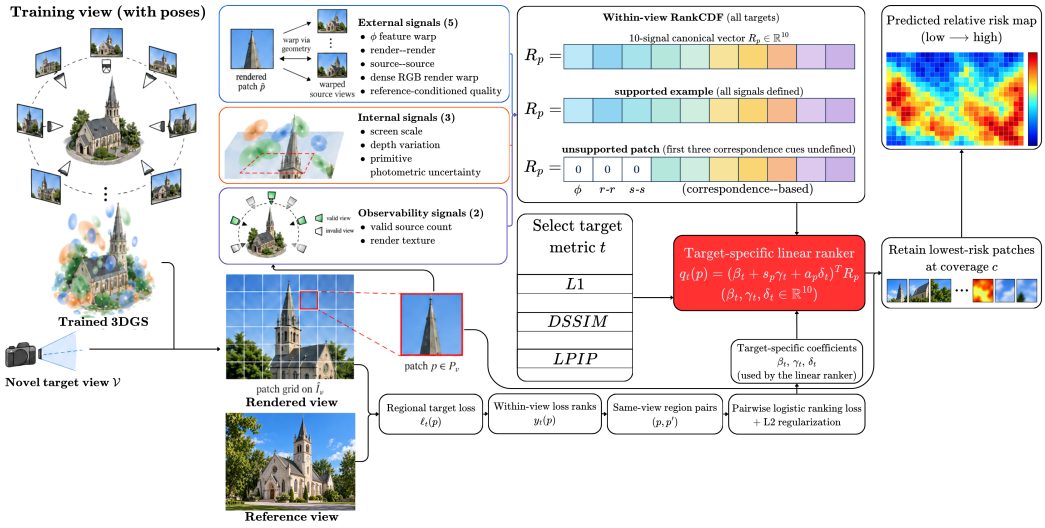


Figure 2: **L2R-GS overview.**

with $w_1 > \cdots > w_{N_v} > 0$. Errors made near the beginning of the retention order therefore matter most. Score ties are broken worst-case, larger loss first, so that no method is credited for the accidental order of regions it scores identically (Appendix B.10).

Sorting by realised target loss is the finite-sample oracle. At deployment, however, ℓ_t is unavailable, and the optimal selector instead orders regions by their conditional target risk $\eta_{t,p}(Z_v) = \mathbb{E}[\ell_t(p) \mid Z_v]$, where Z_v is all inference-time evidence for the view. Because different targets can reverse the ordering of the same pair of regions, the target is part of the problem specification rather than a tuning choice. Formal optimality statements are given in Appendix A.1.

3.2 Pairwise Structure Under Partial Observation

The selective objective depends on which regions are placed before which, rather than on accurate regression of their absolute loss. For a pair (a, b) with $\ell_t(a) < \ell_t(b)$, we call the pair *misordered* if b is retained before a. Let $\mathcal{D}(q)$ be the set of such pairs and define the loss-weighted discordance $W(q) = \sum_{(a,b)\in\mathcal{D}(q)} \big(\ell_t(b) - \ell_t(a)\big)$.

Proposition 1 (Pairwise form of the regret)**.** *There are positional weights $\kappa_{ab} \in [1/(N(N-1)),\, 1/N]$ such that*

$$\mathrm{AURC}(q) - \mathrm{AURC}(\mathrm{oracle}) = \sum_{(a,b)\in\mathcal{D}(q)} \kappa_{ab}\big(\ell_t(b) - \ell_t(a)\big), \tag{3}$$

so that $W(q)/(N(N-1)) \le \mathrm{AURC}(q) - \mathrm{AURC}(\mathrm{oracle}) \le W(q)/N$.

This pairwise view lets us attribute ranking error according to evidence availability, because W is additive over any partition of the pairs while the AURC itself is not. Let S be the regions with at least one valid pixel–source correspondence and U those with none, with $u = |U|/N$. Every regional pair is supported–supported (SS), supported–unsupported (SU), or unsupported–unsupported (UU). Since $\mathcal{D}(q)$ is invariant to rescaling a view's losses, we remove differences in loss level between views with the normalised loss $\tilde{\ell}_t = \ell_t/\bar{\ell}_t$ ($\bar{\ell}_t$ the view's mean) and write $\tilde{W}$ for W under $\tilde{\ell}_t$.

Proposition 2 (Partial-support decomposition)**.** *Under the support-mixture and category-rate stability assumption of Appendix A.3, with $\omega_c(q)$ the expected normalised loss-weighted misorder of one pair of category c,*

$$\mathbb{E}\,\tilde{W}(q) = \frac{N(N-1)}{2}\Big[(1-u)^2\omega_{SS}(q) + 2u(1-u)\,\omega_{SU}(q) + u^2\omega_{UU}(q)\Big]. \tag{4}$$

Equation (4) makes the observation regime explicit: when $u \approx 0$, supported pairs dominate; at intermediate u, cross-support pairs become prominent; and when $u \approx 1$, the ranking is governed

largely by unsupported regions. Applied to the difference between a fused score and a fixed single-source expert, it gives the expected gain as a quadratic in u whose shape—rising, falling or peaked—is fixed by three measurable category improvements. The better of two experts gives the lower envelope of their quadratics. This predicts normalised pairwise gain, rather than raw AURC gain across views. In the empirical test, the mean unsupported share of each view bin is used as a plug-in for u; Appendix A.3 gives the exact fixed-count expression.

Partial support also creates a distinct placement problem. Over the SU pairs with untied losses, let ρ be the fraction in which the unsupported region has the larger target loss and π the fraction in which the score retains the unsupported region first.

Proposition 3 (Placement bound)**.** *The share m of SU pairs that a score misorders satisfies*

$$m \geq |\rho + \pi - 1|, \tag{5}$$

with equality at $\pi = 0$, where $m = 1 - \rho$, at $\pi = 1$, where $m = \rho$, and for the oracle, where $\pi = 1 - \rho$ and $m = 0$.

An external-only score that places all unsupported regions last has $\pi = 0$ and therefore incurs $m = 1 - \rho$ regardless of how accurately it ranks the supported regions. This motivates combining cross-view evidence with correspondence-free internal evidence and exposing support explicitly to the ranker. Full derivations and proofs are provided in Appendix A.

3.3 Evidence Bank

For each region p, we construct a compact bank of 10 frozen scalar signals organised into three information roles, $Z_p = [Z_p^{\text{ext}}, Z_p^{\text{int}}, Z_p^{\text{obs}}]$. The 3DGS reconstruction and all evidence producers remain frozen.

External evidence. External signals ask whether the target rendering is consistent with observations of the same scene. We identify valid source views by geometric reprojection and compute discrepancies between the rendered region and its aligned source observations. For a discrepancy representation j,

$$E_j(p) = \operatorname*{Pool}_{x \in p} \Big[\min_{s \in \mathcal{S}_x} d_j\big(\hat{I}_v(x),\, W_{s \to v}(I_s)(x)\big) \Big], \tag{6}$$

where $\mathcal{S}_x$ is the set of sources valid at pixel x, $W_{s \to v}$ denotes source-to-target warping, and the pool is a mean over pixels where the discrepancy is defined. The bank uses the LPIPS-feature instance of this family [30], which we denote Φ; a cross-view render consistency channel that compares the target render with renders of the same reconstruction from the source cameras, testing its consistency across viewpoints; and a source–source agreement channel that compares warped sources with each other, testing the coherence of the observations themselves. These three are undefined when no pixel in p has a valid source, and source–source agreement also needs two jointly valid sources. Two frozen channels complete the external block: a dense RGB render-warp residual over the training cameras [20] and a learned reference-conditioned quality score [27]. Both retain their released operators, including their native treatment of regions without a valid controlled correspondence.

Internal evidence. Internal signals describe the state of the Gaussian representation and require no source correspondence. For a primitive quantity q_i, we project it to the image using the rendering weights $w_i(x)$,

$$I_q(x) = \frac{\sum_i w_i(x)\, q_i}{\sum_i w_i(x)}. \tag{7}$$

The internal block includes the projected primitive screen scale, the local variation of the rendered log-depth, and a frozen primitive-level predictive photometric uncertainty signal [6]. These cues do not require source correspondence and can provide evidence in unsupported regions when their own producers are available.

Observability. Observability signals describe the regime in which the other evidence is produced rather than risk directly. We use the mean number of valid sources per pixel in a region, $n_{\text{src}}(p) =$

$|p|^{-1}\sum_{x\in p}|\mathcal{S}_x|$, and the local texture energy of the rendered image. Two derived variables are used explicitly by the ranker,

$$s_p = \frac{\log(1+n_{\mathrm{src}}(p))}{\log 9} \in [0,1], \qquad a_p = \frac{\#\{\text{defined external channels at } p\}}{\#\{\text{external channels}\}} \in [0,1], \tag{8}$$

where s_p measures source-support strength, clipped at eight sources, and a_p measures external-evidence availability. It distinguishes the absence of controlled correspondences from regions where render–source or source–source discrepancies can be computed, while the native producers keep their own definitions. The exact correspondence tests, discrepancy definitions, primitive projections, and signal list are given in Appendix B.

Figure 4 in Appendix B.2 illustrates source support in three capture regimes.

3.4 L2R-GS

The evidence channels have incomparable numerical scales, while the downstream task depends only on their ordering within each view. We therefore replace every signal x_j by its within-view rank CDF,

$$R_{p,j} = \begin{cases} \mathrm{AvgRank}_{v,j}(x_j(p))/n_{v,j}, & x_j(p) \text{ is finite,} \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

where $n_{v,j}$ is the number of finite values of channel j in view v and ties receive their average rank. Zero lies outside the range of defined ranks and preserves information about missingness. The resulting representation is invariant to strictly increasing reparameterisations of any evidence channel (Lemma A.5).

Let $R_p \in \mathbb{R}^{10}$ collect the canonicalised evidence for region p. For declared target t, L2R-GS computes

$$q_t(p) = (\beta_t + s_p\gamma_t + a_p\delta_t)^\top R_p, \qquad \beta_t,\gamma_t,\delta_t \in \mathbb{R}^{10}. \tag{10}$$

The three terms have direct roles: β_t gives the target-specific evidence combination applied everywhere; γ_t adjusts that combination with source-support strength; and δ_t adjusts it with external-evidence availability, which is the term that can act on the cross-support pairs of Proposition 3. Through s_p and a_p, observability thus modulates how external and internal evidence are used rather than acting only as another uncertainty score. With ten evidence channels, the model contains 30 parameters per target. Proposition A.4 shows why one shared set of these weights cannot be optimal when two targets reverse a pair's conditional-risk order. Target-specific weights remove that restriction, but do not by themselves guarantee an optimal ordering.

3.5 Pairwise Training and Inference

The training path in Figure 2 uses the realised regional target loss, which we convert to within-view rank labels $y_t(p) = \mathrm{RankCDF}_{\mathcal{P}_v}(\ell_t(p))$. For a pair of regions i, j from the same view, define $z_{ij} = \mathrm{sign}\big(y_t(j) - y_t(i)\big)$. Let $\mathcal{B}_t$ contain the sampled same-view pairs with distinct labels. L2R-GS minimises the regularised pairwise logistic objective

$$\min_{\theta_t} \frac{1}{|\mathcal{B}_t|}\sum_{(i,j)\in\mathcal{B}_t} \log\Big(1+\exp\big[-z_{ij}\big(q_t(j)-q_t(i)\big)\big]\Big) + \lambda\|\theta_t\|_2^2, \qquad \theta_t = (\beta_t,\gamma_t,\delta_t), \tag{11}$$

a tractable surrogate for pair ordering. It does not reproduce the loss-gap and positional weights in the AURC regret of Proposition 1; weighting the pairs by either changes AURC by at most 0.63% (Appendix C.4). For every region, eight partners are drawn uniformly with replacement from the same rendered view, matching the within-view objective; self-pairs and pairs with equal labels are discarded. λ is selected on development data and then frozen.

At inference, no ground truth is required. We render the novel view, compute the available evidence, canonicalise it within the view, evaluate q_t, and sort regions by increasing score. A coverage c retains the first $\lceil cN_v \rceil$ regions of the same ordering, so changing the coverage requires no retraining. For capacity analysis, we also fit boosted trees over exactly the same evidence bank [14]; these serve only as a higher-capacity reference in the experiments.

4 EXPERIMENTS

We test selective ordering, the pairwise predictions under changing support, and extension to human-perceived artifacts.

4.1 EXPERIMENTAL SETUP

Datasets. The primary benchmark has eleven scenes from Mip-NeRF 360 [1], Tanks and Temples [16], and Deep Blending [9]. The final ranker is fitted on two development reconstructions per scene. Three independently trained reconstructions per scene and their held-out views yield 33 test models and 1410 views; Appendix C.2 also tests fits that exclude the scored scene. Additional capture regimes are Replica [22], NeRF-Synthetic [19], DTU [12], LLFF [18], and two Mill-19 scenes [24]. For the human-artifact experiment, we use the observer-annotated artifact masks released with the PuzzleSim method [10] as the target.

Evaluation protocol. Images are divided into 64×64 regions and evaluated under L1, DSSIM, and LPIPS. We use within-view AURC with worst-case score ties, then average views within reconstructions and reconstructions within scenes. Paired two-sided sign-flip tests use scene identities. Mill-19 uses Monte Carlo view-level tests within its two scenes; human-target tests average renders within each scene first. Tie and fold details are in Appendices B.10 and B.7. Every row of the primary comparison is evaluated on the same held-out views.

Baselines. We compare with the released settings of FisherRF [13], Bayes' Rays [7], PPU [6], WarpRF [20], CrossScore [27], PuzzleSim [10], and PRIMU [8], alongside individual controlled channels using the same source-selection and correspondence pipeline. On the primary benchmark, both learned methods are fitted on development reconstructions, without labels from the test reconstructions; Appendix B.7 gives the fitting protocols for all regimes.

4.2 L2R-GS IMPROVES SELECTIVE ORDERING

On the primary benchmark, L2R-GS gives the lowest mean AURC among deployable scores for each target (Table 1). Relative to the strongest published baseline for L1, DSSIM, and LPIPS, it reduces AURC by 4.9%, 9.8%, and 15.8%, respectively. Against the strongest single comparator, the corresponding gains are 4.9%, 9.8%, and 5.8%. For LPIPS, that comparator is the controlled Φ channel, which itself has lower AURC than the published baselines; combining evidence lowers AURC further.

The gain over PRIMU holds in every scene for all three targets, with mean AURC reductions of 9.4%, 13.6%, and 20.8%. The advantage over the strongest single comparator also recurs in each of the three independently trained test suites (Appendix C.10), under AUGRC [23] (Appendix C.11), for scenes held out of development (Appendix C.2), and at fixed coverages of 10, 25 and 50%, where the regions it retains have the lowest mean loss for every target (Appendix C.3); scene-level results are in Appendix C.13. With in-domain fits on four additional capture regimes, L2R-GS leads on all targets except DTU L1, where PRIMU has lower AURC (Table 1, lower panel).

4.3 DOES THE PARTIAL-SUPPORT THEORY PREDICT REAL GAINS?

We test whether the pairwise decomposition predicts where fusion helps and whether cross-support placement contributes to the gain.

Within each capture regime, we fit category rates on even views, predict gain on odd views, and measure the diagnostic gain against the lower-discordance expert in each odd-view bin, on normalised losses. LLFF, NeRF-Synthetic, and DTU each have one bin near $u = 0$; the other four regimes span several support bins. Across 25 domain–bin cells per target, predicted and measured gains correlate with $r = 0.80$, 0.96 and 0.85 for L1, DSSIM and LPIPS ($r_{\mathrm{w}} = 0.87$, 0.97, 0.92; Figure 3); Mill-19 is the least accurate regime (Appendix C.6).

The gain moves from SS pairs at near-complete support to SU and UU pairs as support falls (Figure 6). Its profile need not be monotone: on the primary suites the fitted LPIPS gain peaks at

Table 1: AURC (↓) on the eleven-scene primary benchmark (top) and four additional domains (bottom). "Best published" and "Best individual signal" are per-target minima. Bold and underline mark the best and second-best deployable scores; the oracle is a lower bound.

	Mip-NeRF 360			Tanks and Temples			Deep Blending			All 11		
Method	L1	DSSIM	LPIPS	L1	DSSIM	LPIPS	L1	DSSIM	LPIPS	L1	DSSIM	LPIPS
Oracle	0.022	0.057	0.177	0.048	0.074	0.193	0.015	0.035	0.262	0.026	0.056	0.196
Published methods												
FisherRF	0.041	0.105	0.288	0.078	0.130	0.314	0.026	0.063	0.352	0.045	0.102	0.304
Bayes' Rays	0.040	0.104	0.251	0.073	0.120	0.262	0.029	0.073	0.340	0.044	0.101	0.269
PPU	0.032	0.078	0.250	**0.058**	0.098	0.254	0.021	0.050	0.336	0.035	0.077	0.266
PRIMU	0.030	0.078	0.254	0.066	0.107	0.280	0.020	0.046	0.357	0.035	0.078	0.278
WarpRF	0.028	0.078	0.259	0.064	0.111	0.300	0.019	0.045	0.361	0.033	0.078	0.285
CrossScore	0.034	0.077	0.268	0.069	0.092	0.282	0.022	0.047	0.367	0.038	0.074	0.289
PuzzleSim	0.034	0.085	0.237	0.067	0.094	0.257	0.021	0.049	0.348	0.037	0.080	0.261
Controlled individual evidence												
Visibility	0.033	0.084	0.233	0.072	0.122	0.256	0.027	0.068	0.347	0.039	0.088	0.258
RGB warp	0.029	0.077	0.253	0.065	0.107	0.288	0.025	0.061	0.366	0.035	0.079	0.280
Structural warp	0.029	0.076	0.269	0.066	0.098	0.308	0.026	0.060	0.374	0.035	0.077	0.295
VGG cosine	0.032	0.079	0.209	0.069	0.107	0.235	0.027	0.065	0.335	0.038	0.081	0.237
DINO cosine	0.036	0.091	0.232	0.072	0.122	0.236	0.028	0.069	0.340	0.041	0.093	0.252
LPIPS image warp	0.032	0.078	0.211	0.069	0.107	0.236	0.027	0.066	0.326	0.038	0.081	0.236
Primitive screen scale	0.042	0.108	0.232	0.081	0.147	0.242	0.028	0.077	0.297	0.047	0.110	0.246
Φ feature warp	0.032	0.079	0.206	0.069	0.108	0.230	0.027	0.065	0.333	0.038	0.082	0.233
Ours												
L2R-GS	**0.027**	**0.069**	**0.200**	0.060	**0.087**	**0.217**	**0.018**	**0.042**	**0.291**	**0.031**	**0.067**	**0.220**

Additional capture regimes (in-domain fits)

	Replica			NeRF-Synthetic			DTU			LLFF		
Method	L1	DSSIM	LPIPS	L1	DSSIM	LPIPS	L1	DSSIM	LPIPS	L1	DSSIM	LPIPS
Oracle	0.013	0.014	0.066	0.016	0.031	0.090	0.044	0.080	0.181	0.037	0.085	0.185
Best published	0.018	0.019	0.092	0.018	0.037	0.115	**0.052**	0.095	0.218	0.041	0.095	0.220
Best individual signal	0.021	0.021	0.082	0.018	0.038	0.105	0.055	0.090	0.206	0.041	0.095	0.193
L2R-GS	**0.018**	**0.017**	**0.077**	**0.018**	**0.037**	**0.098**	0.054	**0.090**	**0.204**	**0.040**	**0.090**	**0.191**

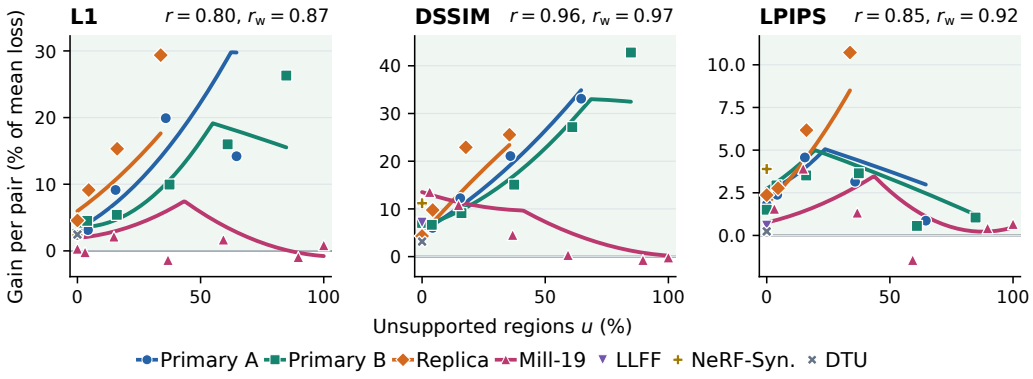


Figure 3: Gain of L2R-GS over the better single cue versus support, per pair in units of the view's mean loss: $\min_h G_h(u)$ fitted on even views (lines) and gains measured on odd views (markers); r and r_w are unweighted and view-weighted correlations over domain–bin cells.

intermediate support while the DSSIM gain grows over most of the observed range, both as fixed by the measured category rates (Appendix C.7).

Cross-support placement. An external-only score that places unsupported regions last attains the placement bound exactly and misorders 35.1%, 39.9%, and 29.7% of SU pairs under L1, DSSIM, and LPIPS. L2R-GS reduces these rates to 24.2%, 23.2%, and 21.5%, respectively (Table 8 in Appendix C.7).

Component ablation. Table 2 separates the evidence roles under a matched target-specific linear learner. External cues are stronger alone than internal cues. Adding internal cues to external ones lowers mean AURC by 1.7%, 3.2%, and 3.7% for L1, DSSIM, and LPIPS; the LPIPS gain varies across scenes ($p = 0.113$), while the other two paired tests give $p \leq 0.005$. Observability cues and then support/availability interactions lower all three mean AURCs further. These steps improve the mean in each test suite; Appendix C.12 gives every tested subset and the paired contrasts.

Table 2: Exploratory evidence-role ablation on the primary test reconstructions (AURC ↓; excess over full L2R-GS in parentheses). E, I, and O denote five external, three internal, and two observability cues. All rows share the training split and pairwise objective; only full uses support/availability interactions.

Variant	L1	DSSIM	LPIPS
External (E)	0.0326 (+3.9%)	0.0702 (+4.8%)	0.2343 (+6.6%)
Internal (I)	0.0356 (+13.5%)	0.0742 (+10.7%)	0.2393 (+8.9%)
External + internal ($E + I$)	0.0321 (+2.2%)	0.0680 (+1.5%)	0.2257 (+2.7%)
+ observability ($E + I + O$)	0.0318 (+1.4%)	0.0676 (+0.8%)	0.2228 (+1.4%)
Full L2R-GS (with interactions)	**0.0314**	**0.0670**	**0.2198**

4.4 Capture Geometry Determines the Opportunity for Fusion

LLFF has almost complete source support: only one of 6,020 regions lacks a valid source. Its strong LPIPS cross-view signal leaves only 4.5% relative room above the oracle, and almost all pairs are SS. Mill-19 has 56.2% unsupported regions in Building and 4.0% in Rubble. Across its 220 views, L2R-GS reduces AURC by 1.9%, 2.8%, and 7.6% for L1, DSSIM, and LPIPS relative to the strongest single signal. The L1 view-level comparison is inconclusive ($p = 0.219$), and two scenes do not establish cross-scene generalisation; the data and reconstruction protocol are in Appendix C.14.

4.5 Beyond Reconstruction Losses: Human-Perceived Artifacts

PuzzleSim [10] released artifact masks from 22 human observers. We use them as a target for a monotone ranker based on external and observability cues; the original 3DGS models are unavailable (Appendix C.15). On 15 confirmation renders from 10 scenes, each excluded during fitting, AURC is 14.0% below PuzzleSim ($p = 0.033$ over scenes) and 13.0% below Φ ($p = 0.105$, inconclusive). Φ is undefined on 10.5% of regions containing 22.6% of the annotated artifact mass.

5 Conclusion

L2R-GS ranks regions of frozen NVS renders for a declared target, combining cross-view and Gaussian-state cues with support-aware weights. It lowers AURC below published methods on the primary benchmark, also for scenes held out of development, and the partial-support decomposition predicts where fusion helps. A separate human-label experiment extends the target test to perceived artifacts.

Limitations. L2R-GS needs labelled development views from the capture regime in which it is deployed; we fit each additional regime in-domain and do not claim transfer across regimes. When source support is almost complete, as on LLFF, a single cross-view cue already comes close to the oracle and fusion has little room to help. The gain profile of Proposition 2 rests on category rates that stay stable as support changes; on Mill-19 it does not predict the L1 and LPIPS profiles. The method selects regions but does not repair them.

Future work. Natural next steps are to act on the ranking, by re-rendering, densifying or refining the regions it rejects or by choosing new views that supply the missing support; to adapt the ranker to a new capture regime from few or no labels; and to apply the same target-declared formulation to other radiance field representations and to feed-forward reconstruction.

AI USE STATEMENT

Generative AI tools assisted with literature discovery, manuscript drafting and editing, figure preparation, and feedback on code, experiments, and mathematical arguments. The authors take responsibility for verifying all AI-assisted material and for the final content of the paper.

ETHICS STATEMENT

This work uses existing research datasets and previously released human annotations. It does not involve new participant recruitment or collection of personal data. The reliability estimates may help users identify uncertain regions in rendered views, but they do not provide a safety guarantee.

REPRODUCIBILITY STATEMENT

The paper and appendix describe the data splits, implementation, evaluation protocol, and theoretical proofs. Code for every reported result, together with the frozen view splits, is provided in the supplementary material.

REFERENCES

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

[2] Sarosij Bose, Arindam Dutta, Sayak Nag, Junge Zhang, Jiachen Li, Konstantinos Karydis, and Amit K. Roy-Chowdhury. Uncertainty-aware diffusion-guided refinement of 3D scenes. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[3] Christopher J. C. Burges. From RankNet to LambdaRank to LambdaMART: An overview. Technical Report MSR-TR-2010-82, Microsoft Research, 2010.

[4] C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1), 1970.

[5] Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11, 2010.

[6] Chamuditha Jayanga Galappaththige, Thomas Gottwald, Peter Stehr, Edgar Heinert, Niko Suenderhauf, Dimity Miller, and Matthias Rottmann. Predictive photometric uncertainty in Gaussian splatting for novel view synthesis. In *European Conference on Computer Vision (ECCV)*, 2026.

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

[8] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[9] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. *ACM Transactions on Graphics*, 37(6), 2018.

[10] Nicolai Hermann, Jorge Condor, and Piotr Didyk. Puzzle similarity: A perceptually-guided cross-reference metric for artifact detection in 3D scene reconstructions. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[11] Youngkyoon Jang and Eduardo Pérez-Pellitero. CoMapGS: Covisibility map-based Gaussian splatting for sparse novel view synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[12] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[13] Wen Jiang, Boshu Lei, and Kostas Daniilidis. FisherRF: Active view selection and mapping with radiance fields using Fisher information. In *European Conference on Computer Vision (ECCV)*, 2024.

[14] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, 2017.

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

[16] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and Temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017.

[17] Yang Liu, Chuanchen Luo, Lue Fan, Naiyan Wang, Junran Peng, and Zhaoxiang Zhang. CityGaussian: Real-time high-quality large-scale scene rendering with Gaussians. In *European Conference on Computer Vision (ECCV)*, 2024.

[18] 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*, 38(4), 2019.

[19] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020.

[20] Sadra Safadoust, Fabio Tosi, Fatma Güney, and Matteo Poggi. WarpRF: Multi-view consistency for training-free uncertainty quantification and applications in radiance fields. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026.

[21] Luca Savant Aira, Diego Valsesia, and Enrico Magli. Modeling uncertainty for Gaussian splatting. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6), 2025.

[22] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[23] Jeremias Traub, Till J. Bungert, Carsten T. Lüth, Michael Baumgartner, Klaus H. Maier-Hein, Lena Maier-Hein, and Paul F. Jäger. Overcoming common flaws in the evaluation of selective classification systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[24] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-NeRF: Scalable construction of large-scale NeRFs for virtual fly-throughs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[25] Mike Van Ness, Tomas M. Bosschieter, Roberto Halpin-Gregorio, and Madeleine Udell. The missing indicator method: From low to high dimensions. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2023.

[26] 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), 2004.

[27] Zirui Wang, Wenjing Bian, and Victor Adrian Prisacariu. CrossScore: Towards multi-view image evaluation and scoring. In *European Conference on Computer Vision (ECCV)*, 2024.

[28] Shangjie Xue, Jesse Dill, Pranay Mathur, Frank Dellaert, Panagiotis Tsiotras, and Danfei Xu. Neural visibility field for uncertainty-driven active mapping. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[29] Shangjie Xue, Jesse Dill, Dhruv Ahuja, Frank Dellaert, Panagiotis Tsiotras, and Danfei Xu. Uncertainty-driven 3D Gaussian splatting active mapping via anisotropic visibility field. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[30] 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)*, 2018.

[31] Han Zhou, Jordy Van Landeghem, Teodora Popordanoska, and Matthew B. Blaschko. A novel characterization of the population area under the risk coverage curve (AURC) and rates of finite sample estimators. In *International Conference on Machine Learning (ICML)*, 2025.

A Theory

This appendix states the results of Section 3.2 formally and proves them. Throughout, a view has $N \geq 2$ equal-area regions $\mathcal{P} = \{p_1, \ldots, p_N\}$; for target t, $\ell_t(p) \geq 0$ is the realised loss of region p, Z_v is the inference-time evidence of the whole view, and $\eta_{t,p}(Z_v) = \mathbb{E}[\ell_t(p) \mid Z_v]$. A selector retaining k regions chooses $K \subseteq \mathcal{P}$ with $|K| = k$ and retained risk $R_t(K) = \frac{1}{k}\sum_{p\in K} \ell_t(p)$.

A.1 Optimality of the Target-Specific Ordering

Proposition A.1 (Oracle top-k optimality)**.** *Let $K_t^\star(k)$ hold the k smallest values of $\{\ell_t(p)\}_{p\in\mathcal{P}}$. Then $R_t(K_t^\star(k)) \leq R_t(K)$ for every K with $|K| = k$.*

Proof. If K contains a region i while excluding a region j with $\ell_t(j) < \ell_t(i)$, exchanging them changes the retained risk by $\frac{1}{k}(\ell_t(j) - \ell_t(i)) < 0$. Repeating such exchanges reaches a set of the k smallest losses. Ties may give several such sets, all with the same risk as $K_t^\star(k)$. □

Proposition A.2 (Conditional-risk optimality)**.** *Conditioned on Z_v, the expected retained risk of a Z_v-measurable set K is $\mathbb{E}[R_t(K) \mid Z_v] = \frac{1}{k}\sum_{p\in K} \eta_{t,p}(Z_v)$. Consequently the set holding the k smallest $\eta_{t,p}(Z_v)$ minimises expected retained risk among all size-k selectors measurable with respect to Z_v.*

Proof. By linearity of conditional expectation, $\mathbb{E}[\frac{1}{k}\sum_{p\in K} \ell_t(p) \mid Z_v] = \frac{1}{k}\sum_{p\in K} \eta_{t,p}(Z_v)$. Applying the exchange argument of Proposition A.1 to the values $\eta_{t,p}(Z_v)$ gives the claim. Independence between regional errors is not required. □

Let a permutation π give the order in which regions are retained and let $a_t(p)$ denote either $\ell_t(p)$ or $\eta_{t,p}(Z_v)$. Then

$$A_t(\pi) = \frac{1}{N}\sum_{k=1}^{N} \frac{1}{k}\sum_{i=1}^{k} a_t(\pi_i) = \sum_{i=1}^{N} w_i\, a_t(\pi_i), \qquad w_i = \frac{1}{N}\sum_{k=i}^{N} \frac{1}{k}, \tag{12}$$

and the weights satisfy $w_1 > w_2 > \cdots > w_N > 0$ with $w_i - w_{i+1} = 1/(Ni)$.

Proposition A.3 (AURC-optimal ordering)**.** *The ordering with $a_t(\pi_1) \leq a_t(\pi_2) \leq \cdots \leq a_t(\pi_N)$ minimises $A_t(\pi)$ over all permutations. With $a_t(p) = \ell_t(p)$ this is the evaluation oracle; with $a_t(p) = \eta_{t,p}(Z_v)$ it is the ordering of minimal expected AURC conditional on Z_v.*

Proof. Equation (12) is a weighted sum whose positional weights decrease monotonically. By the rearrangement inequality the minimum pairs the largest weights with the smallest values, so sorting by increasing a_t minimises every prefix risk and the complete AURC simultaneously. □

Proposition A.4 (Target reversal requires target-dependent weights)**.** *Fix the inference-time evidence Z_v and consider two targets t_1, t_2 and regions a, b such that $\eta_{t_1,a}(Z_v) < \eta_{t_1,b}(Z_v)$ and $\eta_{t_2,a}(Z_v) > \eta_{t_2,b}(Z_v)$. Let $N \geq 2$ be the number of regions. For any parameters $\theta = (\beta, \gamma, \delta)$ shared across the two targets in Eq.* (10)*, at least one target has conditional expected AURC regret of at least*

$$\frac{\min\{\eta_{t_1,b}(Z_v) - \eta_{t_1,a}(Z_v),\, \eta_{t_2,a}(Z_v) - \eta_{t_2,b}(Z_v)\}}{N(N-1)} > 0. \tag{13}$$

The same conclusion holds for any deterministic target-independent ordering.

Proof. With θ fixed, Eq. (10) assigns the same two scores to a, b under both targets. After any target-independent tie break, it must place either a or b first. The first order misorders this pair for t_2; the second misorders it for t_1. Conditional on Z_v, Eq. (12) has the same decreasing positional weights with regional losses replaced by $\eta_{t,p}(Z_v)$. Applying the pairwise regret bound of Proposition 1 to these conditional risks gives the stated lower bound for whichever target has the reversed pair. □

Lemma A.5 (Monotone invariance)**.** *For every strictly increasing g,* $\mathrm{RankCDF}(g(x(p))) = \mathrm{RankCDF}(x(p))$ *for all p.*

Proof. For any two values, $x_i < x_j \iff g(x_i) < g(x_j)$, and equality is preserved, so both the ordering and the tie groups are unchanged. Average ranks and their normalised values are therefore identical. □

The lemma justifies the within-view canonicalisation of Section 3.4: a ranker that reads only rank-CDFs is invariant to any strictly monotone rescaling of a producer, and it deliberately discards magnitude differences between producers whose scales are not comparable.

A.2 Pairwise Form of the AURC Regret

Proof of Proposition 1. Let π be the retention order of q under worst-case ties. Bubble-sort π: repeatedly swap an adjacent pair at positions $(i, i+1)$ with $\ell_{\pi(i)} > \ell_{\pi(i+1)}$. Such a swap changes only the relative order of those two regions, so it removes exactly one misordered pair and creates none, and by (12) it changes the AURC by $-(w_i - w_{i+1})(\ell_{\pi(i)} - \ell_{\pi(i+1)})$ with $w_i - w_{i+1} = 1/(Ni)$. Every misordered pair is removed exactly once, pairs of equal loss are never swapped, and the procedure ends at an ordering with the oracle's AURC. Summing the changes gives (3) with $\kappa_{ab} = 1/(Ni)$ for the position i at which the pair was swapped, $i \in \{1, \ldots, N-1\}$, hence $\kappa_{ab} \in [1/(N(N-1)), 1/N]$. The sandwich follows, and the regret is zero if and only if no pair is misordered. □

The result is useful because W is additive across pair categories even though the AURC itself is not. The pairwise logistic objective (11) targets the order of these pairs, while omitting their loss-gap and positional weights.

A.3 Partial-Support Decomposition

For a realised view, let n_c count pairs of category c, and let $W_c(q)$ be the loss-weighted discordance over those pairs. Define $\bar{\omega}_c(q) = W_c(q)/n_c$ when $n_c > 0$. Then $W(q) = \sum_{c \in \{SS, SU, UU\}} n_c \bar{\omega}_c(q)$ exactly, without any independence assumption. Predicting the gain at a different support level requires a further assumption about these category rates.

The identity holds for any loss, and in particular for the view-normalised loss $\tilde{\ell}_t = \ell_t / \bar{\ell}_t$ of Section 3.2. For $\bar{\ell}_t > 0$, the set $\mathcal{D}(q)$ of misordered pairs is the same under ℓ_t and $\tilde{\ell}_t$, and $\tilde{W}(q) = W(q)/\bar{\ell}_t$. The normalisation therefore changes no ordering, only the unit in which a view's misorder is counted. It matters for the rate assumption below: if the loss level of a view changes with its unsupported share, the unnormalised rates need not remain stable in u even when a score's ordering behaviour is.

Assumption A.1 (Support mixture and category-rate stability)**.** *For each support level u, region support indicators are independent, with probability u of being unsupported. For a score q and pair category c, the expected normalised loss-weighted misorder of a pair conditional on its category*

and the view's unsupported count is $\omega_c(q)$, independent of that count and of u. The score may use evidence from the whole view, including within-view rank canonicalisation.

Proof of Proposition 2. Conditioned on m unsupported regions, there are $\binom{N-m}{2}$ SS pairs, $m(N-m)$ SU pairs, and $\binom{m}{2}$ UU pairs. By the assumed category rates, their expected contributions sum to

$$\mathbb{E}[\tilde{W}(q) \mid |U| = m] = \binom{N-m}{2}\omega_{SS}(q) + m(N-m)\omega_{SU}(q) + \binom{m}{2}\omega_{UU}(q). \tag{14}$$

Under independent support indicators, a random pair is SS, SU or UU with probabilities $(1-u)^2$, $2u(1-u)$ and u^2. Taking expectation over the $N(N-1)/2$ pairs gives (4). □

For a view with observed fraction $\hat{u} = m/N$, the coefficients in (14), divided by $\binom{N}{2}$, differ from the three coefficients in (4) by terms of order $1/N$. The held-out prediction uses the simpler quadratic with the bin mean of $\hat{u}$ as a plug-in; its accuracy is measured rather than assumed.

For a fused score f and an individual expert h, write $\Delta_c^h = \omega_c(h) - \omega_c(f)$ for the improvement of f inside category c. Subtracting (4) for f from (4) for h, the expected gain over h is $\frac{N(N-1)}{2}G_h(u)$ with

$$G_h(u) = (1-u)^2\Delta_{SS}^h + 2u(1-u)\Delta_{SU}^h + u^2\Delta_{UU}^h, \tag{15}$$

and the gain over the better of two experts $h \in \{E, I\}$ is $\min_h G_h(u)$ up to the same factor.

Corollary A.6 (Shape of the gain)**.** *$G_h(0) = \Delta_{SS}^h$ and $G_h(1) = \Delta_{UU}^h$. G_h has an interior maximum on $(0,1)$ if and only if $\Delta_{SU}^h > \max(\Delta_{SS}^h, \Delta_{UU}^h)$, located at $u = (\Delta_{SS}^h - \Delta_{SU}^h)/(\Delta_{SS}^h + \Delta_{UU}^h - 2\Delta_{SU}^h)$; otherwise its maximum on $[0,1]$ is at an endpoint. If each expert is exact in its own domain, $\Delta_{SS}^E = \Delta_{UU}^I = 0$, and the cross terms are negligible, the two branches of $\min_h G_h$ cross at $u^\star = 1/\big(1 + \sqrt{\Delta_{UU}^E/\Delta_{SS}^I}\big)$.*

Proof. Write $A = \Delta_{SS}^h$, $B = \Delta_{UU}^h$, $C = \Delta_{SU}^h$. G_h is a quadratic with $G_h(0) = A$, $G_h(1) = B$ and $G_h'' = 2(A + B - 2C)$. A critical point $u_v = (A - C)/(A + B - 2C)$ is a maximum iff $A + B - 2C < 0$, and lies in $(0,1)$ iff $A - C < 0$ and $B - C < 0$, i.e. iff $C > \max(A, B)$. The crossing point solves $u^2\Delta_{UU}^E = (1-u)^2\Delta_{SS}^I$. □

The profile of the gain in u is therefore not a law but a prediction from three measurable rates; Section 4.3 tests it on held-out views.

A.4 Cross-Support Placement Bound

Proof of Proposition 3. Over the uniform distribution on the SU pairs with untied losses, let A be the event that the unsupported region has the larger loss and B the event that the score retains it first, so $P(A) = \rho$ and $P(B) = \pi$. A pair is misordered iff $(A \wedge B)$ or $(\neg A \wedge \neg B)$; ties in the score are broken worst-case, so the bound is for the realised order. By the Fréchet bounds $P(A \wedge B) \geq \rho + \pi - 1$ and $P(\neg A \wedge \neg B) \geq 1 - \rho - \pi$, and both probabilities are non-negative, so $m \geq \max(0, \rho + \pi - 1) + \max(0, 1 - \rho - \pi) = |\rho + \pi - 1|$. At $\pi = 0$ every SU pair whose supported region is the worse one is misordered and no other, so $m = 1 - \rho$; at $\pi = 1$ symmetrically $m = \rho$; the oracle retains the lower loss first, so $\pi = 1 - \rho$ and $m = 0$. □

The bound formalises why an external score can be accurate within the supported set while still failing systematically across the support boundary: $m - |\rho + \pi - 1|$ is the part of the cross-support error that placement alone cannot remove.

For a score q_E that places every unsupported region after every supported one, the loss-weighted cross-support cost is

$$W_{SU}(q_E) = \sum_{u \in U}\sum_{s \in S}[\ell_t(s) - \ell_t(u)]_+, \tag{16}$$

where $[z]_+ = \max(z, 0)$. Proposition 1 gives $\mathrm{AURC}(q_E) - \mathrm{AURC}(\text{oracle}) \geq W_{SU}(q_E)/(N(N-1))$. Thus the cost of the placement rule depends on the declared target's loss gaps as well as the number of pairs it misorders.

A.5 Scope of the Theory

The pairwise regret identity, the realised-view category decomposition, and the placement bound hold without independent regions. The quadratic gain profile additionally assumes category-rate stability as support changes, stated for view-normalised losses; within-view rank canonicalisation can change those rates with the composition of the view. We test the resulting prediction on disjoint real views within each capture regime (Section 4.3), without assuming that rates transfer across regimes or that L2R-GS recovers the conditional-risk oracle.

B Evidence Pipeline

B.1 Regional Target Losses

The L1 map is the per-pixel mean absolute RGB difference. The DSSIM map is $(1 - \mathrm{SSIM}_{\mathrm{local}})/2$, evaluated with a Gaussian local window and reflection padding. The LPIPS target is its spatial VGG feature-loss map, rather than a whole-image scalar. Each map is computed before cropping to the common evaluation region and mean-pooling over the patch grid. Thus the loss of a patch includes the image context used to compute DSSIM and LPIPS, while coverage remains a count of equal-area patches.

B.2 Source Selection and Correspondence Validity

For each target view we use the eight nearest training cameras as source candidates. A target pixel x and source s form a valid correspondence only when all of the following hold: the reprojected depth is positive; the projection lies inside the source image; the target accumulated alpha is at least 0.5; the warped-source alpha is at least 0.5; the warped-source depth agrees with the target rendered depth within 5% relative error; the back-projected depth is positive; and the forward–backward reprojection error is below 1.5 pixels. Let $\mathcal{S}_x$ contain the candidates passing these tests at pixel x. A region is supported when $\mathcal{S}_x$ is nonempty for at least one pixel $x \in p$; its support strength $n_{\mathrm{src}}(p)$ is the mean of $|\mathcal{S}_x|$ over all pixels in the region, as in Section 3.3. All controlled cross-view signals use the same source candidates, geometry, validity operator and source aggregation, so differences between their outputs are attributable to the discrepancy rather than to different source selection. NeRF-Synthetic and DTU render an object against empty background, where a warp is invalid by construction, so on these two families every method is scored on the same object estimand: a region is evaluated when less than half of its ground-truth pixels are background (NeRF-Synthetic) or more than half lie inside the object mask (DTU), and a view is kept when it holds at least four such regions.

Figure 4 illustrates this support definition on held-out views. The source thumbnail is one of the eight candidate training views; the support mask uses all eight after correspondence validation. Gray regions indicate where controlled cross-view evidence cannot be computed, not necessarily where the render is wrong.

B.3 External Evidence

The controlled channels compute a per-pixel discrepancy between the target render and each valid warped source, take the minimum over the sources valid at that pixel, and mean-pool over pixels with a defined discrepancy in the region, as in (6). The family includes a photometric discrepancy, the ℓ_1 distance between the render and the warped source; a structural discrepancy on fixed Sobel- and Laplacian-derived luminance representations at several scales; and feature-space discrepancies on frozen VGG, DINO and LPIPS-style features [30]. The bank uses the LPIPS-feature channel Φ; the others are reported as individual rows of Table 1. Two further channels complete the controlled block: cross-view render consistency, the LPIPS discrepancy between the target render and the render of the same reconstruction from each source camera warped into the target view, minimum over sources; and source–source agreement, the RGB discrepancy between pairs of warped sources, mean over jointly valid pairs. These three bank channels are undefined when no pixel in the region has a valid source, and source–source agreement also requires at least one pixel with two jointly valid sources. The dense RGB render-warp residual uses the WarpRF operator [20]; the reference-conditioned quality score uses the released CrossScore predictor [27]. Both assign a value to every region under their native definitions. Their parameters are not optimised with L2R-GS.

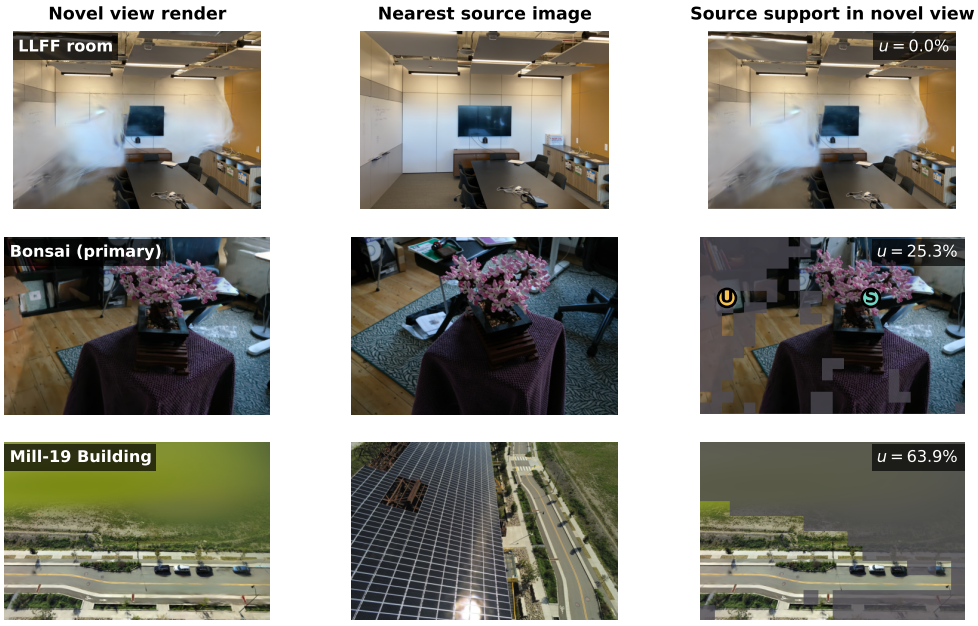


Figure 4: Examples of source support in three capture regimes. Each row shows a held-out render (left), its nearest training source image (middle), and the render with unsupported 64×64 regions shaded gray (right). A region is unsupported only if no pixel has a valid correspondence from any of the eight candidate sources; u is its share of the view's regions. The marked S and U regions in Bonsai form one pair crossing the support boundary. The examples range from near-complete support in LLFF to extensive missing support in Mill-19.

B.4 Internal Evidence

With accumulated alpha $\alpha_{\text{acc}}(x) = \sum_i w_i(x)$ over the Gaussians contributing to pixel x, a primitive quantity is projected as in (7) where $\alpha_{\text{acc}}(x) \geq 10^{-3}$ and is undefined otherwise, then pooled within the region. The bank uses the projected screen scale $q_i^{\text{scale}} = \log(\bar{s}_i / z_i)$, with $\bar{s}_i$ the mean Gaussian scale and z_i the camera-space depth, which increases where large or coarse projected primitives dominate a pixel; the depth variation $E_{\text{depth}}(p) = \text{Pool}_{x \in p} \|\nabla \log D(x)\|$ of the rendered depth D, computed with a Sobel operator; and the frozen primitive-level predictive photometric uncertainty [6], projected through the same blend weights, pooled within the region, and never refitted.

B.5 Observability Signals

The bank uses $n_{\text{src}}(p)$ and the local texture energy $E_{\text{tex}}(p) = \text{Pool}_{x \in p} \|\nabla \hat{I}_v(x)\|$, the gradient magnitude of the render's channel mean, taken from the render and not from the ground truth so that it exists at deployment. The two gating variables of (8) are s_p, clipped to $[0, 1]$, and a_p, the share of the five external channels defined at p. These variables identify the observation regime and are not assumed to be monotone risk estimators.

B.6 Canonicalisation and Missing Evidence

Finite values of each signal are converted to average within-view ranks divided by the number of finite values of that signal in the view, as in (9), so defined values lie in $(0, 1]$. Undefined values receive $R_{p,j} = 0$, which removes the term from the score and preserves missingness without assigning

an artificial high or low risk. For the boosted trees an out-of-range sentinel and a binary missingness indicator are used instead, on which a tree can branch.

B.7 Training Details

For every training view: render the frozen reconstruction; divide the view into the fixed regions; compute the regional target losses; compute the frozen evidence; canonicalise it within the view; convert the target losses to within-view rank labels; sample eight same-view partners per region with replacement, discarding self-pairs and equal-label pairs; and optimise (11) with L-BFGS. The regularisation coefficient is selected from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ by a three-fold cross-validation over the development scene identities inside the training identities of each fold; the selected value, $\lambda = 10^{-4}$, is frozen for every domain. The ten signals were chosen by forward selection on the development reconstructions from a larger pool of frozen signals, adding at each step the signal that most reduced the mean cross-validated AURC over the three targets. No test reconstruction participates in evidence selection or hyperparameter selection.

The primary test uses one model fitted on the two development reconstructions of each scene. For the original additional-domain fits, Replica (8 scenes), NeRF-Synthetic (8 scenes), and DTU (15 scenes) use leave-one-scene-out fitting within each dataset. LLFF (8 scenes) uses leave-one-view-out fitting within each scene, with the evaluated view excluded. Mill-19 uses five folds over views within each of its two scenes. Thus these additional-domain fits measure adaptation within a capture regime, not transfer from the primary benchmark or to an unseen capture regime.

For the primary PRIMU comparison, its primitive representation uses the training images of each frozen reconstruction. We fit one photometric-error regressor on the held-out views of the development reconstructions, sampling at most 20,000 pixels per view (1,076 views and 21.5 million pixels), and apply it without refitting to all 1410 held-out test views. No test-reconstruction label enters the fit. The adapter calls PRIMU's released Gaussian-representation, feature-map, and histogram gradient-boosting regressor code, retaining its 13 pixel features, degree-4 directional representation, concentration parameter 8, and regressor settings. The cross-reconstruction fit and pixel subsampling differ from PRIMU's original per-scene training protocol. The checkpoint is exported with a numerical round-trip check, and the RGB error pass uses the benchmark's frozen render so PRIMU predicts the error scored for every other method. Its released camera path assumes a centered principal point; the adapter supplies the benchmark's intrinsic matrix to its feature renderer and the matching off-centre projection to its backprojector. A render-agreement audit checks the conversion.

B.8 Higher-Capacity Reference

The ablation's gradient-boosted trees [14] use the same ten signals as the linear ranker, with 300 trees and 31 leaves and either a LambdaRank objective [3], with each view as one query group over 32 relevance grades, or regression to the within-view rank label, selected per target on development data. They measure the effect of a higher-capacity interaction function while holding the evidence fixed.

B.9 Inference

Inference requires no target image: render $\rightarrow$ compute evidence $\rightarrow$ canonicalise $\rightarrow$ evaluate $q_t \rightarrow$ sort $\rightarrow$ retain the prefix of length $\lceil cN_v \rceil$. The reconstruction, the producers and the ranker are unchanged as c varies, and the same ranking serves all coverage levels.

B.10 Worst-Case Handling of Score Ties

The deployed action needs only a deterministic ordering, but an evaluation should not credit a method for the accidental array order of regions it scores identically. For every tied group we therefore order regions by decreasing held-out target loss before computing the risk–coverage curve, so a tie is evaluated under its worst admissible ordering. The target loss is used only inside the evaluator and never changes an inference-time score. The policy matters most for unavailable external evi-

Table 3: Cross-domain AURC (↓). Every learned row, L2R-GS and PRIMU alike, is fitted in each domain's own regime with the same folds: leave-one-scene-out on Replica, NeRF-Synthetic and DTU (the last two separately) and leave-one-view-out inside each scene on LLFF; no row reads ground truth of the views it scores (Appendix B.7). On DTU the PPU adapter fails its own render-agreement check because its rasteriser assumes a centred principal point; a dash marks these cells. Bold and underline mark the best and second-best deployable scores; the oracle is a lower bound.

	Replica			NeRF-Synthetic			DTU			LLFF		
Method	L1	DSSIM	LPIPS	L1	DSSIM	LPIPS	L1	DSSIM	LPIPS	L1	DSSIM	LPIPS
Oracle	0.013	0.014	0.066	0.016	0.031	0.090	0.044	0.080	0.181	0.037	0.085	0.185
Published methods												
FisherRF	0.023	0.034	0.143	0.024	0.047	0.115	0.065	0.116	0.250	0.050	0.109	0.236
Bayes' Rays	0.028	0.047	0.148	0.031	0.064	0.137	0.070	0.130	0.267	0.056	0.118	0.250
PPU	0.021	0.024	0.102	0.025	0.054	0.150	—	—	—	—	—	—
PRIMU	0.018	0.020	0.103	0.018	0.037	0.120	**0.052**	0.095	0.222	0.041	0.095	0.228
WarpRF	0.018	0.019	0.095	0.018	0.041	0.129	0.060	0.099	0.234	0.042	0.099	0.237
CrossScore	0.019	0.019	0.099	0.022	0.045	0.137	0.066	0.103	0.229	0.052	0.114	0.271
PuzzleSim	0.020	0.020	0.092	0.020	0.039	0.116	0.061	0.095	0.218	0.045	0.097	0.220
Controlled individual evidence												
Visibility	0.023	0.028	0.104	0.026	0.060	0.157	0.055	0.097	0.218	0.047	0.103	0.226
RGB warp	0.023	0.023	0.101	0.018	0.041	0.134	0.059	0.096	0.229	0.041	0.095	0.224
Structural warp	0.021	0.021	0.103	0.019	0.042	0.141	0.063	0.097	0.235	0.042	0.095	0.246
VGG cosine	0.024	0.024	0.085	0.020	0.039	0.108	0.057	0.090	0.206	0.044	0.095	0.195
DINO cosine	0.025	0.029	0.098	0.025	0.052	0.129	0.056	0.095	0.216	0.046	0.103	0.205
LPIPS image warp	0.024	0.023	0.082	0.020	0.038	0.105	0.058	0.091	0.206	0.044	0.095	0.196
Primitive screen scale	0.023	0.035	0.110	0.032	0.065	0.131	0.058	0.110	0.235	0.058	0.130	0.242
Φ feature warp	0.024	0.024	0.084	0.021	0.040	0.105	0.058	0.091	0.206	0.044	0.095	0.193
Ours												
L2R-GS	**0.018**	**0.017**	**0.077**	**0.018**	**0.037**	**0.098**	0.054	**0.090**	**0.204**	**0.040**	**0.090**	**0.191**

dence: a method that assigns the same undefined state to many unsupported regions is not assumed to have ordered that group favourably.

C Additional Results

C.1 Cross-Domain Benchmark

Table 3 gives every row on the four additional capture regimes, Replica, NeRF-Synthetic, DTU and LLFF; the lower panel of Table 1 summarises it. Each learned row is fitted inside the regime with the folds listed in the caption, so no row reads ground truth of the views it scores. L2R-GS has the lowest AURC among deployable scores for every regime and target except DTU under L1, where PRIMU is lower.

C.2 Scene-Level Holdout on the Primary Benchmark

The test reconstructions of the primary benchmark are independent reconstructions, with their own held-out views, of scenes that also supply development reconstructions. We hold out one scene identity at a time from ranker fitting: L2R-GS and PRIMU are refitted on the development reconstructions of the other ten scenes, then evaluated on the three test reconstructions of the held-out scene. L2R-GS keeps its ten signals, gates, pairwise loss and $\lambda = 10^{-4}$; PRIMU keeps its released features and regressor with the pixel budget of the primary comparison. The signal bank and λ were selected on development data and remain fixed across folds. No test reconstruction informs these choices. Rows without a benchmark-specific fit are unchanged.

Table 4 reports the result. With the scored scene never seen during fitting, L2R-GS lowers AURC relative to the strongest single row by 4.7% (L1, $p < 0.001$), 9.5% (DSSIM, $p < 0.001$) and 5.2% (LPIPS, $p = 0.021$), and relative to PRIMU fitted under the same folds by 12.9–21.6% ($p < 0.001$ for every target). Holding a scene out raises the AURC of L2R-GS by only 0.20–0.65% relative to the fit on all eleven development scenes; the increase is consistent across scenes, so it is significant despite its size.

Table 4: Leave-one-scene-out on the primary benchmark (AURC ↓, identity macro over 11 scenes). L2R-GS and PRIMU are refitted without the scored scene; Δ is the change in AURC of L2R-GS relative to the compared row, with paired sign-flip p. The last two columns compare with L2R-GS fitted on all development scenes (Table 1).

Target	strongest single row	PRIMU	**L2R-GS**	Δ, p vs single	Δ, p vs PRIMU	L2R-GS, all scenes	Δ, p
L1	0.0330 (WarpRF)	0.0361	0.0314	-4.7%, <0.001	-12.9%, <0.001	0.0314	+0.2%, 0.003
DSSIM	0.0743 (CrossScore)	0.0790	0.0673	-9.5%, <0.001	-14.9%, <0.001	0.0670	+0.4%, <0.001
LPIPS	0.2333 (Φ)	0.2822	0.2212	-5.2%, 0.021	-21.6%, <0.001	0.2198	+0.7%, <0.001

C.3 Selective Risk at Fixed Coverages

The AURC averages selective risk over all coverages. Table 5 reports the risk itself, the mean target loss of the retained regions, when a view keeps its first 10, 25 or 50% of regions, with the worst-case tie rule of the AURC and the identity macro of Table 1. L2R-GS has the lowest risk among deployable scores in every cell. Relative to the strongest other row of each cell, which may differ from the rows shown, its risk is 5.6–16.4% lower. At 10% coverage the reductions are 8.0% for L1 (against WarpRF, $p = 0.038$), 13.8% for DSSIM (structural warp, $p < 0.001$) and 5.7% for LPIPS (Φ, $p = 0.087$); at 25 and 50% every paired test gives $p < 0.01$.

Table 5: Selective risk (↓), the mean target loss of the regions retained at 10, 25 and 50% coverage, on the primary test reconstructions (identity macro). Bold marks the lowest deployable risk; the oracle is a lower bound.

	L1			DSSIM			LPIPS		
Method	10%	25%	50%	10%	25%	50%	10%	25%	50%
Random	0.0448	0.0446	0.0445	0.1001	0.1001	0.0998	0.2892	0.2887	0.2885
PRIMU	0.0284	0.0311	0.0345	0.0621	0.0683	0.0772	0.2792	0.2751	0.2740
WarpRF	0.0245	0.0285	0.0332	0.0604	0.0689	0.0792	0.2920	0.2870	0.2818
CrossScore	0.0341	0.0359	0.0378	0.0547	0.0630	0.0746	0.3090	0.2946	0.2812
Structural warp	0.0279	0.0313	0.0352	0.0500	0.0631	0.0790	0.3126	0.2997	0.2885
Φ feature warp	0.0317	0.0344	0.0380	0.0619	0.0714	0.0836	0.1778	0.2050	0.2358
L2R-GS	**0.0226**	**0.0264**	**0.0313**	**0.0431**	**0.0527**	**0.0667**	**0.1676**	**0.1894**	**0.2199**
Oracle	0.0147	0.0189	0.0250	0.0285	0.0387	0.0546	0.1249	0.1558	0.1966

Figure 5 shows what is retained in two test views under DSSIM at 25% coverage. The views are chosen by a fixed rule, among the 1410 test views, as those at the 90th and 5th percentile of the gain of L2R-GS over CrossScore (the strongest single row for DSSIM) in selective risk, normalised by the gap between CrossScore and the oracle. In the first (drjohnson), L2R-GS retains regions with mean loss 0.0216 against 0.0295 for CrossScore and 0.0184 for the oracle. In the second (train), support is almost complete ($u = 1\%$), the regime in which fusion has least to add, and L2R-GS retains regions with mean loss 0.0290 against 0.0259 for CrossScore. L2R-GS retains regions of higher loss than CrossScore at this coverage in 15.0% of test views.

C.4 Pair Weighting of the Training Objective

Proposition 1 weights a misordered pair by its loss gap and a positional factor, whereas Eq. (11) weights every sampled pair equally. Table 6 refits L2R-GS with the same sampled pairs weighted by the loss gap, by the loss gap divided by the view's mean loss, and by the exact regret weight of Proposition 1 at the oracle positions, each rescaled to mean one; signals, design, λ, sampler and seed are unchanged. On development data with each scene held out, and on the test reconstructions with the fit on all development data, no weighting changes AURC by more than 0.63% ($p \geq 0.167$ for every comparison). The uniform surrogate therefore loses nothing measurable against the regret weights, and we keep it.

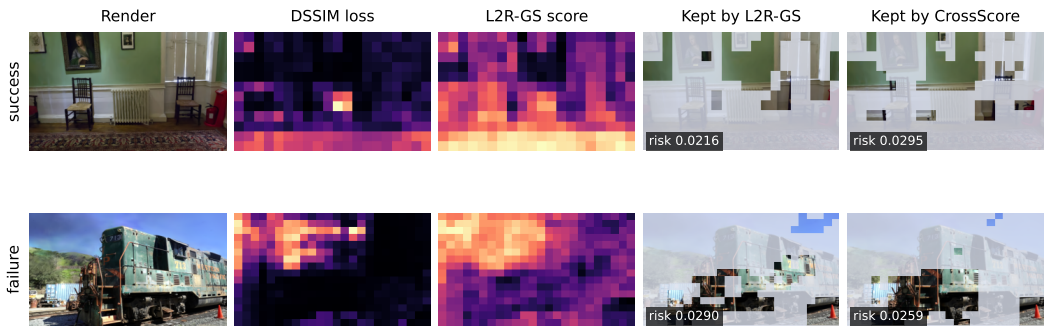


Figure 5: Two test views under DSSIM at 25% coverage, at the 90th (top) and 5th (bottom) percentile of the per-view gain of L2R-GS over CrossScore. From left: render, regional DSSIM loss, L2R-GS score (brighter is retained later), and the regions each score retains (others veiled); "risk" is the mean loss of the retained regions.

Table 6: Pair weighting of the pairwise objective. The first row gives the AURC of L2R-GS (uniform weights); the others give the relative change in AURC from reweighting the same pairs, with paired sign-flip p over scenes.

	development, scene held out			test reconstructions		
Pair weight	L1	DSSIM	LPIPS	L1	DSSIM	LPIPS
uniform (L2R-GS)	0.0345	0.0728	0.2362	0.0314	0.0670	0.2198
loss gap $\lvert\ell_i - \ell_j\rvert$	-0.06% (0.851)	-0.01% (0.952)	+0.02% (0.850)	-0.63% (0.308)	-0.19% (0.278)	-0.06% (0.614)
view-normalised loss gap	-0.20% (0.325)	-0.11% (0.515)	+0.02% (0.798)	-0.59% (0.167)	-0.15% (0.190)	+0.00% (0.956)
positional × loss gap (Prop. 1)	+0.16% (0.423)	+0.08% (0.587)	+0.28% (0.175)	-0.16% (0.433)	-0.07% (0.588)	-0.01% (0.915)

C.5 Target Transfer

The target-swap control isolates the effect of the training labels: the representation, pairwise objective, and regularisation are fixed across all fits. The model trained for the evaluated target has the lowest AURC in every column of Table 7. Reusing a head trained for another loss raises AURC by 4.7–31.2% and can be worse than the strongest single score for the evaluation target.

C.6 Held-Out Test of the Partial-Support Decomposition

Figure 3 tests Proposition 2 on view-normalised losses. For each capture regime and target, category rates are fitted on the even-indexed views; for every support bin of the odd-indexed views with at least five views, the gain is predicted as $\min_h G_h$ at the bin's mean u and compared with the measured gain. Over all 75 domain–bin cells, $r = 0.92$; removing any single regime leaves $r \geq 0.70$ for every target. Within a regime and pooled over targets, $r \geq 0.90$ on both primary suites and $r = 0.86$ on Mill-19. Mill-19 is the exception within a target: its DSSIM profile is predicted ($r = 0.93$ over seven bins) but its L1 and LPIPS profiles are not ($r = -0.08$ and 0.13).

Why the losses are normalised. The registered version of this test used unnormalised losses and gave $r = 0.23$, 0.86 and 0.69 for L1, DSSIM and LPIPS. The weak L1 correlation is consistent with a failure of Assumption A.1; it does not contradict the exact pairwise identity. Unnormalised W scales with a view's mean loss, which for L1 varies widely between views (coefficient of variation 0.5–3.6 across regimes) and on Mill-19 also rises with the unsupported share (Spearman correlation 0.69–0.78 across targets, at most 0.56 elsewhere). Pooling unnormalised category rates therefore gives high-loss views disproportionate weight, and the rates may vary with u. We adopted the normalisation after observing this result. It changes the pairwise diagnostic's units and view weighting; all reported raw AURC comparisons remain unchanged. Mill-19, where the coupling is strongest, remains the least accurate regime.

Table 7: AURC of the linear ranker trained for the row target and evaluated under the column target. Fits use leave-one-scene-out on the development reconstructions; the strongest single score for each column is shown for reference.

Trained for	L1	DSSIM	LPIPS
L1	**0.0345**	0.0762	0.2879
DSSIM	0.0365	**0.0728**	0.2896
LPIPS	0.0435	0.0955	**0.2362**
strongest single row	0.0360 (WarpRF)	0.0801 (CrossScore)	0.2516 (Φ)

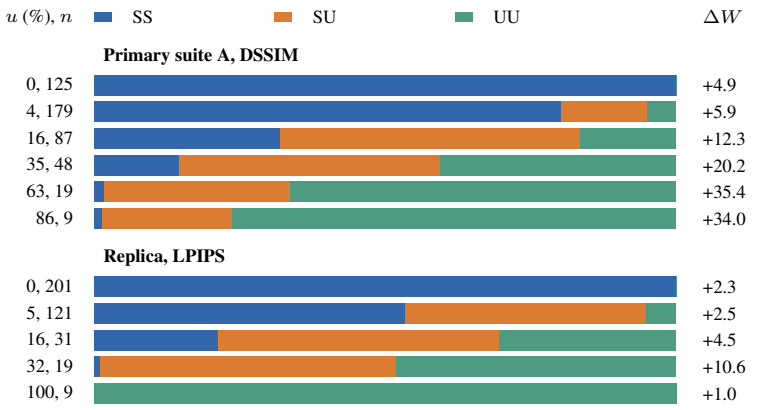


Figure 6: Where fusion reduces loss-weighted misordering. Each bar partitions the measured improvement of L2R-GS over the better external or internal expert into supported (SS), cross-support (SU), and unsupported (UU) pairs. Bins are formed by the unsupported share of a view; u is the bin mean and n its number of views. The right column gives the total improvement per pair in percent of the view's mean loss, before normalising the bar. All bins with at least five views in these two regimes are shown.

C.7 Support-Pair Composition and Placement

Under LPIPS on the primary suites, the fitted gain profile peaks near $u = 0.20$–0.24; the simplified crossing prediction of Corollary A.6 places the peak near $u^\star = 0.32$–0.33. Figure 6 reports both the shares of pair categories and their absolute improvement, since a large share of a small gain can otherwise mislead. Table 8 compares the placement bound with the observed cross-support error.

C.8 Region Size and Fit Stability

Under leave-one-scene-out evaluation on the development reconstructions, L2R-GS improves over the strongest individual signal at 32×32 regions by 4.4%, 12.6%, and 7.0% for L1, DSSIM, and LPIPS. At 128×128 regions the reductions are 3.0%, 9.2%, and 4.2%. Across 11 leave-one-scene-out folds, the effective full-support weight retains its sign for at least 9 of 10 signals for every target. Across five pair-sampling seeds, the spread of the 60 coefficients with magnitude above 0.3 is at most 5.6% of their value.

C.9 Computational Resources

All experiments ran on one server with eight NVIDIA H200 GPUs, 192 CPU cores and 2 TB of memory. Each evidence producer uses a single GPU; the ranker is fitted and evaluated on the CPU. Table 9 reports the wall-clock time of every step of the pipeline on the 11 scenes of one test suite (22–54 evaluated views per scene), as the median over scenes, per scene and per evaluated view. The times were read from the pipeline logs of a run in which several scenes shared the GPUs and include

Table 8: The placement bound on the cross-support pairs of the primary benchmark. m is the share of SU pairs a score misorders and *floor* the bound $|\rho + \pi - 1|$ of Proposition 3, both pooled over 22,722 view–score cells; the external row attains the bound exactly.

Score	L1		DSSIM		LPIPS	
	m	floor	m	floor	m	floor
external row	0.351	0.351	0.399	0.399	0.297	0.297
internal row	0.480	0.172	0.566	0.209	0.276	0.124
no gates (β only)	0.253	0.097	0.242	0.091	0.240	0.098
L2R-GS (gated linear)	0.242	0.107	0.232	0.091	0.215	0.107

process start-up and model loading; the first evidence row also computes the three target-loss maps and controlled channels outside the bank, which the deployed method does not need.

Producing the ten signals of L2R-GS takes about 7.1 s per evaluated view (4.4 min per scene), of which the frozen predictive photometric uncertainty accounts for 3.0 s, against about 31 min to train and render the 3DGS reconstruction itself. The ranker adds little to this cost: fitting its 30 parameters on the development data takes about 14 s per target, and scoring every development view takes 0.004 s.

Table 9: Wall-clock time of each step on the 11 scenes of one test suite (median over scenes). Shared GPUs, including start-up and model loading.

Step	per scene (s)	per view (s)
Reconstruction		
3DGS training (30k steps) and rendering of the held-out views	1848	36.96
Evidence used by L2R-GS		
controlled warps (Φ and the other controlled channels) and target losses	30	0.93
cross-view render, source–source, source count, texture	33	1.05
internal primitive-state channels	15	0.56
WarpRF	33	0.98
CrossScore	13	0.31
predictive photometric uncertainty (PPU)	143	2.98
total for the ten signals	263	7.06
Other baselines of Table 1		
structural warp	18	0.56
FisherRF	23	0.59
Bayes' Rays	41	1.00
PuzzleSim	40	1.12
Ranker (per target, CPU)		
fit on the development data	14 s	
inference on all development views	0.004 s	

C.10 Each Test Reconstruction Separately

Table 10 compares the strongest individual signal with the linear and boosted-tree rankers separately for each test suite. The suites differ in training seed and held-out views; each was read once with the models frozen on the development reconstructions. The strongest single row is chosen inside each suite among the rows produced on all eleven identities, and both the linear and boosted-tree instantiations are compared with it by the paired exact sign-flip test.

C.11 AURC and AUGRC on the Primary Benchmark

Table 11 gives AURC and AUGRC for baseline, controlled-channel, and ranker rows on the full held-out block. AUGRC [23] integrates the generalised risk $c\,R(c)$ and so reads as the loss admitted per region of the whole view. The table also reports the paired relative difference of the linear ranker against each row and the exact sign-flip p-value over the eleven identities. The strongest single row is the same under both metrics for every target; under AUGRC the linear ranker is below it by 3.7% (WarpRF, $p < 0.001$), 6.5% (CrossScore, $p < 0.001$) and 5.0% (Φ, $p = 0.002$).

Table 10: The three test suites separately on all held-out views: AURC of the strongest single row, of the linear ranker and of the boosted trees, with the paired relative difference against that row and the exact sign-flip p-value over the eleven identities.

Suite	**Target**	strongest single row	**L2R-GS** (linear)	Δ, p	L2R-GS (trees)	Δ, p
A	L1	0.0312 (WarpRF)	0.0298	-4.6%, <0.001	0.0299	-4.3%, 0.002
	DSSIM	0.0710 (CrossScore)	0.0629	-11.4%, <0.001	0.0622	-12.4%, <0.001
	LPIPS	0.2300 (Φ)	0.2176	-5.4%, 0.024	0.2141	-6.9%, <0.001
B	L1	0.0353 (WarpRF)	0.0337	-4.6%, 0.002	0.0333	-5.8%, 0.003
	DSSIM	0.0754 (CrossScore)	0.0694	-8.0%, 0.002	0.0690	-8.6%, 0.002
	LPIPS	0.2314 (Φ)	0.2187	-5.5%, 0.007	0.2159	-6.7%, <0.001
C	L1	0.0324 (WarpRF)	0.0307	-5.5%, 0.002	0.0304	-6.1%, <0.001
	DSSIM	0.0765 (CrossScore)	0.0688	-10.2%, <0.001	0.0678	-11.4%, <0.001
	LPIPS	0.2386 (Φ)	0.2230	-6.5%, 0.006	0.2194	-8.0%, <0.001

Table 11: AURC and AUGRC of rows in the controlled-channel analysis on all held-out views of the primary test reconstructions (*All 11*), with the relative difference of the linear ranker against the row (Δ, negative favours L2R-GS) and the exact sign-flip p-value.

(a) AURC

Row	L1	Δ, p	DSSIM	Δ, p	LPIPS	Δ, p
Oracle	0.0255		0.0560		0.1956	
Random	0.0447	+29.7%, <0.001	0.1002	+33.1%, <0.001	0.2888	+23.9%, <0.001
FisherRF	0.0451	+30.4%, <0.001	0.1016	+34.0%, <0.001	0.3043	+27.8%, <0.001
Bayes' Rays	0.0438	+28.4%, <0.001	0.1009	+33.6%, <0.001	0.2695	+18.4%, <0.001
PPU	0.0350	+10.2%, 0.057	0.0768	+12.7%, <0.001	0.2661	+17.4%, <0.001
WarpRF	0.0330	+4.9%, <0.001	0.0783	+14.4%, <0.001	0.2850	+22.9%, <0.001
CrossScore	0.0381	+17.6%, <0.001	0.0743	+9.8%, <0.001	0.2888	+23.9%, <0.001
PuzzleSim	0.0374	+16.2%, <0.001	0.0802	+16.4%, <0.001	0.2610	+15.8%, <0.001
Visibility	0.0392	+20.0%, <0.001	0.0880	+23.8%, <0.001	0.2581	+14.9%, <0.001
RGB warp	0.0349	+10.0%, <0.001	0.0794	+15.6%, <0.001	0.2799	+21.5%, <0.001
Structural warp	0.0355	+11.5%, <0.001	0.0769	+12.8%, <0.001	0.2949	+25.5%, <0.001
VGG cosine	0.0379	+17.2%, <0.001	0.0814	+17.7%, <0.001	0.2370	+7.2%, <0.001
DINO cosine	0.0409	+23.2%, <0.001	0.0927	+27.7%, <0.001	0.2522	+12.8%, <0.001
LPIPS image warp	0.0382	+17.8%, <0.001	0.0811	+17.4%, <0.001	0.2363	+7.0%, 0.002
Primitive screen scale	0.0465	+32.5%, <0.001	0.1095	+38.8%, <0.001	0.2458	+10.6%, 0.002
Φ feature warp	0.0380	+17.4%, <0.001	0.0821	+18.3%, <0.001	0.2333	+5.8%, 0.008
Rank average (label-free)	0.0336	+6.5%, <0.001	0.0733	+8.6%, <0.001	0.2396	+8.3%, <0.001
L2R-GS (boosted trees)	0.0312	-0.6%, 0.374	0.0663	-1.0%, 0.006	0.2165	-1.5%, <0.001
L2R-GS (gated linear)	0.0314		0.0670		0.2198	

(b) AUGRC

Row	L1	Δ, p	DSSIM	Δ, p	LPIPS	Δ, p
Oracle	0.0152		0.0342		0.1130	
Random	0.0224	+21.4%, <0.001	0.0503	+23.1%, <0.001	0.1450	+16.4%, <0.001
FisherRF	0.0224	+21.1%, <0.001	0.0508	+23.8%, <0.001	0.1474	+17.8%, <0.001
Bayes' Rays	0.0221	+20.4%, <0.001	0.0507	+23.7%, <0.001	0.1385	+12.5%, <0.001
PPU	0.0188	+6.1%, 0.161	0.0423	+8.5%, <0.001	0.1355	+10.5%, <0.001
WarpRF	0.0183	+3.7%, <0.001	0.0430	+9.9%, <0.001	0.1422	+14.8%, <0.001
CrossScore	0.0200	+11.8%, <0.001	0.0414	+6.5%, <0.001	0.1422	+14.8%, <0.001
PuzzleSim	0.0198	+11.0%, <0.001	0.0433	+10.7%, <0.001	0.1356	+10.6%, <0.001
Visibility	0.0206	+14.3%, <0.001	0.0465	+16.7%, <0.001	0.1350	+10.2%, <0.001
RGB warp	0.0192	+8.2%, <0.001	0.0440	+12.1%, <0.001	0.1413	+14.2%, <0.001
Structural warp	0.0194	+8.9%, <0.001	0.0434	+10.9%, <0.001	0.1455	+16.7%, <0.001
VGG cosine	0.0203	+12.9%, <0.001	0.0446	+13.3%, <0.001	0.1289	+6.0%, <0.001
DINO cosine	0.0213	+17.1%, <0.001	0.0484	+20.0%, <0.001	0.1338	+9.4%, <0.001
LPIPS image warp	0.0203	+13.2%, <0.001	0.0445	+13.0%, <0.001	0.1289	+6.0%, <0.001
Primitive screen scale	0.0228	+22.8%, <0.001	0.0536	+27.7%, <0.001	0.1296	+6.5%, <0.001
Φ feature warp	0.0203	+13.0%, <0.001	0.0449	+13.8%, <0.001	0.1276	+5.0%, 0.002
Rank average (label-free)	0.0184	+4.4%, 0.002	0.0409	+5.5%, <0.001	0.1282	+5.5%, <0.001
L2R-GS (boosted trees)	0.0175	-0.6%, 0.131	0.0384	-0.9%, <0.001	0.1201	-0.9%, <0.001
L2R-GS (gated linear)	0.0176		0.0387		0.1212	

Table 12: All tested evidence-role subsets on the primary test reconstructions: AURC, increase over full L2R-GS (Δ), and paired sign-flip p-value. The six restricted rows use the same target-specific linear learner, development data, sampled pairs, and frozen λ.

Variant	L1	Δ, p	DSSIM	Δ, p	LPIPS	Δ, p
External (E)	0.0326	+3.9%, <0.001	0.0702	+4.8%, <0.001	0.2343	+6.6%, 0.004
Internal (I)	0.0356	+13.5%, <0.001	0.0742	+10.7%, <0.001	0.2393	+8.9%, 0.002
$E + I$	0.0321	+2.2%, <0.001	0.0680	+1.5%, <0.001	0.2257	+2.7%, 0.002
$E + O$	0.0323	+3.0%, <0.001	0.0693	+3.4%, <0.001	0.2244	+2.1%, 0.002
$I + O$	0.0337	+7.3%, <0.001	0.0717	+7.0%, <0.001	0.2354	+7.1%, 0.002
$E + I + O$	0.0318	+1.4%, <0.001	0.0676	+0.8%, <0.001	0.2228	+1.4%, 0.002
Full L2R-GS	**0.0314**		**0.0670**		**0.2198**	

Table 13: Direct paired contrasts in the evidence-role ladder: AURC reduction relative to the row baseline, number of scene identities improved out of 11, and unadjusted exact two-sided sign-flip p-value (in that order).

Paired contrast	L1	DSSIM	LPIPS
$E + I$ vs. E	1.7% (10/11; 0.005)	3.2% (10/11; 0.002)	3.7% (9/11; 0.113)
$E + I$ vs. I	10.0% (11/11; <0.001)	8.3% (11/11; <0.001)	5.7% (10/11; 0.008)
$E + I + O$ vs. $E + O$	1.5% (10/11; 0.003)	2.5% (11/11; <0.001)	0.7% (8/11; 0.113)
$E + I + O$ vs. $I + O$	5.5% (11/11; <0.001)	5.8% (11/11; <0.001)	5.3% (10/11; 0.003)
$E + I + O$ vs. $E + I$	0.7% (9/11; 0.030)	0.6% (10/11; 0.003)	1.3% (9/11; 0.007)
Full vs. $E + I + O$	1.4% (11/11; <0.001)	0.8% (11/11; <0.001)	1.4% (10/11; 0.002)

C.12 Evidence-Role and Gate Ablations

The evidence-role ablation was added after the primary test results had been inspected; it is exploratory. Each restricted variant was fitted only on the clean development reconstructions, with the same within-view rank transform, pairwise objective, sampled pairs, and target-specific λ as L2R-GS. The full row reuses the frozen L2R-GS weights. We report all six restricted subsets, including $E + O$ and $I + O$ (Table 12). The $E + I + O$ and full rows reproduce the independently computed beta-only and full-model AURCs. The E-only row retains the fixed zero encoding of missing external cues without observability inputs; it is not an optimal external-only bound. Adding internal cues to external ones reduces mean AURC in all three test suites, but the LPIPS scene-paired contrast is inconclusive both without and with observability held fixed ($p = 0.113$ in each comparison; Table 13).

Table 14 ablates the two interaction terms of (10) on the test reconstructions of the primary benchmark, together with a label-free rank average and the boosted-tree capacity reference. The linear variants use the same ten signals, pairwise loss and λ, fitted on the development reconstructions; Δ is the change relative to the full model. Removing the availability term raises the loss-weighted misorder of cross-support pairs by +6.5% and +5.3% under L1 and LPIPS, against +0.6 and +2.4% on supported pairs, while removing the support term raises the supported-pair misorder by +1.4 and +5.2%.

The rank average is the label-free mean of the within-view ranks of the five oriented risk rows. It provides a label-free reference, but its smaller cue set prevents the gap from isolating the effect of learned weights. The matched ten-signal linear variants measure what is gained by letting the fitted weights depend on support and availability.

C.13 Per-Scene Results

Tables 15–17 give per-scene values for the full held-out block behind the primary-benchmark averages in Table 1. Each value averages the three test reconstructions of the scene; the lowest deployable value of each column is in bold. Compared with PRIMU, L2R-GS is lower in every scene for each target (paired exact sign-flip $p < 0.001$ in all three comparisons).

Table 14: Ablation on the primary test reconstructions: AURC, relative change from full L2R-GS, and paired sign-flip p-value. The linear variants use the same ten signals, pairwise objective, and regularisation. The boosted trees provide a capacity reference on the same signals.

Variant	L1	Δ, p	DSSIM	Δ, p	LPIPS	Δ, p
Rank average (label-free)	0.0336	+6.5%, <0.001	0.0733	+8.6%, <0.001	0.2396	+8.3%, <0.001
no gates (β only)	0.0318	+1.4%, <0.001	0.0676	+0.8%, <0.001	0.2228	+1.4%, 0.002
no support gate ($\gamma = 0$)	0.0315	+0.3%, 0.019	0.0672	+0.3%, 0.008	0.2204	+0.3%, 0.067
no availability gate ($\delta = 0$)	0.0315	+0.4%, <0.001	0.0671	+0.2%, 0.020	0.2205	+0.3%, 0.047
boosted trees, same signals (capacity reference)	0.0312	-0.6%, 0.374	0.0663	-1.0%, 0.006	0.2165	-1.5%, <0.001
L2R-GS (gated linear)	0.0314		0.0670		0.2198	

Table 15: Per-scene AURC under L1 on all held-out views of the test reconstructions.

Method	bicycle	bonsai	counter	drjohnson	garden	kitchen	playroom	room	stump	train	truck
Oracle	0.032	0.018	0.025	0.016	0.022	0.017	0.014	0.011	0.032	0.065	0.030
Random	0.060	0.035	0.052	0.029	0.033	0.031	0.023	0.021	0.050	0.104	0.051
FisherRF	0.056	0.033	0.064	0.028	0.034	0.030	0.023	0.020	0.051	0.109	0.046
Bayes' Rays	0.068	0.033	0.048	0.033	0.033	0.027	0.025	0.022	0.047	0.101	0.045
PPU	0.043	0.027	0.045	0.023	0.027	0.024	0.019	0.015	0.046	**0.073**	0.043
PRIMU	0.040	0.026	0.035	0.022	0.027	0.024	0.018	0.015	0.041	0.090	0.041
WarpRF	0.039	0.023	0.039	0.020	0.026	0.022	0.017	0.013	0.036	0.088	0.040
CrossScore	0.049	0.033	0.038	0.024	0.033	0.025	0.020	0.016	0.043	0.093	0.045
PuzzleSim	0.048	0.028	0.045	0.023	0.033	0.024	0.019	0.014	0.043	0.091	0.043
Visibility	0.049	0.030	0.042	0.030	0.029	0.025	0.025	0.017	0.042	0.096	0.048
RGB warp	0.039	0.026	0.037	0.027	0.026	0.023	0.023	0.014	0.039	0.090	0.041
Structural warp	0.039	0.027	0.038	0.028	0.026	0.024	0.024	0.014	0.038	0.092	0.040
VGG cosine	0.044	0.029	0.041	0.029	0.029	0.025	0.024	0.017	0.041	0.091	0.047
DINO cosine	0.059	0.029	0.045	0.031	0.030	0.025	0.025	0.018	0.045	0.095	0.048
LPIPS image warp	0.044	0.029	0.041	0.030	0.028	0.025	0.025	0.017	0.043	0.092	0.047
Primitive screen scale	0.073	0.037	0.044	0.032	0.031	0.028	0.025	0.022	0.058	0.102	0.059
Φ feature warp	0.044	0.029	0.041	0.029	0.029	0.025	0.024	0.017	0.040	0.091	0.047
L2R-GS (gated linear)	**0.036**	0.023	0.034	**0.020**	0.025	0.021	0.017	**0.013**	0.036	0.083	**0.037**
L2R-GS (boosted trees)	0.037	**0.022**	**0.033**	0.020	**0.024**	**0.021**	**0.017**	0.013	**0.035**	0.083	0.037

C.14 Mill-19: Data and Reconstruction

Data. We use the Mega-NeRF release of Building and Rubble [24] at a quarter of the original resolution, with the COLMAP pose results released with CityGaussian [17]; the validation cameras of the release are expressed in the frame of the training cameras, which we checked before merging the two. The Mega-NeRF validation views are the test views; every twentieth training view is also held out. Both sets enter the within-scene view folds, with the evaluated view excluded from model fitting. Building has 1,824 training and 116 evaluated views, Rubble 1,574 and 104.

Reconstruction. The 3DGS recipe of the other domains was applied without change, as registered. It reaches a held-out PSNR of 23.0 dB on Rubble (test views 22.6 dB) and 16.3 dB on Building (test views 13.4 dB): the Building renders miss exposure changes and large parts of the scene. We report the registered recipe rather than one tuned after seeing this.

Amendments. Before any Mill-19 score was computed, PuzzleSim was limited to 250 reference views because the whole training set does not fit in GPU memory; PPU's rasteriser discards everything closer than 0.2 scene units, which on these scenes removes a large share of the visible Gaussians, so its adapter exports the scene under a similarity transform with scale 20, after which its render passes the adapter's render-agreement check.

C.15 Human-Perceived Artifacts as the Declared Target

Data. The PuzzleSim release [10] contains 3DGS renders of Mip-NeRF 360, Tanks and Temples and Deep Blending scenes trained with 60–80% of the views withheld, each with the average of 22

Table 16: Per-scene AURC under DSSIM on all held-out views of the test reconstructions.

Method	bicycle	bonsai	counter	drjohnson	garden	kitchen	playroom	room	stump	train	truck
Oracle	0.102	0.030	0.052	0.043	0.048	0.029	0.027	0.023	0.114	0.091	0.057
Random	0.182	0.057	0.104	0.078	0.085	0.056	0.049	0.047	0.176	0.163	0.099
FisherRF	0.193	0.053	0.120	0.075	0.084	0.055	0.050	0.043	0.184	0.178	0.082
Bayes' Rays	0.199	0.063	0.106	0.089	0.081	0.054	0.057	0.052	0.170	0.157	0.082
PPU	0.132	0.043	0.079	0.062	0.063	0.043	0.038	0.030	0.159	0.113	0.082
PRIMU	0.133	0.045	0.069	0.058	0.069	0.043	0.035	0.036	0.150	0.137	0.077
WarpRF	0.132	0.047	0.083	0.056	0.062	0.046	0.034	0.037	0.141	0.149	0.073
CrossScore	0.136	0.041	0.069	0.058	0.071	0.041	0.037	0.032	0.149	0.107	0.077
PuzzleSim	0.147	0.047	0.078	0.063	0.081	0.048	0.034	0.035	0.161	0.114	0.074
Visibility	0.149	0.051	0.087	0.081	0.069	0.048	0.055	0.037	0.145	0.149	0.096
RGB warp	0.131	0.045	0.075	0.073	0.065	0.044	0.050	0.035	0.141	0.136	0.079
Structural warp	0.132	0.044	0.072	0.071	0.066	0.042	0.048	0.033	0.141	0.123	0.073
VGG cosine	0.130	0.051	0.082	0.078	0.067	0.048	0.052	0.037	0.140	0.126	0.087
DINO cosine	0.171	0.056	0.095	0.083	0.072	0.050	0.056	0.041	0.153	0.150	0.093
LPIPS image warp	0.125	0.050	0.081	0.079	0.064	0.047	0.054	0.036	0.141	0.128	0.087
Primitive screen scale	0.213	0.066	0.103	0.094	0.079	0.055	0.060	0.056	0.184	0.171	0.122
Φ feature warp	0.128	0.052	0.084	0.079	0.066	0.048	0.052	0.038	0.139	0.129	0.088
L2R-GS (gated linear)	0.118	0.037	0.066	0.052	0.059	0.038	0.032	0.029	0.132	**0.106**	0.067
L2R-GS (boosted trees)	**0.117**	**0.037**	**0.065**	**0.052**	**0.058**	**0.037**	**0.032**	**0.028**	**0.131**	0.106	**0.067**

Table 17: Per-scene AURC under LPIPS on all held-out views of the test reconstructions.

Method	bicycle	bonsai	counter	drjohnson	garden	kitchen	playroom	room	stump	train	truck
Oracle	0.205	0.177	0.218	0.285	0.103	0.108	0.239	0.185	0.245	0.229	0.158
Random	0.315	0.286	0.329	0.373	0.155	0.192	0.328	0.280	0.331	0.338	0.243
FisherRF	0.363	0.297	0.362	0.374	0.166	0.191	0.331	0.280	0.357	0.354	0.273
Bayes' Rays	0.309	0.248	0.300	0.374	0.149	0.155	0.306	0.276	0.324	0.303	0.221
PPU	0.264	0.275	0.311	0.349	0.129	0.196	0.323	0.262	0.312	0.285	0.222
PRIMU	0.270	0.294	0.308	0.367	0.143	0.189	0.348	0.280	0.297	0.315	0.244
WarpRF	0.288	0.292	0.314	0.370	0.139	0.182	0.352	0.280	0.317	0.336	0.264
CrossScore	0.273	0.346	0.320	0.379	0.143	0.210	0.354	0.290	0.297	0.322	0.242
PuzzleSim	0.266	0.236	0.304	0.354	0.142	0.160	0.342	0.252	0.303	0.301	0.213
Visibility	0.276	0.254	0.274	0.361	0.133	0.159	0.333	0.249	0.289	0.309	0.203
RGB warp	0.284	0.301	0.296	0.382	0.135	0.187	0.350	0.263	0.305	0.338	0.238
Structural warp	0.294	0.323	0.322	0.391	0.137	0.212	0.358	0.285	0.307	0.348	0.267
VGG cosine	0.242	0.221	0.253	0.348	0.122	0.145	0.322	0.203	0.279	0.286	0.184
DINO cosine	0.281	0.241	0.278	0.357	0.141	0.157	0.322	0.230	0.295	0.277	0.195
LPIPS image warp	0.239	0.229	0.253	0.343	0.120	0.148	0.310	0.205	0.281	0.290	0.183
Primitive screen scale	0.320	0.204	0.259	0.329	0.135	0.147	**0.265**	0.243	0.318	0.281	0.203
Φ feature warp	0.237	0.213	0.251	0.345	0.121	0.143	0.321	0.202	0.275	0.279	0.180
L2R-GS (gated linear)	0.240	0.196	0.238	0.312	0.117	0.136	0.269	0.200	0.276	0.257	0.176
L2R-GS (boosted trees)	**0.232**	**0.193**	**0.236**	**0.311**	**0.114**	**0.133**	0.266	**0.198**	**0.270**	**0.256**	**0.173**

binary masks of regions the annotators perceived as "visually degraded, unnatural, or incongruent", aligned ground truth and the withheld views as references. The 30 renders whose cameras could be registered to our reconstructions are split into a development and a confirmation half of 15 each; the halves share all ten scenes. The target of a region is the mean over its pixels of the fraction of annotators marking the pixel, and the evaluator is the per-view AURC of the main text with this target in place of a reconstruction loss.

Evidence. The reconstructing models are not released, so internal evidence cannot be computed. External signals are computed on the release's render with the release's references as sources, through our geometry of the same scene; observability uses texture and visibility.

Procedure. For a render of scene s, the external slot is chosen by forward selection among Φ, the LPIPS image warp, the RGB warp, the VGG and DINO cosines and PuzzleSim, at most three signals, using leave-one-scene-out AURC on the development renders of the other scenes; a monotone model is then fitted to the human target on those renders. The learner is a monotone LightGBM rank regressor using the selected external signals with texture and visibility; it differs from the linear, pairwise-trained L2R-GS model evaluated on reconstruction losses. On the development half

the whole procedure is nested inside leave-one-scene-out; on the confirmation half it is fitted on the development half only, with scene s excluded, and applied once.

Result. PuzzleSim is selected first in all 10 scene folds, followed by a cross-view signal when forward selection finds one useful. On the development half the ranker reaches 0.1029 against 0.1147 for PuzzleSim and 0.1309 for Φ (10.3% and 21.4% lower, $p = 0.012$ and $p = 0.008$ over scenes); on the confirmation half 0.0803 against 0.0935 and 0.0923 (14.0% and 13.0% lower, $p = 0.033$ and $p = 0.105$ over scenes).