SPLATLEAK: TRACING AND REDUCING MEMBERSHIP LEAKAGE IN 3D GAUSSIAN SPLATTING

Anonymous authors
Paper under double-blind review

ABSTRACT

A 3D Gaussian Splatting (3DGS) asset can be shared without its source photographs. Can its renders reveal which photographs trained it, when neighboring views can reconstruct images that never supervised optimization? We isolate the fidelity gained from a photograph's own supervision using shared-initialization and matched-schedule interventions, even with three synchronized neighboring views already supervising. We audit membership by comparing target errors with those of same-scene references trained without the candidate pool. On eight new test scenes, a photometric log-ratio improves mean normalized partial AUC over 0–1% FPR by 0.1567 relative to PSNR-GAP. We also audit candidates without supplied poses using registered background images and cameras with known candidate intrinsics. Reference-fitted alignment extends this audit across the tested numerical color changes. To reduce source association before release, we distill students from teacher renders. Ordinary distillation retains teacher-source association even without direct supervision from protected photographs. We instead supervise each protected photograph's registered camera using teachers that excluded it. Across four scenes, this lowers the mean per-asset maximum AUC over nine attack configurations from 0.9457 to 0.7364, at a mean held-out PSNR cost of 0.241 dB relative to membership-agnostic distillation. The student retains higher held-out PSNR and SSIM than training on auxiliary photographs alone.

1 INTRODUCTION

3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) turns posed photographs into a scene representation that renders new viewpoints. Efficient variants and encodings make these assets practical to distribute (Lee et al., 2024; Fan et al., 2024). Sharing an asset without its photographs or training manifest leaves a provenance question: can its renders reveal which images supplied photometric supervision? We study individual-image membership within a scene, where members and nonmembers depict the same environment.

High fidelity alone is an ambiguous clue. Nearby photographs may cover the same surfaces well enough to reconstruct a candidate view without its own supervision. A candidate can also affect reconstruction initialization before photometric optimization begins. Interpreting a fidelity advantage therefore requires separating the candidate's own supervision from view difficulty and initialization.

We address this ambiguity with shared-initialization and matched-schedule interventions that identify a candidate's own photometric contribution. With three synchronized neighboring views already supervising, enabling the candidate photograph's supervision improves its reconstruction in all eight preselected view groups across four scenes, by 4.46 dB on average. Membership also remains readable across training assignments, optimizer seeds, model variants and compact encodings (Section 4).

To account for view difficulty, we compare target reconstruction errors with those of reference models trained on separate same-scene photographs, excluding the candidate pool. On eight new test scenes, a photometric log-ratio improves low-FPR ranking over PSNR-GAP and detects 39 more members with two fewer false positives. When candidate poses are unavailable, we estimate them from registered background photographs and cameras with known candidate intrinsics, retaining the readout advantage. Camera matching does not remove numerical RGB differences, so we fit a gamma+RGB transform from each candidate and the mean reference render and apply it to the

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.

target and all references. The aligned readout retains detections across all eight evaluated image conditions while reusing its clean-calibrated cutoff (Sections 5 and 5.3).

To reduce source association before release, we study students trained from teacher renders, with additional same-scene photographs supplying auxiliary views. Ordinary distillation retains teacher-source association even without direct supervision from protected photographs. We instead supervise each protected photograph's registered camera using the pair of teachers that excluded it. Across four scenes, this routing reduces association under both rendered and real auxiliary supervision. With routing fixed, real auxiliary RGB further lowers the mean per-asset maximum attack AUC. Both constructions improve held-out PSNR and SSIM over training on auxiliary photographs alone while deploying a single student (Section 6).

2 Related Work

In neural rendering, MemE uses the different fitting rates of clean and distracting pixels to improve reconstruction (Wang et al., 2025), while image valuation measures training images' contributions to reconstruction quality (Cheimariotis et al., 2026). We connect own-image supervision to fidelity traces observable in released renders.

Membership inference distinguishes training examples through model responses (Shokri et al., 2017; Yeom et al., 2018; Hu et al., 2022). GAN-Leaks and difficulty calibration use references to account for example difficulty (Chen et al., 2020; Watson et al., 2022); GAP applies this principle to registered render fidelity. Likelihoods (Carlini et al., 2022; Brännvall, 2026), quantiles (Bertran et al., 2023; Tang et al., 2024), and population comparisons such as RMIA (Zarifzadeh et al., 2024) provide other readouts. Reference construction also depends on query access: CMIA/PMIA distinguish access before or after shadow training (Du et al., 2026). We examine this distinction through candidate-containing and background-only references.

FedMeNF pairs images with same-view renders in a membership experiment with distinct cars across clients; negatives use a randomly sampled client model from the same partition (Yun et al., 2025). We adapt this paired CNN to same-scene and cross-scene fitting, while within-scene interventions isolate own-image supervision. Adjacent work studies scene recovery from training communication (Zhang et al., 2024) and watermark ownership (Tan et al., 2024; Chen et al., 2025; Li et al., 2026). GSFF's privacy-preserving variant discards color and continuous features for localization (Pietrantoni et al., 2025). Our membership audit retains ordinary RGB rendering.

Synthesized supervision trains 3DGS students (Xiang et al., 2025). Distillation also supports membership protection (Shejwalkar & Houmansadr, 2021). Knowledge cross-distillation and SELENA train a single student using teachers that excluded each example (Chourasia et al., 2022; Tang et al., 2022), whereas MIAShield excludes models at query time (Jarin & Eshete, 2023). We apply teacher exclusion to renders at registered cameras, measuring upstream source association alongside rendering utility in the released 3DGS student.

3 Problem and Audit Construction

3.1 Membership and access

Let G be a released 3DGS asset and I a candidate RGB image with known intrinsics κ. Direct photometric optimization-set membership, $M_{\mathrm{opt}}(I; G)$, equals one exactly when I supervised photometric optimization of G. Images used for SfM or bundle adjustment (Schönberger & Frahm, 2016) need not enter the subsequent photometric optimization set. Our membership label therefore records direct supervision rather than whole-pipeline use or authorship.

Our auditor has candidate RGB and same-scene background photographs B with registered cameras, but does not know the target's training-image list. We restrict target access to renders; the audit uses no target parameters or gradients. A *supplied pose* locates the candidate camera in the asset frame. To *estimate a pose*, we build a 3D map from B and its fixed released cameras, then match the candidate to B for PnP. The map excludes the candidate pool H and nonmember anchors R reserved for optional normalization. Figure 1 summarizes the audit: compare the candidate with target and background-reference renders at a shared pose, then apply a calibrated decision rule.

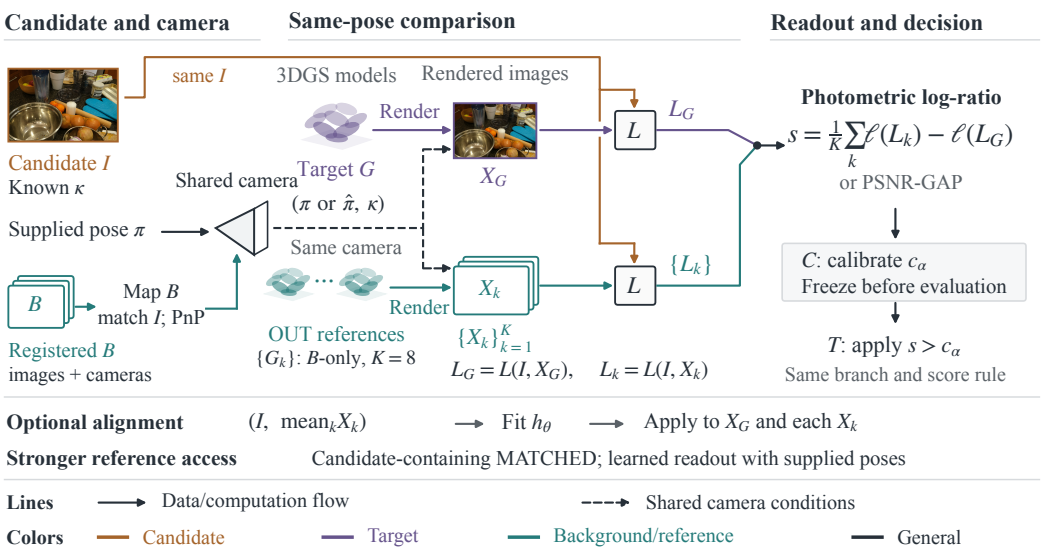


Figure 1: Membership auditing at a shared supplied or estimated camera. Target and OUT reference renders are compared with the same candidate; C freezes the branch-specific cutoff applied to T. Optional alignment shares a reference-fitted transform. MATCHED provides stronger reference access (Section 5.4). Photos show a retained counter query and target render; model icons and image cards are schematic.

3.2 Comparing target and reference reconstruction errors

Other views can reconstruct a candidate well without its own supervision. We therefore compare target fidelity with that of background-only models (Chen et al., 2020; Watson et al., 2022) to measure the target's relative reconstruction advantage. The $K = 8$ OUT references train on B before candidate and anchor RGB access, excluding H and R from initialization and training.

At a shared supplied or estimated pose π, let $X_G = [\mathcal{R}(G; \pi, \kappa)]_0^1$ and $X_k = [\mathcal{R}(G_k^{\text{OUT}}; \pi, \kappa)]_0^1$ be the displayed target and reference renders, clipped to $[0, 1]$. PSNR-GAP measures the target's PSNR advantage over the reference mean:

$$f_G(I) = \text{PSNR}(I, X_G), \qquad s_{\text{GAP}}(I) = f_G(I) - \frac{1}{K}\sum_{k=1}^{K} \text{PSNR}(I, X_k).$$

RAW uses f_G alone. The relative comparison leaves a choice of error measure. Keeping the target and reference renders fixed, we compare the squared RGB error underlying PSNR with a photometric error that combines absolute color differences and local structural agreement. For $L(I, X) = 0.8L_1(I, X) + 0.2[1 - \text{SSIM}(I, X)]$, write $L_G = L(I, X_G)$ and $L_k = L(I, X_k)$. With $\ell(L) = \log \max(L, 10^{-10})$, define

$$s_{\text{log-photo}}(I) = \frac{1}{K}\sum_k \ell(L_k) - \ell(L_G).$$

This is the log ratio of geometric-mean reference error to target error, so a positive score indicates a target advantage. PSNR-GAP has the same relative-error structure in MSE, up to a constant scale. Both scores need candidate RGB, a target render and K reference renders. Anchor-normalized variants additionally use R.

3.3 From scores to audit decisions

Ranking measures how well scores separate members from nonmembers, using AUC and normalized partial AUC (ROC area up to false-positive rate α, divided by α). A decision additionally needs a cutoff. For GAP, it sets how far target fidelity must exceed the candidate's reference mean.

We use C and T for calibration and evaluation roles within each study. In the access and photometric panels of Section 5, they contain queries from disjoint scenes. These panels decide membership by strict $s > c_\alpha$: among n_C^- finite calibration-negative scores, c_α is the jth largest, where $j = \lfloor \alpha(n_C^- + 1) \rfloor$; $j = 0$ gives $c_\alpha = +\infty$. Each readout, pose and calibration condition freezes its cutoff before testing. Earlier studies retain their own rules, including retrospective target-negative quantiles (Appendix A.4).

On T, true- and false-positive counts measure cutoff transfer. Failed queries produce no positive decision but remain in rate denominators. Scenes are independent units; queries, seeds, assignments and teacher banks are nested measurements.

4 Formation and Persistence of the Trace

4.1 Registered optimization views carry a controlled fidelity advantage

We measure the membership signal through registered renders. Eight-shadow LiRA (Carlini et al., 2022) compares candidate-specific IN/OUT fidelity distributions (Appendix A.3). Targets and shadows use member-dependent initialization, which the interventions below control.

Both counter renders in Fig. 2 look plausible, but the member example has lower mean absolute RGB error (0.0154 versus 0.0230). Across six MipNeRF360, two Tanks&Temples, and two Deep-Blending scenes (Barron et al., 2022; Knapitsch et al., 2017; Hedman et al., 2018), LiRA attains pooled AUC 0.9957 on 2,108 pose-balanced queries (Table 1). At the retrospective nominal 1% point, it detects 89.4% of members with 11/1,054 false positives. On the five scenes with raw-PSNR TPR@1% below 0.95, subset-pooled TPR rises from 0.371 for raw PSNR to 0.706 for LiRA. Bonsai remains weak at 0.031. Appendices A.2 and E.2 give pose diagnostics and other fidelity readouts.

Table 1: Original ten-scene comparison on 2,108 pose-balanced queries. TPR uses retrospective target-negative quantiles; LiRA's nominal 1% point has 11/1,054 false positives. Appendix E.2 reports per-scene results and conditional bootstrap intervals.

Readout	AUC	TPR@1%	TPR@5%
Raw PSNR	0.961	0.528	0.870
Render-residual classifier	0.975	0.372	0.935
Pose nearest-member distance	0.537	0.020	0.052
Eight-shadow LiRA	**0.9957**	**0.894**	**0.987**

The advantage is concentrated near the trained camera. Perturbing the camera reduces member PSNR by 3.0–4.8 dB more on average than nonmember PSNR, bringing their margin toward zero (Appendix B.4). This locates the trace; the following interventions test how supervision produces it.

Four targets in each scene s share auxiliary-only initialization to isolate supervision. Each candidate $q \in Q_s$ supervises two targets and is excluded from two. For target r with PSNR Y_{sqr} and assignment M_{sqr}^{opt}, the candidate-centred effect is

$$\widehat{\beta}_s = \frac{1}{|Q_s|} \sum_{q \in Q_s} \frac{1}{2} \sum_{r=0}^{3} (2M_{sqr}^{\mathrm{opt}} - 1) Y_{sqr}.$$

Balanced assignments make this the membership coefficient with candidate and target fixed effects, removing stable view difficulty and target-wide fidelity shifts. Across counter, bonsai, truck and kitchen, effects span 2.003–2.775 dB (mean 2.321 dB). Counter, kitchen and the bonsai–truck macro-average each have randomization $p \leq 1/5000$ in their declared assignment banks. Bonsai's 2.003 dB effect coexists with weak tail detection: supervision leaves a trace even where membership readout struggles (Appendix B.1).

The crossover fixes initialization but changes several candidate assignments together. To isolate one candidate's photometric contribution, we hold the other assignments, auxiliary initialization, camera roster/order, seeds and training clocks fixed and switch only its photometric gradient. Its own-view PSNR rises by 2.42–8.07 dB in eight counter/truck cells, with SSIM improving throughout (Appendix B.2).

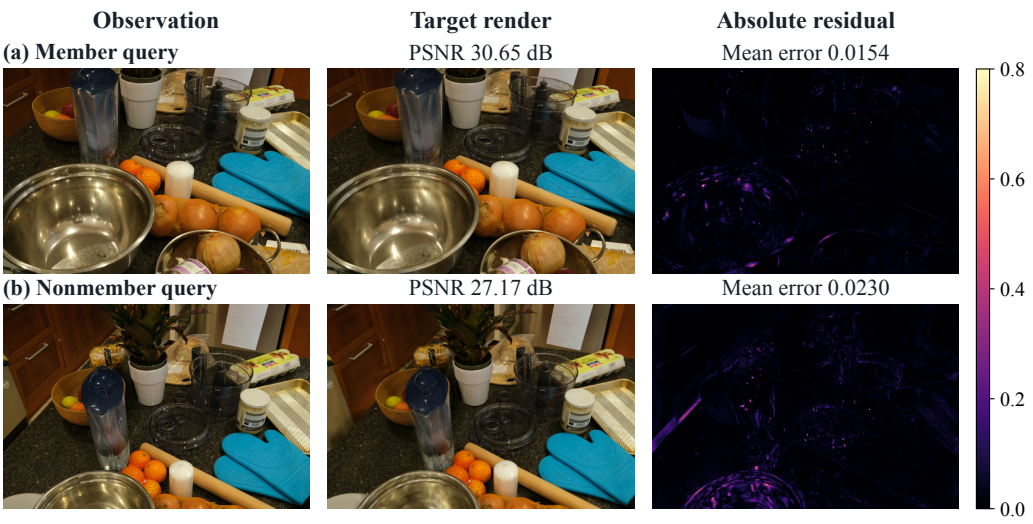


Figure 2: The fidelity observable on two frozen counter candidates. Each observation is compared with the target render at its own registered camera. Residuals average absolute error over RGB channels with intensities in $[0, 1]$; both maps share one color scale. The member example has higher PSNR and smaller mean error; Table 1 quantifies the signal across ten scenes.

To test whether neighboring supervision removes this increment, we repeat the contrast with three synchronized peers already supervising. Enabling the candidate's supervision improves its reconstruction in all eight units across four scenes (Fig. 3). The four-arm comparison gives gains of 0.54–12.39 dB (mean 4.46 dB), exceeding the mean over 16 other views in seven units (Appendix B.3).

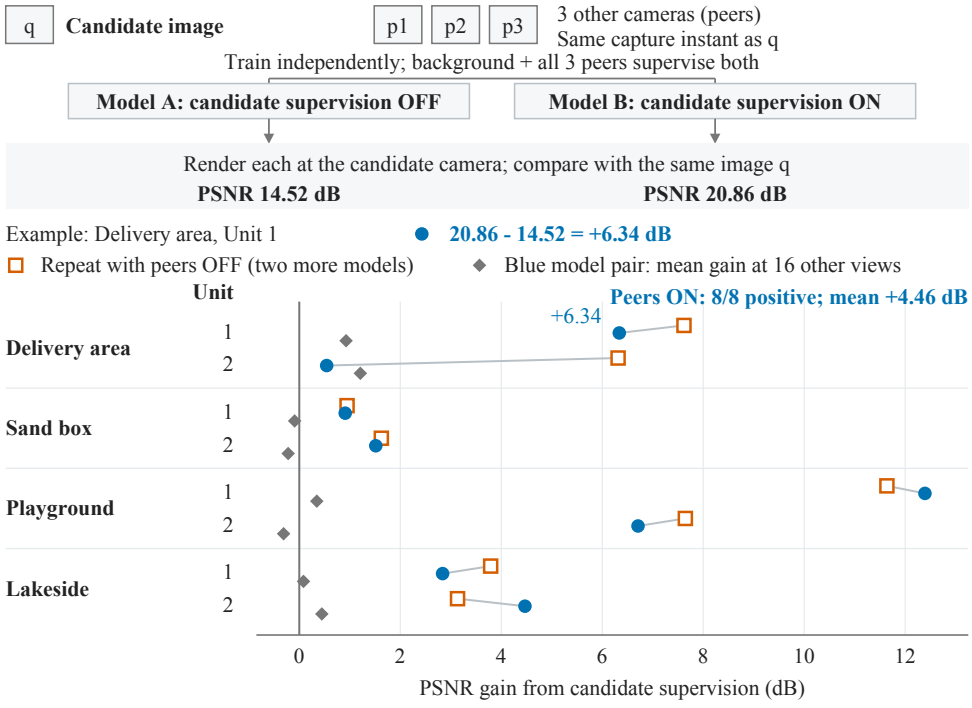


Figure 3: Candidate-supervision gains with peers ON (blue) and OFF (orange), connected within each unit. Gray diamonds show mean gains at 16 other views. All eight peers-ON gains are positive (Appendix B.3).

4.2 THE TRACE SURVIVES TARGET VARIATION AND COMPACT RELEASES

Across six scenes at approximately 80% training coverage, we vary candidate assignments and optimizer seeds while targets and references share each scene's auxiliary-only initialization. With supplied poses and eight candidate-containing references, BaVarIA-n gives equal-scene mean AUC 0.9978 and minimum asset AUC 0.9758 across 48 targets. Strong association therefore survives without candidate-dependent initialization, under this stronger reference access (Appendix C.1).

Scaffold-GS (Lu et al., 2024) also retains the signal with member-dependent initialization (pooled LiRA AUC 0.9949). In a separate four-scene paired study, training appearance embeddings and rendering with each model's mean embedding lowers mean AUC from 0.9797 to 0.9159 relative to disabling embeddings. All eight target panels retain positive association. The configuration costs 1.29 dB mean held-out PSNR (Appendix C.2).

At approximately 90% training coverage on four scenes with member-dependent initialization, eight candidate-containing references leave 356/1,008 queries without OUT support. Twenty restore full support; forty reach 762/905 detections with 0/103 false positives under the earlier 1% calibration rule (Appendix C.3).

SPZ encoding (Niantic Labs, 2026) reduces total file size by 90.37% across nine targets and their candidate-containing references (81 assets). On four test scenes, encoding targets lowers BaVarIA-n mean AUC from 0.9030 to 0.8516 at 0.80 dB mean held-out PSNR cost. Encoding references as well restores AUC to 0.8845, improving every scene (Appendix C.5).

5 MEMBERSHIP AUDITS WITH BACKGROUND REFERENCES

5.1 CHANGING THE ERROR READOUT IMPROVES MEMBERSHIP DISCRIMINATION

We compare log-photo and PSNR-GAP on sixteen new DL3DV scenes (Ling et al., 2024): eight for calibration and eight for evaluation. Each scene has one 30k/native target and eight background-only references. The prespecified primary paired comparison uses clean, supplied-pose, unaligned $K = 8$ scores for 1,226 test candidates (613 members and 613 nonmembers). Target members enter both initialization and optimization; Section 4 separately identifies a downstream supervision contribution. Earlier development panels are not pooled into this test (Appendix G).

Log-photo raises equal-scene nPAUC(.01) by 0.1567, with a paired scene-bootstrap 95% interval of $[0.0274, 0.2540]$ (Table 2). Seven of eight scenes improve; T8 decreases from 0.4000 to 0.1733, while full AUC improves in all eight. At their separately calibrated clean-C cutoffs, log-photo gives 387/11 TP/FP versus PSNR-GAP's 348/13: 39 additional detections and two fewer false positives. Its observed FPR is 1.79%, above the nominal 1% calibration level.

Table 2: Log-photo improves mean ranking in each pair (bold); gray marks the prespecified primary comparison. Eight independent test scenes use clean inputs, eight unaligned references and pose/release-matched 1% clean-C cutoffs (613 test candidates per class). AUC/nPAUC weight scenes equally.

Release / pose	Readout	AUC↑	nPAUC(.01)↑	TP/FP
Native / supplied	PSNR-GAP	0.9583	0.4299	348/13
	log-photo	**0.9743**	**0.5866**	387/11
Native / estimated	PSNR-GAP	0.9253	0.3329	261/12
	log-photo	**0.9404**	**0.4835**	307/12
SPZ / supplied	PSNR-GAP	0.9546	0.4138	347/15
	log-photo	**0.9650**	**0.5349**	343/8

The advantage also persists under pose balancing. All eight T subsets pass the pose battery and retain 1,180/1,226 queries. On this subset, supplied-pose log-photo retains a 0.1560 nPAUC advantage; the original full-H C cutoffs give 376/10 versus 337/13 TP/FP. Two C controls fail and remain in calibration (Appendix G.5).

With 75–79 nonmembers per scene, the primary 0–1% endpoint measures separation before the first observed false positive. Extending the same-score analysis post hoc to 0–5% FPR gives a mean nPAUC gain of 0.1373, with all eight scenes improving (Appendix G.2).

Better ranking need not yield more detections at a transferred cutoff: T4's nPAUC rises from 0.0886 to 0.4557, while TP falls from three to two at zero FP. With the eight T scenes and scores fixed, recomputing the rank cutoff after each C-scene omission gives log-photo 325–443 TP and 5–21 FP. The original all-eight-C rule remains the primary operating point (Appendix G.6).

One unaligned reference already gives log-photo AUC 0.9630, nPAUC 0.5184 and 328/6 TP/FP. The fixed $K = 1, 2, 4, 8$ prefix curve measures this reference-budget tradeoff; aligned fits use eight references.

Log-photo achieves the highest AUC and nPAUC(.01) in both pose settings among the error readouts, render-score RMIA adaptations and linear quantile baseline in Table 55. For L1 and photometric error, log-ratios improve on additive differences. Within the log-ratio form, photometric error also improves on L1 and the MSE underlying PSNR-GAP (Appendix G.4).

5.2 Background-assisted localization retains useful decisions

We next audit candidates without supplied poses, using registered background images and cameras with known candidate intrinsics. We triangulate B alone with HLOC (Sarlin et al., 2019), SuperPoint (DeTone et al., 2018) and LightGlue (Lindenberger et al., 2023), then match candidates to B for PnP, without ground-truth query retrieval or query-to-query bundle adjustment. On the same eight test scenes, log-photo detects 307/613 members with 12/613 false positives, versus PSNR-GAP's 261/613 at the same false-positive count. Their nPAUC difference is 0.1506 (Table 2). Each pose branch has its own C cutoff, so changes from supplied poses combine scores and recalibration.

All clean queries return poses, including some with severe geometric errors. Across eight image conditions, one nonmember is rejected. Errors and rejections remain in the full evaluation denominator, so the counts measure the complete localization-and-scoring pipeline (Appendix G.5).

5.3 Shared photometric alignment adapts the readout

Matching the camera does not match the RGB values. We fit a transform using the candidate and mean reference render, then apply it to the target and all references to retain a common comparison (Table 3).

Table 3: Prespecified gamma+RGB (gray) retains detections across numerical color changes, with a clean-input tradeoff. Eight test scenes are independently relocalized per condition. Each log-photo rule uses eight references and its 1% clean-C cutoff (613 test candidates per class); AUC weights scenes equally.

	None	Scalar	RGB	Gamma+RGB	
Condition	TP/FP	TP/FP	TP/FP	TP/FP	AUC↑
Clean	307/12	277/6	272/5	225/4	0.9201
−0.25 EV	77/1	280/5	279/5	231/4	0.9225
+0.25 EV	58/1	257/5	251/4	208/4	0.9184
$\gamma = 0.8$	82/1	182/3	182/3	216/4	0.9203
$\gamma = 1.25$	92/1	190/3	188/3	246/4	0.9218
WB warm	205/3	183/2	272/5	223/4	0.9226
WB cool	209/3	186/3	277/6	227/4	0.9203
+0.75 EV / clip	0/0	191/4	190/4	154/3	0.9033

With $\mu = \frac{1}{8}\sum_k X_k$, RGB gains a_c and shared exponent γ, we minimize:

$$h_\theta(X)_{u,c} = [a_c X_{u,c}^{\gamma}]_0^1, \qquad E(\theta) = \sum_{(u,c)\in\Omega_4} \big(h_\theta(\mu)_{u,c} - I_{u,c}\big)^2.$$

Here $\theta = (a_1, a_2, a_3, \gamma) \in [0.5, 2]^4$, and Ω_4 samples the mask at stride four along each image axis. The bounded fit starts at identity. Scalar and RGB gain fix $\gamma = 1$ and solve masked least-squares

problems on the full grid. The target render and transformation label never enter fitting. Insufficient support or optimizer failure gives identity; mask and solver details appear in Appendix G.3.

We evaluate numerical EV, gamma, white-balance and clipping changes, independently relocalizing each condition and reusing each readout's own clean-C cutoff. Under moderate EV/gamma shifts, unaligned log-photo falls from 307 clean detections to 58–92. A gamma+RGB rule fixed as a secondary procedure before scoring retains 154–246 detections and 3–4 false positives across all eight conditions, with mean AUC 0.9033–0.9226 (Table 3).

This cross-condition coverage comes with a clean-input tradeoff: gamma+RGB changes TP/FP from 307/12 to 225/4 and AUC from 0.9404 to 0.9201. Scalar gain retains more detections under the tested moderate EV shifts, and RGB gains under white balance. Condition-matched recalibration changes the operating point while leaving each score's ranking unchanged (Appendix G.3).

5.4 Reference information and release matching shape the audit

On the new panel, supplied-pose SPZ log-photo retains a 0.1211 nPAUC advantage over PSNR-GAP, with four fewer detections and seven fewer false positives (Table 2).

The paired log-photo/PSNR-GAP comparisons hold reference access fixed. An earlier DL3DV panel instead varies release matching and reference information across four C and eight T scenes with 610 members and 610 nonmembers, using 7k/30k targets and native/SPZ releases. Matching background references to target checkpoints and encodings raises mean PSNR-GAP AUC from 0.8529 to 0.8937 and changes TP/FP from 117/6 to 126/3. Nonmember anchors adjust score location but leave cross-scene tail mismatch (Appendices A.1 and D.4).

Candidate-containing references support fitting readouts: the eight-model MATCHED bank provides four IN and four OUT observations per query. Same-scene CNNs reach mean AUC 0.9406/0.9485 for two seeds versus BaVarIA-n's 0.9273; cross-scene fits give 0.8135/0.8185 under scene-and-release transfer (Appendix E.3). A separate supplied-pose ETH3D study (Schöps et al., 2017) tests BaVarIA-n calibration transfer, detecting 109/109 and 113/121 members with 0/109 and 2/121 false positives on two test scenes (Appendix D.5).

6 Reducing Source Association in a Deployable Student

A student can inherit source association without reading protected RGB ($M_{\text{opt}} = 0$). To test whether teacher selection reduces it, we audit teacher-source membership M_{src}: whether that RGB supervised the upstream teachers.

Students train at registered auxiliary cameras A and candidate cameras Q, with protected sources $S \subset Q$. Each protected source is excluded from two of four teachers; OUT uses the excluding pair's mean render at that source's camera. GENERIC assigns each candidate a fixed pair without consulting membership. Both use the four-teacher mean at A and the selected pair at Q, sharing teachers, A-only initialization, seeds and 30k schedules. DIRECT trains on real $A \cup S$ RGB.

We evaluate four independent scenes with two target banks and two student seeds per bank. Shadows and attack fitting match each recipe. The nine fixed configurations are RAW, BaVarIA-n, GAP, T, RR8, two CNN seeds, log-L1 and log-photo. Using evaluation labels, we take each asset's maximum AUC over these configurations, A^*_9, then average over seeds, banks and scenes.

Ordinary distillation preserves substantial teacher-source association: GENERIC reaches mean $A^*_9 = 0.9457$. Source-aware routing lowers this to 0.7364 (Table 4), with every configuration decreasing in all sixteen matched asset pairs. Log-photo attains the measured maximum on 13 of 16 OUT assets. The highest per-asset AUC under OUT is 0.8796 (Appendix H.3). The mean held-out PSNR cost relative to GENERIC is 0.241 dB.

ALL4 trades stronger source association for higher fidelity relative to OUT. It uses all four teachers at every candidate camera, reaching mean $A^*_9 = 0.9958$ with 0.406 dB higher mean PSNR. OUT lowers every fixed attack AUC relative to ALL4.

The routing benefit also holds with real auxiliary supervision. GEN-realA and OUT-realA replace the rendered supervision at A with real RGB while retaining their respective routing rules at Q.

Table 4: Source exclusion (gray) lowers association in both routing pairs (bold), with a mean utility cost. Per-asset maximum AUC over nine configurations and common-U utility are averaged over seeds, banks and four scenes. A-only matches four seeds per scene and reports utility only.

Release	$A_9^* \downarrow$	PSNR (dB)↑	SSIM↑	LPIPS↓
GENERIC	0.9457	27.827	0.8873	0.1676
OUT	**0.7364**	27.586	0.8840	0.1720
GEN-realA	0.9155	27.804	0.8865	0.1631
OUT-realA	**0.6255**	27.521	0.8830	0.1678
ALL4 (control)	0.9958	27.991	0.8894	0.1664
A-only (utility only)	—	26.936	0.8705	0.1685

Routing lowers mean A_9^* from 0.9155 to 0.6255. With OUT's routing fixed, replacing auxiliary renders with real RGB further lowers all sixteen asset maxima. Individual frozen-direction attack AUCs show 119 decreases, two ties and 23 increases (Appendix H.3).

Source-excluding students retain reconstruction value beyond A-only releases trained directly on auxiliary RGB with the same initialization and four seeds per scene. OUT and OUT-realA improve held-out PSNR and SSIM in all four scenes, with mean PSNR gains of 0.649 and 0.585 dB, respectively. Their mean LPIPS changes are $+0.0035$ and -0.0007, with mixed scene-level effects (Appendix H.4).

The fidelity cost varies across views. On common held-out U, OUT loses more than 1 dB on 8.02% of matched OUT–GENERIC asset–view pairs (equal-scene mean), with a worst loss of 5.55 dB. OUT-realA adds 0.065 dB mean loss relative to OUT and improves LPIPS, with a worst-pair loss of 6.70 dB. Appendix H.5 shows paired renders and asset–view and averaged-view distributions.

Teacher reuse reduces OUT construction from 119.74 to 23.28 cumulative task minutes on the GENERIC/OUT targets evaluated here (Table 65). Both recipes deploy a smaller, faster single student than DIRECT30k; source-aware routing supplies the additional association reduction.

7 Discussion and Limitations

Rendering quality alone does not describe what a 3DGS asset reveals about its training images. Our controls identify an own-image contribution with synchronized neighbors, but a contribution need not make membership easy to detect: Bonsai combines a controlled fidelity gain with weak low-FPR readout.

Photometric log-ratios improve discrimination under common reference access with both supplied and estimated poses. However, T4's improved ranking yields fewer detections at the transferred cutoff, while omitting one calibration scene at a time changes decisions with test scores held fixed. The readout and its operating point thus contribute differently to audit decisions.

Teacher-source association survives indirect supervision, but source-aware teacher selection reduces it and real auxiliary RGB provides a further reduction. These students retain higher PSNR and SSIM than A-only training, while ALL4 combines higher fidelity with stronger association. Teacher selection and auxiliary supervision change the association–utility tradeoff while preserving single-student deployment.

8 Conclusion

SplatLeak identifies an own-image fidelity contribution that persists with synchronized-neighbor supervision. Background-reference photometric log-ratios improve membership discrimination on new scenes, and shared alignment extends auditing across numerical color changes. Teacher-source association survives indirect supervision in distilled students; source-aware routing reduces it, with real auxiliary supervision providing a further reduction. Controlled supervision, paired auditing and release construction together show how image provenance remains observable through renders and how a single-student release can reduce that association at a measured fidelity cost.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, additional literature discovery, routine code generation, preparation of supplementary reproducibility artifacts, and the organization and drafting of selected portions of the appendix. The main text was written by the authors, with AI assistance limited to improving clarity and readability. AI-assisted literature discovery was used occasionally to surface potentially relevant papers alongside the authors' own literature review; all papers ultimately relied upon or cited were read and assessed from the original sources. Code assistance mainly covered repetitive coding tasks and utility scripts. Assistance with supplementary reproducibility artifacts focused on organizing files, scripts, configurations, and documentation. AI-assisted drafting was limited to selected portions of the appendix based on author-provided experimental results, tables, and notes. The core research ideas, methodology, experimental design, and analysis were developed by the authors, and all AI-assisted content and artifacts were reviewed and verified by the authors.

ETHICS STATEMENT

Membership records photometric supervision rather than image authorship, ownership, or whole-pipeline use. A training manifest remains the direct provenance record. The audit can support release assessment, while unauthorized use could reveal participation in a training collection. Residential imagery, trajectories, and assets remain private; the artifact exposes only the permitted scalar evidence and public-data execution interfaces.

REPRODUCIBILITY STATEMENT

Appendices A–I organize protocols, controlled studies, per-scene evidence, attack comparisons, utility and costs. The Artifact preserves historical study identities and adds the independent sixteen-scene panel and common nine-configuration student evaluation. Saved-scalar replay, saved CNN prediction replay, pixel-level fitting and GPU training have separate interfaces and evidence boundaries. Portable sources document the RGB execution path. Public RGB reruns cover a target and eight references plus a four-arm intervention (13 models), reproducing selected original experiments.

REFERENCES

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5470–5479, June 2022.

Martin Bertran, Shuai Tang, Michael Kearns, Jamie Morgenstern, Aaron Roth, and Zhiwei Steven Wu. Scalable membership inference attacks via quantile regression. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 314–330. Curran Associates, Inc., 2023. doi: 10.52202/075280-0016. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/01328d0767830e73a612f9073e9ff15f-Paper-Conference.pdf.

Rickard Brännvall. Exponential-family membership inference: From LiRA and RMIA to BaVarIA. arXiv preprint arXiv:2603.11799, 2026. URL https://arxiv.org/abs/2603.11799.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, 2022. doi: 10.1109/SP46214.2022.9833649.

Grigorios-Aris Cheimariotis, Antonis Karakottas, Vangelis Chatzis, Angelos Kanlis, and Dimitrios Zarpalas. Image valuation in NeRF-based 3D reconstruction. In Modesto Castrillón-Santana, Carlos M. Travieso-González, Oscar Deniz Suarez, David Freire-Obregón, Daniel Hernández-Sosa, Javier Lorenzo-Navarro, and Oliverio J. Santana (eds.), *Computer Analysis of Images and Patterns*, pp. 375–385, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-04968-1.

Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. GAN-Leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20, pp. 343–362, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370899. doi: 10.1145/3372297.3417238. URL https://doi.org/10.1145/3372297.3417238.

Zixuan Chen, Guangcong Wang, Jiahao Zhu, Jianhuang Lai, and Xiaohua Xie. GuardSplat: Efficient and robust watermarking for 3D gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16325–16335, June 2025.

Rishav Chourasia, Batnyam Enkhtaivan, Kunihiro Ito, Junki Mori, Isamu Teranishi, and Hikaru Tsuchida. Knowledge cross-distillation for membership privacy. *Proceedings on Privacy Enhancing Technologies*, 2022(2):362–377, 2022. doi: 10.2478/popets-2022-0050.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 337–349, June 2018.

Yuntao Du, Jiacheng Li, Yuetian Chen, Kaiyuan Zhang, Zhizhen Yuan, Hanshen Xiao, Bruno Ribeiro, and Ninghui Li. Cascading and proxy membership inference attacks. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, San Diego, CA, USA, February 2026. Internet Society. ISBN 979-8-9919276-8-0. doi: 10.14722/ndss.2026.230661.

Zhiwen Fan, Kevin Wang, Kairun Wen, Zehao Zhu, Dejia Xu, and Zhangyang Wang. LightGaussian: Unbounded 3D gaussian compression with 15x reduction and 200+ FPS. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 140138–140158. Curran Associates, Inc., 2024. doi: 10.52202/079017-4447. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/fd881d3b625437354d4421818f81058f-Paper-Conference.pdf.

Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. *ACM Trans. Graph.*, 37(6): 257:1–257:15, November 2018. ISSN 0730-0301. doi: 10.1145/3272127.3275084. URL https://doi.org/10.1145/3272127.3275084.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Comput. Surv.*, 54(11s): 235:1–235:37, September 2022. ISSN 0360-0300. doi: 10.1145/3523273. URL https://doi.org/10.1145/3523273.

Ismat Jarin and Birhanu Eshete. MIAShield: Defending membership inference attacks via preemptive exclusion of members. *Proceedings on Privacy Enhancing Technologies*, 2023(1):400–416, 2023. doi: 10.56553/popets-2023-0024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, July 2023. ISSN 0730-0301. doi: 10.1145/3592433. URL https://doi.org/10.1145/3592433.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and Temples: Benchmarking large-scale scene reconstruction. *ACM Trans. Graph.*, 36(4):78:1–78:13, July 2017. ISSN 0730-0301. doi: 10.1145/3072959.3073599. URL https://doi.org/10.1145/3072959.3073599.

Joo Chan Lee, Daniel Rho, Xiangyu Sun, Jong Hwan Ko, and Eunbyung Park. Compact 3D gaussian representation for radiance field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21719–21728, June 2024.

Nan Li, Yike Zeng, Qian Zhang, Qi Zhang, Zhiyi Pan, Wei Feng, and Liang Wan. Write where it matters: Policy-guided watermarks for 3D gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6580–6590, June 2026.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17627–17638, October 2023.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. DL3DV-10K: A large-scale scene dataset for deep learning-based 3D vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169, June 2024.

Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-GS: Structured 3D gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20654–20664, June 2024.

Niantic Labs. SPZ: A file format for compressed 3D gaussian splats. Computer software, 2026. URL https://github.com/nianticlabs/spz/tree/v3.0.0. Version 3.0.0, released May 5, 2026.

Maxime Pietrantoni, Gabriela Csurka, and Torsten Sattler. Gaussian splatting feature fields for (privacy-preserving) visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1082–1092, June 2025.

Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12716–12725, June 2019.

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

Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3260–3269, July 2017.

Virat Shejwalkar and Amir Houmansadr. Membership privacy for machine learning models through knowledge transfer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):9549–9557, 2021. doi: 10.1609/aaai.v35i11.17150.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, 2017. doi: 10.1109/SP.2017.41.

Yuqi Tan, Xiang Liu, Shuzhao Xie, Bin Chen, Shu-Tao Xia, and Zhi Wang. WATER-GS: Toward copyright protection for 3D Gaussian Splatting via universal watermarking. arXiv preprint arXiv:2412.05695, 2024. URL https://arxiv.org/abs/2412.05695.

Shuai Tang, Zhiwei Steven Wu, Sergul Aydore, Michael Kearns, and Aaron Roth. Membership inference attacks on diffusion models via quantile regression. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47819–47829. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/tang24g.html.

Xinyu Tang, Saeed Mahloujifar, Liwei Song, Virat Shejwalkar, Milad Nasr, Amir Houmansadr, and Prateek Mittal. Mitigating membership inference attacks by Self-Distillation through a novel ensemble architecture. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 1433–1450, Boston, MA, August 2022. USENIX Association. ISBN 978-1-939133-31-1. URL https://www.usenix.org/conference/usenixsecurity22/presentation/tang.

Yukun Wang, Kunhong Li, Minglin Chen, Longguang Wang, Shunbo Zhou, Kaiwen Xue, and Yulan Guo. Distractor-free novel view synthesis via exploiting memorization effect in optimization. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.),

Computer Vision – ECCV 2024, pp. 477–493, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72949-2.

Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Lauren Watson, Chuan Guo, Graham Cormode, and Alexandre Sablayrolles. On the importance of difficulty calibration in membership inference attacks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=3eIrli0TwQ.

Lintao Xiang, Xinkai Chen, Jianhuang Lai, and Guangcong Wang. Distilled-3DGS: Distilled 3D Gaussian Splatting. arXiv preprint arXiv:2508.14037, 2025. URL https://arxiv.org/abs/2508.14037.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282, 2018. doi: 10.1109/CSF.2018.00027.

Junhyeog Yun, Minui Hong, and Gunhee Kim. FedMeNF: Privacy-preserving federated meta-learning for neural fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2161–2171, October 2025.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58244–58282. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zarifzadeh24a.html.

Bokang Zhang, Yanglin Zhang, Zhikun Zhang, Jinglan Yang, Lingying Huang, and Junfeng Wu. S^2NeRF: Privacy-preserving training framework for NeRF. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, pp. 258–272, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706363. doi: 10.1145/3658644.3690185. URL https://doi.org/10.1145/3658644.3690185.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, June 2018.

APPENDIX

Section A defines the common protocols. Sections B and C develop the controlled fidelity evidence and its persistence. Sections D to F present reference access, the complete W3 photometric study, readout comparisons and localization methods. The independent eight-C/eight-T audit is collected in Section G; student routing, the five-recipe audit, utility and deployment are collected in Section H. Section I gives execution paths and resource requirements. Each supporting study retains its own population and fitting identity; the Artifact indexes its complete records.

CONTENTS

A Methods and experimental protocols 15

B How own-image supervision forms the fidelity trace 19

C Persistence across targets, models and release conditions 24

D Reference access, calibration and photometric alignment 30

E Fidelity metrics and learned readouts 40

F Background-assisted localization and discovery coverage 44

G Independent readout, photometric and localization evaluation 49

H Source exclusion, utility and deployment 56

I Reproduction and resources 68

A METHODS AND EXPERIMENTAL PROTOCOLS

We test whether a registered image contributed photometric supervision to a released scene model. The experiments separate the formation of this trace, its persistence across models and releases, the information available to an auditor, and its propagation through distillation. This section defines the shared measurements and identifies the distinct experimental populations. Each subsequent section gives the design and results for one scientific question.

A.1 MEMBERSHIP, ACCESS AND EXPERIMENTAL POPULATIONS

For a directly optimized asset, $M_{\mathrm{opt}}(q) = 1$ means that candidate RGB I_q supervised its optimization. For a distilled student, $M_{\mathrm{src}}(q) = 1$ instead means that the image supervised the upstream teacher bank. The student receives teacher renders rather than protected source RGB. Registered cameras and intrinsics are fixed inputs unless the experiment explicitly estimates candidate poses. The source-exclusion construction and labels are specified in Section H.

Render-fidelity scoring requires candidate RGB and a target render at the same camera. Same-scene references add either known IN/OUT assignments or background RGB that excludes the candidates. Background-assisted localization additionally uses registered background images, their cameras and candidate intrinsics; it estimates the candidate pose. A bare splat without these inputs is a separate access condition (Section F).

Study	Data and model roles	Unit and initialization
Original ten scenes	Fixed target and $K8$ shadows per scene	Balanced image queries; member-only initialization
Shared-initialization crossover	Four membership assignments per scene	Scene and assignment bank; common A-only initialization
Supplied-pose replication	Four-scene D/S/C/T study and six-scene S/T extension	Targets nested in assignments and scenes; A-only initialization
Synchronized supervision	Four arms for eight rig-frame units	One query image and its three peers; common background initialization
Appearance and high coverage	Paired APP0/APP32-FM; nested $K8/20/40$ references	Scene and model realization; member-only initialization
Reference access and encoding	Earlier three-C/four-T DL3DV panel	One target and same-scene banks; independent target/reference assignments
Access/localization panel (W2)	Four C scenes and eight T scenes; one target, eight OUT and eight MATCHED references per scene	Scene AUC and full-roster decisions; model-specific allowed RGB
Photometric confirmation (W3)	Eight new C scenes and four new T scenes; one target and eight OUT references per scene	Scene AUC and full-roster decisions; model-specific allowed RGB
Independent readout panel (W4)	Eight new C scenes and eight new T scenes; one target and eight OUT references per scene (Section G)	Scene ranking and full-roster decisions; model-specific allowed RGB
Source exclusion	Four scenes, two target banks and two student seeds; original A_7^* (Section H.2) and five-recipe A_9^* (Section H.3)	Scene, source/teacher bank and nested optimizer seed; shared A-only initialization

Table 5: Experimental populations. C and T denote calibration and evaluation roles within the named study; they are not interchangeable pools across studies. Detailed scene counts appear with their results.

The four-C/eight-T access panel (W2). Twelve unused DL3DV scenes (Ling et al., 2024) are selected by fixed identifier order, registered-image counts and released OPENCV metadata. Each has 407–429 registered images. We reserve 32 nonmember anchors R before constructing membership, then partition the remaining population into candidates H and background B. Membership is balanced within H, and the target uses $\lfloor (N_0 - 32)/2 \rfloor$ training images. Anchors provide neither initialization observations nor photometric supervision.

Each scene contributes one target, eight background-only OUT models trained before candidate or anchor RGB is accessed, and eight MATCHED models with four IN and four OUT observations for each candidate. Feature pairs, tracks and initialization colors use each model's allowed training images. All 204 trajectories run to 30k and retain their 7k checkpoint. Native and SPZ releases of these checkpoints form four paired states. Each C scene has one primary state, with two T scenes assigned to each state. Primary evaluation therefore averages eight scene AUCs and pools decisions over 610 members and 610 nonmembers; C supplies 304 negatives. Additional states are paired measurements of the same trajectories.

Table 6: W2 four-C/eight-T access and localization design. Each scene has one primary recipe, 32 excluded anchors, and balanced optimization membership in H. Q_b is the number per class in the metadata-balanced query subset. The last two columns give the strongest pose-only classifier AUC and its permutation p value. Scene aliases are distinct from the earlier C/T panel in Section D.2.

Scene	Release	B	H	R	H_+	Q_b	Battery	A_1	p
W2C1	7k/native	225	150	32	75	70	Pass	0.5506	0.4428
W2C2	7k/SPZ	228	150	32	75	73	Pass	0.5977	0.1343
W2C3	30k/native	237	156	32	78	73	Pass	0.5466	0.4229
W2C4	30k/SPZ	228	152	32	76	75	Pass	0.6133	0.0597
W2T1	7k/native	227	150	32	75	70	Pass	0.5759	0.2537
W2T2	7k/SPZ	227	150	32	75	73	Pass	0.5646	0.3134
W2T3	30k/native	231	152	32	76	76	Fail	0.6607	0.0149
W2T4	30k/SPZ	238	158	32	79	77	Pass	0.5665	0.3731
W2T5	7k/native	239	158	32	79	76	Pass	0.6151	0.0697
W2T6	7k/SPZ	225	150	32	75	73	Pass	0.5939	0.1393
W2T7	30k/native	229	150	32	75	71	Pass	0.5100	0.9353
W2T8	30k/SPZ	228	152	32	76	70	Fail	0.6706	0.0050

The W3 population, scalar alignment, calibration sensitivity and complete localization results are presented together in Section D.6.

A.2 Query construction, camera controls and rendering

The original `random50_seed0` split assigns each sorted frame independently to member or non-member with probability one half. A MILP selects equal-cardinality groups subject to per-dimension standardized mean differences at most $\tau = 0.15$ over whitened camera center (3D), view direction (3D) and nearest-member-camera distance. Member queries exclude their zero-distance self entry when computing the nearest camera. The edge variables optimize cardinality without pairwise-distance cost, so the selected groups, rather than arbitrary solver edges, define the comparisons.

The four-layer battery checks raw SMD (L1), maximum-SMD permutation significance (L2), energy-distance permutation significance (L3), and pose-classifier discrimination (L4). L1 is a reported diagnostic, while the MILP uses the separate $\tau = 0.15$ construction constraint above. L2 and L3 each pass when their 200-permutation plus-one p-value is strictly greater than 0.05. L4 uses standardized multivariate pose covariates, fixed five-fold out-of-fold logistic predictions and the same permutation convention. A split fails L4 when both $\mathrm{A1best} = \max(\mathrm{AUC}, 1 - \mathrm{AUC}) > 0.55$ and $p_{\mathrm{L4}} < 0.05$. Inclusion requires both L2 and L3 to pass and no L4 failure. The post-matching nearest-distance score evaluated with the membership baselines is a separate univariate diagnostic.

The benchmark pool contains nine MipNeRF360, two Tanks&Temples and two DeepBlending scenes. Garden and room are matcher-development scenes and are excluded irrespective of their later outcomes. Treehill fails the frozen battery before training. The remaining ten scenes constitute the full primary roster. Their L4 values span 0.5028–0.5906 with no significant permutation result (Table 7).

Scene	n/class	L4 A1best	p	Scene	n/class	L4 A1best	p
bicycle	87	0.5414	0.5473	stump	56	0.5906	0.2587
bonsai	131	0.5391	0.4279	drjohnson	118	0.5442	0.3881
counter	109	0.5337	0.4428	playroom	102	0.5028	0.9652
flowers	77	0.5527	0.4080	train	135	0.5070	0.9204
kitchen	126	0.5142	0.7761	truck	113	0.5108	0.8607

Table 7: Frozen L4 pose-classifier diagnostics for the original ten-scene `random50_seed0` queries. Each entry comes from the selected-query manifest before render-based scoring; p is the corresponding permutation-test value. The selected sets contain 1,054 views per class in total.

Initialization and model configuration. The original targets and shadows use `member_retri`: member-only DLT retriangulation and member median RGB at the released poses. The strict-SfM variant and the membership-independent A-only crossover test different parts of this construction in Section B. Default training uses 30k iterations, a black background and NVIDIA RTX 3090 GPUs. The downscale factor is 2/4 for MipNeRF360 indoor/outdoor and 1 for Tanks&Temples and DeepBlending. Scaffold configurations are specified separately in Section C.

The eight original shadow seeds are 200–207. The generator restarts on sorted image order in each scene, so equal cross-scene ordinals share membership vectors. They are not independent shadow-assignment replicates. The first all-IN/all-OUT ordinals are 585/589, beyond the evaluated scene sizes, explaining complete two-sided coverage. Loaded names and member-only initialization lineage are checked against each shadow's split before scoring.

Parity and full-canvas measurements. The render canvas equals the effective training image size, $\mathrm{round}(\text{file size}/r)$, while field of view uses the full-resolution COLMAP intrinsics. Candidate resizing follows PIL bicubic and the training `PILtoTorch` path. Target and reference renders use identical cameras, canvas and black background. In the additional DL3DV panel, released OPENCV images are converted to centered PINHOLE images at 960×540, with linear interpolation and a constant black border; training and scoring retain that full canvas.

The additional metrics consume float32 renders clamped to $[0, 1]$. SSIM (Wang et al., 2004) follows the 3DGS evaluation implementation: an 11×11 Gaussian window, standard deviation 1.5, padding 5, constants 0.01^2 and 0.03^2, and averaging over channels and pixels. LPIPS (Zhang et al., 2018) uses AlexNet v0.1 in evaluation mode, its standard $[0, 1]$ normalization interface and no additional crop or resize. SPZ uses release v3.0.0, format 4, 12 fractional position bits, 5 first-order and 4 remaining SH bits, without scale normalization.

A.3 Fidelity scores and reference estimators

For RGB residual $r(q) = \frac{1}{3}\sum_c |\mathcal{R}(G; q) - I_q|_c$, let $e_G(q)$ be mean squared RGB error floored at 10^{-10}. PSNR is $-10 \log_{10} e_G(q)$. The original eight-feature residual vector is

$$\big[\mathrm{PSNR}, -\bar{r}, -\mathrm{MSE}, -r_{50}, -r_{90}, -r_{99}, -\overline{|\nabla^2 r|}, -\mathrm{Gini}(\tilde{r})\big].$$

Here ∇^2 is the four-neighbour Laplacian and $\tilde{r}$ is the 16×16 average-pooled residual. For sorted patches, $\mathrm{Gini}(\tilde{r}) = \sum_i (2i - n - 1)\tilde{r}_{(i)} / [n \sum_i \tilde{r}_{(i)}]$. Every direction is member-oriented. The original `RR_logreg` combines these features with logistic regression. Its ten-scene per-scene AUCs are cross-validation fold means; the perturbation and pose-recovery studies use pooled out-of-fold predictions, differing by at most 0.004 AUC.

Two-sided LiRA. Following the online IN/OUT construction of LiRA (Carlini et al., 2022), reference PSNRs at each query give means $\mu_{\mathrm{in}}(q), \mu_{\mathrm{out}}(q)$. The primary estimator pools within-class deviations across queries and reference models to estimate global standard deviations $\sigma_{\mathrm{in}}, \sigma_{\mathrm{out}}$:

$$\Lambda(q) = \log \mathcal{N}(t(q); \mu_{\mathrm{in}}(q), \sigma_{\mathrm{in}}) - \log \mathcal{N}(t(q); \mu_{\mathrm{out}}(q), \sigma_{\mathrm{out}}).$$

The third argument here is standard deviation. The per-query variant instead estimates it from each query's class values, falling back to the global value when that side has fewer than three references. A missing IN or OUT side produces an unscored query recorded in coverage. Hinge-IN and the one-sided OUT-tail are additional diagnostics; hinge-IN removes the penalty for exceeding the IN mean.

BaVarIA-n with actual reference support. BaVarIA-n (Brännvall, 2026) uses a Gaussian likelihood with Bayesian-stabilized variance and a scene/class empirical-Bayes prior. For class c and query q, let n_{cq}, μ_{cq} and SS_{cq} be its reference count, mean and sum of squared deviations from μ_{cq}, respectively. With raw reference PSNR y,

$$\begin{aligned}
&\kappa_0 = 1, \quad \alpha_0 = 2, \quad \mu_{0c} = \mathrm{mean}\{y_{kq} : m_{kq} = c\},\\
&\beta_{0c} = \max(\mathrm{var}\{y_{kq} : m_{kq} = c\}, 10^{-12})(\alpha_0 - 1),\\
&\alpha_{cq} = \alpha_0 + n_{cq}/2,\\
&\beta_{cq} = \beta_{0c} + \tfrac{1}{2} SS_{cq} + \frac{\kappa_0 n_{cq}(\mu_{cq} - \mu_{0c})^2}{2(\kappa_0 + n_{cq})}, \qquad v_{cq} = \frac{\beta_{cq}}{\alpha_{cq} - 1}.
\end{aligned}$$

The score is $\log \mathcal{N}(t; \mu_{1q}, v_{1q}) - \log \mathcal{N}(t; \mu_{0q}, v_{0q})$, now parameterized by variance. Balanced MATCHED banks use $n_{cq} = 4$. The appearance, high-coverage and ETH3D studies retain their actual class counts and specified support rules; unsupported queries remain visible in their coverage summaries.

Background-only scores. For release r, eight OUT models give mean $\mu_r(q)$ and unbiased variance $v_r(q)$. Fixed GAP is $t(q) - \mu_{30k,\mathrm{native}}(q)$, applying reference-based difficulty calibration (Watson et al., 2022) to registered rendering fidelity. Fixed T uses the same mature native references in

$$s_r(q) = -\log \overline{F}_{m-1}\left(\frac{t(q) - \mu_r(q)}{\sqrt{\max(v_r(q), 10^{-12})(1 + 1/m)}}\right), \qquad m = 8,$$

where $\overline{F}$ is the Student-t survival function. Recipe-known GAP/T uses the declared target release. The four-state mixture averages tail probabilities: $s_{\mathrm{mix}} = -\log[\frac{1}{4}\sum_{r\in\mathcal{R}} \exp(-s_r)]$. Its paired snapshots do not constitute independent reference models. MATCHED GAP/T instead uses the query's four OUT members of the candidate-containing bank, so it belongs to the stronger-access group.

For another oriented metric f_m, RAW is $f_m(I_q, G)$ and GAP subtracts its mean over the eight OUT references. The additional comparison includes negative L1, SSIM, negative LPIPS and the negative derived photometric reconstruction loss $-[0.8\ell_1 + 0.2(1 - \mathrm{SSIM})]$. The loss weight follows the retained model records. Metric choice and reference subtraction are evaluated separately in Section E.

A.4 Calibration rules and statistical interpretation

Each study retains its original calibration rule. A nominal level describes how a cutoff was constructed; evaluation TP/FP reports its realized operating point. The relevant rules differ as follows.

The original nominal 0.1% point uses two pooled negatives, $2/1054 = 0.19\%$ without ties, and is a sensitivity diagnostic. Query bootstrap intervals condition on the fixed target/reference assets and retain their cross-scene seed dependence. In W2, W3 and W4, AUC averages scenes equally; TP/FP pools integer decisions on each study's original class rosters. Normalized partial AUC at β is $\beta^{-1}\int_0^\beta \mathrm{TPR}(u)\,du$ under the retained ROC interpolation. Missing required observations produce abstentions within the full denominator.

A fixed GAP cutoff corresponds to a candidate-dependent raw-PSNR requirement: the reference mean plus the calibrated offset,

$$\mathrm{GAP}(q) > c_\alpha \quad \Longleftrightarrow \quad t(q) > \mu_{\mathrm{OUT}}(q) + c_\alpha.$$

R32 additionally centers scores using separate nonmember anchors; its measured tail behavior is reported in Section D.

Study	Cutoff and decision	Interpretation
Original ten-scene LiRA/features	$k = \lceil \alpha n_- \rceil$-th largest labeled target-negative score; $s \geq \tau$	Retrospective target-labeled summary; nominal 1% is $11/1054$ without ties
Four-scene supplied-pose study and its extension	$b = \lfloor \alpha n_- \rfloor$; rank $b+1$ among C negatives; strict $s > \tau$	At most b C exceedances; T behavior measured separately
W2 four-C/eight-T access panel	$j = \lfloor \alpha(N_C^- + 1) \rfloor$; rank j, strict $s > \tau$; $j = 0$ gives $+\infty$	$N_C^- = 304$, ranks 3/15 at 1%/5%; separate rule per readout/pose/CNN seed
W3 eight-C/four-T photometric panel	Same strict order-statistic rule, fitted separately for the declared branch, pose and target panel	$N_C^- = 615$, ranks 6/30 at 1%/5%; all8 primary and eleven fixed C pools
W4 eight-C/eight-T readout panel	Same strict order-statistic rule, fitted separately per pose, release, alignment and readout	Primary $N_C^- = 620$, rank 6 at 1%; leave-one-C sensitivity in Section G.6
Real post-release capture	Beta-bound-selected anchor rank; strict $s > \tau$	Requires independent exchangeable null units for its risk interpretation; Section F gives the observed assumptions

Table 8: Threshold conventions. Earlier reference-access and ETH3D transfer retain their explicitly specified C pools and ranks in Section D. A cutoff from one study is not silently substituted for another.

B How own-image supervision forms the fidelity trace

The mechanism experiments progressively control initialization, training order and synchronous visual coverage. Together they identify a contribution from a candidate's own photometric supervision, while retaining the effects of subsequent optimization and topology changes.

B.1 A crossover with membership-independent initialization

Counter is the assay, bonsai and truck are prespecified replications, and kitchen is an analysis-specific extension with known earlier results. Each scene uses a bit-identical initialization constructed from auxiliary A RGB alone. Candidate Q features, tracks, colors and assignments never enter it. Released full-reconstruction poses and intrinsics remain fixed inputs. The partitions and roughly 80% training coverage in Table 9 belong to this mechanism study.

scene	role	$\lvert V \rvert$	$\lvert A \rvert$	$\lvert Q \rvert$	Q-IN/target	train/target
counter	assay	240	144	96	48	192 (80.00%)
bonsai	replication	292	176	116	58	234 (80.14%)
truck	replication	251	151	100	50	201 (80.08%)
kitchen	analysis ext.	279	167	112	56	223 (79.93%)

Table 9: Scene roles, partitions, and per-target training coverage. Each target uses $A \cup Q_{\mathrm{IN}}$. The approximately 80% coverage belongs to this mechanism experiment and is unrelated to LiRA calibration or operational-FPR coverage.

Each candidate receives one of six two-IN/two-OUT four-bit codes. Thus every candidate supervises two targets and each target trains on half of Q. A 5,000-state finite assignment bank preserves row/column totals, code counts, complement symmetry and target-wise marginal pose SMD at most 0.15. One committed uniform draw determines the observed assignment, and all Q candidates are evaluated.

The matched counter comparator changes the shared A-only initialization to a target-specific fixed-pose initialization from $A \cup Q_{\mathrm{IN},r}$. Training sets, camera metadata, normalization, optimizer schedule, renderer and target-seed mapping remain matched.

This study retains its original raw-RGB endpoint $Y_{sqr} = -10\log_{10}(\mathrm{MSE}_{sqr} + 10^{-12})$. Its candidate-centred effect is

$$D_{sq} = \frac{1}{2}\sum_{r=0}^{3}(2M_{sqr}^{\mathrm{opt}} - 1)Y_{sqr}, \qquad \widehat{\beta}_s = \frac{1}{|Q_s|}\sum_{q\in Q_s} D_{sq}.$$

Balanced candidate and target margins make this the membership coefficient of a candidate- plus target-fixed-effect regression. Counter and kitchen enumerate all 5,000 bank states, including ties. The two-scene replication averages scenes equally and draws 5,000 independent pairs of bank indices with replacement, using a plus-one Monte-Carlo p-value.

evidence set	scene/arm	initialization	$\widehat{\beta}$ (dB)	MSE IN−OUT	inference
assay	counter/shared	A-only shared	2.223166	-5.56724×10^{-4}	exact $1/5000$
assay	counter/control	target-specific	2.238110	-5.53492×10^{-4}	exact $1/5000$
replication	bonsai/shared	A-only shared	2.002902	-3.10804×10^{-4}	component; no scene p
replication	truck/shared	A-only shared	2.775057	-1.53904×10^{-3}	component; no scene p
replication	macro	A-only shared	2.388980	—	MC plus-one $1/5001$
analysis ext.	kitchen/shared	A-only shared	2.281053	-2.43647×10^{-4}	exact $1/5000$

Table 10: Thirty-thousand-iteration crossover results. Each prespecified evidence set satisfies its practical-effect and one-sided randomization criterion. The equal-scene mean across the four shared-initialization scenes is 2.320545 dB, reported descriptively without a new pooled test. Negative MSE IN−OUT denotes lower error when a candidate is assigned IN.

Shared-initialization effects are 2.00–2.78 dB, with descriptive four-scene mean 2.320545 dB (Table 10). Membership-dependent initialization therefore cannot alone explain the fidelity trace. Assignment changes several candidates simultaneously, so the effect includes own supervision, peer exposure, densification and their interactions. The maximum ridge pose-only A1best values are 0.720/0.739/0.628/0.753; their corresponding bank upper-tail proportions are 0.5708/0.1376/0.9872/0.1440. These diagnostics compare the draw with its constrained bank and do not trigger reassignment.

Strict member-only feature matching. A separate five-scene sensitivity rebuilds member-only SIFT matches and triangulates them at fixed published poses, giving zero mixed tracks. Target-only RR changes by at most 0.0043, within the predefined 0.05 tolerance; the complete five-scene table is retained in the Artifact. Both sides retain member-dependent initialization; the crossover above supplies the separate initialization-independent test.

B.2 An own-gradient intervention under matched training order

The matched-schedule study (M1-S) holds other image assignments fixed and switches the candidate's photometric gradient. Within each frozen 12-candidate roster on counter and truck, a pose-only rule selects the candidate nearest to and farthest from its nearest A camera, breaking ties by filename. The broad design precedes natural add-one outcomes; the exact pose-based instance is materialized after those outcomes exist. Two nested optimizer seeds per stratum yield eight comparison cells.

ON and OFF share the $A + q$ roster, serialized 30k camera schedule, A-only initialization, normalization and optimizer/densification clocks. OFF executes the candidate's forward/backward and scheduled statistic/optimizer slots with its photometric scalar multiplied by zero. Sixteen models produce 320 renders over the twelve candidates and eight disjoint outer views. The primary endpoint is q-own-view PSNR(ON)−PSNR(OFF).

Natural add-one and matched-schedule PSNR changes agree in direction in all eight cells. SSIM improves and MSE falls in every cell; q-minus-median-peer gains are 2.44–8.04 dB. A descriptive maximum-peer check also leaves q ahead by at least 2.13 dB. Peer-distance correlations range from -0.227 to 0.700 without a repeated direction. These observations support own photometric supervision as a contributor to the selected effects.

Table 11: Natural add-one and matched-schedule q-own-view effects at 30k. Natural is $A{+}q - A$; M1-S is ON−OFF under the identical serialized $A{+}q$ schedule. Higher PSNR/SSIM and lower MSE are favorable. Peer med. is the median M1-S PSNR change over the other 11 frozen candidate views.

			Natural	Matched schedule (M1-S)				
Scene	Stratum	Seed	ΔPSNR	ΔPSNR	ΔSSIM	ΔMSE	Peer median ΔPSNR	$q-$peer median
counter	near	0	2.4091	2.5283	0.041196	−0.000769	0.0084	2.5199
counter	near	1	3.0591	2.4169	0.040909	−0.000693	−0.0264	2.4434
counter	far	0	3.1134	3.0864	0.050306	−0.001067	0.0137	3.0727
counter	far	1	2.8011	3.0467	0.049397	−0.001110	0.0683	2.9785
truck	near	0	4.0544	3.8702	0.043726	−0.002012	−0.1747	4.0450
truck	near	1	3.9794	3.6939	0.039125	−0.001928	0.0237	3.6702
truck	far	0	7.9833	6.4472	0.097079	−0.005408	0.0563	6.3909
truck	far	1	7.1392	8.0666	0.117098	−0.006648	0.0314	8.0352

Outer-view utility and topology remain heterogeneous. Three outer8 PSNR changes are negative; LPIPS improves in all four counter cells and worsens in all four truck cells. ON has 580–36,042 more Gaussians. These topology responses form part of the gradient intervention rather than an independently identified mediator. The original companion measurements are retained in Artifact `docs/FULL_RESULTS.md`, record U. Scene is the highest independent unit ($n = 2$), with seeds and near/far strata nested within it.

B.3 Supervision beyond synchronized peer coverage

Four ETH3D scenes (Schöps et al., 2017), delivery area, sand box, playground and lakeside, provide two preselected rig-frame units each. Of 237/278/240/266 complete four-camera frames, a fixed stream selects 118/139/120/133 background frames. The two remaining canonical-camera positions nearest one third and two thirds of trajectory arc length define the units. The smallest camera ID is canonical. Each scene's two units share fresh background-only matching, tracks, points, colors and normalization.

For a unit, q is the canonical image and P its three synchronous peers. Four native 3DGS arms share point cloud, model seed, camera roster and serialized 30k schedule, using SH3, full resolution and a black background. Off slots have zero photometric gradient while optimizer steps and densification clocks continue. Let Y_{ab} enable own-image supervision a and peer supervision b. The query effects are

$$\Delta_{q|P=0} = Y_{10}(q) - Y_{00}(q), \qquad \Delta_{q|P=1} = Y_{11}(q) - Y_{01}(q).$$

Peer effects are $Y_{01} - Y_{00}$ and $Y_{11} - Y_{10}$; their interaction is $Y_{11} - Y_{10} - Y_{01} + Y_{00}$. All four query values and conditional contrasts are retained in Tables 12 and 13.

Synchronization fixes capture time, not identical visible surfaces. These contrasts measure the added contribution of q under the three observed peer viewpoints, including content they do not fully cover.

Table 12: All four-arm query PSNRs (dB) for the eight preselected synchronized-view units. Y_{ab} switches own-view supervision a and the three synchronous peers b off (0) or on (1). Frames 1/2 are the pose-selected trajectory positions near one/two thirds. Each row contains four 30k models with common background initialization and schedule.

Scene	Frame	Y_{00}	Y_{10}	Y_{01}	Y_{11}
delivery area	1	10.9258	18.5445	14.5189	20.8565
delivery area	2	13.0564	19.3705	17.8497	18.3920
sand box	1	18.7153	19.6614	18.7016	19.6109
sand box	2	20.8913	22.5161	20.1538	21.6685
playground	1	11.4901	23.1280	13.2944	25.6858
playground	2	11.5158	19.1620	12.0812	18.7897
lakeside	1	17.1186	20.9090	15.9919	18.8300
lakeside	2	15.5911	18.7254	13.3742	17.8434

Table 13: Conditional effects on the query and the 16 fixed other views (dB). Every own-view contrast is positive. Own and Peers identify the supervision switched ON minus OFF; the second header line gives the other supervision's fixed state. The first five effects are measured at the candidate view; I is the interaction. The final two columns average own-supervision effects over the same 16 other views. Means weight the eight units equally, equivalently averaging within and then across the four scenes.

		Candidate-view effects (dB)					Other 16-view mean	
Scene	Frame	Own peers OFF	Own peers ON	Peers own OFF	Peers own ON	I	Own peers OFF	Own peers ON
delivery area	1	+7.6187	+6.3376	+3.5931	+2.3120	-1.2811	+0.3770	+0.9265
delivery area	2	+6.3141	+0.5423	+4.7933	-0.9786	-5.7718	-0.5378	+1.2088
sand box	1	+0.9461	+0.9093	-0.0137	-0.0504	-0.0368	+0.5558	-0.0923
sand box	2	+1.6248	+1.5148	-0.7375	-0.8476	-0.1100	+0.3158	-0.2208
playground	1	+11.6379	+12.3914	+1.8043	+2.5578	+0.7535	-0.0898	+0.3467
playground	2	+7.6462	+6.7085	+0.5654	-0.3722	-0.9377	-0.4311	-0.3111
lakeside	1	+3.7904	+2.8382	-1.1267	-2.0789	-0.9522	-0.6087	+0.0832
lakeside	2	+3.1343	+4.4692	-2.2169	-0.8820	+1.3349	+0.1774	+0.4465
Mean		**+5.3391**	**+4.4639**	**+0.8327**	**-0.0425**	**-0.8752**	**-0.0302**	**+0.2984**

With peers on, own-image supervision improves all eight query PSNRs by 0.5423–12.3914 dB, with mean 4.4639 dB. With peers off, the mean is 5.3391 dB. The 16 other canonical views are selected before training, evenly in timestamp order after excluding the intervention frames. Their mean changes are reported for every unit; they may include background training views. The own-view effect exceeds this mean in seven of eight units. In the retained delivery-area frame-2 exception, q gains 0.5423 dB while the other-view mean gains 1.2088 dB.

Peer effects have mixed signs, with four positive values when own supervision is off and two when it is on. Six interactions are negative and two positive. Synchronous scene coverage therefore leaves an additional own-image contribution whose interaction with peer supervision varies by unit. Released reconstruction cameras remain fixed throughout. This image-level intervention differs from the joint rig-frame membership study in Section D. The scalar replay retains all eight units and all 160 unit–view pairs with their four-arm values.

B.4 Exact-view decay and statistical readout

For each imposed camera perturbation δ, define the PSNR drop $d_i(\delta) = \mathrm{PSNR}_i(0) - \mathrm{PSNR}_i(\delta)$. The group contrast $\mathrm{DiD} = \overline{d}_{\mathrm{member}} - \overline{d}_{\mathrm{nonmember}}$ equals the reduction of the member/nonmember PSNR margin. It uses aggregate-balanced groups rather than the solver's arbitrary edge assignments. Members lose 3.0–4.8 dB more on average, and their margin approaches zero (Table 14). This locates the advantage near the trained view, complementing the supervision interventions.

scene	perturb	Δ_{mem}	Δ_{non}	mean gap	marg. ret.
counter	rot 0.5°	13.76	9.50	$+4.26$	-0.03
counter	rot 2°	18.02	13.67	$+4.35$	-0.06
counter	trans $0.1\times$	11.06	7.13	$+3.92$	0.05
counter	trans $0.25\times$	14.70	10.05	$+4.66$	-0.13
bonsai	rot 0.5°	15.41	12.10	$+3.31$	0.01
bonsai	rot 2°	19.23	16.03	$+3.20$	0.04
bonsai	trans $0.1\times$	11.68	8.67	$+3.01$	0.10
bonsai	trans $0.25\times$	14.66	11.68	$+2.98$	0.11
truck	rot 0.5°	11.60	7.09	$+4.50$	0.05
truck	rot 2°	15.08	10.33	$+4.75$	-0.01
truck	trans $0.1\times$	9.54	5.18	$+4.35$	0.08
truck	trans $0.25\times$	12.41	7.88	$+4.54$	0.04

Table 14: Group-mean member-vs-non-member PSNR drop (dB) under the same perturbation conditions on aggregate-balanced selected sets (counter 109, bonsai 131, truck 113 views per class). Translation magnitudes are multiples of the median inter-camera spacing. Δ: per-view drop from exact pose, averaged by class; mean gap $= \overline{\Delta}_{\text{mem}} - \overline{\Delta}_{\text{non}}$; marg. ret. $=$ group-margin(δ)/group-margin(0) (exact-pose margins 4.12/3.34/4.72 dB).

Bonsai combines weak retrospective nominal-1% LiRA detection (0.031; Section E.2) with a positive shared-initialization effect (2.003 dB; Table 10). The controlled effect and the statistical tail readout therefore answer different questions. The Artifact retains their paired visualization.

C PERSISTENCE ACROSS TARGETS, MODELS AND RELEASE CONDITIONS

These studies vary target realization, model family, training coverage and release representation. Their protocols and units remain distinct: seed replication tests stability within a scene, appearance conditioning changes a training/serving bundle, and encoding compares paired releases of the same trained model.

C.1 TARGET ASSIGNMENTS AND OPTIMIZER SEEDS

The four-scene supplied-pose study uses A-only initialization and 32 evaluation targets under 32 independent assignments. Its equal-scene AUC and normalized pAUC to FPR .05 are 0.9789 and 0.8511. Its lowest asset AUC is 0.8847 on bonsai target 4, also the minimum across all 80 assets in this study and the extension below. Estimator selection and the accompanying calibration outcomes are given in Section D.

The six-scene extension adds bicycle, flowers, kitchen, stump, train and playroom, all previously represented in the original ten-scene measurement. Each scene has eight reference models and eight evaluation models. Four evaluation columns come from one fixed joint assignment matrix, each trained under two optimizer seeds. Every candidate is four-IN/four-OUT across each eight-model bank. The extension trains 96 models: 48 references and 48 evaluation targets under 24 linked assignments.

The pose-only restricted randomization directly constrains camera center, direction and distance to the nearest camera in $A \cup Q_{\mathrm{IN}}$, with member self-exclusion, to SMD at most 0.15. The scene partitions, score, recipe and endpoints remain unchanged. All models use one shared A-only initialization per scene, 30k iterations, SH3 and parity-fixed rendering.

Endpoints are computed for each asset before averaging the two seeds within assignment, then four assignments within scene, then scenes equally. Table 15 retains all six scene summaries and assignment ranges. All 4,160 evaluation scores are finite. The minimum individual-asset AUC is 0.975765 on kitchen $(a = 2, r = 0)$; the minimum raw margin is 2.123819 dB on another kitchen asset $(a = 0, r = 1)$.

None of the 24 seed pairs changes the direction of AUC relative to 0.5 or of the raw PSNR margin relative to zero. Equal-scene means of the within-scene median absolute changes are 0.001894 AUC, 0.121 dB and 0.01885 pAUC. The largest AUC change, 0.01754, and pAUC change, 0.2883, both occur on kitchen assignment 2. The lowest nominal-1% decision agreement is 0.875 on kitchen assignment 1. Thus overall direction persists while the weakest scene retains tail sensitivity. Complete asset and assignment records are in Artifact records Q/S.

Together these two studies provide 80 evaluation assets under 56 assignments. Their different assignment and seed structures are preserved in the separate summaries. Under the transported 1% rule, train exceeds its FP budget by one (5 versus 4), while kitchen is exactly at its budget. These calibration outcomes are distinct from the strong continuous ranking.

C.2 SCAFFOLD-GS AND APPEARANCE-CONDITIONED SERVING

The all-ten Scaffold-GS reproduction disables per-image appearance embeddings (`appearance_dim=0`). Targets and shadows each use their own member-only split and initialization, and both sides use the Scaffold renderer. All 10×8 shadows provide complete two-sided query support. Pooled LiRA AUC is 0.9949, with the same intermediate kitchen and weak-tail bonsai pattern (Table 16). Earlier target-only Scaffold RR is 0.905/0.866 on kitchen/bonsai, compared with vanilla RR 0.934/0.873 in that comparison.

The separate appearance study compares APP0 with APP32-FM. APP32 trains a 32-dimensional appearance table and its color MLP, then serves every camera with the model-local mean vector. It receives neither camera identity lookup nor target membership at inference. Each of counter, bonsai, truck and playroom has eight paired shadow identities and two paired target realizations, for 80 models. Within an identity, the two configurations share roster, initialization, serialized schedule and seed.

Table 15: Six-scene ranking and fixed-cutoff decisions. Ranking endpoints are computed per asset, averaged over two seeds within each assignment, then equally over assignments and scenes; ranges span four assignment means. The lower panel preserves the complete integer denominators and historical TP floors/FP budgets. The equal-scene 1% TPR/FPR is 0.9631/0.0032 (5%: 0.9826/0.0100); pooled 1% TPR is $1980/2080 = 0.9519$. These are different weighting conventions. The nominal levels describe transported cutoff sources, and the historical criterion is not calibrated false-positive control.

Ranking

Scene	AUC (assignment range)	Δ_{PSNR} dB (range)	nPAUC$_{.05}$
bicycle	1.0000 (1.0000–1.0000)	7.049 (6.660–7.324)	1.0000
flowers	1.0000 (1.0000–1.0000)	5.867 (5.680–5.996)	1.0000
kitchen	0.9887 (0.9845–0.9941)	2.417 (2.133–2.943)	0.8801
stump	1.0000 (1.0000–1.0000)	11.059 (10.743–11.490)	1.0000
train	0.9980 (0.9969–0.9986)	4.677 (4.391–4.820)	0.9694
playroom	1.0000 (1.0000–1.0000)	8.054 (7.156–8.701)	1.0000
equal-scene	**0.9978**	**6.520**	**0.9749**

Decisions on the full population

scene	N_+/N_-	TP floor / FP budget	1% TP/FP	criterion	5% TP/FP
bicycle	312/312	156/3	312/0	yes	312/2
flowers	280/280	140/2	280/0	yes	280/0
kitchen	448/448	224/4	364/4	yes	406/9
stump	200/200	100/2	200/0	yes	200/0
train	480/480	240/4	465/5	no	475/16
playroom	360/360	180/3	359/0	yes	360/0
pooled accounting	**2080/2080**	–	**1980/9**	**5/6**	**2033/27**

scene	n_{valid}/n_-	LiRA AUC	TPR@1%	TPR@5%	one-sided excl.	shadow cov.
counter	218/109	0.9957	0.9174	0.9908	0	8/8
kitchen	252/126	0.9785	0.6984	0.9127	0	8/8
bonsai	262/131	0.9316	0.0382	0.7099	0	8/8
bicycle	174/87	1.0000	1.0000	1.0000	0	8/8
flowers	154/77	1.0000	1.0000	1.0000	0	8/8
stump	112/56	1.0000	1.0000	1.0000	0	8/8
truck	226/113	0.9998	0.9912	1.0000	0	8/8
train	270/135	0.9964	0.9333	0.9852	0	8/8
drjohnson	236/118	1.0000	1.0000	1.0000	0	8/8
playroom	204/102	1.0000	1.0000	1.0000	0	8/8
pooled	2108/1054	**0.9949**	**0.9156**	**0.9791**	0	10×8

Table 16: All-ten Scaffold-GS LiRA with the global-$K8$ two-sided estimator and the retrospective empirical-quantile convention from Section A.4. The model family changes while the leakage pattern remains: kitchen is intermediate, and bonsai is again the weak-tail scene at the nominal 1% threshold.

The primary global-σ LiRA uses the actual class-specific K8 counts, ranging from one to seven. All eight APP32-FM target panels have AUC above 0.5 and positive raw score margins and pass the specified exact one-sided rank-sum test. They remain technical target realizations within four scenes. Tables 17 and 18 separate attack attenuation from held-out fidelity.

Table 17: Matched Scaffold-GS direct-membership results. L is the primary global-σ LiRA AUC, and B is the secondary variable-count BaVarIA-n AUC. Δ is APP32-FM minus APP0 for L, and p_{32} is the exact one-sided APP32-FM rank-sum value. The two target realizations per scene are technical replicates.

scene	target	L APP0	L APP32	Δ L	B APP0	B APP32	p_{32}
counter	z0	0.996465	0.953960	-0.042505	0.996297	0.952024	2.30×10^{-41}
counter	z1	0.998317	0.972561	-0.025755	0.998064	0.969615	3.66×10^{-47}
bonsai	z0	0.932755	0.758814	-0.173941	0.934794	0.740924	3.90×10^{-14}
bonsai	z1	0.910320	0.733873	-0.176447	0.906882	0.714411	9.83×10^{-12}
truck	z0	1.000000	0.958650	-0.041350	0.999922	0.958180	3.33×10^{-44}
truck	z1	1.000000	0.949644	-0.050356	1.000000	0.950192	1.27×10^{-41}
playroom	z0	1.000000	1.000000	0.000000	1.000000	0.998078	6.97×10^{-61}
playroom	z1	1.000000	1.000000	0.000000	1.000000	0.998462	6.97×10^{-61}
equal-panel mean	–	0.979732	0.915938	-0.063794	0.979495	0.910236	–

Table 18: Held-out utility under matched APP0 and APP32-FM targets. PSNR columns are absolute scene means, and all Δ columns are APP32-FM minus APP0. PSNR is in dB; SSIM and LPIPS are unitless.

scene	APP0 PSNR	APP32 PSNR	ΔPSNR	ΔSSIM	ΔLPIPS
counter	28.817992	27.858847	-0.959145	-0.009249	$+0.010191$
bonsai	32.383791	29.401977	-2.981814	-0.016232	$+0.013221$
truck	24.741831	23.512739	-1.229093	-0.016850	$+0.016469$
playroom	29.363040	29.362847	-0.000193	-0.001317	-0.000534
equal-scene mean	28.826664	27.534103	-1.292561	-0.010912	$+0.009836$

Mean primary AUC changes from 0.979732 to 0.915938. Playroom remains at AUC 1.0, while bonsai falls most. The equal-scene held-out changes are -1.292561 dB PSNR, -0.010912 SSIM and $+0.009836$ LPIPS. This is a comparison of the complete APP32 training and fixed-mean serving bundle with APP0; native camera-indexed appearance serving is outside the tested configuration.

C.3 Training fraction and reference support

The high-coverage study trains one target and 40 references in each of counter, bonsai, truck and playroom using $\lfloor 0.9N \rfloor$ images. All 1,008 registered views are evaluated: 905 members and 103 nonmembers. Each target uses a random stream independent of its bank. The 40 distinct reference masks are arranged in two balanced 20-model blocks, supplying 2–3 OUT observations per query at $K20$ and 4–5 at $K40$. The fixed first eight and first twenty models define the smaller budgets. The study contains 164 trained models and 41,328 render records.

For a bank drawn from the same N images, training each reference on $\lfloor fN \rfloor$ images leaves $d = N - \lfloor fN \rfloor$ OUT slots per reference. Supplying at least r OUT observations to every candidate therefore requires $Kd \geq rN$, or $K \geq \lceil rN/d \rceil$. This capacity bound is necessary; how the masks allocate those slots determines per-query coverage. The balanced blocks above distribute the available support across the full roster.

BaVarIA-n uses actual class counts and priors fitted to the supported reference pool. Missing IN or OUT support produces an abstention within the original denominator. The study transports the earlier MATCHED C cutoffs 0.9655193026881035/0.5525160502861346 at 1%/5%, and RAW transports 33.01089132544482/30.891366108148407 dB. No new calibration scene is introduced.

Table 19: Full registered-population ranking and fixed decisions under approximately 90% training coverage. The 1%/5% labels identify the calibration levels inherited from Section D.2. Each decision cell is TP/N_+; FP/N_-. Scored and abstained columns give positive/negative counts. AUC uses scored queries; decision denominators include abstentions. $K20$ and $K40$ score every query.

Scene	Bank	AUC	Scored $+/-$	Abst. $+/-$	1% TP; FP	5% TP; FP
counter	RAW	0.7340	216/24	0/0	9/216; 1/24	87/216; 4/24
	$K8$	1.0000	142/11	74/13	78/216; 0/24	135/216; 0/24
	$K20$	0.9956	216/24	0/0	180/216; 0/24	207/216; 0/24
	$K40$	0.9954	216/24	0/0	188/216; 0/24	204/216; 0/24
bonsai	RAW	0.7085	262/30	0/0	193/262; 14/30	218/262; 19/30
	$K8$	0.9176	170/17	92/13	16/262; 0/30	112/262; 0/30
	$K20$	0.9569	262/30	0/0	74/262; 0/30	240/262; 4/30
	$K40$	0.9660	262/30	0/0	155/262; 0/30	245/262; 5/30
truck	RAW	0.9152	225/26	0/0	0/225; 0/26	2/225; 0/26
	$K8$	0.9982	143/19	82/7	133/225; 0/26	139/225; 0/26
	$K20$	0.9995	225/26	0/0	212/225; 0/26	221/225; 0/26
	$K40$	1.0000	225/26	0/0	219/225; 0/26	222/225; 0/26
playroom	RAW	0.9494	202/23	0/0	191/202; 8/23	202/202; 12/23
	$K8$	1.0000	139/11	63/12	135/202; 0/23	136/202; 0/23
	$K20$	1.0000	202/23	0/0	199/202; 0/23	201/202; 0/23
	$K40$	1.0000	202/23	0/0	200/202; 0/23	202/202; 0/23

At $K8$, 356 queries lack OUT support, giving 64.68% coverage; its AUC describes that supported population. Both $K20$ and $K40$ score every query. Their equal-scene AUCs are 0.987980/0.990350, and the inherited 1% cutoffs detect 665/762 of 905 members with zero of 103 false positives. Bonsai supplies 81 of the 97 extra detections; counter's AUC decreases slightly. At the 5% cutoffs, detections are 869/873 and false positives 4/5, all in bonsai.

The secondary balanced subsets have 24/30/26/23 queries per class. Counter and bonsai pass the pose battery, while truck and playroom fail L4; only bonsai also meets the 30-per-class floor. All outcomes remain available without filtering the primary pool. For $K40$, their AUCs are 0.9965/0.9867/1.0000/1.0000 and their 1% TP/FP counts are 22/0, 17/0, 25/0 and 23/0. Full support, pose and retrospective operating-point records are retained in Artifact record AA. One false positive among 23–30 scene nonmembers changes FPR by 3.33–4.35 percentage points, so the observed counts are more informative than treating the inherited nominal label as a population guarantee.

C.4 Training endpoint and spherical harmonics

Counter, bonsai and truck compare 30k/SH3 with stopping at 7k or reducing SH to degree 1. Target-only RR, the PSNR membership margin and held-out PSNR/SSIM are measured on the same evaluation populations (Table 20). Early stopping lowers the margin and the readout, with held-out costs of 0.7–2.1 dB PSNR and 0.017–0.020 SSIM. Lower SH has a smaller effect, particularly on truck.

The LiRA comparison uses 7k targets with 7k shadows and recomputes the 30k baseline through the same path. On counter+truck, pooled nominal-1% TPR changes from 0.969 to 0.613 and AUC from 0.997 to 0.970. Bonsai's weak 1% tail is read alongside its declining AUC and 5% result. Residual leakage remains substantial in these early-stopped assets.

scene	training knob	RR AUC	marg. dB	uPSNR	uSSIM
counter	none (30k/sh3)	0.961	4.12	28.35	0.909
counter	early-stop 7k	0.824	1.95	27.08	0.889
counter	lower-SH (sh1)	0.919	3.34	27.65	0.899
bonsai	none (30k/sh3)	0.877	3.34	31.66	0.951
bonsai	early-stop 7k	0.780	2.04	29.57	0.934
bonsai	lower-SH (sh1)	0.844	2.95	30.61	0.945
truck	none (30k/sh3)	0.988	4.72	24.17	0.864
truck	early-stop 7k	0.932	2.18	23.45	0.844
truck	lower-SH (sh1)	0.985	4.28	23.97	0.861

Table 20: Early stopping and reduced SH degree under target-only RR_logreg. Membership readouts are AUC and the member-minus-nonmember PSNR margin in dB; held-out nonmember PSNR/SSIM measure utility by rerendering the retained models. Early stopping lowers AUC and roughly halves the margin, with lower reconstruction quality; reducing SH degree has smaller effects. SSIM uses the canonical 3DGS Gaussian window; LPIPS is not measured. Table 21 evaluates early stopping with matched LiRA references.

scene	AUC (30k→7k)	@1% (30k→7k)	@5% (30k→7k)
counter (primary)	0.991 → 0.932	0.688 → 0.294	0.982 → 0.624
truck (primary)	1.000 → 0.994	0.991 → 0.858	1.000 → 0.982
bonsai†	0.936 → 0.815	0.031 → 0.038	0.695 → 0.260
pooled (counter+truck)	0.997 → 0.970	0.969 → 0.613	0.996 → 0.838

Table 21: Early stopping at 7k versus 30k under global-$K8$ two-sided LiRA. Targets and shadows use the same stopping point, with $K{=}8$ seeds 200–207, and both columns use the same computation path. The @1%/@5% columns give TPR under the same retrospective empirical-quantile convention. Early stopping reduces counter+truck TPR while retaining pooled @1% TPR of 0.613 and truck @5% TPR of approximately 0.982. † Bonsai's baseline @1% TPR is already 0.031, so its AUC and @5% result provide the more informative attenuation comparisons. Reduced SH degree was not evaluated with LiRA.

C.5 SPZ ENCODING AND MATCHED REFERENCE RELEASES

The earlier reference-access panel supplies 81 native 30k assets, one target and eight MATCHED models in each of nine scenes. Native/native (NN), encoded/encoded (CC) and encoded/native (CN) comparisons preserve the trained models, four-IN/four-OUT assignments and cameras. CN is evaluated with both the native C cutoff and its own recalibrated cutoff, using identical scores in the two cases. The preplanned codec follow-up occurs after native T results are known; its C cutoffs precede T codec scoring and joint label association.

SPZ v3.0.0 (Niantic Labs, 2026) uses the settings in Section A, with antialiasing disabled, native SfM scale and unchanged Gaussian counts. The 78,095,768 Gaussians occupy 19,367,874,520 native bytes and 1,865,398,028 encoded bytes, a 90.3686% reduction. The 10.3827-fold ratio uses summed sizes. T contains 316 candidates per class; the pose-matched subset contains 310 per class, and every condition has complete coverage.

Table 22: SPZ ranking across all nine scenes. Each AUC cell gives full H / pose-matched subset. n_H and n_Q are counts per class. NN uses native target/native references; CC encodes both; CN encodes the target and retains native references. The T mean weights the four within-scene AUCs equally. All conditions have full coverage and zero abstention.

Scene	Role	n_H	n_Q	NN: H / Q	CC: H / Q	CN: H / Q
counter	D	48	44	0.9913 / 0.9897	0.9878 / 0.9861	0.9831 / 0.9804
bonsai	D	58	56	0.9174 / 0.9238	0.8653 / 0.8721	0.8977 / 0.9082
C1	C	82	80	0.9966 / 0.9966	0.9954 / 0.9953	0.9951 / 0.9953
C2	C	82	82	0.9833 / 0.9833	0.9772 / 0.9772	0.9429 / 0.9429
C3	C	82	79	0.9578 / 0.9545	0.9512 / 0.9474	0.9392 / 0.9345
T1	T	82	81	0.9393 / 0.9403	0.9304 / 0.9305	0.9198 / 0.9288
T2	T	80	80	0.8994 / 0.8994	0.8416 / 0.8416	0.7391 / 0.7391
T3	T	74	72	0.8915 / 0.8883	0.8996 / 0.8968	0.8976 / 0.8945
T4	T	80	77	0.8817 / 0.9003	0.8664 / 0.8826	0.8500 / 0.8644
T mean				0.9030 / 0.9071	0.8845 / 0.8879	0.8516 / 0.8567

Table 23: Fixed decisions at the 1% C calibration level on all four T populations. Each cell is TP/N_+; FP/N_-. CN stress and recalibration use the same scores with different C cutoffs. Every query remains in the denominator.

Scene	NN, native C	CC, own C	CN, native C	CN, own C
T1	31/82; 2/82	29/82; 1/82	8/82; 1/82	32/82; 3/82
T2	24/80; 4/80	16/80; 3/80	0/80; 0/80	4/80; 0/80
T3	26/74; 4/74	24/74; 4/74	28/74; 6/74	68/74; 20/74
T4	24/80; 1/80	23/80; 2/80	6/80; 0/80	21/80; 1/80
Total	105/316; 11/316	92/316; 10/316	42/316; 7/316	125/316; 24/316

Mean T AUC is 0.902986 for NN, 0.884482 for CC and 0.851614 for CN. Encoding the references improves ranking on every encoded T target. Development bonsai has the opposite CC/CN ordering. At the 1% C level, CC detects 92/316 with 10 false positives. CN's own-C recalibration raises detection from 42 to 125 but increases false positives from 7 to 24, including 20/74 in T3. Its 5% T3 count is 31/74. Matching the release representation therefore improves ranking without removing calibration-transfer sensitivity.

Paired utility uses the complete registered training complement, comprising 1,679 views across nine scenes and 796 in T. Equal-scene T changes are -0.801235 dB PSNR, -0.008846 SSIM and $+0.009231$ LPIPS. T2 loses 2.144335 dB; T3 gains 0.041404 dB while losing SSIM and LPIPS quality. All nine scenes lose SSIM and LPIPS quality. The full per-scene utility, storage, cutoff and 5% matrices remain in Artifact record AB.

Paired release states in W2. The newer panel crosses 7k/30k with native/SPZ on every scene. Table 24 keeps all nine scalar ranking families; anchor centering leaves their within-scene AUC unchanged. Fixed GAP reaches 0.7734/0.7523 at 7k native/SPZ and 0.9384/0.9246 at 30k. Recipe-known GAP recovers part of the early mismatch at 0.8435/0.8355. This paired comparison retains the original mixed-primary C cutoffs. Its fixed-reference definition differs from the uniform-state metric analysis in Section E.

Table 24: All nine scalar ranking families across the four paired target releases on the eight W2 test scenes, at supplied poses. Every cell is equal-scene mean AUC on the same 610 members and 610 nonmembers. Groups identify reference construction; recipe-known methods additionally use the target recipe. Anchor-centered variants have the same within-scene AUC as their uncentered family; their different decisions remain in the full records. Full nPAUC and 1%/5% decisions are in Artifact record AD.

Readout	7k/native	7k/SPZ	30k/native	30k/SPZ
No reference 3DGS bank				
RAW	0.7585	0.7460	0.9046	0.8999
Background-only references				
Fixed GAP	0.7734	0.7523	0.9384	0.9246
Fixed T	0.7560	0.7307	0.9371	0.9226
Recipe-known GAP	0.8435	0.8355	0.9384	0.9343
Recipe-known T	0.8443	0.8388	0.9371	0.9343
Mixture	0.8216	0.8042	0.9419	0.9352
Candidate-containing reference construction				
BaVarIA-n	0.8889	0.8833	0.9828	0.9816
Matched GAP	0.8655	0.8585	0.9532	0.9490
Matched T	0.8778	0.8724	0.9650	0.9650

D Reference access, calibration and photometric alignment

Reference normalization separates candidate difficulty from the asset's fidelity trace. Its effectiveness depends on which images and training states the references can represent. The studies below distinguish score ranking from decisions made with a previously selected cutoff. They cover independent targets, background-only access, release knowledge and photometric processing under each study's own calibration population.

D.1 Prospective estimator selection and independent targets

The earlier supplied-pose study uses counter, bonsai and truck for development, with four development targets, eight reference models, eight calibration targets and eight evaluation targets per scene. Drjohnson has eight models in each of the latter three roles and contributes no estimator-selection labels. The 108 trained models comprise 76 development/reference/calibration models and 32 evaluation targets. They use auxiliary-image-only initialization, the full released-pose setting and 30k direct 3DGS. Each target's fixed roster has 48/58/50/53 queries per class on counter/bonsai/truck/drjohnson, giving 384/464/400/424 per class over its eight evaluation targets.

The selection statistic J averages normalized partial AUC over FPR $[0, 0.05]$, first over the four development targets and then equally over the three scenes. All four candidate estimators appear in Table 25. BaVarIA-n is the sole method within the prespecified 0.005 band around the best J; the complexity tie-break is unused. The development-label permutation test repeats selection in each of 9,999 randomizations, giving the selection-adjusted $(0 + 1)/(9999 + 1) = 10^{-4}$ result. This test concerns the development association.

estimator	J	worst-scene $\text{pAUC}_{.05}$	within 0.005 of best
global-σ LiRA	0.786402	0.401308	no
per-query LiRA	0.747769	**0.619798**	no
BaVarIA-n	**0.799323**	0.444411	yes
BaVarIA-t	0.755266	0.376040	no

Table 25: Complete prospective estimator selection. BaVarIA-n (Brännvall, 2026) is the only estimator within the prespecified 0.005 equivalence band around the best J; the complexity tie-break is therefore unused. A higher worst-scene value alone does not override the primary mean-pAUC statistic.

Each development scene uses its separate calibration targets to select a scene-specific cutoff. Drjohnson uses the pooled development-scene cutoff. The historical rule takes descending rank $\lfloor \alpha n_C^- \rfloor + 1$

and predicts with strict exceedance. At 1%, the counter/bonsai/truck pools contain 384/464/400 negatives, ranks 4/5/5, and cutoffs 0.433055/0.524370/-0.405600. The pooled development pool has 1,248 negatives, rank 13 and cutoff 0.433055. Full-precision cutoffs, 5% rules and the held-out-scene/five-fold candidate-excluded sensitivities remain in Artifact record Q.

scene	macro AUC	macro $\text{pAUC}_{.05}$	1% TP/FP	5% TP/FP	1% boundary	verdict
counter	0.990560	0.880642	337/6	376/26	TP$\geq$ 192, FP$\leq$ 3	FAIL
bonsai	0.925312	0.525490	169/10	327/18	TP$\geq$ 232, FP$\leq$ 4	FAIL
truck	0.999950	0.999000	400/10	400/24	TP$\geq$ 200, FP$\leq$ 4	FAIL
drjohnson	0.999956	0.999110	422/0	423/0	TP$\geq$ 212, FP$\leq$ 4	PASS

Table 26: Aggregate evaluation-target results. pAUC is normalized over FPR $[0, 0.05]$ and macro-averaged over targets; 5% results are secondary. Drjohnson meets the held-out boundary, while the three development scenes do not establish reliable threshold transfer.

The ranking is strong on counter, truck and the held-out scene, but three development-scene decisions fail their joint requirement of at least 50% member detection and at most $\lfloor 0.01N_{-} \rfloor$ false positives. Bonsai also misses the detection floor. Drjohnson meets both requirements with 422 detections and no false positives. Across the four evaluation scenes, the primary decisions total 1,328 TP and 26 FP; the development-scene subtotal is 906/26.

The six-scene extension in Section C.1 transports the same pooled development cutoff to independent targets and optimizer seeds. Table 15 preserves the complete cell-count outcomes. Five scenes meet the historical joint criterion; train exceeds its four-FP budget by one, and kitchen reaches its budget. Targets sharing an assignment are paired optimizer repeats, not independent membership assignments, and these scene-level criteria do not turn correlated query images into independent null trials.

D.2 Candidate-containing, auxiliary and background-only references

The earlier reference-access panel has counter/bonsai development scenes and a distinct three-C/four-T DL3DV panel. Calibration/test scenes are the first metadata entries satisfying the fixed 150–240-second duration rule, selected before fidelity inspection. Their registered-view counts are 410, 412, 411, 410, 404, 373 and 403. A seeded permutation assigns $2\lfloor N/5 \rfloor$ views to the candidate pool H and the rest to background B. A target trains on half of H and $\lfloor |B|/2 \rfloor$ background views, totaling $\lfloor N/2 \rfloor$.

- **MATCHED** uses eight independent assignments with four IN and four OUT observations for every candidate and the target's training-set size.
- **AUX50** fixes half of B before outcomes and combines it with the same candidate assignments. It preserves the per-model training count while granting 69.71–70.00% aggregate RGB access across the bank.
- **PRETRAINED-OUT** trains eight models solely on B, with the target's image count, before candidate RGB is introduced. Its query-local Student-t score uses the target and the current candidate's reference scores. The $K4$ condition takes the fixed first four models.
- **EARLY-REF** and **AUX50-EARLY-REF** use the corresponding 7k bank against 30k targets. **MATCHED-OUT-ONLY** takes the four OUT models from MATCHED and replaces its complete estimator with the query-local tail score.

All models use SH3 and permitted-image-only initialization; released camera poses remain fixed upstream inputs. BaVarIA-n in the first, auxiliary and early conditions estimates scene priors from the complete reference-scored H pool. The OUT-tail conditions do not require other candidates' reference scores. Consequently, MATCHED-OUT-ONLY changes reference provenance, prior use and the estimator together; it does not isolate removal of the IN likelihood.

Each C scene contributes 82 negatives, yielding 246 and descending ranks 2/12 at the 1%/5% levels. Cutoffs are fixed before T fidelity, and all T predictions are saved before joint label association. T has 82/80/74/80 candidates per class, or 632 queries in total; all eight conditions have complete

coverage. A separately recorded pose-matched subset has 81/80/72/77 per class and does not filter this primary population.

	T1 (82/class)		T2 (80/class)		T3 (74/class)		T4 (80/class)	
condition	AUC	TP/FP	AUC	TP/FP	AUC	TP/FP	AUC	TP/FP
Raw PSNR	0.7426	5/0	0.7355	72/56	0.8271	0/0	0.7555	7/1
Matched	0.9393	31/2	0.8994	24/4	0.8915	26/4	0.8817	24/1
Early ref.	0.4365	4/4	0.5105	0/0	0.3475	0/0	0.6409	12/3
Aux50	0.9142	15/2	0.8730	15/1	0.9131	17/1	0.8588	17/1
Aux50 + early	0.6457	5/0	0.5489	0/0	0.4973	0/0	0.6300	5/1
Pretrained OUT-$K8$	0.8682	1/0	0.8648	0/0	0.8716	14/2	0.8177	4/0
Pretrained OUT-$K4$	0.8514	1/3	0.8470	3/1	0.9001	6/1	0.8078	1/0
Matched OUT-only	0.9148	18/1	0.8870	15/1	0.9447	32/0	0.8681	10/1

Table 27: All eight reference conditions on the complete H candidate pools. Counts use each condition's own C-derived 1% cutoff. Scores retain their member-oriented direction, including the below-chance EARLY rows. Coverage is 100% and abstention is zero in every class and condition.

Condition	Mean T AUC	Mean T nPAUC$_{.05}$
Raw PSNR	0.7651	0.1686
Matched	0.9030	0.2953
Early ref.	0.4839	0.0684
Aux50	0.8897	0.2514
Aux50 + early	0.5805	0.1457
Pretrained OUT-$K8$	0.8556	0.1622
Pretrained OUT-$K4$	0.8516	0.0978
Matched OUT-only	0.9037	0.3838

Table 28: Reference-condition ranking on the four complete T scenes, with equal-scene weighting. nPAUC integrates ROC area over FPR 0–5% and divides by 0.05. From 246 C nonmembers, the 1% and 5% cutoffs are the second- and twelfth-largest C scores, respectively, with strict $s > \tau$. Full-precision cutoffs, each in its own score scale, are retained in the Artifact.

MATCHED improves each test scene over RAW; AUX50 retains much of that ranking at reduced candidate-containing access. PRETRAINED-OUT ranks above RAW in all four scenes but detects only 19 of 316 members at its primary cutoff. EARLY-REF loses the ranking on T1 and T3, with mean AUC 0.4839, while T4 remains above chance. Early references therefore create a substantive mismatch despite using the same rendering statistic and candidate poses.

The OUT-only calibration tail is set by C3: its four largest scores have target-minus-reference gaps of 3.68–5.29 dB, at the 94th–100th within-scene percentiles, whereas their 0.47–0.80 dB predictive scales lie at the 29th–79th percentiles. Large gaps, rather than uniformly tiny variances, characterize these extremes. The underlying training cause is unresolved. Transferring MATCHED-OUT-ONLY's cutoff 3.3135 to PRETRAINED-OUT increases FP from 2 to 44. The $K8/K4$ comparison also changes the Student-t degrees of freedom from 7 to 3, so its tail behavior is not an isolated reference-provenance experiment. Full 5% results and both cutoff-transfer stress tests are retained in record Y.

D.3 Why early references can reverse the likelihood readout

The likelihood-response analysis evaluates RAW, OUT-GAP, standardized OUT residual, predictive OUT-tail, BaVarIA-n, a tied-variance likelihood ratio and hinge-IN. Tied variance averages the two BaVarIA posterior variances while retaining the query means; hinge-IN uses the original per-query Gaussian implementation and its fewer-than-three-observations fallback. These alternatives diagnose how reference mismatch changes the score.

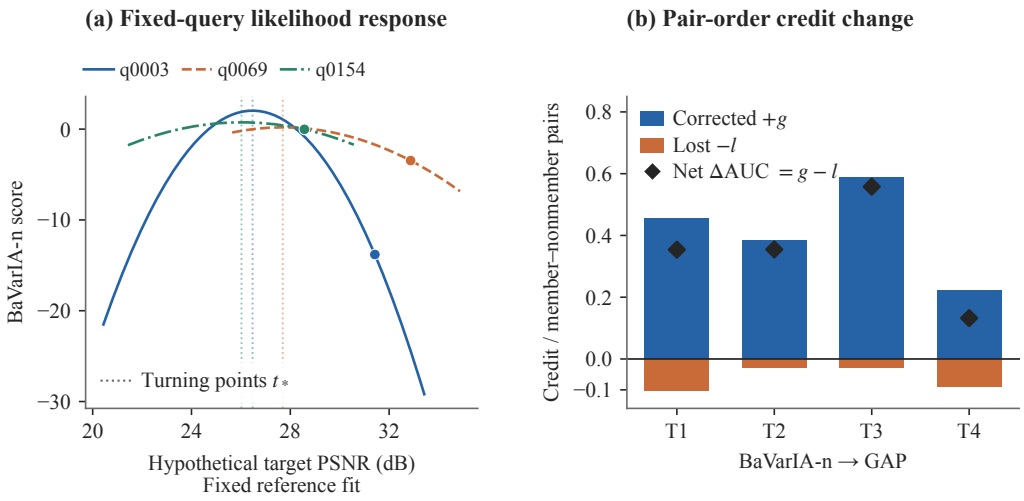


Figure 4: Early-reference likelihood extrapolation. Left: three positive T1 queries selected evenly in query order from the negative-slope subset; only target PSNR varies, with each query's fitted references held fixed. Dots mark the observed target score; t_* marks each turning point. Right: normalized corrected credit g and lost credit l when replacing BaVarIA-n with GAP, on common finite support in all four earlier T scenes. Bars show $+g$ and $-l$; diamonds show $\Delta\mathrm{AUC} = g - l$. Ties contribute half credit.

For a fixed query, hold its fitted IN/OUT means μ_1, μ_0 and variances v_1, v_0 fixed. The Gaussian log-likelihood ratio has derivative

$$s'(t) = \frac{t - \mu_0}{v_0} - \frac{t - \mu_1}{v_1}.$$

When $v_1 < v_0$, it reaches a maximum at $t_* = (\mu_1 v_0 - \mu_0 v_1)/(v_0 - v_1)$ and decreases for $t > t_*$. Exceeding both means need not cross this turning point. Cross-query ranking also depends on each query's fitted parameters.

For the earlier EARLY-REF bank, 596 of 632 target values exceed both reference means, including all 316 members. The LLR's local slope is negative at 299 observed queries, compared with 35 under MATCHED, where 172 values exceed both means and 31 lie beyond an upper turning point. The 294 observations beyond an upper turning point also have lower IN than OUT predictive variance. Figure 4 displays the response and all four scenes' pair-order accounting. GAP improves EARLY-REF mean AUC from 0.483855 to 0.833574, with more corrected than lost member/nonmember pair credits in every scene.

Improved ranking does not itself transport a cutoff. The original C-calibrated EARLY-REF GAP rule detects only 2/316 members with 1/316 false positives. Its 8.460445 dB cutoff is far above T3's 1.141831 dB positive median: T3 has AUC 0.905405 and zero detections. Figure 5 makes the calibration/test location mismatch visible. Under MATCHED references, mean AUC is 0.9030/0.8580/0.9037/0.9209 for BaVarIA-n/GAP/OUT-tail/hinge-IN, respectively; GAP is not the strongest ranking score in that condition.

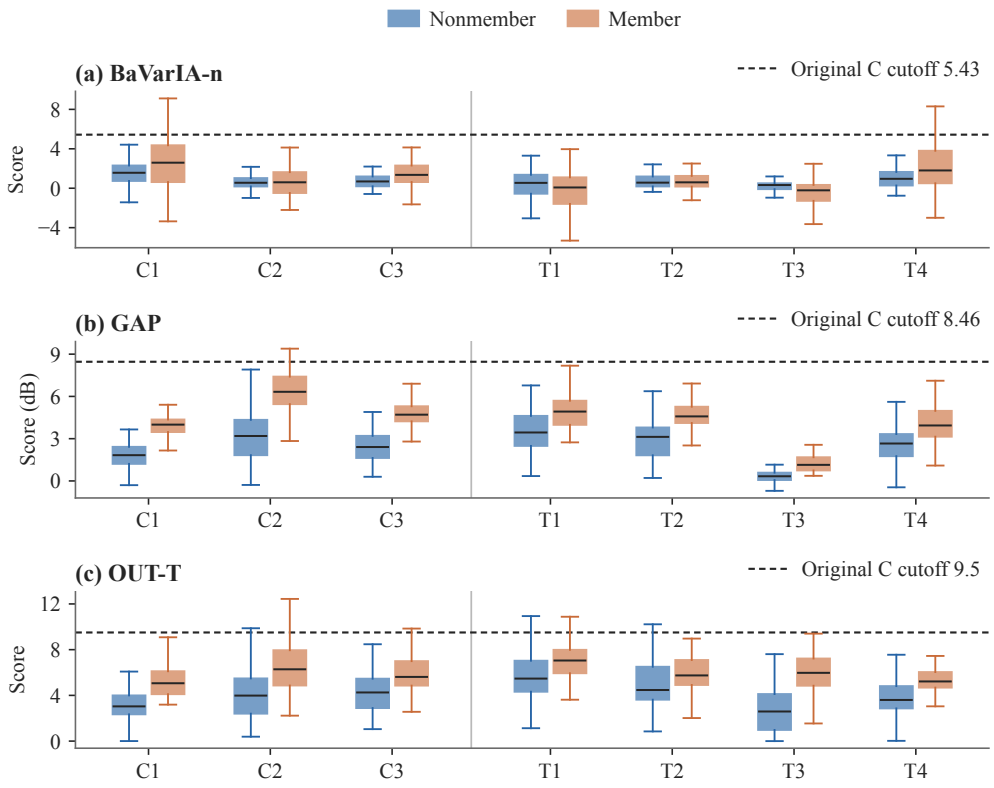


Figure 5: The same earlier EARLY-REF panel under BaVarIA-n, GAP and OUT-T. Each row retains its own score scale and original C-derived cutoff; boxes use all finite scores in the relevant C or T scene. Boxplots omit outlier markers for readability; cutoff estimation includes every finite calibration negative. GAP restores useful ordering while the calibration tail still limits detection.

Two further checks preserve this dependence on regime. In the original ten-scene leave-one-scene-out analysis, BaVarIA-n yields AUC 0.989639 and 997 TP/18 FP, versus GAP's 0.976771 and 662 TP/10 FP, each out of 1,054 per class. At 90% training coverage with only eight references, BaVarIA-n/GAP share support on 594/905 positives and 58/103 negatives, whereas standardized and predictive-tail scores need two OUT observations and cover only 157/905 and 13/103. The full reference-budget study in Section C.3 resolves the support deficit at $K40$, where BaVarIA-n gives 762 TP and zero FP at the inherited 1% cutoff.

D.4 Release knowledge and anchor adaptation in W2

The W2 four-C/eight-T access study uses the roles and primary release states in Section A.1. Every scalar family has its own C rule from 304 negatives, with descending ranks 3/15. Across scalar, pose and CNN configurations there are 66 saved rules. Table 29 retains all 15 supplied-pose scalar configurations, including candidate-containing references; Section F reports the paired estimated-pose decisions.

Table 29: Complete supplied-pose scalar comparison on the eight W2 test scenes. AUC and nPAUC average scenes equally; Min. is the lowest scene AUC. Decision cells are TP; FP, out of 610 per class, under separately fitted 1%/5% C rules. Groups distinguish reference-model and known-nonmember-anchor access; recipe-known methods additionally use the target recipe. All methods score the full primary population.

Readout	AUC	Min.	nPAUC .01	nPAUC .05	1% C TP; FP	5% C TP; FP
No reference 3DGS bank						
RAW	0.8320	0.6389	0.2748	0.3559	303; 145	360; 248
Background-only references						
Fixed GAP	0.8529	0.6084	0.3536	0.4356	117; 6	206; 19
Fixed T	0.8404	0.6054	0.2780	0.3721	66; 1	215; 13
Recipe-known GAP	0.8937	0.7324	0.3701	0.4570	126; 3	255; 18
Recipe-known T	0.8879	0.7193	0.2347	0.3602	67; 1	252; 18
Mixture	0.8730	0.6819	0.2664	0.3734	71; 2	245; 17
No reference 3DGS bank + known-nonmember anchors						
RAW–R8	0.8320	0.6389	0.2748	0.3559	3; 0	49; 0
RAW–R16	0.8320	0.6389	0.2748	0.3559	3; 0	30; 0
RAW–R32	0.8320	0.6389	0.2748	0.3559	0; 0	11; 0
Background-only references + known-nonmember anchors						
GAP–R8	0.8529	0.6084	0.3536	0.4356	103; 5	240; 32
GAP–R16	0.8529	0.6084	0.3536	0.4356	84; 4	221; 25
GAP–R32	0.8529	0.6084	0.3536	0.4356	137; 8	244; 29
Candidate-containing reference construction						
BaVarIA-n	0.9273	0.7573	0.5817	0.6414	383; 11	485; 40
Matched GAP	0.9108	0.7715	0.3102	0.4653	224; 11	260; 21
Matched T	0.9159	0.7652	0.3522	0.5078	260; 7	388; 26

Fixed GAP uses eight 30k/native background-only models throughout; recipe-known GAP/T matches the declared training/codec state, while the mixture averages four predictive tails with equal weights. These forms adapt the expected reference response or uncertainty without selecting a new cutoff from T membership labels. Recipe-known GAP raises mean AUC from 0.8529 to 0.8937 and changes primary TP/FP from 117/6 to 126/3. Fixed T detects fewer members with fewer false positives (66/1). BaVarIA-n reaches 0.9273 AUC with 383/11, using the stronger candidate-containing reference access. No single method leads every ranking and decision comparison.

R32 subtracts the finite anchor-score median separately for each asset; R16/R8 use the fixed first 16/8 anchors. The minimum finite counts are 16/8/4, respectively; all measured assets have all 32. Centering changes the transferred decision without changing within-scene AUC. It is a location adjustment: 32 anchors cannot supply a finite 1% rank under $j = \lfloor 0.01(32+1) \rfloor = 0$. The rank-zero rule would return $+\infty$.

RAW-R32 makes no primary 1% detection or false positive in either pose setting. C3 has the three largest centered negative values, 12.505294, 11.422240 and 9.359395 dB; the largest centered T-positive value is only 9.120474 dB. That C scene's candidate-negative median is 3.174625 dB above its anchor center. Thus, anchors do not remove this candidate-versus-anchor tail mismatch. GAP-R32 changes TP/FP from 117/6 to 137/8, with R16/R8 giving 84/4 and 103/5; the budget response is not monotone. All anchor values and per-scene decisions remain in record AD.

D.5 Scene-held-out calibration on synchronized ETH3D capture

Six ETH3D scenes (Schöps et al., 2017) are assigned in their pre-existing order to development, calibration and test pairs before training or score inspection. Membership is at the rig-frame level: a canonical-camera query is a member when all four synchronized images of its frame enter training. Each scene has one `random50_seed0` target and eight seeds 200–207, trained for 30k with direct 3DGS, `member_retri` initialization, the MILP-matched rosters and parity-fixed rendering. The 54 models yield 12,744 exact-view renders and 1,416 query scores.

BaVarIA-n retains each query's actual 1–7 observations per reference class. Each C scene contributes 116 negatives; centered systematic positions in its fixed lexicographic order remove 14 of electro's 130 and none of forest's 116. The pooled 232 negatives give ranks 3/12, cutoffs 0.393653518975009/-0.412270247368713, and strict exceedance. All 460 T predictions precede the joint label reveal.

Table 30: ETH3D separation across the D/C/T roles. AUC is tie-aware, and $\Delta\bar{s}$ is the member-minus-nonmember mean score gap. All six scenes show positive separation with AUC between 0.9866 and 1.0000.

role	scene	N_+	N_-	AUC	$\Delta\bar{s}$
D	delivery_area	107	107	0.986636	17.712537
D	sand_box	125	125	0.998272	18.050996
C	electro	130	130	0.999231	21.512546
C	forest	116	116	0.994872	4.391864
T	playground	109	109	1.000000	33.584112
T	lakeside	121	121	0.993716	15.952918

Table 31: Global C thresholds and sealed T outcomes. Ranks are descending among 232 balanced C nonmembers. Prediction uses strict $s > \tau_C$, with the single boundary tie at each level classified as nonmember. The 1% row is primary and the 5% row is supporting.

				playground		lakeside	
level	C rank	τ_C	C FP/N_C^-	TP/FP	TPR/FPR	TP/FP	TPR/FPR
1%	3	0.393654	2/232	109/0	1.000/0.000	113/2	0.934/0.017
5%	12	-0.412270	11/232	109/0	1.000/0.000	120/10	0.992/0.083

Mean test AUC is 0.996858 with complete coverage. Playground identifies all 109 members with no false positives at the primary cutoff. Lakeside identifies 113/121, exceeding its 61-TP requirement by 52, but its two false positives exceed the one-FP allowance by one. At 5%, lakeside has 120 detections and ten false positives. This study therefore combines strong scene-held-out separability with a narrow primary tail-calibration miss on one test scene.

D.6 Photometric alignment across calibration, ranking and localization (W3)

Twelve further DL3DV scenes supply eight calibration scenes and four evaluation scenes. Each scene has one target and eight background-only OUT models, with model initialization and training restricted to their allowed RGB. The target retains its 7k and 30k native checkpoints and their SPZ encodings; all references remain at 30k/native. The primary comparison fixes every target to 30k/native. Four single-state panels and a separate mixed-state bridge preserve the remaining declared release comparisons. The bridge cycles 7k/native, 7k/SPZ, 30k/native and 30k/SPZ across

scene identifiers, with two C scenes and one T scene per recipe. These snapshots are paired states of the same models.

The C population has 1,230 candidates, including 615 nonmembers. The four T scenes contribute 79/75/78/75 candidates per class, totaling 307 members and 307 nonmembers. All candidates remain in the scoring and decision rosters. The gain definition is fixed before C scoring, each C cutoff is saved before T scoring, and T predictions precede joint membership-label association. W3 estimates its own rules and effects; the W2 population and mixed primary recipes remain the separate access experiment.

D.6.1 Reference-only photometric alignment

Shared gain alignment adjusts target and reference renders to the candidate's numerical RGB scale before comparing their fidelity. Let X_t and $X_1, \dots, X_K$ denote the renderer's displayed outputs, already clamped to $[0, 1]$, at one scoring camera, with $K = 8$ and $\mu = K^{-1} \sum_k X_k$. The fit uses the candidate I and reference mean alone. Over RGB channel elements, define

$$\mathcal{M}(I, \mu) = \{u : 0.05 < \mu_u < 0.95,\ 0.02 < I_u < 0.98\},$$
$$a_{\text{raw}}(I) = \frac{\sum_{u \in \mathcal{M}} \mu_u I_u}{\sum_{u \in \mathcal{M}} \mu_u^2}, \qquad a(I) = \text{clip}(a_{\text{raw}}(I), 0.5, 2).$$

If fewer than 1% of all channel elements are valid or the denominator is zero, the retained fallback is $a = 1$. The fit accumulates sums in float64 over float32 pixels. One shared scalar corrects the target and all references, yielding

$$s_{\text{gain}}(I) = \text{PSNR}(I, \text{clip}(aX_t, 0, 1)) - \frac{1}{K} \sum_{k=1}^{K} \text{PSNR}(I, \text{clip}(aX_k, 0, 1)) .$$

PSNR uses full-canvas float32 MSE and the 10^{-10} floor from Section A.3. The target render does not enter the gain fit. The transformation of a candidate is $I^{(\text{EV})} = \text{clip}(2^{\text{EV}} I, 0, 1)$ in numerical RGB, for $\text{EV} = 0, \pm 0.25$. Exactly these transformed float32 pixels feed localization and scoring. In the estimated-pose branch, each condition independently extracts query features, matches the fixed B map and solves PnP.

Why the shared scalar cancels a common scale change. Fix the scoring camera, displayed renders X_j and mean μ. For $g > 0$, compare I with gI under an unchanged fit mask, no fallback, inactive gain bounds, no clipping of the transformed candidate or corrected renders, and an inactive MSE floor. The least-squares numerator scales by g while its denominator stays fixed, so

$$a(gI) = g\, a(I), \qquad \text{MSE}(gI, g\, a(I) X_j) = g^2 \,\text{MSE}(I, a(I) X_j).$$

Every target and reference PSNR therefore shifts by the same $-20 \log_{10} g$, which cancels in their difference. This calculation starts from the fixed displayed renders and requires no recovery of unclamped renderer values. Mask changes, clipping and relocalization can change the relation in the implemented pipeline; W3 measures the complete path under those operations.

Separating score alignment from calibration processing. Branch A uses original GAP with its clean-C cutoff. Branch B keeps exactly the same query GAP scores but uses C processed under the matching numerical RGB transformation. Branch C uses the shared-gain score with its own clean-C cutoff. Within each pose/target panel, C reuses that cutoff across all three candidate conditions. A, B and C have separately estimated numerical cutoffs, and A/B have identical rankings. A clean RAW control is also retained with its own C rule.

D.6.2 Calibration across eight independent scenes

The primary all8 pool contains 615 negative candidates from eight new C scenes, with 76/77/77/76/77/80/77/75 negatives in identifier order. At levels $\alpha = 0.01/0.05$, the strict order-statistic rule uses descending ranks 6/30, giving at most 5/29 C exceedances in the absence of ties. Rules are separate for the two pose modes, the four individual target states and the mixed-state bridge. All required C negative scores are finite. The primary target and reference state is 30k/native.

Eleven pools were fixed before T outcomes: all8, the first four C scenes, the last four C scenes, and eight leave-one-C-scene-out (LOCO) pools. The two halves contain 306 and 309 negatives; each uses ranks 3/15 at the two levels. Their difference reflects both scene composition and negative support. Every score still uses the same eight OUT reference models per scene, so changing the C pool does not change the reference budget. Table 32 reports the primary rule alongside the ranges from the frozen alternatives.

Table 32: Frozen C-pool sensitivity at 30k native and $\alpha = .01$. All8 is the primary rule; LOCO varies eight leave-one-C-scene-out pools, and all 11 adds first4 and last4. Ranges describe saved alternative C calibrations on the same 307/307 T roster; no pool is selected using T. All8 has 615 scored C negatives; exact cutoffs accompany the Artifact.

Condition	Readout	All8 TP/FP	LOCO TP	LOCO FP	All 11 TP	All 11 FP
Supplied poses						
Clean	B: matched	162/7	162–162	7–8	162–162	7–8
Clean	C: gain	153/7	152–164	7–8	147–164	6–8
-0.25 EV	B: matched	79/9	75–85	6–11	70–85	6–11
-0.25 EV	C: gain	148/6	148–160	6–8	143–160	5–8
$+0.25$ EV	B: matched	80/11	79–92	11–15	61–107	6–16
$+0.25$ EV	C: gain	153/7	149–166	6–9	148–166	6–9
Relocalized poses						
Clean	B: matched	121/8	121–121	8–8	121–121	8–8
Clean	C: gain	114/5	106–127	4–6	105–127	4–6
-0.25 EV	B: matched	63/8	56–67	6–9	56–67	6–9
-0.25 EV	C: gain	117/6	110–126	4–7	108–126	4–7
$+0.25$ EV	B: matched	64/15	53–86	11–15	36–88	9–15
$+0.25$ EV	C: gain	120/5	109–125	4–7	107–125	4–7

Shared gain retains useful detections across the fixed C pools, with a transferred operating point that depends on the calibration population. Under relocalization, the primary all8 rule gives 114/5, 117/6 and 120/5 TP/FP for clean, -0.25 EV and $+0.25$ EV, respectively, out of 307 candidates per class. The first-four pool gives fewer detections and false positives: 105/4, 108/4 and 107/4. The last-four pool matches all8 in these three conditions. Table 32 gives the complete supplied-pose primary results and the LOCO and all-pool ranges for both pose modes.

At the primary C level, the six all8 gain conditions realize pooled T FPR between $5/307 = 1.63\%$ and $7/307 = 2.28\%$. Thus the nominal 1% rule and observed T error rates are distinct. The scene and pool ranges are descriptive quantities, without confidence-interval or significance interpretation.

D.6.3 Ranking recovery and clean-candidate tradeoffs

Reference-only shared gain restores GAP ranking under both numerical RGB shifts. Table 33 presents the primary 30k/native target state with eight mature native OUT references and the all8 C pool. It includes both pose branches, all three candidate conditions and all A/B/C comparisons. AUC and normalized partial AUC are averaged equally over the four T scenes, while decision counts pool the 307 members and 307 nonmembers.

Table 33: W3 primary readouts at 30k native, calibrated on all eight C scenes at $\alpha = .01$. AUC and normalized partial AUC are equal-scene means; TP/FP and FPR use the full 307/307 pooled roster. A and B share GAP scores and rankings, using clean and condition-matched calibration, respectively. Shaded C rows use reference-only shared gain with its own clean cutoff, reused across candidate conditions within each pose branch. Both pose branches use the same four T scenes.

		Ranking			Decisions	
Condition	Readout	AUC	nPAUC$_{.01}$	nPAUC$_{.05}$	TP/FP	FPR (%)
Supplied poses						
Clean	A: GAP	0.9413	0.3572	0.5240	162/7	2.28
Clean	B: matched	0.9413	0.3572	0.5240	162/7	2.28
Clean	C: gain	0.9393	0.3451	0.5484	153/7	2.28
-0.25 EV	A: GAP	0.8278	0.1009	0.2427	11/0	0.00
-0.25 EV	B: matched	0.8278	0.1009	0.2427	79/9	2.93
-0.25 EV	C: gain	0.9380	0.3419	0.5605	148/6	1.95
$+0.25$ EV	A: GAP	0.8460	0.1344	0.2245	24/3	0.98
$+0.25$ EV	B: matched	0.8460	0.1344	0.2245	80/11	3.58
$+0.25$ EV	C: gain	0.9423	0.3279	0.5426	153/7	2.28
Relocalized poses						
Clean	A: GAP	0.9217	0.2140	0.3859	121/8	2.61
Clean	B: matched	0.9217	0.2140	0.3859	121/8	2.61
Clean	C: gain	0.9201	0.2481	0.4338	114/5	1.63
-0.25 EV	A: GAP	0.8074	0.1348	0.1919	9/0	0.00
-0.25 EV	B: matched	0.8074	0.1348	0.1919	63/8	2.61
-0.25 EV	C: gain	0.9214	0.2515	0.4601	117/6	1.95
$+0.25$ EV	A: GAP	0.8315	0.0929	0.1634	16/4	1.30
$+0.25$ EV	B: matched	0.8315	0.0929	0.1634	64/15	4.89
$+0.25$ EV	C: gain	0.9236	0.2350	0.4501	120/5	1.63

Under supplied poses, original GAP AUC is 0.8278/0.8460 at $-0.25/+0.25$ EV, versus 0.9380/0.9423 after shared gain. Under relocalization, the corresponding values are 0.8074/0.8315 and 0.9214/0.9236. Every one of the sixteen scene-by-pose-by-exposure AUC differences is positive. Across the four pose/exposure combinations, macro AUC gains range from 0.0920 to 0.1140. This is agreement across paired conditions in four scenes; repeated poses and transformations do not create additional scene replicates.

Table 34: Per-scene W3 ordering and low-FPR tails with relocalized poses at 30k native. GAP denotes the common A/B score. Gain improves exposure AUC in every scene; the full tail outcomes retain, among others, the W3T1 negative-exposure nPAUC decrease. Normalization divides partial ROC area by .01.

Scene	Condition	GAP AUC	Gain AUC	GAP nPAUC	Gain nPAUC	ΔnPAUC
W3T1	Clean	0.9276	0.9258	0.1646	0.1646	0.0000
W3T1	-0.25 EV	0.8806	0.9276	0.3418	0.1772	-0.1646
W3T1	$+0.25$ EV	0.8859	0.9353	0.2025	0.2152	0.0127
W3T2	Clean	0.8777	0.8597	0.0533	0.0533	0.0000
W3T2	-0.25 EV	0.6910	0.8652	0.0133	0.0533	0.0400
W3T2	$+0.25$ EV	0.7390	0.8564	0.0400	0.0533	0.0133
W3T3	Clean	0.9234	0.9400	0.3846	0.3077	-0.0769
W3T3	-0.25 EV	0.7965	0.9359	0.0641	0.2821	0.2179
W3T3	$+0.25$ EV	0.8314	0.9458	0.1026	0.2179	0.1154
W3T4	Clean	0.9580	0.9547	0.2533	0.4667	0.2133
W3T4	-0.25 EV	0.8613	0.9568	0.1200	0.4933	0.3733
W3T4	$+0.25$ EV	0.8699	0.9570	0.0267	0.4533	0.4267

The clean comparison and low-FPR tail identify where alignment changes the readout rather than uniformly improving every metric. On clean candidates, supplied-pose AUC changes from 0.9413 to 0.9393 and TP/FP from 162/7 to 153/7. Relocalized AUC changes from 0.9217 to 0.9201, TP/FP from 121/8 to 114/5, and nPAUC at 0.01 increases from 0.2140 to 0.2481. For W3T1 under relo-

calized -0.25 EV, gain improves AUC by 0.0469 while nPAUC at 0.01 falls by 0.1646. The full per-scene table keeps this tail reversal alongside the exposure gains.

D.6.4 PHOTOMETRIC ALIGNMENT THROUGH LOCALIZATION

W3 repeats localization for clean and both transformed candidates before scoring, using the background-map construction and query registration in Section F.1. The B map is fixed within a scene; condition-specific query features, matches and PnP results are separate. Each of the three conditions returns poses for all 614 candidates. Median rotation error relative to the registered camera is $0.0338°$, $0.0337°$ and $0.0345°$ for clean, -0.25 EV and $+0.25$ EV, respectively. Maximum errors are $2.3400°$, $0.6924°$ and $0.6953°$. The single clean candidate above $1°$ remains in the denominator and score population.

Table 35: W3 relocalization and member identities at 30k native. Rotation and center errors are pooled medians over 614 queries per condition. Rotation is in degrees; each query's center error is first divided by its scene's B-camera RMS radius. A/B/C share each condition's returned pose. Gained/lost compare B to gain among members; retention is the percentage of gain's 114 clean-detected members retained after exposure.

Condition	Returned	Rot. med. (°)	Center med.	Gained	Lost	Retention (%)
Clean	614/614	0.0338	0.000570	13	20	100.00
-0.25 EV	614/614	0.0337	0.000552	68	14	94.74
$+0.25$ EV	614/614	0.0345	0.000577	67	11	93.86

Original GAP and shared gain consume the same camera in each condition. Their paired decisions therefore compare scoring and calibration after identical localization. Relative to condition-matched calibration B, gain detects 68/67 additional members and loses 14/11 B detections at $-0.25/+0.25$ EV, giving net gains of 54/56. False positives fall from 8/15 to 6/5. Gain detects 117/307 and 120/307 members overall, or 38.11% and 39.09%. Of its 114 clean detections, it retains 108 and 107, respectively, giving clean-detection retention of 94.74% and 93.86%. These retention rates use the clean detected set as denominator; they differ from total member recall.

The gain estimates track the imposed RGB scale. At relocalized cameras, their median values are 0.9994, 0.8412 and 1.1809 for clean and the two EV conditions, with p10–p90 ranges $[0.9612, 1.0465]$, $[0.8101, 0.8852]$ and $[1.1302, 1.2375]$. No T fit requires fallback or reaches a gain bound. C has eight fallbacks (five in C5 and three in C8) and no bound hits; the same fit is reused across the four target states.

The pose records report camera-center error normalized by the RMS radius of the B-camera centers, rather than meters. On matched queries, the median rotation between the clean and transformed estimated poses is $0.0092°/0.0115°$. Gain changes the fidelity readout after these cameras have been computed. Its recovery of detections therefore establishes the benefit of photometric alignment within the complete audit, without attributing that benefit to more accurate pose estimation.

The artifact retains all four target states and the mixed-state bridge, RAW controls, nPAUC at 0.05, both calibration levels, all eleven C pools, full-precision cutoffs, and per-scene and per-query scores, poses, gains and decision transitions. These are paired measurements of the same W3 models.

E FIDELITY METRICS AND LEARNED READOUTS

E.1 OBSERVABLE CHOICE AND REFERENCE NORMALIZATION

Fidelity-based membership readout extends beyond PSNR. We separate the choice of image metric from reference normalization on the existing four-C/eight-T panel. M1 compares five RAW readouts at each scene's original assigned primary target state. M2 fixes every target and its eight background-only OUT references to 30k/SPZ, providing complete metric-matched reference scores in a common encoded condition. Both use the original H roster, memberships, supplied cameras and geometric controls. Their settings were specified before additional scoring, and each C cutoff was saved before the new T metric execution. These are additional sensitivity analyses of the already analyzed panel.

The member-oriented observables are PSNR, negative L1, SSIM, negative LPIPS and the derived photometric loss $-[0.8\ell_1+0.2(1-\mathrm{SSIM})]$. For each observable, GAP subtracts the mean of its eight directed OUT values. Each metric/readout has its own cutoff under the same C rule. Sections A.3 and A.4 give the metric implementations, signs and calibration definitions. Measurements use the same full 960×540 float32 render, clamped to $[0, 1]$. All 608 C and 1,220 T queries have finite scores.

Table 36: All fifteen predefined metric configurations on the same eight test scenes, with supplied poses. AUC and nPAUC over FPR $[0, .01]$ are equal-scene means. TP; FP each use denominator 610 at the 1% C level. L1, LPIPS and the derived photometric loss receive negative signs before scoring.

Metric	Readout	AUC	nPAUC .01	TP; FP
M1: original primary target states				
PSNR	RAW	0.8320	0.2748	303; 145
ℓ_1	RAW	0.8407	0.3074	342; 150
SSIM	RAW	0.7921	0.3126	431; 259
LPIPS	RAW	0.7240	0.1271	209; 65
Photo. loss	RAW	0.8272	0.3213	392; 209
M2: target and eight OUT references at 30k/SPZ				
PSNR	RAW	0.8999	0.4236	273; 59
PSNR	GAP	0.9343	0.4014	199; 9
ℓ_1	RAW	0.9136	0.3955	434; 124
ℓ_1	GAP	0.8979	0.1643	34; 4
SSIM	RAW	0.9007	0.4171	454; 159
SSIM	GAP	0.9186	0.3122	125; 2
LPIPS	RAW	0.7943	0.1650	206; 58
LPIPS	GAP	0.8764	0.2477	57; 3
Photo. loss	RAW	0.9147	0.4273	479; 141
Photo. loss	GAP	0.9115	0.2284	46; 2

At the original primary states, L1 improves mean RAW AUC from PSNR's 0.8320 to 0.8407. SSIM and LPIPS also read out membership, with AUC 0.7921 and 0.7240. Under the common 30k/SPZ condition, PSNR-GAP attains the largest mean AUC among the predefined metric/readout combinations, 0.9343. GAP raises mean AUC by 0.0344 for PSNR, 0.0179 for SSIM and 0.0821 for LPIPS, improving six, six and eight scenes, respectively. L1 and the derived loss instead lose 0.0157 and 0.0031 mean AUC after subtraction. Table 37 retains the positive and negative changes in every scene, including both pose-control failures.

Table 37: Every test scene's AUC changes. The first column compares L1 with PSNR in M1; the remaining columns compare GAP with RAW for each M2 metric. Positive values favor the first method. † marks T3 and T8, which fail the pose-control battery and remain in all H results.

	M1	M2: GAP−RAW				
Scene	L1−PSNR	PSNR	ℓ_1	SSIM	LPIPS	Photo. loss
T1	+0.0148	+0.1051	+0.0476	+0.0807	+0.0532	+0.0517
T2	+0.0434	+0.0414	-0.0384	+0.0388	+0.0782	-0.0208
T3†	+0.0023	-0.0261	-0.0800	-0.0500	+0.0199	-0.0642
T4	-0.0154	+0.0328	+0.0260	+0.0197	+0.1336	+0.0151
T5	+0.0155	+0.0713	-0.0104	-0.0151	+0.0716	-0.0111
T6	-0.0075	-0.0366	-0.0652	+0.0041	+0.0683	-0.0354
T7	+0.0133	+0.0556	-0.0100	+0.0169	+0.0261	+0.0057
T8†	+0.0038	+0.0313	+0.0045	+0.0485	+0.2059	+0.0339

Overall ranking, ROC-tail area and fixed decisions respond differently. PSNR-GAP improves M2 AUC while its nPAUC up to FPR .01 falls from RAW's 0.4236 to 0.4014. At the 1% C level, the counts change from 273 TP/59 FP to 199 TP/9 FP, each out of 610. SSIM-GAP gives 125 TP/2 FP and LPIPS-GAP 57 TP/3 FP. These are distinct realized operating points. The six control-passing scenes' unchanged balanced subsets retain PSNR-GAP AUC 0.9242 and 141/440 TP with 6/440 FP

(Table 38). This auxiliary analysis changes both scene inclusion and query support, so its difference from H-all does not estimate a causal pose contribution.

The two control-failing scenes remain visible: M2 PSNR-GAP gives T3 AUC 0.9590 with 47/76 TP and 2/76 FP, and T8 AUC 0.9654 with 7/76 TP and 0/76 FP at the same C level.

Table 38: The unchanged balanced query subsets of six control-passing scenes (T1, T2, T4–T7). AUC is the equal-scene mean; M2 GAP counts use 440 queries per class and the same full-C 1% cutoffs. All five metrics and both M2 readouts remain visible.

Metric	M1 RAW AUC	M2 AUC: RAW→GAP	GAP TP; FP
PSNR	0.7825	0.8738→0.9242	141; 6
ℓ_1	0.7935	0.8917→0.8905	33; 4
SSIM	0.7315	0.8784→0.9091	95; 2
LPIPS	0.6819	0.7781→0.8574	41; 3
Photo. loss	0.7754	0.8937→0.9020	43; 2

The artifact retains the single-feature RR8 breakdown (M0). Feature directions stay fixed; negative MSE and PSNR have identical orderings and decisions under their monotonic relation.

The metric-sensitivity entry of the artifact's `docs/FULL_RESULTS.md` links all 120 per-scene M1/M2 rows, both C levels, full-precision cutoffs, nPAUC .05, controlled/FAIL rows and per-query scores and decisions.

E.2 Original ten-scene residual comparison

The original `random50_seed0` study compares simple residual statistics, the eight-feature `RR_logreg` combiner and global-σ LiRA using the same balanced queries and parity-matched renders. Table 39 retains all ten scenes. The first six columns report TPR at the historical nominal 1% empirical-quantile threshold, which is selected from labeled target nonmembers. The pose-only column instead reports AUC for the univariate nearest-member-distance diagnostic on the selected queries. It is distinct from the multivariate L4 inclusion control in Section A.2.

Table 39: Original ten-scene comparison. Six residual-score columns give retrospective TPR at the nominal 1% level; the last column gives pose $\mathrm{A1best} = \max(\mathrm{AUC}, 1 - \mathrm{AUC})$. LiRA uses the predefined $K{=}8$ global-σ estimator. Residual percentiles and Gini use the member-oriented signs in Section A.3.

	Retrospective TPR at nominal 1%						Pose control
Scene	PSNR	$-\ell_1$	$-r_{90}$	$-$Gini	RR	LiRA	AUC
counter	0.220	0.257	0.257	0.119	0.284	0.688	0.533
kitchen	0.476	0.540	0.532	0.135	0.452	0.683	0.531
bonsai	0.313	0.366	0.405	0.206	0.206	0.031	0.503
bicycle	0.816	0.954	0.954	0.069	0.954	1.000	0.552
flowers	1.000	1.000	1.000	0.000	1.000	1.000	0.541
stump	1.000	1.000	1.000	0.321	1.000	1.000	0.561
truck	0.956	0.982	0.965	0.018	0.991	0.991	0.548
train	0.815	0.800	0.785	0.000	0.926	1.000	0.530
drjohnson	1.000	1.000	1.000	0.424	1.000	1.000	0.502
playroom	1.000	1.000	1.000	0.765	1.000	1.000	0.526

Simple fidelity statistics already separate members from nonmembers, while LiRA provides the strongest pooled nominal 1% readout (Table 40). Pooling uses per-scene standardization for the simple scores and raw Λ for LiRA. The 1,000 query-level bootstrap resamples yield intervals conditional on the fixed target and shadow assets. They quantify query-resampling variation rather than variation across scenes or target training runs. Shared shadow-membership patterns at equal cross-scene filename ordinals remain a source of dependence.

Table 40: Pooled original ten-scene results and conditional bootstrap 95% intervals. The nominal levels use the historical retrospective convention, with 1,054 queries per class.

Readout	AUC [95% CI]	TPR@1% [95% CI]	TPR@5%
PSNR	0.961 [.952,.969]	0.528 [.448,.682]	0.870
`RR_logreg`	0.975 [.968,.982]	0.372 [.121,.777]	0.935
Pose nn-distance	0.537 [.513,.561]	0.020 [.012,.029]	0.052
LiRA $K8$	0.9957 [.994,.997]	0.894 [.830,.965]	0.987

Reference budget and normalization away from saturation. Extending every scene to 16 shadows (seeds 200–215) gives complete query coverage for both the global- and per-query-variance estimators. Recomputing $K8$ through the same implementation reproduces AUC 0.9957 and TPR@1%/5% of 0.894/0.987. The two $K16$ variants reach pooled AUC 0.9970, with respective TPR@1% of 0.955/0.949 and TPR@5% of 0.986/0.981. The predefined headline remains global-σ $K8$; the larger budgets measure sensitivity to reference count.

On the five scenes with PSNR TPR@1% below 0.95 (counter, kitchen, bonsai, bicycle and train), subset-pooled TPR rises from PSNR's 0.371 to LiRA-$K8$'s 0.706, and to 0.832/0.869 with global/per-query $K16$. The corresponding AUCs are 0.922/0.988/0.991 for PSNR/$K8$/$K16$ global. Excluding bonsai gives TPR 0.470/0.856/0.932/0.941 for these four readouts. This analysis shows where reference normalization adds to the raw fidelity trace.

Bonsai is the $K8$ global-variance exception: LiRA gives TPR 0.031 versus PSNR's 0.313. Increasing to $K16$ raises global/per-query TPR to 0.244/0.657 and AUC from 0.936 to 0.950/0.958. The per-query TPR is 0.092/0.657/0.725/0.771 at $K{=}8/16/24/32$, exceeding raw PSNR once $K \geq 16$. Eight-of-sixteen shadow-subset resampling without replacement gives mean/SD TPR 0.259/0.171 (bonsai global), 0.203/0.170 (bonsai per-query), 0.782/0.101 (counter global) and 0.525/0.273 (counter per-query). This conditional variation links the weak bonsai tail to reference budget and variance estimation. Complete budget and subset records remain in the artifact's archived `tab_k16.tex`, `tab_k16pool.tex` and `tab_desat.tex` and original Sections B and L.

E.3 Learned residual-pair readout and transfer

On the W2 four-C/eight-T panel, we compare the scalar scores with an eight-feature logistic classifier and a FedMeNF-style paired-classifier adaptation (Yun et al., 2025), using supplied query poses. RR8 standardizes the residual features using its training shadows, then fits logistic regression with $C = 1$ and a maximum of 1,000 iterations. The CNN consumes six channels,

$$[(I_q - 0.5)/0.25, \quad (G(q) - I_q)/\sigma_{\text{train}}],$$

where σ_{train} is the training-shadow residual RMS without label conditioning, floored at 10^{-6}. It preserves the native image canvas and exposes the candidate/render residual. Renders are clamped in float32 before conversion to float16 pair caches; RR8 is computed before conversion. The retained parity records quantify the PSNR change from caching.

The architecture is a $6 \to 32$ convolution (kernel/stride/padding $5/2/2$), ReLU, a $32 \to 64$ convolution $(3/2/1)$, ReLU, adaptive average pooling to 8×8, a $4096 \to 128$ linear layer, ReLU and a scalar output. Adam uses learning rate 10^{-3}, weight decay 10^{-4}, batch size 4 and at most 40 epochs. We report both fixed seeds, 90851 and 90852. The selected epoch minimizes validation binary cross-entropy; the classifier is then refitted from the same initialization for that count.

Training and selection roles. Same-scene selection fits six earlier counter MATCHED shadows and validates on the other two. Its two selected counts are 37 and 38 epochs. Each new scene then fits all eight of its own MATCHED shadows at its assigned primary recipe, recomputing the residual scale from those training shadows. Candidate images recur across shadow model groups with known memberships, making this access setting transductive. Cross-scene selection instead trains on all eight earlier counter shadows, validates on all eight earlier bonsai shadows and refits on both scenes, selecting one epoch for each seed. It performs no attack fitting on the new scenes. Earlier development assets use their mature native checkpoints. Every method and seed has a separate four-C cutoff, with no target-label or test-scene selection.

Effective same-scene readout and different transfer tails. The two same-scene CNN seeds attain mean AUC 0.9406 and 0.9485 versus MATCHED BaVarIA-n's 0.9273. Their scene-level changes are positive in five and eight of the eight scenes, respectively. At the 1% C level, CNN counts are 388 TP/8 FP and 351 TP/2 FP, compared with BaVarIA-n's 383 TP/11 FP (610 per class). Tables 41 and 42 retain both seeds and every test scene.

Table 41: All learned configurations and their MATCHED BaVarIA-n comparator at the original primary states, with supplied poses. AUC/nPAUC are eight-scene means; Min. is the lowest scene AUC. TP; FP each use denominator 610 at the indicated C level. Both fixed CNN seeds are reported.

Readout	AUC	Min.	nPAUC .01	1%	5%
BaVarIA-n	0.9273	0.7573	0.5817	383; 11	485; 40
RR8 same-scene	0.8524	0.6444	0.3279	181; 2	289; 18
RR8 cross-scene	0.8334	0.6487	0.2455	328; 102	379; 196
CNN same, 90851	0.9406	0.8052	0.5379	388; 8	505; 35
CNN same, 90852	0.9485	0.8095	0.7118	351; 2	498; 24
CNN cross, 90851	0.8135	0.6461	0.2131	12; 1	298; 78
CNN cross, 90852	0.8185	0.6478	0.2553	97; 12	264; 66

Table 42: Primary-state AUC in every test scene. Same/cross columns are the two fixed CNN seeds; † identifies the two metadata-control failures. The two same-scene seeds improve on BaVarIA-n in five and eight scenes, respectively.

Scene	BaVarIA-n	Same 90851	Same 90852	Cross 90851	Cross 90852
T1	0.7573	0.8204	0.8331	0.6788	0.6848
T2	0.9787	0.9900	0.9829	0.8135	0.7980
T3†	0.9995	1.0000	1.0000	0.9797	0.9924
T4	0.9962	1.0000	1.0000	0.8797	0.8830
T5	0.7688	0.8052	0.8095	0.6461	0.6478
T6	0.9232	0.9202	0.9632	0.7621	0.7776
T7	0.9966	0.9916	0.9996	0.9253	0.9307
T8†	0.9979	0.9972	1.0000	0.8229	0.8336

Cross-scene CNN AUC remains 0.8135/0.8185, with 12 TP/1 FP and 97 TP/12 FP at the 1% C level. At 5%, the counts become 298/78 and 264/66. Cross-scene RR8 gives 328 TP with 102 FP at the primary level. Scene transfer thus retains ranking signal while changing the realized fixed-threshold operating points. On the six control-passing scenes' original balanced subsets, same-scene CNN AUC is 0.9201/0.9318 versus BaVarIA-n's 0.8999.

All new targets are also evaluated at all four release states. Same-scene attacks stay fitted at their assigned primary recipe, so non-primary target states introduce a recipe mismatch; the cross-scene early and encoded states likewise combine scene and release shift. The artifact retains the complete release-state and validation tables, model-group assignments, residual scales, cache-parity measurements and saved predictions. The separate M_{src} study in Section H reports the earlier raw-RGB-pair CNN's constant-logit failures; its different model-bank and utility design precludes attributing the new panel's improvement solely to residual scaling.

F Background-assisted localization and discovery coverage

Registered background images connect an unposed candidate to the asset's coordinate frame. We evaluate that complete path from localization to calibrated membership decisions, retaining every candidate in its original denominator. T1–T8 below denote W2T1–W2T8 of the four-C/eight-T panel in Section A.1. Earlier registration studies use their own references, query populations and decision rules.

F.1 A BACKGROUND MAP SUPPLIES THE SCORING CAMERA

The auditor receives background RGB B, their released cameras and the candidate intrinsics. HLOC (Sarlin et al., 2019) combines SuperPoint (DeTone et al., 2018) and LightGlue (Lindenberger et al., 2023). SuperPoint uses at most 4,096 keypoints, NMS radius 3, fixed sampling, grayscale images and a forced maximum dimension of 1,600. Map construction selects 12 background neighbors with a 30° rotation threshold and triangulates B observations at their fixed released cameras. Neither candidate H nor anchor R observations enter this map.

Each query is matched exhaustively to B. PnP requires 12 inliers at an 8-pixel error threshold, with focal-length and extra-parameter refinement disabled. The pipeline uses no ground-truth query retrieval, joint H–H bundle adjustment or fallback pose. The estimated camera then renders the target and every OUT reference. Anchors follow the same path. Invalid localization would yield an abstention in the original class denominator; all 1,220 test candidates and 256 anchors return poses, and all primary scores are finite.

Scene	Set	Returned	Rot. med.	p90	Max.	$> 1°$	Trans. p90
T1	H	150/150	0.020	89.861	179.845	18	1.66116
T1	R	32/32	0.019	90.041	179.839	5	1.67933
T2	H	150/150	0.044	0.128	104.918	2	0.00197
T2	R	32/32	0.042	0.112	89.830	1	0.00434
T3	H	152/152	0.043	0.190	159.426	3	0.00294
T3	R	32/32	0.044	0.082	0.296	0	0.00237
T4	H	158/158	0.033	0.087	2.080	2	0.00046
T4	R	32/32	0.042	0.092	0.713	0	0.00052
T5	H	158/158	0.057	0.132	2.708	2	0.00208
T5	R	32/32	0.040	0.114	0.628	0	0.00238
T6	H	150/150	0.019	0.050	0.596	0	0.00039
T6	R	32/32	0.018	0.032	0.085	0	0.00042
T7	H	150/150	0.033	0.112	0.891	0	0.00190
T7	R	32/32	0.033	0.086	0.124	0	0.00159
T8	H	152/152	0.028	0.077	0.392	0	0.00078
T8	R	32/32	0.032	0.063	0.075	0	0.00074

Table 43: Complete test-scene pose coverage and error tails. H contains candidates and R contains 32 nonmember anchors per scene. Rotation statistics are in degrees; translation p90 is normalized by the RMS radius of background-camera centers. Every returned pose enters scoring, including all 27 H queries with rotation error above $1°$.

Full pose return coexists with a long error tail (Table 43). T1 has median rotation error $0.020°$, p90 $89.861°$ and maximum $179.845°$; 18 of its 150 candidates exceed $1°$. Across the eight scenes, 27 candidates exceed that error, and all remain in scoring. On T1, RAW AUC changes from 0.6389 to 0.5444, whereas GAP changes from 0.6734 to 0.6670. The relative target/reference discrepancy is empirically less sensitive to these measured camera errors than absolute fidelity.

Test-scene maps take 42.72–86.49 seconds to construct. The complete H+R batches contain 182–190 queries and take 693.79–1012.29 seconds, or 3.77–5.33 seconds per query when amortized over the batch. These times cover feature extraction, matching and PnP. Maps are reusable; reference training and rendering are separate components. Complete C/T map and timing records accompany the full results.

F.2 PAIRED CAMERAS RETAIN USEFUL DECISIONS AND SCENE DIFFERENCES

Table 44 compares identical H populations at supplied and estimated cameras. It retains all eight scenes and four readout roles: absolute fidelity, background-reference subtraction, variability-adjusted OUT-T, and release-matched subtraction. Each pose branch uses its own frozen C cutoff. Estimated GAP attains equal-scene AUC 0.8483 and detects 104/610 members with 6/610 false positives at 1% C. At 5% C, its counts are 219/610 and 19/610. Ranking changes by -0.0046 for GAP, compared with -0.0338 for RAW using the unrounded recorded means. The learned attacks and additional SSIM/LPIPS comparisons use supplied cameras in Section E; their estimated-pose outcomes were not measured.

Scene	Pose	N	RAW	GAP	OUT-T	Rcp. GAP	GAP TP; FP
T1	S	75	0.6389	0.6734	0.6574	0.7579	13; 3
T1	E	75	0.5444	0.6670	0.6487	0.6923	13; 3
T2	S	75	0.7808	0.8958	0.8809	0.9216	4; 1
T2	E	75	0.7580	0.8905	0.8720	0.9113	3; 1
T3†	S	76	0.9874	0.9706	0.9650	0.9706	50; 2
T3†	E	76	0.9692	0.9683	0.9621	0.9683	46; 2
T4	S	79	0.9639	0.9897	0.9896	0.9968	17; 0
T4	E	79	0.9356	0.9857	0.9856	0.9939	14; 0
T5	S	79	0.6574	0.6084	0.6054	0.7324	1; 0
T5	E	79	0.6401	0.6166	0.6195	0.7334	1; 0
T6	S	75	0.7959	0.7668	0.7541	0.8512	0; 0
T6	E	75	0.7879	0.7563	0.7465	0.8484	0; 0
T7	S	75	0.8972	0.9540	0.9173	0.9540	30; 0
T7	E	75	0.8364	0.9396	0.9100	0.9396	25; 0
T8†	S	76	0.9340	0.9643	0.9538	0.9654	2; 0
T8†	E	76	0.9133	0.9623	0.9519	0.9643	2; 0
All	S	610	0.8320	0.8529	0.8404	0.8937	117; 6
All	E	610	0.7981	0.8483	0.8370	0.8814	104; 6

Table 44: Paired supplied (S) and estimated (E) cameras at each scene's primary release. Four readout columns give AUC; GAP counts are TP; FP at the branch's frozen 1% C cutoff, with N queries per class. GAP and OUT-T use the default 30k/native references. Rcp. GAP denotes recipe-known GAP, which matches the target state. † marks metadata-control failures, retained in all-scene results. All-row AUCs are the original equal-scene means, with pooled counts.

Query identities reveal which discoveries persist. At 1% C, the branches share 101 true detections and all six false positives; 16 supplied-pose detections are lost and three are gained. At 5%, 196 true detections and 17 false positives coincide, with 10/23 true detections and 2/2 false positives lost/gained. These are changes between complete calibrated pipelines. Discoveries remain uneven: T3 and T7 contribute 71 of the 104 estimated-GAP detections, while T6 has none. Estimated OUT-T yields 43 true detections and zero false positives; recipe-known GAP gives 111 and three. The mixture and GAP–R32 give 59/2 and 132/8, respectively. Their reference and anchor behavior is developed in Section D.

The metadata battery tests pose–membership association separately from localization accuracy. The unchanged MILP matcher uses $\tau = 0.15$ and at least 30 matched pairs. T3 and T8 fail the strongest classifier test, with AUC/p equal to 0.660665/0.014925 and 0.670612/0.004975. Their full H results stay in the primary population. On the six passing scenes' common balanced subset, GAP retains AUC 0.8171/0.8106 at supplied/estimated poses and 63/440 versus 54/440 true detections, with three false positives in both branches (Table 45). The failing subset is shown alongside it. Because this auxiliary comparison changes scene inclusion and query support together, its difference from the full pool does not isolate a causal pose contribution.

The artifact retains the counter-H96 development run, which transported earlier registered-C thresholds. The W2 comparison above calibrates supplied and estimated branches separately.

Readout	AUC S	AUC E	S TP; FP	E TP; FP	E nPAUC
Six passing scenes; $N_+ = N_- = 440$					
RAW	0.7825	0.7428	145; 76	135; 77	0.1559
GAP	0.8171	0.8106	63; 3	54; 3	0.2619
OUT-T	0.8024	0.7986	44; 1	27; 0	0.2062
Recipe-known GAP	0.8666	0.8493	68; 1	57; 1	0.2848
T3/T8, failing; $N_+ = N_- = 146$					
RAW	0.9562	0.9350	146; 69	144; 67	0.5013
GAP	0.9674	0.9654	52; 2	48; 2	0.4771
OUT-T	0.9616	0.9591	22; 0	16; 0	0.5741
Recipe-known GAP	0.9680	0.9664	56; 2	52; 2	0.5128

Table 45: Common pose-balanced subsets under the original supplied/estimated C rules. AUC is averaged across the scenes in each group. Decision columns are TP; FP at 1% C; the last column gives estimated-pose nPAUC .01. The passing group uses T1/T2/T4/T5/T6/T7. These subsets change scene and query support jointly; all eight full H populations remain the primary evaluation.

F.3 Other reference access and real-capture discovery coverage

Other registration access. An earlier three-scene study builds its registration bank from target optimization images while removing each member query's own observations and rebuilding tracks. It measures transductive-reference recovery: nearly all queries return poses, with severe drjohnson errors and one correspondence failure retained. The B-only map in Section F.1 instead excludes candidate and anchor observations at construction. Full procedures and outcomes for this earlier bank remain in the artifact.

Without released camera metadata, the tested fixed rendered-reference search localizes 0/6 queries within tolerance; an adaptive search also yields 0/24 acceptable poses across three scenes. Ground-truth poses enter after search for evaluation. These failures do not establish metadata-free localization from a bare asset. The artifact retains the search and post-reveal scale diagnostics, including cases with many inliers but incorrect poses.

Residential development data. Desk and Music test acquisition-to-decision coverage. Targets are fixed before collecting calibration anchors and prospective nonmembers, making both genuine post-release nonmembers. Table 46 retains all 802 role rows: 216 fail localization and 586 receive finite scores. The 520 candidates are separate from the 282 calibration anchors and produce 44 and 37 discoveries in Desk and Music, respectively.

Table 46: Residential coverage and decisions at each scene's frozen anchor cutoff. Every role row remains in its full denominator; failed localization is an abstention. Positive calls are TP for Q^+ and FP for Q^-. The 520 candidate rows exclude 282 calibration anchors. The historical one-sided FPR bound describes the test cohort: anchor/test exchangeability was not established. Both scenes fail the required 50% TPR and 25% discovery-coverage floors.

Scene	Role	Full	Scored	Abstentions	Positive calls/full
Desk	Q^+	119	75	44	43/119
Desk	A_{cal}	136	83	53	–
Desk	Q^-	136	80	56	1/136
Music	Q^+	119	108	11	35/119
Music	A_{cal}	146	122	24	–
Music	Q^-	146	118	28	2/146
Total		**802**	**586**	**216**	–

Scene	Full-roster TPR	FPR upper 95% bound	Discovery coverage
Desk	36.13%	3.44%	17.25%
Music	29.41%	4.25%	13.96%

The calibration rule takes the largest acceptable anchor exceedance rank whose one-sided 95% binomial upper bound is at most 5%. With 83/122 finite anchors, it selects their highest/second-highest score, giving PSNR cutoffs 30.058709/31.306119 dB. Predictions use strict comparison. Metadata SMD and the historical directional role-classifier check did not establish anchor/test exchangeability: directional AUCs were 0.265211/0.081273, and the one-sided rule is orientation-sensitive. The corresponding test-cohort FPR bounds are therefore descriptive risk arithmetic, rather than transferable anchor-based guarantees.

Both scenes miss the required 50% full-denominator TPR and 25% discovery coverage despite meeting the historical 5% FPR-bound component (Table 46). Desk detects 43/119 members on the full roster (43/75 among scored members), leaving 44 localization abstentions and 32 scored members below threshold. Music detects 35/108 scored members (35/119 on the full roster); even adding all 11 localization failures would give only 38.7% full-denominator TPR. The operational outcome remains failure in both scenes. Anonymous role/score/decision records support numerical replay, while residential imagery, filenames, geometry and model assets remain private.

F.4 Sensitivity to the scoring camera and candidate image

Imposed pose and resolution changes. The earlier perturbation study changes the comparison camera or resolution on fixed trained assets, isolating input sensitivity from localization performance. Table 47 retains the three-scene render-residual grid. Halving comparison resolution leaves AUC 0.870/0.743/0.970 on counter/bonsai/truck, whereas a 0.5° rotation gives 0.608/0.602/0.910. The member-specific fidelity decay and supervision interventions are discussed in Section B. The earlier LiRA Tier2 grid remains an archived historical result, with its original-output gap stated in Section I.

Perturbation levels	counter (.961)	bonsai (.877)	truck (.988)
Resolution .9/.75/.5	.957/.943/.870	.879/.853/.743	.987/.986/.970
Rotation .5/2/5°	.608/.488/.479	.602/.483/.474	.910/.728/.580
Translation .1/.25/.5×	.757/.651/.516	.718/.679/.574	.938/.803/.715

Table 47: Original pooled out-of-fold render-residual AUC under imposed query perturbations on fixed trained models. Parentheses give the exact-camera baseline. Each cell follows the three parameter levels in its row. Resolution ratios keep focal length and field of view unchanged; translation is a multiple of median inter-camera spacing.

Earlier W2 candidate transformations with clean calibration. The eight-scene test panel also evaluates five transformations of candidate RGB at supplied poses. Supervision labels, clean R anchors and the original C cutoffs stay fixed. JPEG uses Pillow RGB roundtrips at quality 90/75, uint8 conversion, chroma subsampling 0 and no optimization. Exposure multiplies numerical sRGB values by 2^{EV} and clips to $[0, 1]$ at $\mathrm{EV} = \pm 0.25$. CROP80 takes the central 80% and bicubically resizes to 960×540; focal lengths scale by 1.25, the centered principal point stays fixed, and target/reference renders use these updated intrinsics. These are candidate-input transformations; the served-render output interventions in Section H change a different interface.

	RAW		GAP		BaVarIA-n	
Condition	AUC	TP; FP	AUC	TP; FP	AUC	TP; FP
Clean	0.8320	303; 145	0.8529	117; 6	0.9273	383; 11
JPEG90	0.8319	303; 144	0.8530	114; 6	0.9272	384; 11
JPEG75	0.8318	303; 143	0.8527	107; 6	0.9272	384; 11
CROP80	0.7978	385; 253	0.8170	140; 10	0.9004	332; 8
-0.25 EV	0.7234	0; 2	0.7483	4; 1	0.8409	191; 20
$+0.25$ EV	0.6914	0; 0	0.7361	1; 0	0.8331	209; 23

Table 48: Candidate-image transformations at supplied poses, with clean anchors and unchanged clean C calibration. All eight primary releases retain 610 queries per class. Each method has paired AUC and 1% TP; FP columns. RAW, background-only GAP and MATCHED BaVarIA-n represent the three reference-access roles; all 15 scalar readouts and both calibration levels remain in the full results.

JPEG90/75 preserves GAP AUC at 0.8530/0.8527, close to 0.8529 clean, with 114/6 and 107/6 primary decisions. CROP80 gives lower AUC 0.8170 while increasing discoveries and false positives to 140/10. Exposure changes reduce GAP AUC to 0.7483/0.7361 and decisions to 4/1 and 1/0. BaVarIA-n also loses ranking under exposure. The measured dependence on photometric consistency therefore extends beyond raw fidelity, while the tested JPEG transformations retain the readout. Learned classifiers were evaluated on clean releases; these conditions contain scalar readouts. W3 tests the shared-gain response on a new population with its own calibration and within-panel A/B/C pairing (Section D.6).

The full tables, methods and available per-query records are indexed by original sections C, G, J, R, AE and AI in the Artifact's `docs/FULL_RESULTS.md`. For the newer panel, these include both calibration levels, calibration-scene pose records and every transformed scalar condition. Each study retains its original population and access conditions.

G Independent readout, photometric and localization evaluation

G.1 Population, readouts and primary comparison

The independent DL3DV panel (W4 in the Artifact) has eight calibration scenes and eight test scenes, each with one target and eight OUT references trained for 30k iterations. The references use their respective background subsets for initialization and training. Targets use their member background subset and member queries; nonmember queries and R32 anchors remain excluded. The primary contrast fixes native release, supplied pose, clean RGB, no photometric alignment and eight references. Neither the calibration nor test scores select the readout formula. The calibration scenes determine cutoffs; the eight new test scenes determine the reported effect.

The native renderer's training-image exposure correction is disabled: the saved W4 and student-study training configurations set `train_test_exp=False`, and scoring renders do not enable `use_trained_exp`. The score therefore does not retrieve an exposure transform indexed by a training image's name.

PSNR-GAP subtracts the arithmetic mean of reference PSNRs from target PSNR. For each model, the alternative photometric error is $L_{\text{photo}} = 0.8L_1 + 0.2(1 - \text{SSIM})$. Log-photo averages the logarithms of reference errors and subtracts the logarithm of target error, flooring each error at 10^{-10}. Thus the reference numerator is a geometric mean. Applying a logarithm after an additive GAP, or after an arithmetic mean of errors, would define different scores. PSNR-GAP itself is a log-ratio readout in MSE; the primary comparison tests the error domain within this common log-ratio structure.

Each calibration cutoff uses the pooled finite nonmember scores of all eight C scenes. With $j = \lfloor 0.01(n_C^- + 1) \rfloor$, the cutoff is the jth largest negative score; decisions require strict score greater than cutoff. The pose, release, alignment and readout each have their own rule. A missing score abstains and remains in the decision denominator. Ranking uses finite scores, retaining tied score

groups. nPAUC(.01) is the ROC area on $[0, .01]$ divided by .01, with linear interpolation at the boundary and no random-baseline rescaling.

Table 49: Every primary test scene, supplied pose, clean native release, eight references, no alignment. The last column is log-photo minus PSNR-GAP.

Scene	GAP AUC	Log-photo AUC	GAP nPAUC	Log-photo nPAUC	ΔnPAUC
T1	0.9388	0.9716	0.2800	0.5200	0.2400
T2	0.9887	0.9970	0.6494	0.8961	0.2468
T3	0.9868	0.9968	0.6400	0.8667	0.2267
T4	0.9122	0.9273	0.0886	0.4557	0.3671
T5	0.9189	0.9432	0.0260	0.1039	0.0779
T6	0.9789	0.9906	0.7143	0.8182	0.1039
T7	0.9776	0.9921	0.6410	0.8590	0.2179
T8	0.9646	0.9756	0.4000	0.1733	-0.2267

The primary equal-scene difference is $+0.156701$ nPAUC(.01), with a 95% paired scene-bootstrap interval $[+0.027383, +0.253997]$. The companion AUC difference is $+0.015926$, interval $[+0.011155, +0.021759]$. The bootstrap resamples the eight paired scene differences 10,000 times with seed 9192026. All primary queries have finite scores for both methods, so common-support and method-specific results coincide. T8 is the negative tail-ranking result despite higher AUC. Because its primary cameras are supplied, estimated-pose error cannot explain this primary decrease.

Ranking and cutoff transfer answer different questions. T4's nPAUC improves from 0.088608 to 0.455696 while its TP count decreases from three to two, with zero FP for both scores. Pooled native/supplied FPR is $11/613 = 1.7945\%$ for log-photo and $13/613 = 2.1207\%$ for PSNR-GAP. Equal-scene FPR is 1.7883% and 2.1365%, respectively. The primary ranking interval does not estimate uncertainty in calibration-pool composition.

G.2 Finite-tail resolution and wider-tail support

Each test scene has 75–79 nonmembers, so its first positive empirical FPR is $1/n_s^- = 1.2658$–1.3333%. The two primary readouts have no cross-label score ties. Under the stated ROC integration, the TPR is constant between zero and this first false positive. Consequently, for these scores,

$$\mathrm{nPAUC}_s(.01) = \frac{\#\{i : y_i = 1,\ s_i > \max_{j:y_j=0} s_j\}}{n_s^+}.$$

The primary endpoint therefore measures separation above the largest observed negative score in each scene. This empirical ranking summary uses evaluation labels; it neither defines nor replaces the C-calibrated decision rule.

Table 50: Ranking on the same eight native/supplied-pose test scenes with unaligned $K = 8$ scores. Differences are log-photo minus PSNR-GAP, averaged equally over scenes; intervals use paired scene bootstrap. The wider-tail summary is post-hoc and does not replace the primary endpoint or calibration.

Metric	PSNR-GAP	log-photo	Difference	95% interval	Improved
nPAUC(.01), primary	0.4299	0.5866	+0.1567	[+0.0274, +0.2540]	7/8
nPAUC(.05), post-hoc	0.6006	0.7379	+0.1373	[+0.0920, +0.1827]	8/8
AUC	0.9583	0.9743	+0.0159	[+0.0112, +0.0218]	8/8

A post-hoc companion analysis of the same fixed scores integrates to 5% FPR. Log-photo improves nPAUC(.05) by 0.137332 on average, with all eight scenes improving and a paired scene-bootstrap interval of $[0.092024, 0.182731]$. Together with the full-AUC result, this shows that the readout advantage extends beyond the first observed false positive. Five percent still spans only a few negative-rank steps per scene. The original .01 endpoint, its T8 decrease and all calibrated decisions remain unchanged.

Table 51: Sensitivity of the primary log-photo minus PSNR-GAP nPAUC(.01) difference to omitting each negative once from both methods. The formal analysis keeps all queries; these ranges neither select a subset nor change a cutoff.

Scene	Original	Minimum after omission	Maximum after omission
T1	+0.2400	+0.0667	+0.2400
T2	+0.2468	+0.1429	+0.2857
T3	+0.2267	+0.1333	+0.2267
T4	+0.3671	+0.2152	+0.4557
T5	+0.0779	+0.0779	+0.2727
T6	+0.1039	-0.0130	+0.1039
T7	+0.2179	+0.1795	+0.2179
T8	-0.2267	-0.2267	+0.1200

Omitting each negative once from both readouts characterizes sensitivity to an extreme observation. Across the 613 omissions, the scene difference can cross zero in T6 and T8 (Table 51). The formal result retains every query; the omission analysis is not used to choose scenes, samples or calibration rules.

G.3 Photometric conditions and shared fits

All transformations act on float32 RGB and clip outputs to $[0, 1]$. The eight conditions are clean, multiplication by $2^{-0.25}$ or $2^{0.25}$, elementwise powers 0.8 or 1.25, RGB multipliers $(1.1, 1.0, 0.9)$ or $(0.9, 1.0, 1.1)$, and multiplication by $2^{0.75}$ with clipping. They are numerical image transformations. Every estimated-pose condition performs its own background-assisted localization; the clean R32 anchors retain their original condition within the corresponding pose/release branch.

Let μ be the arithmetic mean of the eight individually clamped reference renders. Fits use RGB elements satisfying $0.05 < \mu < 0.95$ and $0.02 < I < 0.98$. Scalar gain minimizes squared discrepancy on this mask; RGB gain solves the corresponding per-channel problems. Gains lie in $[0.5, 2]$. The gamma+RGB candidate fits $[a_c \mu_c^{\gamma}]_0^1$ with $\gamma \in [0.5, 2]$ and three gains in the same bounds. Its least-squares fit uses stride four in each spatial dimension, unit initialization, the trust-region reflective solver, at most 64 function evaluations, and 10^{-6} tolerances. Scalar gain requires at least 1% valid RGB elements overall; RGB gain requires this fraction within each channel, and gamma+RGB within each channel of the sampled grid. Insufficient support or a failed nonlinear fit gives the identity transform. The fitted transform is then shared by the target and every reference. Target renders, membership labels and the imposed transformation label do not enter fitting.

The four candidate rules and their numerical settings were frozen before new C/T scoring. Gamma+RGB therefore has a fixed algorithm across conditions while fitting query-specific parameters. Its complete eight-condition result is a secondary analysis of a pre-frozen candidate, distinct from the primary unaligned comparison. Choosing scalar on one T condition, RGB on another and gamma+RGB on a third would describe a different system from any one evaluated rule.

Table 52: All four log-photo rules across eight numerical conditions, supplied poses, native release and eight references. Gray tracks the prespecified secondary gamma+RGB procedure, including its clean-input tradeoff. AUC/nPAUC are equal-scene means. Clean-C and condition-matched C change the cutoff for the same row scores. Each decision denominator is 613 per class; a failed query remains in that denominator.

		Ranking		Decisions: TP/FP	
Condition	Alignment	AUC	nPAUC(.01)	Clean-C	Matched-C
Clean	None	0.9743	0.5866	387/11	387/11
	Scalar	0.9618	0.5159	357/7	357/7
	RGB	0.9617	0.5191	355/5	355/5
	Gamma+RGB	0.9513	0.5372	290/4	290/4
-0.25 EV	None	0.8674	0.4012	134/1	265/9
	Scalar	0.9627	0.5159	358/8	358/8
	RGB	0.9626	0.5207	357/6	357/6
	Gamma+RGB	0.9524	0.5422	302/4	302/4
$+0.25$ EV	None	0.8831	0.3010	116/1	254/9
	Scalar	0.9605	0.5029	346/7	347/7
	RGB	0.9604	0.5126	340/6	341/6
	Gamma+RGB	0.9508	0.5242	265/3	265/3
$\gamma = 0.8$	None	0.8701	0.2992	152/3	264/14
	Scalar	0.9250	0.3924	269/4	336/8
	RGB	0.9255	0.4055	271/3	334/8
	Gamma+RGB	0.9520	0.5078	287/4	288/4
$\gamma = 1.25$	None	0.8694	0.3596	161/1	275/9
	Scalar	0.9304	0.4601	272/3	342/10
	RGB	0.9304	0.4634	277/3	343/10
	Gamma+RGB	0.9507	0.5422	309/6	309/5
WB warm	None	0.9582	0.4671	285/4	344/13
	Scalar	0.9515	0.4424	268/3	326/8
	RGB	0.9618	0.5224	354/5	354/5
	Gamma+RGB	0.9517	0.5355	294/4	294/4
WB cool	None	0.9599	0.4823	283/3	353/13
	Scalar	0.9554	0.4569	264/3	332/8
	RGB	0.9616	0.5224	354/6	354/6
	Gamma+RGB	0.9514	0.5324	288/4	288/4
$+0.75$ EV / clip	None	0.8001	0.2058	1/0	179/14
	Scalar	0.9510	0.4199	248/5	246/5
	RGB	0.9504	0.4150	247/5	244/4
	Gamma+RGB	0.9459	0.4472	201/3	212/3

Table 53: All four log-photo rules across eight numerical conditions, estimated poses, native release and eight references. Gray tracks the prespecified secondary gamma+RGB procedure, including its clean-input tradeoff. AUC/nPAUC are equal-scene means. Clean-C and condition-matched C change the cutoff for the same row scores. Each decision denominator is 613 per class; a failed query remains in that denominator.

		Ranking		Decisions: TP/FP	
Condition	Alignment	AUC	nPAUC(.01)	Clean-C	Matched-C
Clean	None	0.9404	0.4835	307/12	307/12
	Scalar	0.9296	0.4179	277/6	277/6
	RGB	0.9294	0.4228	272/5	272/5
	Gamma+RGB	0.9201	0.4391	225/4	225/4
-0.25 EV	None	0.8374	0.3126	77/1	216/16
	Scalar	0.9327	0.4244	280/5	280/6
	RGB	0.9324	0.4227	279/5	279/5
	Gamma+RGB	0.9225	0.4506	231/4	232/4
$+0.25$ EV	None	0.8466	0.2037	58/1	205/15
	Scalar	0.9282	0.4014	257/5	260/6
	RGB	0.9283	0.4097	251/4	255/4
	Gamma+RGB	0.9184	0.4181	208/4	209/4
$\gamma = 0.8$	None	0.8361	0.2172	82/1	200/14
	Scalar	0.8937	0.3071	182/3	249/7
	RGB	0.8937	0.3056	182/3	248/7
	Gamma+RGB	0.9203	0.4194	216/4	218/5
$\gamma = 1.25$	None	0.8415	0.3041	92/1	239/19
	Scalar	0.9017	0.3702	190/3	291/16
	RGB	0.9018	0.3687	188/3	294/16
	Gamma+RGB	0.9218	0.4555	246/4	250/4
WB warm	None	0.9299	0.3593	205/3	273/13
	Scalar	0.9240	0.3478	183/2	260/10
	RGB	0.9331	0.4244	272/5	272/5
	Gamma+RGB	0.9226	0.4326	223/4	222/4
WB cool	None	0.9280	0.4108	209/3	298/15
	Scalar	0.9241	0.3605	186/3	269/10
	RGB	0.9299	0.4211	277/6	276/5
	Gamma+RGB	0.9203	0.4376	227/4	227/4
$+0.75$ EV / clip	None	0.7545	0.1499	0/0	129/19
	Scalar	0.9078	0.3516	191/4	191/4
	RGB	0.9074	0.3500	190/4	191/4
	Gamma+RGB	0.9033	0.3344	154/3	202/5

Clean-C calibration reuses each rule's clean threshold across all conditions. Condition-matched calibration changes the threshold but not the scores or AUC. The full matrices show the tradeoff: alignment recovers detections under mismatch and can lower clean discrimination. On relocalized candidates, gamma+RGB preserves AUC 0.9033–0.9226 and 154–246 TP with 3–4 FP, while scalar/RGB gain can retain more detections for their corresponding linear transformations. The earlier W3 scalar-gain experiment has eight C and four T scenes, a different target population and PSNR-GAP; its ideal unclipped scale-cancellation calculation is not a theorem for the SSIM-containing loss used here.

The saved fits quantify the safeguards. Across 9,808 T query–condition observations, supplied poses have no fallback; scalar, RGB and gamma+RGB hit parameter bounds 6, 10 and 439 times. Estimated poses have one abstention without a fit. Among the remaining 9,807 observations, each adapter has 17 identity fallbacks, with 2, 6 and 502 bound hits, respectively. A failed optimizer-success flag can also mean insufficient mask/grid support before optimization; it is not by itself evidence that the numerical solver ran and failed. Repeated conditions are not additional independent queries.

G.4 Reference budget and baseline adaptations

The reference budget uses fixed construction-order prefixes, OUT00 through OUT07. No reference subset is selected by test performance. The $K = 1, 2, 4, 8$ log-photo curve uses unaligned scores. All photometric fits use the eight-reference mean, so the low-K curve does not measure low-reference photometric adaptation. The background localization map and optional R32 queries are separate access costs. On these fixed prefixes, nPAUC peaks at K4 in both pose branches; K8 gives more detections and false positives at its calibrated cutoff.

Table 54: Unaligned log-photo with fixed reference prefixes. These budgets constrain score aggregation; alignment experiments all use eight references. B-map images and optional R32 anchors are separate access requirements.

Pose	K	AUC	nPAUC(.01)	TP/FP
supplied	1	0.9630	0.5184	328/6
supplied	2	0.9735	0.5678	347/8
supplied	4	0.9746	0.6082	351/8
supplied	8	0.9743	0.5866	387/11
estimated	1	0.9272	0.4071	226/6
estimated	2	0.9388	0.4678	272/8
estimated	4	0.9394	0.4870	272/8
estimated	8	0.9404	0.4835	307/12

Table 55: Error readouts and render-score baseline adaptations, clean native release. Each method uses its own C cutoff. RMIA vote and QR use clean R32 anchors; the continuous ratio does not depend on their values.

Pose	Readout	AUC	nPAUC(.01)	TP/FP
supplied	RAW	0.8846	0.4779	34/0
supplied	PSNR-GAP	0.9583	0.4299	348/13
supplied	L1 difference	0.9160	0.2253	160/12
supplied	Photo difference	0.9332	0.3068	196/8
supplied	log-L1	0.9652	0.5099	358/11
supplied	log-photo	0.9743	0.5866	387/11
supplied	RMIA render / vote	0.8698	0.4218	162/0
supplied	RMIA render / ratio	0.9302	0.4646	257/9
supplied	QR / R32	0.8846	0.4779	301/51
estimated	RAW	0.8219	0.3975	26/0
estimated	PSNR-GAP	0.9253	0.3329	261/12
estimated	L1 difference	0.8905	0.1447	171/11
estimated	Photo difference	0.9060	0.2097	162/8
estimated	log-L1	0.9326	0.4051	283/10
estimated	log-photo	0.9404	0.4835	307/12
estimated	RMIA render / vote	0.8096	0.3457	131/0
estimated	RMIA render / ratio	0.8693	0.3857	213/9
estimated	QR / R32	0.8219	0.3975	225/49

The RMIA comparison is a render-score adaptation. It forms $p = \exp[-10^{-\mathrm{PSNR}/10}/T]$, averages reference probabilities arithmetically, and uses $p_{\mathrm{ref}} = ((1 + a)\overline{p}_{\mathrm{OUT}} + 1 - a)/2$. Candidate and anchor normalized log ratios use the same correction. The vote is the fraction of 32 clean anchors for which the candidate-to-anchor ratio exceeds γ strictly; tied votes are retained. Development on the twelve earlier W2 scenes selected $T \approx 0.003061$ and $a = 0.5$ from fifteen settings, with $\gamma = 2$ fixed; the Artifact retains the full-precision setting. The continuous normalized ratio is a separate ablation. Formal voting requires all 32 anchors and the specified K references to be valid. The K8 vote's cutoff is 0.90625; its supplied and estimated counts are 162/0 and 131/0. It does not saturate at an unusable top cutoff in this experiment.

The QR baseline is a low-capacity cross-scene adaptation of quantile prediction: a linear 0.99 quantile of target PSNR from 24 fixed image statistics, scene-balanced pinball loss and L1 regularization.

Its covariates contain RGB/intensity/gradient statistics, not target fidelity or membership. Development uses the earlier W2 scenes, with leave-one-scene-out nonmember pinball loss selecting regularization from $\{0, .001, .01, .1\}$. The selected .1 model has all 24 slopes zero. With R32 location adaptation, its score differs from RAW-R32 by a common constant, yielding the same ranking and independently calibrated decisions: 301/51 supplied and 225/49 estimated. This result characterizes the fitted linear adaptation; it is not a reproduction of the original image-encoder QMIA model. RAW/GAP/QR location adaptation requires at least 16 valid anchors.

G.5 Localization and geometric controls

All clean T candidates return estimated poses. Across eight conditions, the single rejection is a T3 nonmember under $+0.75$ EV clipping; it remains in the full denominator. Returned camera centers can nevertheless be extremely inaccurate. The diagnostic uses COLMAP world-to-camera poses, center $-R^\top t$, and center distance divided by the RMS radius of registered B-camera centers. Distances are in scene coordinates before normalization, not meters. PnP reprojection statistics use the returned inlier correspondences and HLOC's COLMAP $+0.5$ pixel convention.

A different T3 nonmember passes the finite-pose/minimum-inlier rule under $+0.75$ EV clipping but has normalized center error 1.0798×10^{15} and rotation error $61.4330°$, versus 0.001138 and $0.4252°$ on clean RGB. The extreme error originates in the returned pose, not a degenerate normalization radius. Despite this error, the stored mean reprojection residual is 1.61×10^{-12} pixels. Without the aligned raw correspondences, this value cannot establish camera accuracy or identify the failure's cause. The query retains its original log-photo score, approximately -6.53×10^{-7}, and remains in the full decision denominator. The Artifact retains the same-query geometry comparison.

The original seven pose features, MILP matcher with $\tau = .15$, and four-layer battery construct secondary Q_{bal} subsets after predictions are saved. All eight T subsets pass and retain 1,180 of 1,226 queries, balanced at 590 per class. C5 returns no pairs and no optimal-solution certificate, so its battery is unavailable. The helper maps an unsuccessful solver result or missing solution to this same record and did not retain the original solver status; it cannot distinguish infeasibility from a time limit here. C8's optimal 72-pair subset passes the permutation and energy layers but fails the pose-classifier layer (best classifier AUC 0.663194, permutation $p = .014925$). Both scenes remain in the original calibration pool, and full H remains the primary evaluation population.

Table 56: Fixed geometry-controlled test subsets, with 590 members and 590 nonmembers across the original eight T scenes. All use the original full-H eight-C clean cutoffs; the subset changes neither calibration nor the readouts.

Release / pose	Readout	AUC	nPAUC(.01)	TP/FP
native / supplied	PSNR-GAP	0.9580	0.4303	337/13
native / supplied	log-photo	0.9740	0.5863	376/10
native / estimated	PSNR-GAP	0.9248	0.3354	252/12
native / estimated	log-photo	0.9398	0.4851	299/11
SPZ / supplied	PSNR-GAP	0.9535	0.4153	336/15
SPZ / supplied	log-photo	0.9642	0.5362	332/8

On fixed Q_{bal}, the supplied-pose paired nPAUC difference is $+0.155979$, with scene-bootstrap 95% interval $[+0.022558, +0.258371]$. The companion AUC difference is $+0.015995$, interval $[+0.010507, +0.022616]$. Scores and cutoffs are inherited unchanged from the original full-H eight-C analysis; only the evaluation roster becomes the existing subset. Every selected query remains in its subset denominator. This provides geometry-controlled support on the selected population rather than replacing the full-H result.

A post-hoc descriptive check uses the saved localization fields of all 1,226 clean T candidates. Taking smaller rotation error, smaller normalized center error and higher B-camera density as member-oriented scores gives equal-scene AUCs of 0.5028, 0.4974 and 0.4961, respectively. All relevant fields are finite. These marginal summaries show little overall label discrimination; averaging can conceal scene-specific effects, and camera-center density measures sampling rather than surface visibility.

Pairing scores by query connects localization to the audit decision. From supplied to estimated poses, log-photo loses 82 member detections and gains two (387 to 307 TP); PSNR-GAP loses 87 and gains none (348 to 261). Their FP counts change from 11 to 12 and from 13 to 12. These transitions use each pose branch's own clean-C cutoff: log-photo's threshold changes from 0.479026 to 0.464330. The transitions therefore combine score changes and recalibration. The equal-scene mean paired log-photo score change is -0.144456 for members and -0.005914 for nonmembers. The Artifact includes the saved-data calculation for these descriptive summaries.

G.6 Calibration-scene sensitivity

Table 57: Calibration-pool sensitivity for the primary native/supplied clean scores. Each leave-one-C-out cutoff is applied to the same fixed 1,226 T queries; no fold selects a model or replaces the all-eight-C result. The last column is log-photo equal-scene FPR.

Omitted C	GAP cutoff	GAP TP/FP	Log-photo cutoff	Log-photo TP/FP	FPR
All eight C	3.31813	348/13	0.47903	387/11	1.788%
C1	3.70330	321/12	0.59508	325/5	0.816%
C2	3.31813	348/13	0.47903	387/11	1.788%
C3	3.70330	321/12	0.59508	325/5	0.816%
C4	3.70330	321/12	0.59508	325/5	0.816%
C5	3.70330	321/12	0.59508	325/5	0.816%
C6	3.03575	369/18	0.39176	443/21	3.442%
C7	3.70330	321/12	0.59508	325/5	0.816%
C8	3.70330	321/12	0.59508	325/5	0.816%

Calibration-pool sensitivity instead changes the C roster. Each fold omits one C scene, recomputes its finite-negative rank cutoff and applies it to the same full eight-scene T population. For primary log-photo, these eight folds give 325–443 TP and 5–21 FP, with equal-scene FPR 0.816–3.442%. Thus useful ranking coexists with material dependence of the decision point on calibration composition. No fold is selected as the primary rule, and the held-out C scene is not the evaluation set in this analysis. The Artifact retains the corresponding tables for other pose/release/readout branches and the per-query fit and localization records.

The full primary C pool has 620 finite negatives and rank index $j = 6$. Leaving one scene out gives 540–545 negatives and $j = 5$. Omitting C2 retains the original cutoff; omitting C6 lowers it to 0.391757, whereas the other six omissions yield 0.595084. The change therefore combines upper-tail composition with the discrete rank rule, rather than identifying a uniquely unsuitable calibration scene.

H Source exclusion, utility and deployment

Source-aware teacher routing reduces association in a single released student. We define the routing mechanism and the common source/utility population, then present the five-recipe audit, reconstruction value and deployment costs. Matched IN controls, independent teacher-bank replications and the local-pose probe provide supporting comparisons with their own populations and fitting identities.

H.1 Protected source membership and teacher routing

For a directly optimized asset, $M_{\mathrm{opt}}(q)$ records whether candidate RGB I_q provided photometric supervision. A distilled student receives teacher renders instead of protected candidate RGB, so its audited label is $M_{\mathrm{src}}(q)$: whether I_q supervised the upstream teacher bank. The registered roster, poses, intrinsics and auxiliary set A are fixed, unprotected inputs. The protected source set is $S \subset Q$.

A balanced code $E(q) \subset \{1, 2, 3, 4\}$ of size two assigns each source to two excluding teachers. Teacher G_k trains on A and sources satisfying $k \notin E(q)$. At a camera p associated with query q, the

three student targets are

$$\widetilde{I}_{\mathrm{OUT}}(p;q) = \tfrac{1}{2} \sum_{k \in E(q)} \mathcal{R}(G_k; p),$$
$$\widetilde{I}_{\mathrm{IN}}(p;q) = \tfrac{1}{2} \sum_{k \notin E(q)} \mathcal{R}(G_k; p),$$
$$\widetilde{I}_{\mathrm{G}}(p;q) = \tfrac{1}{2} \sum_{k \in P(q)} \mathcal{R}(G_k; p).$$

OUT routes each source through teachers that omit its RGB; the complementary IN route is a source-aware positive control whose selected teachers saw that source. Nonsource candidates supply no RGB to any teacher. GENERIC assigns a fixed pair $P(q)$ without reading source codes, labels or outcomes. This control separates the contribution of source-aware routing from ordinary two-teacher distillation. The deployed asset is one student PLY; teachers are used during construction.

H.2 Common source and utility populations

The common-utility comparison spans playroom, truck, kitchen and drjohnson. Before model construction, camera-center farthest-point selection fixes held-out U, followed by candidate Q, using camera metadata. All scenes have $|Q| = 72$. In that scene order, $|U| = 45/50/55/52$ and $|A| = 108/129/152/139$. Each target bank has 36 protected sources $S \subset Q$ and 36 nonsources. Every model within a scene shares its A-only initialization. U is excluded from initialization, photometric supervision, pseudo-target generation, routing and attack selection. Released camera calibration remains shared metadata.

DIRECT7k, DIRECT30k, GENERIC30k and OUT30k use the same protected sources. DIRECT receives original source RGB and is evaluated against M_{opt}. Students receive teacher targets and are evaluated against M_{src}. GENERIC/OUT share the exact $A \cup Q$ camera roster, initialization, optimizer seed and 30k schedule. Their A-camera target averages all four teachers; Q-camera targets average the generic or excluding pair. Pseudo-targets average individually clamped float32 renders and are stored without quantization.

Each scene contains eight shadow banks and two target banks, with four teachers in each bank. Two target optimization seeds are nested within each target bank. The target banks vary both the source assignment and teacher realization; they are paired GENERIC/OUT realizations within a scene. The 7k and 30k DIRECT endpoints belong to one trajectory.

H.3 Five student recipes under a common audit

The common-U design in Section H.2 evaluates five recipes on each of four scenes, two target banks and two nested student seeds: 80 assets in total. GENERIC and OUT retain their existing targets and receive newly matched reference banks and attack fits. ALL4 uses the mean of all four teacher renders at both auxiliary and candidate cameras. GEN-realA and OUT-realA replace the auxiliary A slots with real RGB while preserving their corresponding GENERIC or source-excluding rule at Q. All use auxiliary-only initialization and the shared 30k schedule. Individual teacher renders are clamped before their mean is formed.

Each asset has nine configurations: RAW, BaVarIA-n, GAP, T, RR8, CNN seeds 90851 and 90852, log-L1 and log-photo. The two CNN seeds are separate configurations of one family. Each recipe has matched attack fitting: six shadows for development, two for validation and eight for final refitting; validation BCE selects CNN epochs. Target AUC does not select the fit. The 720 asset–configuration measurements each contain 36 members and 36 nonmembers. Maxima are taken within assets before averaging student seeds, banks and scenes. Consequently A_9^* summarizes a fixed measured set and is not an attack selectable without evaluation labels.

Table 58: Common nine-configuration maximum AUC by scene. Each value averages the two student seeds and then the two banks, after taking the configuration maximum separately for each asset. Column groups pair teacher routing under rendered or real auxiliary supervision; ALL4 is the control.

	Rendered auxiliary RGB		Real auxiliary RGB		Control
Scene	GENERIC	OUT	GEN-realA	OUT-realA	ALL4
playroom	0.9172	0.6966	0.9030	0.6501	1.0000
truck	0.9531	0.7492	0.9236	0.5988	1.0000
kitchen	0.9655	0.8590	0.8953	0.6379	0.9830
drjohnson	0.9470	0.6408	0.9400	0.6154	1.0000

Table 59: Fixed-configuration AUC under all five recipes, averaged within student seeds and banks and then equally across four scenes. Each recipe fits its own matched attacks. Columns use the same routing pairs as Table 58; these means differ from taking a maximum within each asset.

	Rendered auxiliary RGB		Real auxiliary RGB		Control
Configuration	GENERIC	OUT	GEN-realA	OUT-realA	ALL4
RAW	0.7344	0.5448	0.7211	0.5311	0.7626
BaVarIA-n	0.9045	0.6265	0.8821	0.5191	0.9950
GAP	0.9285	0.6800	0.8986	0.5854	0.9905
T	0.9226	0.6852	0.8993	0.5907	0.9847
RR8	0.7413	0.5568	0.7298	0.5222	0.7808
CNN / 90851	0.8570	0.5332	0.7823	0.5348	0.9634
CNN / 90852	0.8033	0.5137	0.8155	0.5013	0.9823
log-L1	0.9167	0.6609	0.8849	0.5747	0.9859
log-photo	0.9440	0.7314	0.9076	0.6102	0.9922

The equal-scene A_9^* values are 0.9457, 0.7364, 0.9958, 0.9155 and 0.6255 for GENERIC, OUT, ALL4, GEN-realA and OUT-realA. OUT lowers every fixed configuration AUC relative to both GENERIC and ALL4 in all sixteen asset pairs. OUT-realA versus OUT has 119 decreases, two ties and 23 increases among the 144 fixed-configuration comparisons, while all sixteen asset maxima decrease. Against GEN-realA, OUT-realA has 141 decreases and three increases, all three from the same CNN configuration in kitchen. The four scenes, rather than these nested comparisons, are the independent units.

Table 60: Bank variation under the common nine-configuration audit. Each upper-panel entry averages two per-asset maxima within one target bank; it does not maximize an already averaged attack AUC. All five recipes share the same bank and nested-seed identities. The lower panel identifies the highest-residual OUT and OUT-realA assets within each scene and their attaining configurations. Banks and seeds are nested within the four independent scenes.

Bank-mean maximum AUC

Scene	Bank	GENERIC	OUT	GEN-realA	OUT-realA	ALL4
playroom	00	0.9236	0.6485	0.8997	0.5883	1.0000
	01	0.9109	0.7446	0.9062	0.7118	1.0000
truck	00	0.9348	0.7095	0.9279	0.5972	1.0000
	01	0.9715	0.7890	0.9194	0.6003	1.0000
kitchen	00	0.9660	0.8468	0.8997	0.6281	0.9900
	01	0.9649	0.8711	0.8908	0.6478	0.9761
drjohnson	00	0.9163	0.6671	0.9213	0.6296	1.0000
	01	0.9776	0.6146	0.9587	0.6011	1.0000

Highest residual asset in each scene

Scene	Recipe	Maximum AUC	Bank	Seed	Attaining configuration
playroom	OUT	0.7569	01	1	log-photo
playroom	OUT-realA	0.7153	01	1	GAP
truck	OUT	0.7986	01	1	log-photo
truck	OUT-realA	0.6188	00	0	log-photo
kitchen	OUT	0.8796	01	0	log-photo
kitchen	OUT-realA	0.6636	01	1	log-photo
drjohnson	OUT	0.6728	00	0	log-photo
drjohnson	OUT-realA	0.6350	00	1	T

Table 60 resolves the bank variation behind the scene means and identifies the highest-residual OUT and OUT-realA assets in each scene. Every bank mean uses the same two nested seeds as its paired recipes; the attack maximum is taken within each asset first.

The realA variants replace the all-teacher renders at A while preserving their respective routing rules at Q. Teachers selected at another candidate camera q' can have trained on q, leaving a possible path for source influence through other views. The matched realA comparison measures the effect of changing auxiliary supervision on the resulting student.

The old seven-configuration means, GENERIC 0.9356 and OUT 0.6995, use their original reference realization and fits. Reevaluating the same retained targets with new matched fits gives $A_7^* = 0.9334$ and 0.7038; adding log-L1/log-photo gives $A_9^* = 0.9457$ and 0.7364. Log-photo attains the measured maximum on 13 of 16 OUT assets. OUT's kitchen scene mean rises from 0.7762 to 0.8590 when the two readouts enter, and its largest asset AUC is 0.8796. The earlier two-scene RR40 local-response experiment is preserved in Section H.8; RR40 is not part of the new common nine-configuration set.

Individual configurations can move against the asset maximum: on playroom bank01/seed0, CNN seed90851 rises from 0.351080 to 0.581790 under OUT-to-OUT-realA, while the asset maximum falls from 0.732253 to 0.708333 and that CNN's nPAUC(.01) falls from 0.027778 to zero. These recipes fit different networks and retain their frozen score directions.

Direction sensitivity. For each asset, a separate post-hoc diagnostic takes $\max_m \max(a_m, 1 - a_m)$, where a_m is the saved AUC of configuration m. It permits either score orientation using evaluation labels while keeping every fitted model and measurement fixed.

Table 61: Post-hoc direction sensitivity of the same nine-configuration student audit. The symmetric summary permits either orientation of each saved score; both columns use evaluation labels, asset-wise maxima and equal-scene averaging. The primary score directions and fitted models remain fixed.

Recipe	Primary A_9^*	Direction-symmetric maximum
GENERIC	0.9457	0.9457
OUT	0.7364	0.7390
ALL4	0.9958	0.9958
GEN-realA	0.9155	0.9155
OUT-realA	0.6255	0.6295

The direction-symmetric mean is 0.739005 for OUT and 0.629533 for OUT-realA (Table 61). GENERIC-to-OUT and GEN-realA-to-OUT-realA both still decrease in all sixteen asset pairs. OUT-to-OUT-realA decreases in fifteen pairs: on drjohnson bank01/seed0, OUT's maximum is 0.611883, while OUT-realA's BaVarIA-n AUC of 0.351852 becomes 0.648148 when reversed. Its original fixed-direction maximum is 0.604167. This exception belongs to the direction-symmetric diagnostic; the sixteen decreases in the primary fixed-direction asset maxima remain unchanged.

The per-asset tail point uses a strict threshold above the maximum target-negative score. Thus all reported FP counts at that point are zero by construction. For example, OUT kitchen bank00/seed1 log-L1 detects 16/36 members, while OUT-realA playroom bank01/seed0 BaVarIA-n detects 10/36. These reveal residual separability under a retrospective rule. Thirty-six negative observations and interpolated nPAUC(.01) do not establish a stable prospective 1% operating point.

H.4 A-ONLY RECONSTRUCTION REFERENCE

An A-only release measures the reconstruction utility available from auxiliary RGB without using protected sources or a teacher bank. We fix the four actual optimization seeds used by the target students in each scene, giving sixteen models. Each uses the same auxiliary-only initialization, scene normalization, calibrated cameras and native 3DGS optimization settings as its matched students. Training samples uniformly without replacement from the A camera pool, refilling the pool when exhausted, for 30k native iterations. Every sampled camera provides real A supervision; the schedule contains no empty Q slots. The Artifact records the actual seeds, configurations and sampled camera sequence.

Evaluation uses the same 45/50/55/52 held-out U views for playroom/truck/kitchen/drjohnson, with PSNR, SSIM and LPIPS-Alex v0.1. Pairing by scene, seed and view gives 808 measurements per strategy comparison. We average within each asset, then over the four seeds within each scene, and finally equally over the four scenes. GENERIC, OUT, OUT-realA and DIRECT30k retain their original measurements on these views. This comparison evaluates complete release strategies: A-only trains on A, DIRECT on A and protected source RGB, and the students on their respective $A \cup Q$ targets. The matched GENERIC–OUT comparison separately isolates teacher routing. A-only supplies a utility reference, with no teacher-source membership AUC assigned.

Table 62: Common-U reconstruction utility of complete release strategies, including the DIRECT7k early-stopping reference; all other rows use 30k iterations. Rows average the retained seeds and banks within one scene, using the same 45/50/55/52 U views for playroom/truck/kitchen/drjohnson. A-only trains on real A RGB with ordinary A sampling; DIRECT trains on real A and protected sources. The student recipes train on their respective A and Q targets. PSNR is in dB; LPIPS uses AlexNet v0.1. Scenes, rather than individual assets or views, are the four independent units.

Scene	Release	PSNR	SSIM	LPIPS
playroom	A-only	27.378	0.8920	0.1741
	GENERIC	28.583	0.9076	0.1694
	OUT	28.344	0.9051	0.1730
	OUT-realA	28.206	0.9037	0.1708
	DIRECT30k	28.799	0.9070	0.1598
	DIRECT7k	28.392	0.9015	0.1946
truck	A-only	23.277	0.8343	0.1352
	GENERIC	24.296	0.8587	0.1391
	OUT	24.085	0.8551	0.1415
	OUT-realA	24.031	0.8551	0.1340
	DIRECT30k	24.325	0.8611	0.1203
	DIRECT7k	23.092	0.8266	0.1893
kitchen	A-only	30.078	0.9220	0.1002
	GENERIC	30.568	0.9266	0.1025
	OUT	30.359	0.9255	0.1034
	OUT-realA	30.336	0.9254	0.1009
	DIRECT30k	30.621	0.9260	0.0972
	DIRECT7k	28.727	0.9052	0.1271
drjohnson	A-only	27.013	0.8338	0.2645
	GENERIC	27.860	0.8564	0.2592
	OUT	27.555	0.8504	0.2702
	OUT-realA	27.511	0.8479	0.2657
	DIRECT30k	27.846	0.8517	0.2426
	DIRECT7k	26.923	0.8321	0.3551

Table 63: Equal-scene utility differences relative to A-only on the same model-seed/U pairs. Each comparison uses 808 measurements across four scenes. Positive PSNR/SSIM and negative LPIPS favor the listed release. Matching seeds and initialization does not equate the strategies' different training-camera populations.

Release minus A-only	ΔPSNR (dB)	ΔSSIM	ΔLPIPS
GENERIC	+0.891	+0.0168	-0.0009
OUT	+0.649	+0.0135	+0.0035
OUT-realA	+0.585	+0.0125	-0.0007
DIRECT30k	+0.961	+0.0159	-0.0135

DIRECT7k supplies the early-stopping utility reference in Table 62. Relative to it, OUT gains 0.802 dB PSNR and 0.01767 SSIM and reduces LPIPS by 0.04447 in equal-scene means; playroom PSNR is 0.049 dB lower. The original A7 audit gives DIRECT7k maxima 0.9996/0.9954/0.9448/1.0000 for playroom/truck/kitchen/drjohnson. Those fitted attacks retain their historical identity and are not relabeled as the current A9 audit. DIRECT/student comparisons include their different training-camera populations and real versus synthetic supervision.

A-only reaches equal-scene PSNR 26.936 dB, SSIM 0.8705 and LPIPS 0.1685. OUT improves PSNR by 0.649 dB and SSIM by 0.01349; OUT-realA improves them by 0.585 dB and 0.01251. Both gains occur in all four scenes for both metrics. OUT-realA's scene-level PSNR gains range from 0.258 dB on kitchen to 0.828 dB on playroom. Perceptual changes differ: OUT increases mean LPIPS by 0.00354, whereas OUT-realA reduces it by 0.00066, with improvements on playroom and truck and increases on kitchen and drjohnson. GENERIC and DIRECT30k exceed A-only by 0.891 and 0.961 dB, respectively. These complete-strategy comparisons establish reconstruction gains beyond auxiliary-only training alongside the students' measured residual source association.

H.5 UTILITY DISTRIBUTIONS AND PAIRED VISUAL EVIDENCE

Held-out U lists contain 45 playroom, 50 truck, 55 kitchen and 52 drjohnson views. Pairing the same scene, bank, seed and view yields 808 actual asset–view comparisons for each recipe contrast. A second summary first averages the four bank/seed differences per camera, giving 202 camera observations. Every equal-scene statistic weights the four scenes equally, regardless of their U counts.

Table 64: View-level PSNR changes for matched OUT minus GENERIC. Bank/seed deltas are averaged for each U view before its scene distribution is summarized. The small mean cost coexists with spatially heterogeneous changes; the final column counts views with positive PSNR change. Full SSIM/LPIPS and asset/bank distributions are retained in the Artifact.

Scene	N_U	Mean	Min.	P10	P90	Max.	Improved
playroom	45	-0.240	-2.095	-1.341	0.399	4.269	10
truck	50	-0.211	-0.847	-0.637	-0.002	0.225	5
kitchen	55	-0.208	-2.787	-0.539	0.085	0.164	17
drjohnson	52	-0.305	-1.823	-0.619	0.007	0.132	6

OUT minus GENERIC costs 0.241294 dB in equal-scene mean PSNR. Losses above 1 dB occur in 62/808 actual pairs (7.6733% pooled, 8.0166% equal-scene). After averaging realizations by view, the equal-scene fraction is 6.1393%. The worst actual pair loses 5.5536 dB in kitchen (Fig. 6); the worst averaged view loses 2.7875 dB. ALL4 gains 0.4057 dB over OUT while retaining stronger source association.

OUT-realA minus OUT changes equal-scene PSNR by -0.064634 dB, SSIM by -0.000979 and LPIPS by -0.004198. Its 18/808 actual pairs with losses above 1 dB correspond to 2.2277% pooled or 2.3598% equal-scene. The three such averaged views give 1.4851% pooled or 1.5657% equal-scene. These two weighting conventions are kept distinct.

The worst actual OUT-realA minus OUT pair (Fig. 7) and the worst averaged camera are both playroom/`DSC05773.jpg`. In bank00/seed0, bank00/seed1, bank01/seed0 and bank01/seed1, the PSNR changes are -6.3729, $+0.4359$, -6.7046 and -1.8515 dB, averaging -3.6233 dB. Three of the four realizations lose utility at this camera; one improves. A small mean recipe cost therefore coexists with large localized changes.

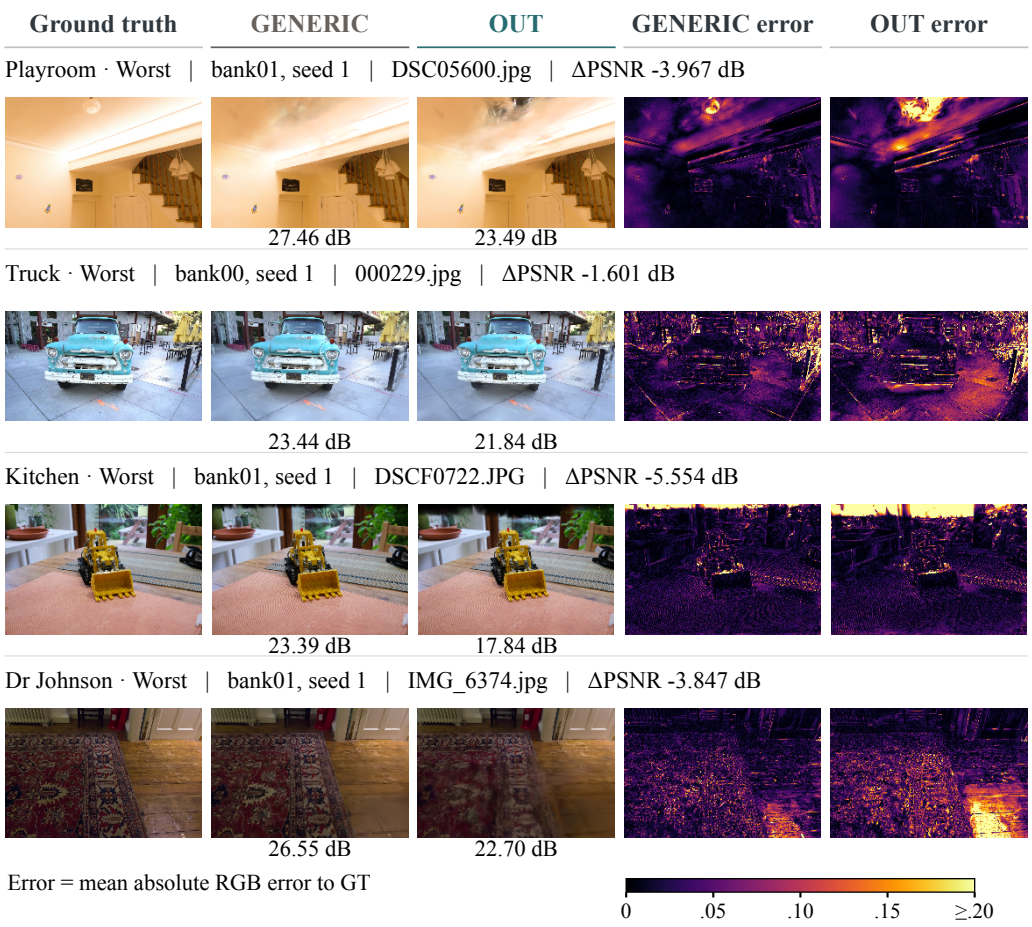


Figure 6: Local PSNR losses under source-exclusion routing. Each row shows the actual bank/seed/U-view pair with the smallest OUT-minus-GENERIC PSNR difference in that scene, retaining matched GT and student renders as complete, uncropped frames. Mean absolute RGB residuals share a $[0, 0.20]$ scale, saturated at 0.20. These cases complement the mean utility result; the Artifact retains the full fixed-quantile selection.

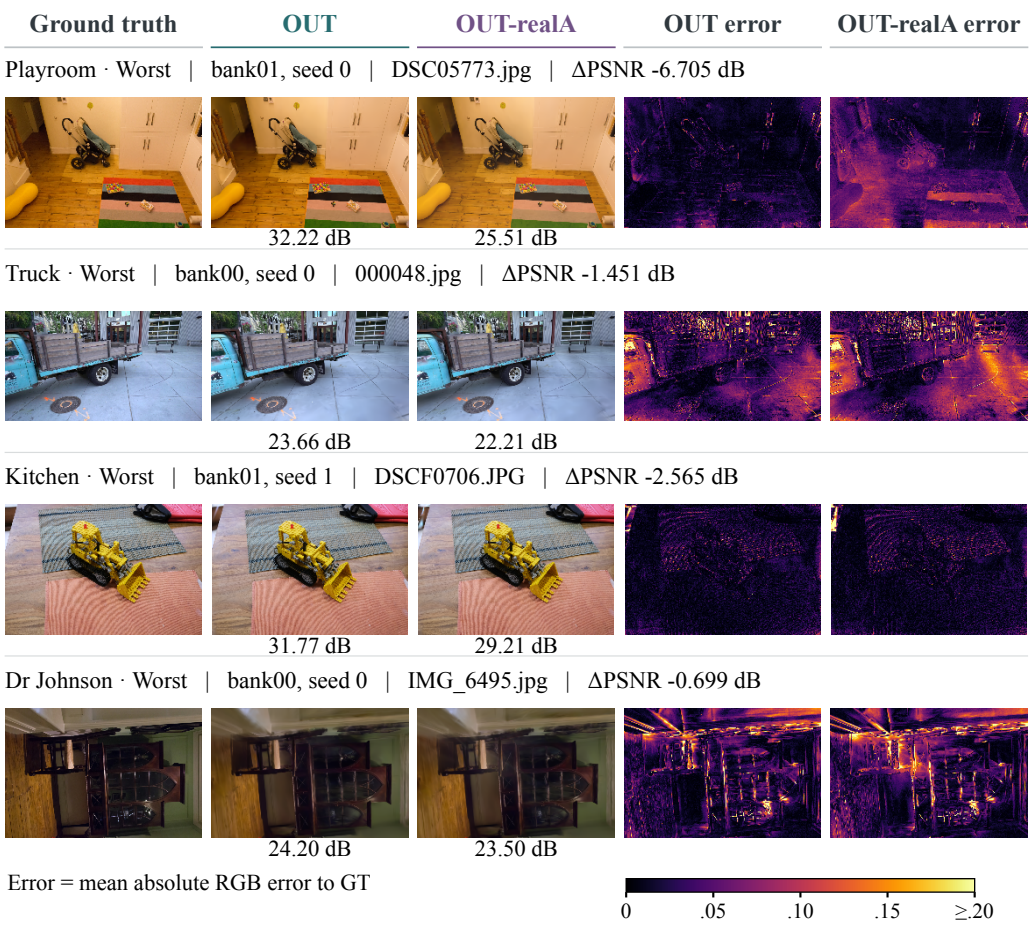


Figure 7: Local PSNR losses when replacing rendered auxiliary supervision with real RGB. Each row shows the actual bank/seed/U-view pair with the smallest OUT-realA-minus-OUT PSNR difference in that scene. Full frames and the shared $[0, 0.20]$ mean absolute RGB error scale are retained, with saturation at 0.20. A small mean PSNR cost can thus coexist with larger individual losses; the Artifact retains the full fixed-quantile selection.

H.6 CONSTRUCTION AND DEPLOYMENT

The four-scene mean OUT construction time is 119.74 minutes from scratch and 23.28 minutes when its teachers are available. From-scratch time includes four teacher training tasks, pseudo-image generation and student training; teacher-reused time includes pseudo-image generation and student training. The from-scratch cost is 4.92 times DIRECT30k. Joint GENERIC/OUT pseudo-generation is charged in full to each standalone release and shared when constructing both, so summing the two standalone rows would double-count that shared work. Task times can overlap and do not equal queue elapsed time.

OUT remains a single standard student PLY, averaging 415.85 MB and 138.1 fixed-query CUDA FPS on the recorded RTX 3090 hardware. Here MB is 10^6 bytes and minutes are recorded seconds divided by 60. Comparing the four-scene means with DIRECT30k gives a file-size reduction of 16.3% and FPS increase of 12.2%. Relative to GENERIC, the corresponding changes are 1.9% and 1.7%. Most of the deployment benefit is shared by the distilled releases; source-aware routing adds the matched reduction in source association. The artifact retains construction components, Gaussian counts, hardware identity and per-asset utility/cost rows.

Table 65: Original four-scene construction and deployment measurements, averaged over seeds, banks and scenes. Scratch/reuse includes/excludes teacher training; both student totals include the full joint pseudo-image cost. Task minutes can overlap and are not elapsed campaign time. MB is 10^6 bytes. CUDA FPS uses RTX 3090 at each asset's first metadata-ordered query camera, with three warmups and 30 timed renders, excluding image I/O.

Release	U PSNR (dB)	Scratch/reuse (min)	MB	FPS
DIRECT30k	27.898	24.32/24.32	496.79	123.1
GENERIC30k	27.827	119.84/23.38	423.98	135.7
OUT30k	27.586	119.74/23.28	415.85	138.1

H.7 ROUTING CONTROLS AND INDEPENDENT-BANK REPLICATION

In the initial and independent-bank studies, the student uses real A images and teacher targets at each registered non-outer Q camera plus 1.0° rotation and 0.2 median-inter-camera-spacing translation perturbations. All three arms share the A-only initialization, ordered camera roster, seed, 30k schedule, SH degree 3 and black background. GENERIC permutes canonical query names and cycles the six teacher pairs, reusing the map across shadows and target seeds. Exact and nearby cameras share the same pair. These supporting studies use real auxiliary RGB and perturbed Q cameras, unlike the common-U recipes in Section H.2.

Initial matched comparison. Counter is the development scene and bonsai the prespecified confirmation, each with two paired target seeds. Counter has 144 auxiliary views, 40 scored sources, 40 scored nonsources, eight additional sources and eight outer views, yielding 408 student cameras. Bonsai has 176 auxiliary views, 48 scored sources, 48 scored nonsources, 12 additional nonsources and eight outer views, yielding 500 cameras. The additional queries preserve the training recipe but do not enter AUC.

Each arm fits its own S8/BaVarIA-n reference from eight defense-matched shadow assignments, with each scored candidate source-member in four assignments and nonsource in four. Exact-camera PSNR is the primary endpoint. The local endpoint averages PSNR at the exact pose and held-out 0.5° rotation and 0.1-spacing translation poses before fitting. These correlated poses form one supporting endpoint.

Table 66: Initial source-association AUCs, with two paired target seeds per scene. Exact columns use arm-matched S8/BaVarIA-n; local G/OUT average the three correlated evaluation poses before fitting. IN is the source-aware positive control and G is GENERIC. Full local IN and operating-point records are archived with the original tables.

Scene	Seed	N_+/N_-	IN exact	G exact	OUT exact	G local	OUT local
counter	0	40/40	0.969375	0.892500	0.634375	0.865625	0.595000
counter	1	40/40	0.985625	0.905000	0.636250	0.876875	0.577500
bonsai	0	48/48	0.896701	0.831163	0.631944	0.798611	0.619792
bonsai	1	48/48	0.901042	0.852431	0.643229	0.818576	0.631076

The equal-scene exact IN-to-OUT AUC change is $0.938186 \rightarrow 0.636450$. All four GENERIC-minus-OUT effects are positive, with scene means 0.263438 on counter and 0.204210 on bonsai, exceeding the prespecified 0.05 increment. At the retrospective at-most-5% empirical-FPR point, GENERIC-to-OUT TPR changes from $0.625/0.775$ to $0.225/0.175$ on counter and from $0.458/0.521$ to $0.167/0.146$ on bonsai. One false positive is 2.5% and 2.083%, respectively. Exact OUT AUC remains 0.632–0.643; the four two-sided Mann–Whitney results have uncorrected $p = 0.03854/0.03586/0.02569/0.01528$, and none passes Holm correction at 0.05.

On eight fixed outer views per scene, the equal-scene OUT-minus-GENERIC changes are -0.0752 dB PSNR, -0.002285 SSIM and $+0.001196$ LPIPS. These pass the frozen aggregate bands of -0.10 dB, -0.003 and $+0.002$. Individual changes vary: both counter PSNR effects exceed the aggregate loss band, and counter seed 1 also exceeds the SSIM and LPIPS bands. OUT-minus-IN averages -0.0483 dB/-0.00253/$+0.00137$ for outer PSNR/SSIM/LPIPS, with perceptual degradation in all four pairs. These are incremental routing comparisons on the same distilled recipe.

Independent teacher/codebook banks. Playroom provides a development replication and truck a cross-dataset stress case. Each scene has one shadow-fit bank with eight assignments and two independent target banks, each containing four teachers and two nested student seeds. The camera population and A-only initialization stay fixed across banks. The shadow-fit bank supplies attack training rather than an additional target replication.

Playroom has 135 A and 90 Q views, including 48 target sources, 34 nonsources and eight outer views; truck has 151 A and 100 Q, including 48 sources, 44 nonsources and eight outer views. Their student rosters contain 381 and 427 cameras. The scored cohorts contain 30/30 and 40/40 source/nonsource queries. Arm-matched S8/BaVarIA-n fitting precedes target scoring. Scene is the highest independent unit; teacher banks are realizations within scene and student seeds are nested within bank.

Table 67: Independent target-bank replication: every exact-view S8/BaVarIA-n AUC. Seeds are nested within a teacher/codebook bank. Residual marks the original two-sided, same-direction OUT association flag; no OUT row is significantly reversed.

Scene	Bank	Seed	IN	G	OUT	Residual
playroom	0	0	1.0000	0.8867	0.4978	no
playroom	0	1	1.0000	0.8822	0.5089	no
playroom	1	0	1.0000	0.8956	0.4567	no
playroom	1	1	1.0000	0.8889	0.4978	no
truck	0	0	1.0000	0.8906	0.6163	no
truck	0	1	0.9981	0.9025	0.6438	yes
truck	1	0	1.0000	0.9794	0.6450	yes
truck	1	1	1.0000	0.9669	0.6594	yes

GENERIC-minus-OUT exact AUC differences are positive for all eight target pairs. Bank-mean reductions are 0.3811/0.4150 on playroom and 0.2666/0.3209 on truck, giving scene means 0.3981 and 0.2938. All IN controls meet the frozen one-sided 5% exact Mann–Whitney rule and have positive source-minus-nonsource margins. Three truck OUT rows retain same-direction association under the two-sided rule. The local endpoint agrees in every pair, with scene-mean reductions 0.3728/0.3219. At the retrospective at-most-5% FPR point, OUT TPR spans 0–0.067 on playroom

and 0.100–0.275 on truck. Thresholds maximize TP within the FP budget, retain the smallest FP count on ties and never split equal-score groups; no exact panel has tied scores.

The independent-bank utility cost is larger than in the first study. Across banks and nested seeds, OUT-minus-GENERIC outer PSNR/SSIM/LPIPS is -0.397326 dB/-0.001717/$+0.005490$ on playroom and -0.238792 dB/-0.002993/$+0.001643$ on truck. Neither scene passes all three original utility bands. PSNR worsens on 25/32 playroom and 28/32 truck paired view measurements, with worst losses of 1.3803 and 0.7807 dB. The eight fixed outer views remain shared across banks and seeds.

Complementary attack diagnostics. The earlier raw-pair CNN on playroom/truck uses candidate and same-pose render RGB with seeds 90801/90802. Five of eight scene–arm–seed fits give constant logits and AUC 0.5; the remaining fits do not establish stable GENERIC discrimination. RR8 has a discriminating GENERIC control, while truck GAP/T retains residual AUC around 0.75. The normalized candidate/residual CNN in the common-U study has different target-bank and utility designs, so these studies do not isolate representation alone. The Artifact preserves the complete raw-score screen, early-stopping comparisons and learned-attack diagnostics.

H.8 Local-pose residual responses as an additional attack

On kitchen and drjohnson, RR40 extends RR8 with a fixed five-pose response: the base camera and yaw/pitch rotations of $\pm 0.25^\circ$ about the camera's local axes. Camera center and intrinsics stay fixed, and each render is compared with the same original candidate RGB. Concatenating the eight residual statistics in pose-major order gives forty features. A training-shadow standardizer and logistic regression with $C = 1$ and at most 1,000 iterations use the same six/two/all-eight fit, validation and refit partition. Probe choice, feature order and fitting recipe are fixed before target evaluation.

Table 68: Fixed local-response AUC comparison on kitchen and drjohnson using the earlier seven-configuration fits; Δ is RR40 minus RR8. RR40 increases GENERIC discrimination and is effective on DIRECT positive controls, while OUT remains attenuated. A_8^* is the per-asset maximum AUC including RR40 and equals A_7^* here. Entries average seeds and banks; this two-scene extension is separate from the current nine-configuration audit in Section H.3.

Release	RR8	RR40	Δ	A_8^*
kitchen				
DIRECT7k	0.7130	0.7610	0.0480	0.9448
DIRECT30k	0.8989	0.9408	0.0419	1.0000
GENERIC30k	0.6848	0.7122	0.0274	0.9495
OUT30k	0.5870	0.5667	-0.0203	0.7766
drjohnson				
DIRECT7k	0.9624	0.9823	0.0199	1.0000
DIRECT30k	0.9985	1.0000	0.0015	1.0000
GENERIC30k	0.8030	0.9028	0.0997	0.9323
OUT30k	0.5965	0.5694	-0.0270	0.6372

The added local responses improve the GENERIC positive controls: RR8-to-RR40 mean AUC rises from 0.6848 to 0.7122 on kitchen and from 0.8030 to 0.9028 on drjohnson. OUT gives RR40 AUC 0.5667 and 0.5694, below GENERIC in all eight matched target assets. Including RR40 in the per-asset maximum leaves the two-scene A_8^* equal to A_7^*: 0.9409 for GENERIC and 0.7069 for OUT. RR40 augments the earlier seven-configuration fits on these two scenes; it is separate from the common nine-configuration audit in Section H.3.

H.9 Output perturbations at a served-render interface

Output perturbation is applied by a rendering service. We rerender the fixed targets and shadows under 13 output/query-budget configurations on each of counter and truck, giving the 26 scene–configuration evaluations in Table 69. Models are not retrained. Global-$K8$ LiRA has full query and shadow coverage. Jitter level h scales random-axis rotation to $0.5h^\circ$ and translation to $0.10h$

median inter-camera spacing. Repeated queries average per-draw PSNR before LiRA fitting and scoring, with budget N. Downsampling uses bilinear resize down and back to the comparison canvas; quantization uses the stated bit depth, and additive Gaussian noise uses standard deviation 0.05 before clamping to $[0, 1]$. This interface differs from candidate-image transformations in Section F.

Table 69: Thirteen output/query-budget configurations on each of counter and truck, giving 26 scene–configuration evaluations with global-$K8$ LiRA. Columns give AUC, retrospective TPR at nominal 1% FPR (@1%), and nonmember render PSNR in dB. The historical @1% rule uses descending negative rank $\lceil 0.01N_{-} \rceil$ and $\geq$ comparison. N is the query budget. SSIM and the nominal 5% point accompany the complete original table in the Artifact.

	Counter			Truck		
Configuration	AUC	@1%	PSNR	AUC	@1%	PSNR
none	0.991	0.688	28.35	1.000	0.991	24.17
pose-jitter 0.25	0.962	0.532	23.46	0.990	0.965	20.78
pose-jitter 0.5 (N=1)	0.910	0.330	21.09	0.980	0.761	18.37
pose-jitter 0.5 (N=4)	0.934	0.468	21.07	0.987	0.929	18.47
pose-jitter 0.5 (N=16)	0.948	0.422	21.02	0.987	0.947	18.47
pose-jitter 1.0	0.835	0.184	18.78	0.927	0.319	16.55
downsample 0.75	0.989	0.679	27.80	0.998	0.991	24.04
downsample 0.5	0.987	0.679	27.18	0.995	0.982	23.55
downsample 0.25	0.979	0.624	25.10	0.991	0.982	21.81
quant 4-bit	0.991	0.670	27.25	1.000	0.991	23.76
quant 3-bit	0.989	0.633	24.87	0.996	0.991	22.54
quant 2-bit	0.987	0.560	19.64	0.991	0.991	18.83
noise 0.05 (sanity)	0.989	0.615	24.08	0.994	0.991	22.01

Pose jitter attenuates the readout at a substantial utility cost. At jitter 0.5, increasing the query budget from 1 to 16 raises counter AUC from 0.910 to 0.948 and truck AUC from 0.980 to 0.987. Downsampling and quantization preserve stronger membership discrimination, especially on truck. These interventions modify API responses; a directly released PLY or checkpoint can be rendered without them.

I Reproduction and Resources

I.1 Executable Paths and Retained Scientific Inputs

The Artifact provides numerical replay from saved evidence and scientific entry points from public RGB. Running `python -B replay.py` from the package root reconstructs applicable reference fits, calibrated decisions and summaries under `outputs/`. The public-RGB path uses frozen splits, public image identities, ordered query crosswalks and released cameras; `method/README.md` describes dataset preparation and the compatible GPU implementations. `method/run.py` exposes target/reference, four-arm, panel, localization, learned-attack and source-exclusion execution.

Saved model-level PSNR, L1, SSIM and photometric errors support score, ROC, threshold and decision reconstruction. W3 calibration is reconstructed from saved C scores, retaining its eleven pools, five target panels and two pose modes. Pixel-level gain fits and condition-specific poses retain their measured values. W4 has no CNN readout; the separate student's learned-readout replay evaluates saved predictions rather than refitting networks. Fit parameters, mask support and fallback records describe the executed image fits without rerunning their pixel optimization. Native RGB, model weights and feature caches must be obtained or generated separately; numerical replay needs none of them.

The public path adapters have passed CPU input, design and original-schedule checks; a second-machine GPU rerun of these adapters has not been performed. Fresh reproduction of the final student study constructs 400 models, including 64 earlier assets reused by the original 336-trajectory extension. The A-only comparison adds sixteen models. These counts identify fresh-run requirements; the public-RGB executions below cover selected existing settings.

Table 70: What the retained evidence supports. The full-results guide records these calculation levels alongside the complete historical tables.

Evidence	Executable calculation and retained scope
Reference/metric scalars	Refit applicable PSNR reference estimators and reconstruct C rules, RAW/GAP/log-photo rankings and decisions. Stored pixel metrics determine the implemented errors; PSNR implies the floored effective MSE used by RMIA.
W3 photometric evidence	Rebuild C rules from saved C scores and query labels, then reconstruct T rankings, fixed decisions and paired summaries. Gain fits and condition-specific poses retain their measured GPU-pipeline values.
Four-scene source exclusion	Reconstruct attack metrics from saved predictions and analytic scores, then reduce paired utility and measured costs. Take the per-asset maximum before seed/bank/scene averaging; learned predictions retain their saved fitting identity.
Learned and original scores	Re-evaluate saved CNN/RR8 predictions and original fitted $K8$ scores. The original ten-scene single-feature replay retains its $k = \lceil \alpha n_{-} \rceil$ threshold and $s \geq c$ convention.
Initial matched GENERIC	Saved fitted scores and raw utility reproduce the initial comparison in Section H.7; its original generic-shadow PSNR matrix is unavailable.
Original L4 records	Check ten saved pose-classifier statistics and their 1,054-per-class total. This does not refit the classifier or permutations, or recover the different high-fraction manifests.
Historical output gaps	Original LiRA pose-perturbation Tier2 shadow outputs remain unrecovered. The Artifact records the additional source gaps for migrated high-fraction and parameter/attribution studies, with their historical values and evidence levels.
RGB execution	Run the published initialization/training/rendering/localization methods with external RGB, cameras, checkpoints where required and a GPU stack. The CPU package contains measurements, not those large pixel/model inputs.

The Artifact's `docs/EVIDENCE_GUIDE.md` maps claims to inputs and replay outputs, and `docs/FULL_RESULTS.md` indexes complete historical records. `docs/APPENDIX_MIGRATION.md` locates the detailed tables, diagnostics and resource records moved from the paper. Those records retain their original populations, fitting identities and source coverage. Private residential imagery, filenames, geometry and model assets remain outside the release; the anonymous role/score ledger supports the coverage and decision calculations in Section F.3.

I.2 PUBLIC-RGB EXECUTIONS

The public-input reproduction executes thirteen training trajectories: one counter target, eight MATCHED references and a four-arm intervention on one preselected delivery-area rig unit. These runs exercise initialization, training and rendering in two existing experimental settings. The counter target from the first public-input pass is paired with eight subsequently trained references. Starting from public RGB, released cameras and the frozen split, the target constructs feature matches, member-only tracks, colors and 150,954 initialization points from its 120 training images. It trains to 30k, saves 7k/30k states and renders all 96 H queries.

Mean counter PSNR is 29.571163 dB versus the original target's 29.566276 dB, a difference of 0.004887 dB; the largest individual-query difference is 1.064765 dB. Reproduced BaVarIA-n AUC is 0.993490, with 44/48 TP and 1/48 FP at the inherited registered-C primary level. Table 71 retains every readout and both calibration levels.

Table 71: Public-RGB counter reproduction with one target and eight MATCHED references. Decision cells give TP; FP, each out of 48. The two levels reuse the earlier registered-C cutoffs of the reference-access study in Section D.

Readout	AUC	1% TP; FP	5% TP; FP
RAW	0.9201	8; 1	35; 2
BaVarIA-n	0.9935	44; 1	46; 2
GAP-OUT	0.9544	28; 1	43; 5
T-OUT	0.9657	24; 0	37; 3

In the public four-arm intervention, enabling the query's own supervision improves its PSNR by 5.798877 dB with the three peers off and 7.610250 dB with them on. Peer responses have mixed signs. The sixteen other fixed views average 0.4274/0.7351 dB own-image effects in those two conditions, with negative and positive individual changes (Table 72). The complete measurement population is one query, three peers and sixteen other views in the same rig unit. These views describe the spatial response of that intervention, not sixteen additional independent interventions. Exact public image names and rig/frame identities accompany the four-arm results.

Table 72: Public delivery-area reproduction on one preselected rig unit. Y_{ab} switches query supervision with a and three-peer supervision with b. Own columns are $Y_{10} - Y_{00}$ and $Y_{11} - Y_{01}$; values are PSNR or differences in dB. The final row averages all sixteen other fixed views. Their individual own-effects range from -1.912 to 4.051 (peers off) and -1.497 to 4.291 (peers on).

View	n	Y_{00}	Y_{10}	Y_{01}	Y_{11}	Own, off	Own, on
Query	1	14.337	20.136	11.261	18.871	5.799	7.610
Peer 1	1	13.325	15.295	20.590	17.659	1.971	-2.931
Peer 2	1	12.204	10.401	18.063	17.707	-1.803	-0.356
Peer 3	1	11.104	10.507	17.264	19.733	-0.597	2.469
Other views, mean	16	17.862	18.290	17.993	18.728	0.427	0.735

I.3 Construction and query costs

Reference construction is amortized across candidates. With pretrained background-only references, a supplied-pose query uses its RGB and one target plus eight reference renders. Estimated-pose queries additionally use the registered B-image map and the localization implementation in Section F. In the measured counter H96 execution, the first map build takes 205.78 seconds including initial model-weight preparation, followed by 461.60 seconds for query features, matching and PnP. Rendering the target and eight OUT references under the estimated and supplied cameras adds 60.22 seconds. These component boundaries differ from a per-query latency measurement.

The public counter initialization–training–rendering entry point takes 1,703.562 seconds, including 1,449.748 seconds in its training loop. The four public rig training loops sum to 0.672678 GPU-hours. The independent readout panel uses 144 trained models across its scoring conditions. The final student extension records 146.3150 cumulative GPU task-hours, including CPU/I/O and controller-observation boundaries; one pseudo-generation task lacks timing. Shared teacher and pseudo-image work is charged once at its construction boundary. Concurrent task sums do not equal campaign elapsed time or pure CUDA time.

The sixteen A-only models take 6.4275 cumulative training task-hours, and U evaluation takes 27.96 seconds. Their final PLY files total 7.7195 GB. Single-release teacher reuse, file size and serving FPS are reported with the student results in Section H.6. The Artifact retains per-model and component timings, historical batches, SPZ conversion costs and replay agreement records.