APPEARANCE-INDUCED SURFACE BIAS AND SCORE SEPARATION IN GAUSSIAN SPLATTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Gaussian splatting reconstructs surfaces by fitting geometry and a finite appearance model jointly to multi view images, a coupling that the prevailing account interprets as an ambiguity in which several shapes explain the images equally well. We show that finite appearance induces a strict preference for physically worse surfaces, a phenomenon we term appearance induced surface bias. On seven DTU scenes, allowing the geometry of 2D Gaussian splatting to move raises test PSNR by 2.9 dB while the reconstructed surface degrades in every scene. We propose *Score Separation*, a geometry inference framework that retains a regularized estimator for appearance and selects geometry with an independent score. Two mechanisms underlie the bias. Near the true surface, an appearance residual aligned with the image change of a surface displacement shifts the preferred surface, and the penalty that regularizes appearance also enters the ranking of candidate surfaces. Variable projection characterizes the first mechanism, and primal and dual certificates establish strict preferences for worse surfaces in 19 of 23 reconstructed scenes independently of the optimizer. Score Separation reduces controlled recovery error from 6.09 to 4.75 mm, lowers error on unseen materials by 13.7% through its complexity corrected score, and improves reconstructed surfaces where the penalty dominates the descent. As a secondary diagnostic, the contribution depth error reveals surface deterioration that image quality and mean depth conceal.

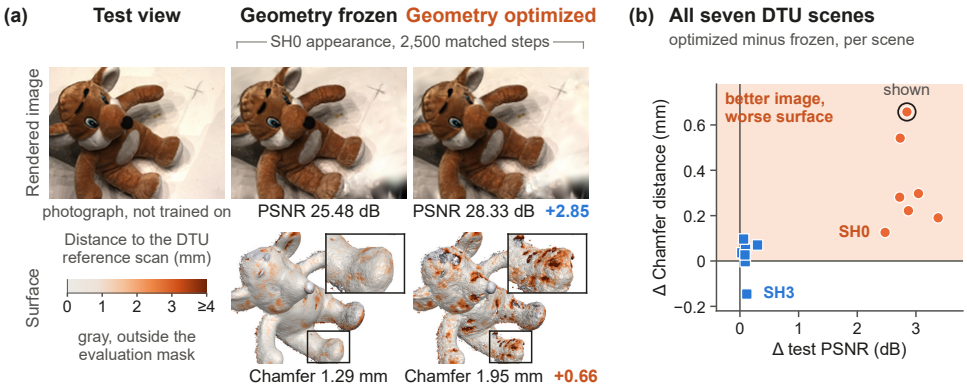

Figure 1: **Better test images, worse surfaces.** (a) Continuing a 2DGS reconstruction of DTU scan105 for 2,500 steps with diffuse (SH0) appearance and free geometry raises test PSNR by 2.85 dB, while the surface departs from the reference scan (color encodes the distance, and insets show the largest change). (b) Across all seven scenes (optimized minus frozen, Table 3 and Appendix Figure A2), SH3 nearly eliminates both changes.

1 INTRODUCTION

Gaussian splatting (Kerbl et al., 2023) and its surface variants (Huang et al., 2024; Guédon & Lepetit, 2024; Yu et al., 2024b; Chen et al., 2025) have become a leading approach to recovering geometry from photographs. Because none of these methods observes the surface directly, each renders a candidate surface, fits a finite appearance model such as spherical harmonics, textures, or a neural

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.

field to the training images, and retains the surface whose fit best reproduces them, so that geometry is judged through appearance. The prevailing account describes this coupling as an *ambiguity* (Zhang et al., 2020), in which many combinations of shape and appearance explain the training views equally well and the limited capacity of the appearance model breaks the tie in favor of the correct shape. Recent theory makes this account precise for appearance that is independent per point, where bounded angular capacity excludes incorrect parallax explanations (Wang & Miller, 2026). Under this view, a better fit to unseen views never comes at the expense of the surface.

Our experiments contradict this expectation. When trained 2DGS reconstructions of seven DTU scenes are continued with diffuse color, releasing the geometry raises test PSNR by 2.9 dB on average, whereas the extracted surface degrades in every scene and its Chamfer distance increases by 0.33 mm (Figure 1). The images improve *because* the surface moves to absorb what the appearance model cannot represent. We formalize the effect through the *geometric preference*, the ordering of candidate surfaces after appearance is refitted to each. Ambiguity corresponds to a tie in this ordering, whereas *appearance induced surface bias* is a strict preference for a physically worse surface. Such a preference excludes the true geometry from the minimizer set of the fitted objective.

We trace the bias to two mechanisms. The first stems from limited capacity, since a finite appearance model leaves a residual that refitting cannot absorb, and variable projection (Golub & Pereyra, 1973) shows that near the true surface the orientation of this residual relative to the image change of a surface displacement determines the preferred surface. The second stems from regularization, because the penalty that stabilizes the appearance estimate also enters the joint objective that ranks candidate surfaces, so that a surface with a worse data fit prevails whenever its penalty saving is larger. Richer appearance models (Rong et al., 2025) and multi view evidence that bypasses appearance (Chen et al., 2025) mitigate the first mechanism, yet many current surface reconstruction pipelines use the same coupled objective to estimate appearance and update geometry, which leaves them exposed to the second mechanism whenever appearance is regularized. Both mechanisms call for separate objectives for appearance estimation and geometry selection.

We therefore propose *Score Separation*, a geometry inference framework that decouples regularized appearance estimation from geometry selection. A regularized estimator fits the appearance of each candidate surface exactly as in joint fitting, while an independent score, evaluated at this fitted appearance, selects the geometry. We instantiate this score with the data term alone (DATA) and with its complexity corrected form derived from generalized cross validation (GCV) (Golub et al., 1979).

Controlled recovery against joint fitting and published regularizers (Tables 1 and 2), primal and dual certificates on reconstructed scenes, and continued training of 2DGS and PGSR hosts support these claims, and a causal intervention confirms the orientation mechanism. Finally, the contribution depth error (CDE) retains the spread of contributions along each ray that mean depth discards.

In summary, we make three contributions.

- **Appearance induced surface bias and its mechanisms.** We formalize the bias as a strict geometric preference, certify it with primal and dual bounds in 19 of 23 reconstructed scenes, and attribute it to residual orientation (Proposition 1) and to regularization dependent geometry ranking.
- **Score Separation.** A geometry inference framework that retains regularization for appearance estimation and selects geometry with an independent score, reducing controlled recovery error from 6.09 to 4.75 mm at a shared estimator and improving reconstructed surfaces where the penalty dominates the descent.
- **Validation and diagnosis.** Causal interventions, certified states, and experiments on two reconstruction hosts establish when the bias arises, and the contribution depth error exposes surface deterioration that images and mean depth conceal.

2 Related work

Surfaces from radiance fields. Neural implicit surfaces (Wang et al., 2021) and Gaussian splatting (Kerbl et al., 2023) recover geometry through differentiable rendering. Surface variants flatten Gaussians into oriented surfels (Huang et al., 2024), extract opacity level sets (Yu et al., 2024b), align meshes (Guédon & Lepetit, 2024), impose continuity (He et al., 2026), add multi view

photometric and planar constraints (Chen et al., 2025), or refine opacity (Yong et al., 2025), while further work improves fidelity (Yu et al., 2024a; Xue et al., 2026). Closest to our work, per primitive textures (Rong et al., 2025) and dedicated geometry Gaussians (Zhou & Lähner, 2026) decouple the *representation* of appearance from that of geometry. These methods commonly optimize geometry and appearance within a coupled objective to which they add geometric or appearance constraints. We examine how finite appearance estimation itself alters the ranking of candidate geometry and decouple the *decision* on geometry. The formulation acts at the level of geometry scoring and requires no change to the underlying scene representation.

Shape and appearance ambiguity. Zhang et al. (2020) identify the ambiguity between shape and radiance and attribute the avoidance of incorrect shapes to the limited capacity of NeRF (Mildenhall et al., 2020). Verbin et al. (2022) counter spurious highlights with structured directional appearance, Fang et al. (2023) reduce the ambiguity with colors estimated in closed form, and Wang & Miller (2026) prove that bounded angular capacity excludes incorrect parallax explanations for opaque surfaces with independent appearance per point. These analyses regard limited capacity as a safeguard, which holds when appearance is free per point but fails once appearance is shared across the surface, as in spherical harmonic fields, textures, and neural fields. In this shared regime, limited capacity *creates* strict preferences for incorrect surfaces, which we certify on reconstructed scenes.

Nuisance estimation and model selection. Joint estimation of a nuisance can favor an incorrect explanation, as in blind deconvolution, where joint maximization over kernel and image prefers the blurred explanation (Levin et al., 2009). Variable projection (Golub & Pereyra, 1973) eliminates linear nuisances exactly, bilevel differentiation (Franceschi et al., 2018) tracks an inner solution, and generalized cross validation (Golub et al., 1979) scores a regularized fit by its predictive error. We apply these tools to the question of which surface a finite appearance model prefers and use them to separate the objective that estimates appearance from the score that selects geometry.

3 Methodology

Our methodology proceeds from diagnosis to remedy, defining the bias (Section 3.1), explaining it through residual orientation (Section 3.2) and regularization (Section 3.3), and deriving Score Separation (Section 3.4) together with a diagnostic (Section 3.5).

Native rendering and appearance estimation. For N training RGB observations $Y \in \mathbb{R}^{N\times 3}$, the native operator $A(g) \in \mathbb{R}^{N\times m}$ of geometry g maps appearance coefficients $a \in \mathbb{R}^{m\times 3}$ to composited pixels $\widehat{y}_r = \sum_l w_{rl}(g)\, f_a(X_{rl}(g)) + b_r(g)$, with contributor weights w_{rl}, material coordinates X_{rl}, and a fixed exterior b. Appearance in a finite class $\mathcal{A}_b$ is fitted by minimizing $J(g,a) = L(g,a) + \lambda R(g,a)$, with image error $L = \|A(g)a + b(g) - Y\|_F^2/(6N)$ and penalty $R = \mathrm{tr}(a^\top \mathcal{R}(g) a)/6$.

3.1 Appearance induced surface bias

The profile $\Phi_b(g) = \min_{a\in\mathcal{A}_b} J(g,a)$ ranks candidate surfaces after appearance is refitted to each. For the true geometry $g^\star$, we define the *preference margin* as

$$\Delta_b(g; g^\star) = \Phi_b(g) - \Phi_b(g^\star). \tag{1}$$

Ambiguity corresponds to ties, $\Delta_b = 0$, whereas a strictly negative margin at a physically worse candidate constitutes a difference in kind, which we call appearance induced surface bias. Such a margin excludes the true geometry from the minimizer set of the fitted objective, whichever optimizer is used.

Strict preferences appear in the native profiles. Our controlled witnesses render natively under a shared bilinear appearance layer. Plane and curved candidates displaced by $+34$ and $+36$ mm attain margins of -3.2×10^{-5} and -4.2×10^{-5} (Figure 2b), and the displaced plane even lowers the test image MSE from 1.900 to 1.839×10^{-3}, whereas warping the observations without any appearance model selects the truth. Phase shifted grids and volumetric 3DGS compositing exhibit the same effect, while affine controls consistently select the truth (Appendices B.3 and B.4). The preference persists as capacity grows, because the ranking depends on how the approximation error varies *across* surfaces, and a 13×13 grid still strictly prefers an incorrect surface in 35 of 36 configurations.

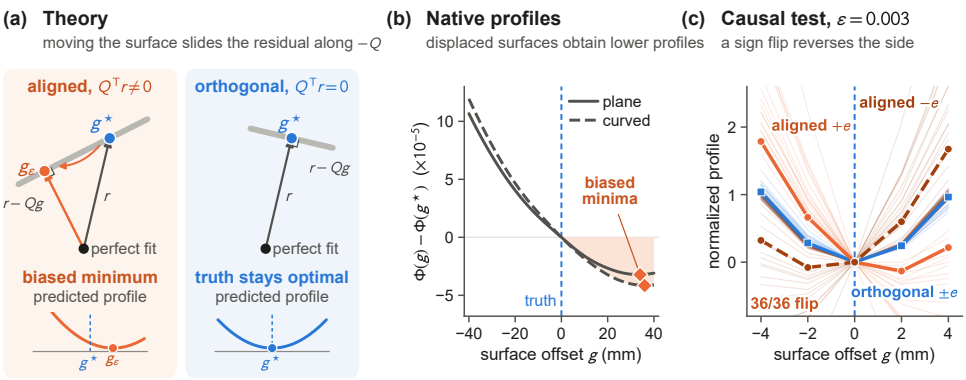

Figure 2: **Residual orientation determines the preferred surface.** (a) Proposition 1 in residual space. Moving the surface by g shifts the residual from r to $r - Qg$, so an aligned residual moves the profile minimum to $g_\epsilon \neq g^\star$, whereas an orthogonal residual leaves it at the truth. (b) Measured native profiles score incorrect surfaces at $+34$ and $+36$ mm below the truth. (c) Reversing an aligned residual of equal energy moves the minimum to the opposite side in all 36 contexts (means drawn thick), while the orthogonal residual preserves the truth.

3.2 Residual orientation explains the preference

A surface perturbation induces a direction of image change. Refitting absorbs the component of this change that lies in the range of the appearance model, and the remainder represents image variation that geometry can explain and appearance cannot. Whenever the current appearance residual has a component along this remainder, moving the surface lowers the fitting objective even as the surface departs from the truth. Near the truth, the *orientation* of the residual therefore determines which surface is preferred. We stack the RGB observations and write the native profile as $F(g; y) = \min_a \|A(g)a + b(g) - y\|^2/(2n)$. Starting from a shared material a_0 that is exactly realizable at the true surface $g = 0$, we add a physical material residual ϵe, so that $y_\epsilon = A_0 a_0 + b_0 + \epsilon z$ with $z = \mathcal{L}_0 e$ retaining the full native contributor mixture. Let P_{A_0} denote the projection onto the appearance range at the truth, and let $Q = (I - P_{A_0})\,\partial_g(Aa_0 + b)|_0$ collect the image changes of surface displacements that refitting cannot absorb.

Proposition 1 (Local residual orientation)**.** *On a smooth branch with fixed visibility, persistent full column rank, and inactive bounds,*

$$\nabla_g F(0; y_\epsilon) = -\frac{\epsilon}{n} Q^\top z + O(\epsilon^2), \qquad g_\epsilon = \epsilon (Q^\top Q)^{-1} Q^\top z + O(\epsilon^2), \tag{2}$$

where the second expression, for the nearby strict minimizer, additionally requires $Q^\top Q \succ 0$.

To first order, the truth therefore remains stationary only when the residual is orthogonal to Q. Any aligned component, $Q^\top z \neq 0$, creates a strictly preferred nearby incorrect surface that lowers the profile by $\frac{\epsilon^2}{2n} z^\top Q (Q^\top Q)^{-1} Q^\top z + O(\epsilon^3)$, and residuals of equal energy produce opposite preferences (Figure 2a). Appendix B states the assumptions and gives the proof.

3.3 Regularization alters geometry ranking

Regularized appearance estimation opens a second route to bias. With the fitted appearance $a_\lambda(g) = \arg\min_a J(g, a)$, the profile decomposes into a data term and a penalty term,

$$D(g) = L(g, a_\lambda(g)), \qquad P(g) = \lambda R(g, a_\lambda(g)), \qquad F(g) = D(g) + P(g), \tag{3}$$

where F is the regularized profile that joint fitting minimizes. A candidate with a worse data fit, $\Delta D > 0$, still prevails whenever its penalty saving exceeds this loss, $-\Delta P > \Delta D$. The penalty thus contributes directly to the geometry ranking, and in our matched kernel experiment it overturns the data in 11 of 36 signed comparisons. Since the penalty exists to stabilize the *estimation* of appearance, this observation motivates a geometry score that excludes it.

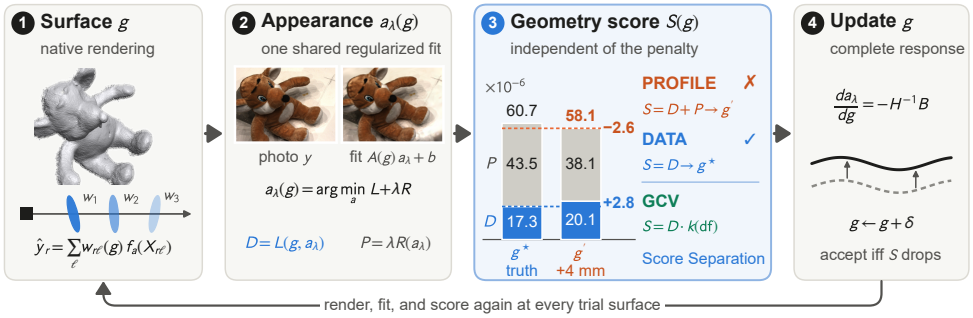

Figure 3: **Score Separation keeps the penalty in appearance estimation and removes it from geometry selection.** (1) The native renderer retains every surfel contribution $w_{r\ell}(g)$ to a pixel ray. (2) A single regularized fit estimates the appearance for every score. (3) The geometry score is a separate decision, and in the measured bars (truth $g^\star$ against g' at $+4$ mm) $D + P$ prefers g' because its penalty saving outweighs a worse data fit, whereas D retains $g^\star$. (4) Updates follow the complete response.

3.4 Score Separation for Geometry Inference

We propose Score Separation as a geometry inference framework that uses a regularized estimator for appearance and an independent score for geometry selection (Figure 3). For a candidate geometry g, the estimator returns the fitted appearance $a_\lambda(g)$, and a score $S(g) = S(g, a_\lambda(g))$ evaluates the candidate at this appearance, with joint fitting recovered as the special case $S_F = D + P$ (PROFILE) of Eq. (3).

One appearance estimate for every score. All scores are evaluated at the same appearance, which we fit once per candidate with the estimator defined above. The material f_a combines 512 Matérn kernels with an unpenalized quadratic trend, $\mathcal{R}$ penalizes only the kernels, and with $M = A^\top A + N\lambda\mathcal{R}$ the coefficients solve $Ma = A^\top(Y - b)$ on all training rows (Appendix C). The penalty thereby retains its role in estimation, where it stabilizes the fit.

An independent geometry score. DATA ranks geometry by the data term alone, $S_D = D$, so that regularization shapes $a_\lambda(g)$ while its value is excluded from geometry selection. GCV (Golub et al., 1979) divides the data term by a complexity factor, $S_G = D/(1 - \mathrm{df}\,/N)^2$ with $\mathrm{df} = \mathrm{tr}(M^{-1}A^\top A)$, which approximates the error on held out rows and thereby rewards prediction. DATA and GCV are the two instantiations of Score Separation that we evaluate. As comparators, we consider two predictive scores that refit on fold complements and score the omitted rows, namely Camera-PGP with one training camera per fold and Pixel-CV with one of six balanced pixel folds.

Complete response. Because geometry changes the fitted appearance, the update accounts for this dependence, and on a smooth branch with invertible $H = J_{aa}$ and $B = J_{ag}$, implicit differentiation (Golub & Pereyra, 1973; Franceschi et al., 2018) gives $da_\lambda/dg = -H^{-1}B$ and

$$\nabla_g S = S_g - B^\top H^{-1} S_a. \tag{4}$$

Since the outer score differs from the inner objective, the stationarity of J leaves the response term in place, and we retain it together with every geometry dependence of the renderer.

Optimization and integration. Controlled experiments differentiate through the native operator, whereas reconstructed scenes use central differences with full refitting of native SH, and each trial is rendered, refitted, and accepted when the selected score decreases. Because Score Separation modifies only the rule that ranks geometry, it integrates into any splatting pipeline that fits appearance with a penalty, and in our benchmark DATA has negligible overhead per step, while GCV adds 3.9% per geometry step and 17.4% per complete run (Appendix Table A8).

3.5 Contribution Depth Diagnostic

The contribution depth error (CDE) evaluates geometric deterioration in saved states and remains outside the optimization of Score Separation. Images provide no audit of the surface, and mean depth retains only the first moment of the contributions along a ray. On fixed valid reference rays $\mathcal{Q}$ with

Table 1: **Score Separation leads a controlled benchmark in which only the geometry score or the appearance model changes.** Height RMSE (mm) with shared observations, start (10.67 mm), geometry family, budget, and estimator ($\lambda = 10^{-5}$, fixed before any outcome). Closed form color ($\lambda_{\mathrm{cf}} = 100$, SH degree 2) and multi view NCC ($\lambda_{\mathrm{ncc}} = 1$, 5×5 grid patch) add their terms to F with weights selected on development materials. Rank averages over materials, Mat. counts materials improved over PROFILE, bold marks the best, underline the second, and · settings not run (conditions in Appendix F).

	Default recovery				Robustness			Preservation
Method	Mean↓	P90↓	Rank↓	Mat.↑	Weak parallax↓	Noise↓	Long budget↓	From truth↓
PROFILE (Huang et al., 2024)	6.09	9.76	7.9		8.24	9.25	7.89	4.74
Spectral, order 12	4.87	6.46	4.0	9/9	·	·	**4.94**	·
Bilinear grid 49^2 (Rong et al., 2025)	5.65	8.45	6.0	6/9	·	·	5.77	·
Image matching (Schönberger et al., 2016)	8.83	12.65	9.4	2/9	·	·	7.38	·
Closed form color (Fang et al., 2023)	6.33	9.37	7.7	5/9	8.23	5.27	5.39	2.20
Multi view NCC (Chen et al., 2025)	5.35	6.60	6.0	5/9	**6.41**	6.38	5.96	**1.38**
Camera-PGP (Stone, 1974)	5.03	7.15	4.7	8/9	6.91	4.75	6.02	3.03
Pixel-CV (Stone, 1974)	4.75	6.00	3.9	9/9	·	·	·	·
DATA (ours)	4.75	6.50	3.4	9/9	6.61	6.37	5.87	2.63
GCV (ours)	**4.32**	**5.32**	**2.0**	9/9	6.60	**4.39**	5.15	2.66

normalized weights $\widetilde{w}_{rl}$ and mean depth $\bar{z}_r = \sum_l \widetilde{w}_{rl} z_{rl}$, CDE retains the second moment relative to the reference, which adds the spread of the contributions to the error of the mean depth,

$$\mathrm{CDE}^2 = \frac{1}{|\mathcal{Q}|} \sum_{r \in \mathcal{Q}} \sum_l \widetilde{w}_{rl} (z_{rl} - z_r^\star)^2 = \mathrm{RMSE}^2_{\mathrm{mean}} + \frac{1}{|\mathcal{Q}|} \sum_{r \in \mathcal{Q}} \sum_l \widetilde{w}_{rl} (z_{rl} - \bar{z}_r)^2. \tag{5}$$

4 Experiments

Data. Controlled scenes texture analytical surfaces with natural photographs, so that the true height is known, and scene reconstructions use DTU photographs (Jensen et al., 2014) and rendered Replica RGB images with reference scans or meshes. **Baselines.** All methods share observations, initialization, geometry family, estimator, and budget. The baselines comprise joint fitting (PROFILE), richer appearance, image matching, closed form color (Fang et al., 2023), and multi view NCC (Chen et al., 2025), with predictive scores as comparators and, on reconstructed scenes, native SH refitting and a PGSR host. **Metrics.** Controlled scenes report the height RMSE of the recovered surface, and reconstructions report Chamfer distance and accuracy under the official DTU protocol together with the native ray distance to the reference. References serve only to evaluate saved states, and the statistical units are scenes or materials (Appendix A).

4.1 Score Separation leads a controlled benchmark

Controlled scenes isolate the penalty route, since every score uses the identical regularized estimator (Table 1).

Removing the penalty from the score improves recovery in every setting. With the shared estimator, DATA reduces the mean height error of PROFILE from 6.09 to 4.75 mm and improves all nine materials. GCV reaches 4.32 mm and attains the best mean, P90, and rank among all ten methods, with a paired advantage over every other comparator whose interval excludes zero and a lower error than spectral appearance on seven of nine materials. Neither published regularizer differs reliably from PROFILE, as closed form color reaches 6.33 mm and the multi view NCC term of PGSR 5.35 mm, and GCV outperforms each of them on eight of nine materials. NCC leads under weak parallax and from a correct start, GCV remains first under noise, where Camera-PGP overtakes DATA, and under the long budget richer spectral appearance slightly outperforms GCV, a first indication of the capacity route.

Table 2: **Tuned on one grid and tested on unseen materials, complexity corrected separation improves every material.** Each score selects λ on development materials (10^{-5} for Camera-PGP, 10^{-6} otherwise, with $\lambda_{\mathrm{cf}} = 1$, SH degree 2, and $\lambda_{\mathrm{ncc}} = 1$ for the published regularizers) and is frozen before six unseen DTD materials are rendered. Height RMSE (mm), paired difference Δ from PROFILE with 95% bootstrap interval, and materials improved.

	Unseen DTD material									
Score	Band.	Cryst.	Honey.	Perf.	Sprink.	Zigzag.	Mean↓	P95↓	Δ [95% CI]↓	Mat.↑
PROFILE (Huang et al., 2024)	6.06	4.19	5.54	2.77	3.76	4.31	4.44	7.00		
Camera-PGP (Stone, 1974)	6.03	3.77	**4.17**	2.55	3.83	5.91	4.38	6.96	−0.06 [−0.75, 0.72]	4/6
Closed form (Fang et al., 2023)	8.09	**3.00**	5.81	5.23	5.20	4.84	5.36	9.31	0.92 [−0.09, 1.85]	1/6
NCC (Chen et al., 2025)	6.66	4.40	6.98	4.13	4.81	5.45	5.40	9.77	0.96 [0.60, 1.29]	0/6
DATA (ours)	5.36	3.78	6.03	**2.43**	3.80	4.43	4.30	6.96	−0.14 [−0.45, 0.19]	3/6
GCV (ours)	**4.70**	3.50	5.08	2.58	**3.68**	**3.46**	**3.83**	**5.30**	**−0.61** [−0.97, −0.28]	6/6

Table 3: **In both hosts, free geometry yields better test images and worse surfaces unless appearance free evidence constrains the surface.** Mean change after 2,500 matched steps (optimized minus frozen geometry) on seven DTU scenes, host Chamfer distance (CD, mm) in parentheses, and scenes that gain PSNR or lose surface quality (7/7 gives $p = 0.016$). NC is the 2DGS normal term, NCC the multi view patch term of PGSR, and orange a worsening above 0.05 mm (every scene in Appendix Tables A4 and A5).

		Mean change			Scenes		
Host (CD)	Setting	PSNR↑	CD↓	Acc.↓	PSNR better	CD worse	Acc. worse
2DGS (1.69)	SH0, photometric	+2.87	+0.33	+0.63	7/7	7/7	7/7
(Huang et al., 2024)	SH0, +NC	+2.32	+0.12	+0.22	7/7	5/7	7/7
	SH3, photometric	+0.11	+0.02	+0.10	7/7	5/7	6/7
PGSR (1.57)	SH0, photometric	+3.33	+0.39	+0.51	7/7	7/7	6/7
(Chen et al., 2025)	SH0, full PGSR loss	+2.56	+0.02	+0.01	7/7	4/7	4/7
	SH3, photometric	+0.17	+0.06	+0.07	7/7	5/7	6/7

4.2 The lead persists under fair tuning and transfers to unseen materials

Fair tuning determines whether the lead reflects the score or the regularization weight. We tune every score on the same grid of development materials, on which PROFILE improves from 6.99 to 4.69 mm, freeze all settings, and evaluate on six unseen DTD materials (Cimpoi et al., 2014) (Table 2).

GCV lowers the mean error from 4.44 to 3.83 mm, a **13.7% reduction**, improves every material, and reduces the P95 error from 7.00 to 5.30 mm. After tuning, DATA and Camera-PGP perform on par with PROFILE, whereas closed form color (5.36 mm) and multi view NCC (5.40 mm) fall behind it. Because DATA and GCV select the same regularization and share the estimator at matched geometry, the 10.9% gain of GCV over DATA (-0.47 mm, $[-0.80, -0.14]$) isolates the effect of the complexity correction. Test PSNR lies between 31.33 and 31.41 dB for every score, so image quality cannot distinguish these surfaces. A prespecified replication on 12 new materials confirms the complete response (Appendix Table A7).

4.3 Freeing geometry trades surface accuracy for image fit

We continue 2DGS hosts of seven DTU scenes (SH3, 15k iterations, width of 160 pixels). After 500 steps that refit SH0 or SH3 appearance alone, 2,500 matched steps run with geometry either frozen or free, every eighth view is held out, and meshes extracted with a single TSDF recipe are evaluated with the official DTU protocol. Since neither 2DGS nor PGSR regularizes appearance, PROFILE and DATA coincide on these hosts, and the continuations therefore isolate the *capacity* route.

Under diffuse appearance, releasing the geometry raises test PSNR by 2.47 to 3.38 dB in every scene (Table 3), whereas the Chamfer distance increases by 0.13 to 0.66 mm while completeness remains nearly unchanged (Appendix D), so the surface is displaced with its coverage intact. SSIM and LPIPS

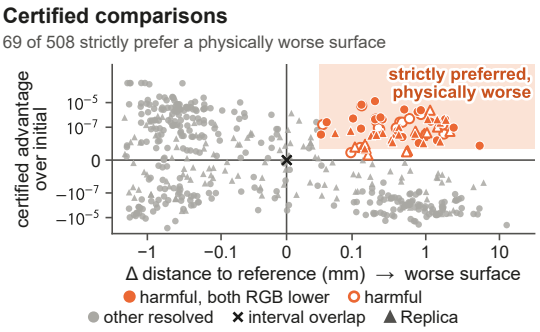

Figure 4: **Scene reconstructions strictly prefer physically worse surfaces.** Certified advantage over the initial reconstruction against the measured change in distance to the reference, for all 508 comparisons. States with $L_{g_0} - U_g > 0$ are strictly preferred.

Table 4: **Harmful preferences arise in 19 of 23 scenes and replicate on new ones.** Harm counts strict preferences for a surface more than 0.05 mm farther from the reference, RGB those that also lower training and test RGB error, and Scenes the scenes with a harmful preference.

Cohort	Pairs	Strict	Harm	RGB	Scenes
DTU, first	224	137	25	13	6/8
Replica, first	48	26	12	7	2/2
DTU, confirm.	108	37	11	8	6/7
Replica, confirm.	128	58	21	10	7/8
Total	508	258	**69**	**38**	**19/23**

confirm the image gain in all seven scenes of both hosts (Appendix Table A6). Normal consistency attenuates the effect, and SH3 removes most of it. *Image gain and surface loss thus scale together with the appearance deficit that geometry absorbs.*

The bias transfers to a strong multi view host. In PGSR (Chen et al., 2025), releasing the geometry under SH0 and the photometric loss raises test PSNR by 3.33 dB while CD increases by 0.39 mm in all seven scenes. With its full loss, PGSR retains a 2.56 dB gain at unchanged CD, and a post hoc ablation (Appendix Table A5) attributes this protection to the multi view NCC term, which compares observed patches without the fitted appearance, whereas the geometric regularizers alone leave the damage in place (+0.32 mm). Evidence that bypasses the appearance model therefore protects the surface.

4.4 Certificates and a causal test locate the preference in the objective

Comparing *optimal* appearance fits removes the optimizer from the question (Appendix E). We freeze each initial 2DGS reconstruction and every refinement endpoint, comprising 312 states in 40 regions of ten reconstructions and, in a prespecified confirmation, 295 states in 59 regions of 15 new reconstructions. Over a convex class of native SH of degree 3 with nonnegative contributor colors and exact compositing, independent primal and dual witnesses bound the optimal regularized profile of each frozen geometry g by $[L_g, U_g]$. A strict preference over the initial reconstruction g_0 requires $U_g + 10^{-12} < L_{g_0}$, and harm denotes a measured increase in the native ray distance to the reference of more than 0.05 mm.

Of the 508 comparisons, 505 resolve without interval overlap (Figure 4). Among them, 69 are strict preferences for a *physically worse* surface in 19 of the 23 scenes, and 38 of these also lower both training and test image error (Table 4). The confirmation reproduces this finding with 32 harmful strict preferences in 13 of 15 new scenes, with margins at least $12.9\times$ the interval width (Appendix E).

Reversing the residual flips the preferred surface. In 36 native contexts, we realize residuals of equal energy aligned with or orthogonal to Q (Section 3.2). At $\epsilon = .003$, reversing the sign of the aligned residual flips the ± 4 mm preference in all 36 contexts, whereas the orthogonal residual suppresses the linear response by a factor of 1,857 (Figure 2c and Appendix Table A2).

4.5 Penalty dominance governs the gain on reconstructed scenes

We refine regions of every reconstruction from Section 4.4 with PROFILE, DATA, GCV, and native SH refitting (Appendix G) and split the descent of PROFILE into data change ΔD and penalty change ΔP on the common head. The *penalty dominance* $\pi = -\Delta P/(|\Delta D| + |\Delta P|)$ exceeds $1/2$ when penalty savings dominate the descent.

Table 5: **Score Separation is most effective where the penalty dominates the objective.** Share is the median penalty share at the start, $\pi > \frac{1}{2}$ counts penalty dominated regions, and changes in distance to the reference (mm) are averaged over regions, then scenes, with orange for drift over 0.05 mm and n/r for values not recorded.

		Mechanics		Change in distance to the reference					
Cohort (scenes)	Regions	Share	$\pi > \frac{1}{2}$	Initial	PROFILE	DATA	GCV	SH refit	DATA − PROFILE
DTU, first (8)	32	n/r	n/r	12.48	−0.049	+0.014	+0.107	−0.278	+0.063
Replica, first (2)	8	65%	5/8	4.29	+0.898	+0.134	+0.384	−0.006	**−0.764**
DTU, confirmation (7)	27	6%	0/27	28.09	+0.095	+0.139	+1.256	−0.014	+0.044
Replica, confirmation (8)	32	33%	4/32	13.99	−1.295	−1.295	−0.812	−0.097	+0.001

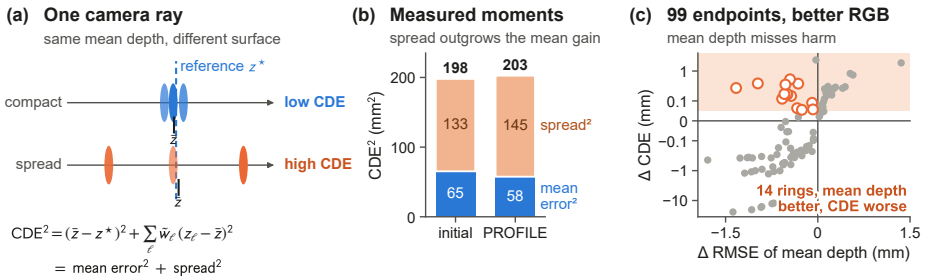

Figure 5: **CDE exposes deterioration that images and mean depth hide.** (a) Equal mean depth $\bar{z}$, different spread. (b) In one region of DTU scan37, PROFILE lowers the mean depth error but raises the spread more. (c) Among 99 first cohort endpoints with better RGB, rings mark the 14 whose mean depth improves while CDE worsens by over 0.05 mm.

Separation improves the surface where the penalty drives the descent. Across the 67 regions with a recorded decomposition, the advantage of DATA grows with penalty dominance (Spearman $\rho = -0.26$, $p = 0.035$, Appendix Figure A5a). In the nine penalty dominated regions, all in Replica scenes, DATA ends on average 0.40 mm closer to the reference and is closer in seven, whereas the mean difference in the 58 data driven regions is −0.03 mm. In the first Replica cohort of Table 5, the penalty accounts for 65% of the objective and DATA removes 85% of the 0.90 mm drift of PROFILE, while at shares of 6% and 33% the two scores nearly coincide.

GCV requires fixed material support. At realistic operating points its complexity factor reaches eight, so GCV ends farther from the reference in all seven new DTU scenes (Appendix Figure A5b).

Capacity governs the remaining bias. With one shared SH field, DATA outperforms JOINT in five of six 3DGS regimes, and free SH per Gaussian brings both to 0.26 mm (Appendix Table A11).

4.6 Contribution depth exposes hidden deterioration

Of 99 DTU endpoints with better RGB, CDE flags 43 as deteriorated and mean depth 27, and 14 of the 17 flagged only by CDE improve mean depth by over 0.05 mm (Figure 5c, Appendix Table A10).

5 Conclusion

Appearance induced surface bias is a strict geometric preference produced by finite appearance fitting, which we trace to residual orientation and regularization dependent geometry ranking. Score Separation retains regularized appearance estimation while evaluating geometry with an independent score, and controlled recovery, certified reconstructed states, and experiments at scene scale show that it improves geometry when regularization contributes substantially to optimization. CDE exposes deterioration that image metrics and mean depth miss, and penalty dominance defines a measurable operating regime for prospective evaluation.

Reproducibility statement

Section 3.4 and Appendix A specify the geometry families, native appearance operator, normalization, score derivatives, update rules, material and camera splits, reference support, and statistical units. Appendix D details the protocol at scene scale, Appendix E the construction and verification of every primal and dual certificate, and Appendix G all refinement cohorts on scene reconstructions, including the prespecified confirmation. The supplementary package runs on a CPU in under a minute without datasets. It recomputes every number in the main tables from the endpoint records and checks it against the manuscript source, rebuilds all 508 certified comparisons from the saved bounds of all 607 frozen states, replays complete certificate problems with saved primal and dual witnesses for a sample of regions using an independent checker, and regenerates every data figure. The complete certificate problems exceed the size limit and will be released with the code and the native verifier.

Use of large language models

We used large language models in two ways. First, they aided and polished writing by suggesting edits to drafted text for clarity, grammar, and concision. Second, they supported retrieval and discovery by helping locate related work and check bibliographic details. The authors reviewed all suggestions and take full responsibility for the content of the paper.

References

Danpeng Chen, Hai Li, Weicai Ye, Yifan Wang, Weijian Xie, Shangjin Zhai, Nan Wang, Haomin Liu, Hujun Bao, and Guofeng Zhang. PGSR: Planar-based Gaussian splatting for efficient and high-fidelity surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, 31(9):6100–6111, 2025. doi: 10.1109/TVCG.2024.3494046.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014. URL https://openaccess.thecvf.com/content_cvpr_2014/html/Cimpoi_Describing_Textures_in_2014_CVPR_paper.html.

Qihang Fang, Yafei Song, Keqiang Li, and Liefeng Bo. Reducing shape-radiance ambiguity in radiance fields with a closed-form color estimation method. In *Advances in Neural Information Processing Systems*, volume 36, pp. 58902–58920, 2023. doi: 10.52202/075280-2570. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/b87738474533cab76c7bee4e08443aca-Abstract-Conference.html.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1568–1577, 2018. URL https://proceedings.mlr.press/v80/franceschi18a.html.

Gene H. Golub and Victor Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on Numerical Analysis*, 10(2):413–432, 1973. doi: 10.1137/0710036.

Gene H. Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979. doi: 10.1080/00401706.1979.10489751.

Antoine Guédon and Vincent Lepetit. SuGaR: Surface-aligned gaussian splatting for efficient 3D mesh reconstruction and high-quality mesh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5354–5363, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Guedon_SuGaR_Surface-Aligned_Gaussian_Splatting_for_Efficient_3D_Mesh_Reconstruction_and_CVPR_2024_paper.html.

Bing He, Jingnan Gao, Yunuo Chen, Ning Cao, Gang Chen, Zhengxue Cheng, Li Song, and Wenjun Zhang. SurfSplat: Conquering feedforward 2D gaussian splatting with surface continuity priors. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3a10c46572628d58cb44fb705f25cbbf-Abstract-Conference.html.

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2D gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH Conference Papers*, 2024. doi: 10.1145/3641519.3657428. URL https://surfsplatting.github.io/.

Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 406–413, 2014. URL https://openaccess.thecvf.com/content_cvpr_2014/html/Jensen_Large_Scale_Multi-view_2014_CVPR_paper.html.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. doi: 10.1145/3592433. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Anat Levin, Yair Weiss, Frédo Durand, and William T. Freeman. Understanding and evaluating blind deconvolution algorithms. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1964–1971, 2009. doi: 10.1109/CVPR.2009.5206815. URL https://people.csail.mit.edu/billf/papers/deconvLevinEtalCVPR09.pdf.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. doi: 10.1007/978-3-030-58452-8_24.

Victor Rong, Jingxiang Chen, Sherwin Bahmani, Kiriakos Kutulakos, and David Lindell. GStex: Per-primitive texturing of 2D gaussian splatting for decoupled appearance and geometry modeling. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 3508–3518, 2025. URL https://openaccess.thecvf.com/content/WACV2025/html/Rong_GStex_Per-Primitive_Texturing_of_2D_Gaussian_Splatting_for_Decoupled_Appearance_WACV_2025_paper.html.

Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, pp. 501–518, 2016.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, 36(2):111–147, 1974.

Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5491–5500, 2022.

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Songhe Wang and David Miller. When 3D Gaussian Splatting recovers real surfaces. In *European Conference on Computer Vision (ECCV)*, 2026. URL https://eccv.ecva.net/virtual/2026/poster/5350. Official ECCV 2026 poster 5350; preprint arXiv:2608.30054.

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

Weiyi Xue, Fan Lu, Chi Zhang, Tianhang Wang, Sanqing Qu, Zehan Zheng, Boyuan Zheng, Junqiao Zhao, and Guang Chen. Signal structure-aware gaussian splatting for large-scale scene reconstruction. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10010766.

Silong Yong, Venkata Nagarjun Pudureddiyur Manivannan, Bernhard Kerbl, Zifu Wan, Simon Stepputtis, Katia Sycara, and Yaqi Xie. OMG: Opacity matters in material modeling with gaussian splatting. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/58b286aea34a91a3d33e58af0586fa40-Abstract-Conference.html.

Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3D gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19447–19456, 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yu_Mip-Splatting_Alias-free_3D_Gaussian_Splatting_CVPR_2024_paper.html.

Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes. *ACM Transactions on Graphics*, 43(6), 2024b. doi: 10.1145/3687937.

Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. NeRF++: Analyzing and improving neural radiance fields. arXiv:2010.07492, 2020. URL https://arxiv.org/abs/2010.07492.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Hongyu Zhou and Zorah Lähner. Geometry gaussians: Decoupling appearance and geometry in gaussian splatting. arXiv:2606.05124, 2026. URL https://arxiv.org/abs/2606.05124.

ORGANIZATION OF THE APPENDIX

The appendix follows the order of the main text. Appendix A gives the shared protocols; Appendix B proves Proposition 1 and reports the causal test and the volumetric transfer (Section 3); Appendix C details the estimator and its derivatives (Section 3.4); Appendices D to H complete the experiments of Sections 4.3 to 4.5, and the contribution depth protocol of Section 4.6 closes Appendix G.

A EXPERIMENTAL PROTOCOLS

The experiments distinguish preference among fixed candidates, recovery from an erroneous surface, and preservation of an existing reconstruction. Controlled observations texture analytical surfaces with natural photographs; native refinement uses real DTU photographs or rendered Replica RGB with independently evaluated reference geometry. Physical references evaluate saved states and never select an update.

Table A1: Core protocols. A geometry opportunity proposes a step and evaluates its backtracking trials.

	Controlled recovery	Native local refinement
Geometry	Six coherent quadratic height coefficients	Six coherent height coefficients of a 2DGS region selected from training data
Training / test	Six / two cameras; 13×13 pixels each (1,014 training rows)	Six / two cameras; at most 128 training rows each
Update	Native gradient through the coefficient solve; BFGS	Central differences with full refitting (± 0.05 mm); BFGS
Budget / step cap	60 opportunities (40 from truth); 0.5 mm	40 opportunities; 0.5 mm center displacement
Statistical unit	Material, after averaging shape/start repeats	Scene, after averaging its regions

Inputs and tuning. Controlled training cameras are 0, 1, 2, 4, 5, 6; cameras 3 and 7 are held out. Nine materials derived from DTU give 36 conditions of shape and start. Six previously unused DTD images (banded, crystalline, honeycombed, perforated, sprinkled, zigzagged) form the cohort of unseen materials. Every score receives $\lambda \in \{10^{-6}, 10^{-5}, 10^{-4}\}$ on development materials 2, 5, 7; development height RMSE selects one global λ per score before any unseen outcome (10^{-6} for PROFILE, DATA, GCV; 10^{-5} for Camera-PGP). Test RGB never enters selection or fitting.

Updates and uncertainty. Proposals use backtracking on the actual score at scales $1, \frac{1}{2}, \ldots, \frac{1}{16}$; acceptance requires a decrease above 10^{-13}, and a failed search halves the trust radius. Runs stop at the budget or at a radius below 0.0039 mm; no checkpoint is selected by reference error. Material intervals use 5,000 percentile bootstrap resamples of material means. Leaving out one material at a time retains a negative GCV$-$PROFILE difference in every split (-0.71 to -0.46 mm).

Native references. DTU local distance is the directed distance from ray intersections, weighted by native compositing, to the complete official STL points, distinct from the official score for the whole object; Replica uses the exact distance from points to triangles of the complete mesh. Regions are chosen before any method runs, from opacity, normal coherence (at least 0.70), spatial separation, and training support.

B RESIDUAL ORIENTATION, PROOF AND CAUSAL TEST

B.1 ASSUMPTIONS AND PROOF OF PROPOSITION 1

Let $g \in \mathbb{R}^d$ be a locally injective chart of physical geometry with truth 0 in its interior, and

$$F(g; y) = \min_a \frac{1}{2n} \|A(g)a + b(g) - y\|^2, \qquad y_\epsilon = A_0 a_0 + b_0 + \epsilon z, \quad z = \mathcal{L}_0 e,$$

where RGB channels are stacked and $\mathcal{L}_0$ evaluates one shared material function at all native contributors and composites them. Assume A, b are C^3 near 0 with fixed renderer support, visibility,

and ordering; A has persistent full column rank; and the appearance optimum is interior. Define $P = A_0 A_0^\dagger$, $J = \partial_g (A a_0 + b)|_0$, $Q = (I - P)J$, and $K = Q^\top Q$. Then

$$\nabla_g F(0; y_\epsilon) = -\epsilon Q^\top z/n + O(\epsilon^2), \qquad \nabla_g^2 F(0; y_0) = K/n, \tag{6}$$

and, if $K \succ 0$, there is a unique nearby strict local minimizer with

$$g_\epsilon = \epsilon K^{-1} Q^\top z + O(\epsilon^2), \qquad F(g_\epsilon; y_\epsilon) - F(0; y_\epsilon) = -\frac{\epsilon^2}{2n} z^\top Q K^{-1} Q^\top z + O(\epsilon^3). \tag{7}$$

Proof. At $g = 0$, the orthogonality of least squares gives $a_\epsilon = a_0 + \epsilon A_0^\dagger z$ and residual $r_\epsilon = -\epsilon(I - P)z$. The envelope derivative $r_\epsilon^\top \partial_g (A a_\epsilon + b)/n$ yields the gradient. At the realizable base the residual vanishes and its total geometry derivative after refitting is Q, which gives the Hessian. A nonzero gradient admits a strict descent direction; when $K \succ 0$ the implicit function theorem applied to $\nabla_g F = 0$ yields the minimizer, and a Taylor expansion the objective decrease. □

Residuals e and $-e$ of equal energy reverse the leading displacement, and $Q^\top \mathcal{L}_0 e_\perp = 0$ removes the linear term. Rank or visibility changes and active bounds leave the stated regime; the result is a local specialization of variable projection (Golub & Pereyra, 1973), not a global identifiability theorem.

B.2 Constructed intervention

Across two shapes, two camera baselines, three bilinear capacities, and three phases (36 contexts), we remove 32 Fourier functions from the appearance space on a fixed 97×97 source grid. Whitening by source energy yields an aligned residual maximizing $\langle Q, \mathcal{L}_0 e \rangle$ and a deterministic orthogonal residual of equal energy; an analytic amplitude bound keeps the base material plus either signed residual inside $[0, 1]$ without clipping. Each context has two orientations, two signs, seven amplitudes, and five heights in $\{-4, -2, 0, 2, 4\}$ mm (5,040 native profile values). The construction uses known geometry to test sufficiency.

Table A2: Amplitude sweep of the constructed intervention (36 contexts). Accuracy is the fraction of signed aligned residuals whose actual preference matches the predicted direction.

ϵ	.0001	.0003	.001	.003	.01	.03	.1
± 2 mm accuracy (%)	83.3	94.4	98.6	**100**	97.2	93.1	77.8
± 4 mm accuracy (%)	58.3	79.2	97.2	**100**	97.2	91.7	77.8
Sign pairs reversing the preference	24	32	35	**36**	34	31	20

The paired odd component of the finite slope has the predicted sign in all 252 pairs of context and amplitude, with median relative error 0.01%. At $\epsilon = .003$ the orthogonal control suppresses the mean odd response by $1{,}857\times$. Very small residuals compete with effects of finite comparisons, and large ones engage the even response and active bounds, which limits accuracy at the extremes. Source orthogonality holds to 7.9×10^{-18} and residual energies agree to relative 10^{-12}. Applied to natural textures, a predictor that sees only the source image reaches 48% direction accuracy over 648 configurations. Refitting and active constraints reshape natural residuals, so the mechanism explains preferences, and forecasting them from the photograph alone remains open.

B.3 Further mechanism evidence

Witnesses and capacity. Controlled witnesses use 25 surfels and a shared 7×7 bilinear grid (25 active columns at truth and at the witnesses). Profiles scan $[-40, 40]$ mm at 2 mm spacing; a strict preference requires $|g| \geq 4$ mm and a margin below -10^{-7}, and boundary minima are censored. At matched numbers of active columns and matched ranks, grids shifted in phase select heights differing by more than 4 mm in 244 plane and 502 curved pairs (for example 38 versus 30 mm and 38 versus 22 mm), while affine materials select the truth at every phase.

Penalty reversals and the complexity factor. At one fitted estimator $\Delta F = \Delta D + \Delta P$; for $\Delta D > 0$ the joint score still prefers the candidate exactly when $-\Delta P > \Delta D$. This occurs in 11 of 36 comparisons from truth to ± 4 mm, covering six of nine materials (Figure A1). For GCV, $G = Dk$ with $k = (1 - \mathrm{df}\,/N)^{-2}$ and $\Delta G = k_0 \Delta D + D_1 \Delta k$ separates the data change from the complexity change (closing to 3.4×10^{-19} over 84 states); at the selected $\lambda = 10^{-6}$ the complexity factor raises the minima that contain the truth from 8 to 10 of 12 scans of material and shape.

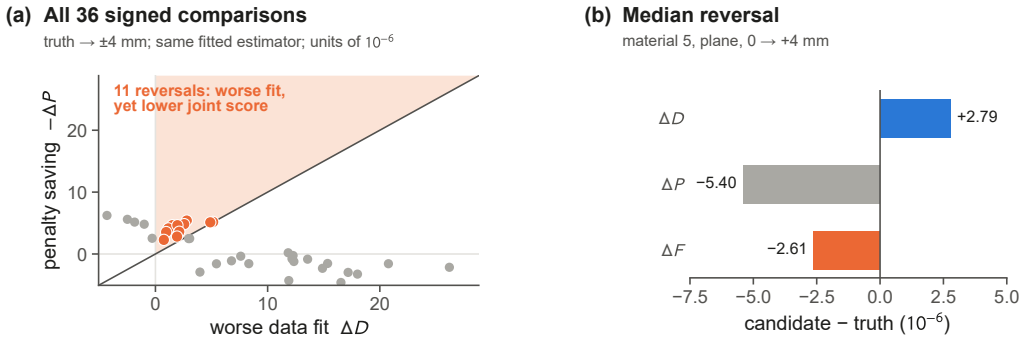

Figure A1: **The penalty value overturns the data ranking.** (a) All 36 signed comparisons under one fitted estimator. In the shaded region candidates fit the data worse ($\Delta D > 0$) yet save more penalty than they lose in fit ($-\Delta P > \Delta D$), so the joint score prefers them; 11 comparisons fall there. (b) Signed terms of the median reversal by $Q = -\Delta P/\Delta D$.

B.4 Volumetric 3DGS transfer with shared appearance

Repeating the witness construction with native volumetric compositing (25 Gaussians, scales 0.1 m, opacity 0.8) and one shared material from a natural photograph sampled at primitive centers yields 2,548 bounded profile values over two camera baselines, four phases, three grid sizes, and seven displacements. With a coarse 3×3 shared grid, 58 of 96 profiles strictly prefer a wrong surface and 56 of these also lower test RGB error; phase alone changes the sampled minimum in 16 of 24 contexts. A 5×5 grid leaves 20 of 96 strict wrong preferences, and a 7×7 grid none; affine materials and independent colors per Gaussian select the truth in every profile. Rendering directly again reproduces every condition to 3.1×10^{-7} in RGB and 1.5×10^{-10} in objective, with no change of any sampled minimum.

C Estimator and complete derivatives

Native design. For native material coordinates $x_{r\ell}$ and compositing weights $w_{r\ell}$, $P_{rq} = \sum_\ell w_{r\ell}[p_2(x_{r\ell})]_q$, $B_{rj} = \sum_\ell w_{r\ell} k_\Sigma(x_{r\ell}, z_j)$, and $A = [P \;\; B]$. The quadratic trend in six dimensions is unpenalized; the Matérn kernel with smoothness $3/2$, $(1 + \sqrt{3}d/\ell)e^{-\sqrt{3}d/\ell}$ uses $d^2 = (x-z)^\top \Sigma^{-1}(x-z)$ with $\ell = 0.14$, where the covariance Σ weighted by contributions and up to 512 centers (contribution barycenters of sensor rows chosen by farthest point sampling) follow the geometry. For $C = 3$ channels and N rows, $a = [\theta^\top \;\; \beta^\top]^\top$ solves

$$J(g,a) = \frac{\|Aa + b - Y\|_F^2}{2NC} + \frac{\lambda}{2C}\operatorname{tr}(\beta^\top K \beta), \qquad Ma = A^\top(Y - b), \tag{8}$$

with $M = A^\top A + N\lambda \operatorname{diag}(0_6, K)$. At matched geometry and regularization, PROFILE, DATA, and GCV use exactly the same estimate fitted on all training rows; predictive folds share the basis conditioned on geometry but exclude the RGB of the omitted fold from the fit.

Geometry and controls. The controlled height increment is $\delta z = h_0 + h_1 x/0.3 + h_2 y/0.3 + h_3 x^2/0.09 + h_4 xy/0.09 + h_5 y^2/0.09$ (x, y in meters, coefficients in millimeters); mirrored initial coefficients give 10.671 mm height RMSE. The spectral control uses a shared spectral model of order 12, Fine49 a 49×49 bilinear grid, and direct image matching compares observations without an appearance model, all with the same geometry and budget.

Geometry response. With $B = J_{ag}$, coefficient stationarity, and invertible $H = J_{aa}$ on a fixed rendering branch, $da_\lambda/dg = -H^{-1}B$ and $\nabla_g S = S_g - B^\top H^{-1} S_a$ (Golub & Pereyra, 1973; Franceschi et al., 2018). For PROFILE stationarity gives the envelope derivative $\nabla_g F = J_g$; for DATA, $\nabla_g D = L_g + \lambda B^\top H^{-1} R_a$; for GCV with $c = 1 - \mathrm{df}\,/N$,

$$\nabla_g S_{\mathrm{GCV}} = c^{-2}\nabla_g D + \frac{2D}{Nc^3}\nabla_g \,\mathrm{df}, \qquad \mathrm{df} = \operatorname{tr}(M^{-1}A^\top A). \tag{9}$$

The denominator is never clipped. Native updates on reconstructed scenes render and refit both finite difference offsets again, retaining visibility and ordering changes.

D CONTINUATION AT SCENE SCALE ON SEVEN DTU SCENES

Protocol. The cohort comprises all seven standard DTU scenes that are disjoint from the eight scenes of the first refinement cohort (scans 97, 105, 106, 110, 114, 118, 122), which later also form the DTU confirmation cohort; no scene was selected by outcome. Each scene uses a 2DGS host trained on RGB alone, with SH of degree 3, for 15k iterations from its COLMAP initialization at a width of 160 pixels, with every eighth view held out. On the fixed host geometry, SH of degree 0 or 3 receives 500 warmup steps that update appearance only. Two branches then continue for 2,500 steps from that identical state, with the same optimizer moments and camera order, either appearance only with all geometric parameters frozen (checked for bitwise equality) or joint updates of every center, rotation, scale, opacity, and SH coefficient. Topology is fixed. The photometric objective is $0.8L_1 + 0.2(1 - \mathrm{SSIM})$; the branch with normal consistency adds the 2DGS term with weight 0.05. Test PSNR averages the PSNR of the eight test views. Meshes are extracted by TSDF fusion with one fixed recipe and evaluated with the official DTU protocol (0.2 mm sampling, observation mask, 20 mm cutoff, no fitted alignment); frozen surfaces are extracted directly from the frozen checkpoints and coincide with the host surface.

Outcomes. Table A3 lists every scene and Figure A2 shows every test render and surface. Under SH0 and the photometric objective, all seven scenes improve test PSNR ($+2.47$ to $+3.38$ dB) and worsen Chamfer distance ($+0.13$ to $+0.66$ mm) and accuracy ($+0.33$ to $+1.06$ mm); exact sign flip tests over scenes give $p = 0.016$ for each, and the scene bootstrap interval of the Chamfer change is $[+0.21, +0.48]$ mm. Training PSNR rises by 5.8 dB. Completeness changes little (mean $+0.03$ mm), so the damage is surface placed away from the reference while coverage remains. Normal consistency reduces the Chamfer increase to $+0.12$ mm, yet accuracy still worsens in all seven scenes. Under SH3, test PSNR changes by $+0.11$ dB (photometric) or -0.20 dB (normal consistency), and Chamfer by $+0.02$ or -0.01 mm. Because the SH0 branch restricts appearance on geometry fitted with SH3, the experiment measures how much of a finite appearance deficit geometry absorbs; it contrasts trajectories with finite budgets at reduced resolution, and profiled optima are the subject of Appendix E.

Table A3: **Every scene of the experiment at scene scale.** Test PSNR and official DTU surface metrics (mm) after 2,500 matched steps. The first six value columns compare SH0 with geometry frozen (F) or optimized (O) under the photometric objective; the last three give the Chamfer change for the remaining conditions.

	SH0, photometric						SH0, normal	SH3, photo.	SH3, normal
Scan	PSNR F	PSNR O	Chamfer F	Chamfer O	Acc. F	Acc. O	ΔChamfer	ΔChamfer	ΔChamfer
97	23.69	26.56	1.79	2.02	1.55	1.90	+0.14	+0.07	0.00
105	25.48	28.33	1.29	1.95	1.19	2.17	+0.26	+0.04	0.00
106	30.14	32.86	1.50	1.79	1.38	1.97	−0.03	+0.05	−0.02
110	28.33	31.70	2.89	3.08	2.89	3.28	+0.23	−0.15	−0.02
114	27.01	29.48	1.47	1.60	1.34	1.67	0.00	0.00	−0.01
118	31.28	34.33	1.59	1.89	1.47	2.21	+0.06	+0.03	0.00
122	31.02	33.75	1.28	1.83	1.24	2.30	+0.21	+0.10	0.00

PGSR host. PGSR (Chen et al., 2025) (official code, commit de24f1a) is trained for 15k iterations of its default schedule on the identical 160 pixel images, split, and cameras, with SH3, its full native loss, and NCC on the training resolution (patch 7×7), starting from the same sparse points after its own reprojection filter. The continuation repeats the 2DGS protocol with inherited optimizer moments, the same camera permutation, fixed topology, and a bitwise frozen geometry check. The photometric arm uses $0.8L_1 + 0.2(1 - \mathrm{SSIM})$, and the full arm adds PGSR's scale, single view normal, multi view geometric, and multi view NCC terms at their late training weights. Because these terms have no gradient with respect to SH, one frozen branch serves both arms. Meshes fuse PGSR's planar depth with the TSDF recipe of the 2DGS experiment, a choice registered before any evaluation, and the verbatim evaluator reproduces the stored 2DGS host of scan105 (1.2905 mm). PGSR's own TSDF

Table A4: **Every scene of the continuation at scene scale, in all four settings.** Entries give the change after 2,500 matched steps (optimized minus frozen geometry). Frozen surfaces equal the host, whose Chamfer distance (CD, mm) comes first. Acc. is accuracy, NC the 2DGS normal consistency term, and orange marks worsening above 0.05 mm. The last row counts scenes with better PSNR or a worse surface, and 7/7 gives $p = 0.016$ in an exact sign flip test.

	Host	SH0, photometric			SH0, +NC			SH3, photometric			SH3, +NC		
Scan	CD	PSNR↑	CD↓	Acc.↓	PSNR	CD	Acc.	PSNR	CD	Acc.	PSNR	CD	Acc.
97	1.79	+2.87	+0.22	+0.34	+2.36	+0.14	+0.19	+0.30	+0.07	+0.14	−0.02	0.00	+0.04
105	1.29	+2.85	+0.66	+0.98	+2.22	+0.26	+0.37	+0.03	+0.04	+0.10	−0.23	0.00	0.00
106	1.50	+2.72	+0.28	+0.59	+2.47	−0.03	+0.06	+0.10	+0.05	+0.12	−0.17	−0.02	−0.01
110	2.89	+3.38	+0.19	+0.39	+2.71	+0.23	+0.22	+0.12	−0.15	−0.06	−0.22	−0.02	−0.03
114	1.47	+2.47	+0.13	+0.33	+1.94	0.00	+0.07	+0.09	0.00	+0.05	−0.23	−0.01	−0.02
118	1.59	+3.04	+0.30	+0.73	+2.30	+0.06	+0.17	+0.09	+0.03	+0.13	−0.21	0.00	+0.01
122	1.28	+2.73	+0.54	+1.06	+2.23	+0.21	+0.42	+0.06	+0.10	+0.21	−0.33	0.00	−0.01
Mean	1.69	**+2.87**	**+0.33**	**+0.63**	+2.32	+0.12	+0.22	+0.11	+0.02	+0.10	−0.20	−0.01	0.00
Scenes		7/7	**7/7**	**7/7**	7/7	5/7	**7/7**	7/7	5/7	6/7	0/7	3/7	3/7

extraction agrees, with $+0.53$ mm CD under the photometric arm and $+0.04$ mm under the full loss. The two NCC ablations were added after the pilot and are labeled post hoc.

Table A5: **Every scene of the PGSR host under the protocol of Table 3.** Host Chamfer distance (mm) and change after 2,500 matched steps (optimized minus frozen geometry) in test PSNR (dB), Chamfer distance, and accuracy (mm), with the preregistered TSDF recipe. The last two rows give seven scene means of the post hoc SH0 ablations. Orange marks worsening above 0.05 mm.

	Host	SH0, photometric			SH0, full PGSR loss			SH3, full loss	
Scan	CD	PSNR	CD	Acc.	PSNR	CD	Acc.	PSNR	CD
97	1.67	+3.45	+0.20	+0.24	+3.10	−0.01	−0.01	+0.04	+0.01
105	1.33	+2.72	+0.47	+0.42	+2.32	+0.04	+0.03	+0.06	−0.01
106	1.14	+2.75	+0.46	+0.59	+2.00	−0.03	−0.02	−0.14	−0.03
110	3.08	+4.90	+0.03	−0.05	+3.91	+0.07	+0.02	−0.18	+0.01
114	0.98	+3.26	+1.22	+1.73	+2.30	+0.05	+0.06	−0.27	−0.01
118	1.72	+3.15	+0.03	+0.23	+2.32	−0.01	−0.01	−0.22	−0.01
122	1.09	+3.05	+0.34	+0.41	+1.96	+0.01	+0.02	−0.31	+0.02
Photometric + NCC		+2.86	−0.01	+0.01					
Full loss without NCC					+2.81	+0.32	+0.26		

Image metrics. We render the eight test views of every saved endpoint of both hosts without further training and add SSIM (Wang et al., 2004) with the 11×11 Gaussian window of the host code and LPIPS (Zhang et al., 2018) with VGG and AlexNet features to PSNR. Recomputed PSNR and SSIM agree with the stored values for all 1,120 test views. In every setting where freeing geometry raises PSNR in all seven scenes, SSIM and VGG LPIPS improve in all seven scenes as well, with one exception in scan105 under SH3 and the photometric loss (-0.0005 SSIM). In the two SH3 settings where PSNR falls, VGG LPIPS still improves in six of seven scenes (Table A6).

E Primal and dual certificates for preferences on reconstructed scenes

Appearance class. For each frozen state we retain the geometry, sensor identities, fixed exterior, and every native compositing contribution. For one channel let $c_0 \in \mathbb{R}^K$ be the common SH anchor of degree 3 ($K = 16P$ for P primitives) and $c = c_0 + \delta$. Whenever every retained training contributor has a nonnegative color before clamping, the training prediction is $Bc + b$, where B holds native SH evaluations and compositing weights and b the fixed exterior and color offset. With C evaluating every pair of primitive and direction that contributes to training, we impose $C(c_0 + \delta) + \frac{1}{2}\mathbf{1} \succeq 0$;

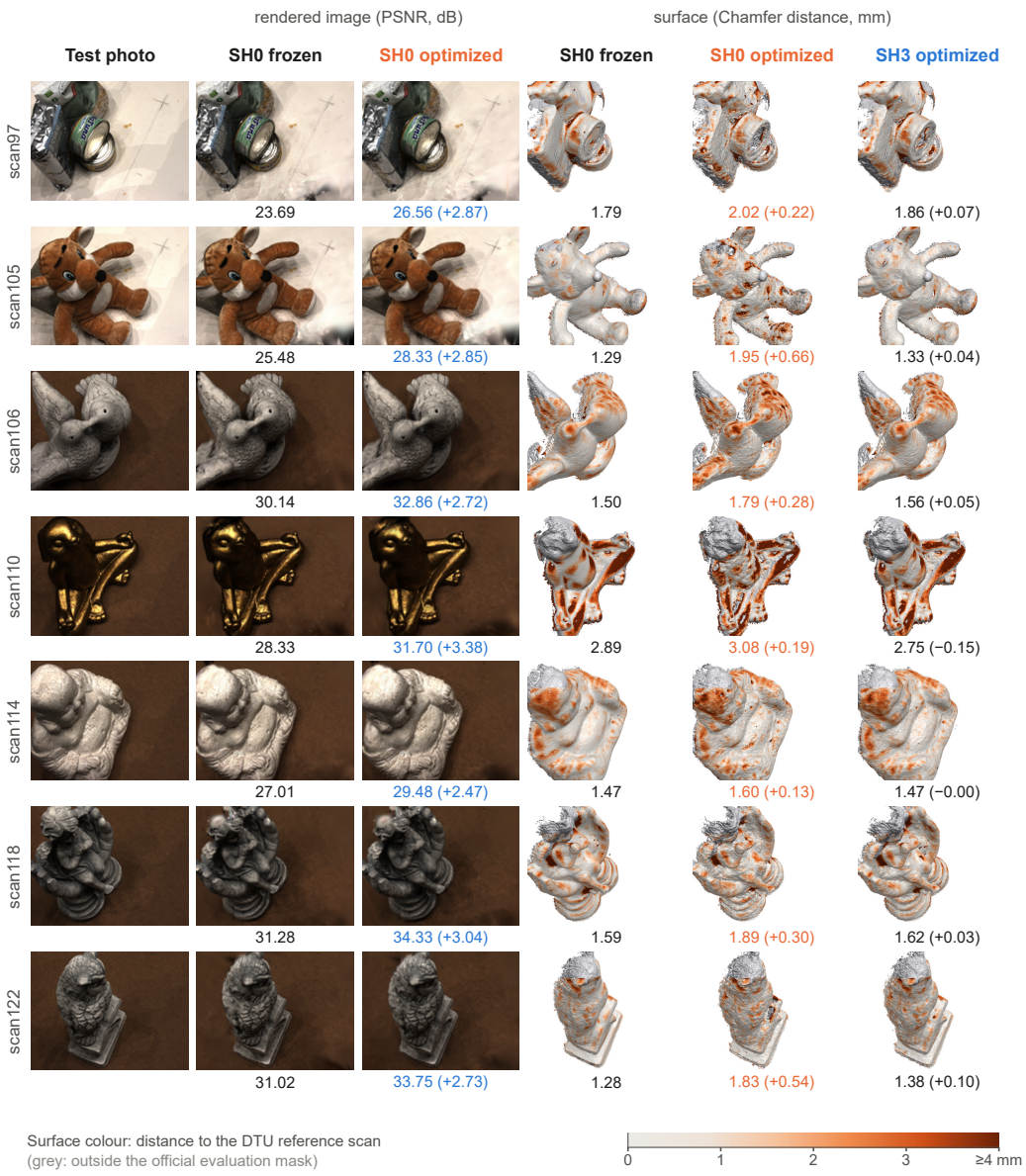

Figure A2: **All seven scenes of the experiment at scene scale, without selection.** Each row shows the test photograph at evaluation resolution, SH0 renders with geometry frozen and optimized (test PSNR below), and the extracted surfaces together with the SH3 optimized surface (Chamfer distance below; change from frozen in parentheses). Surface color is the distance to the DTU reference scan inside the official evaluation mask. Every SH0 row shows the same pattern of better images and worse surfaces.

Table A6: **Image metrics agree with PSNR on both hosts.** Mean change over seven DTU scenes after 2,500 matched steps (optimized minus frozen geometry) in test PSNR (dB), SSIM, and LPIPS with VGG and AlexNet features, each averaged over the eight test views, and scenes that improve (higher PSNR and SSIM, lower LPIPS). Seven of seven gives $p = 0.016$ in an exact sign flip test.

Host	Setting	PSNR↑		SSIM↑		LPIPS, VGG↓		LPIPS, Alex↓	
		Δ	Scenes	Δ	Scenes	Δ	Scenes	Δ	Scenes
2DGS	SH0, photometric	+2.87	7/7	+0.0152	7/7	−0.0144	7/7	−0.0104	7/7
	SH0, +NC	+2.32	7/7	+0.0112	7/7	−0.0087	7/7	−0.0073	7/7
	SH3, photometric	+0.11	7/7	+0.0017	6/7	−0.0067	7/7	−0.0025	6/7
	SH3, +NC	−0.20	0/7	−0.0002	5/7	−0.0024	6/7	−0.0003	4/7
PGSR	SH0, photometric	+3.33	7/7	+0.0169	7/7	−0.0153	7/7	−0.0124	7/7
	SH0, full PGSR loss	+2.56	7/7	+0.0123	7/7	−0.0100	7/7	−0.0092	7/7
	SH3, photometric	+0.17	7/7	+0.0018	7/7	−0.0073	7/7	−0.0016	7/7
	SH3, full PGSR loss	−0.14	2/7	−0.0001	2/7	−0.0027	6/7	+0.0005	2/7
Post hoc	SH0, photometric +NCC	+2.86	7/7	+0.0139	7/7	−0.0121	7/7	−0.0108	7/7
	SH0, full without NCC	+2.81	7/7	+0.0136	7/7	−0.0106	7/7	−0.0103	7/7

there is no upper bound and no constraint on unused pairs, and test predictions keep the true clamp. With $h = y - b - Bc_0$, $\ell = -\frac{1}{2}\mathbf{1} - Cc_0$, and $\alpha = N\lambda/K$ ($\lambda = 10^{-5}$), the channel problem is

$$p^\star = \min_{\delta,r} \frac{\alpha}{2}\|\delta\|^2 + \frac{1}{2}\|r\|^2, \qquad B\delta - r = h, \quad C\delta \succeq \ell, \tag{10}$$

and summing channel energies divided by $3N$ gives half the training MSE plus $\frac{\lambda}{2}\,\mathrm{mean}(\delta^2)$.

Bounds. Let $z = (\delta, r)$, $Q = \mathrm{diag}(\alpha I, I)$, $A = \begin{bmatrix} B & -I \\ C & 0 \end{bmatrix}$, $d = (h, \ell)$, and $u = (\eta, \nu)$ with $\nu \preceq 0$. Any feasible primal and any such dual give

$$\begin{aligned} L(u) &= -\tfrac{1}{2}(A^\top u)^\top Q^{-1}(A^\top u) - d^\top u \;\le\; p^\star \;\le\; U(z) = \tfrac{1}{2} z^\top Q z, \\ U - L &= \frac{\|\alpha\delta + B^\top\eta + C^\top\nu\|^2}{2\alpha} + \frac{1}{2}\|r - \eta\|^2 - \nu^\top s, \end{aligned} \tag{11}$$

with slack $s = C\delta - \ell$; neither stationarity nor a solver flag is needed. OSQP proposes a primal and dual; a DC repair per primitive makes the primal strictly feasible, and multipliers from sign projection and nonnegative least squares tighten the dual. An independent implementation reconstructs the design by direct SH evaluation, reloads the saved witnesses, and checks feasibility, the native objective, and the gap identity in extended floating point precision. A strict preference requires $U_g + 10^{-12} < L_{g_0}$; overlapping intervals abstain.

Outcomes. All 607 states pass verification, 256 and 56 states in the first DTU and Replica cohorts and 135 and 160 in the confirmation cohorts (per cohort outcomes in Table 4). The smallest ratio of a harmful margin to the combined width of its two intervals is $34.2\times$ and $12.9\times$ in the first DTU and Replica cohorts and $96.3\times$ and $14.4\times$ in the confirmation cohorts. The total of 19/23 scenes counts distinct scene identities (room0 and office0 appear in both Replica cohorts). Bound widths stay below 4×10^{-8}, the native objective and the saved primal agree to 2.2×10^{-19}, and the gap identity closes to 4.5×10^{-22}. Seventeen adversarial witness controls behave as specified. Corrupted provenance, wrong dual signs, and infeasible primals are rejected, while valid but weak duals only widen the interval. The bounds are exact for the stated convex subclass of clamped SH, and the initial reconstruction serves as a reference state, with physical truth given by the reference scan or mesh. Fitting uncertainty in the unrestricted shader is large by comparison, since raising native SH fitting from 60 to 300 Adam steps lowers the median objective by 52.9%, and L-BFGS polishing by a further 16.1%.

F Controlled recovery and replication

Closed form color baseline. We adapt the regularizer of Fang et al. (2023) to the benchmark as $S_{\mathrm{CF}} = F + \lambda_{\mathrm{cf}} L_{\mathrm{cf}}$. At every native render sample, SH coefficients follow from the transmittance

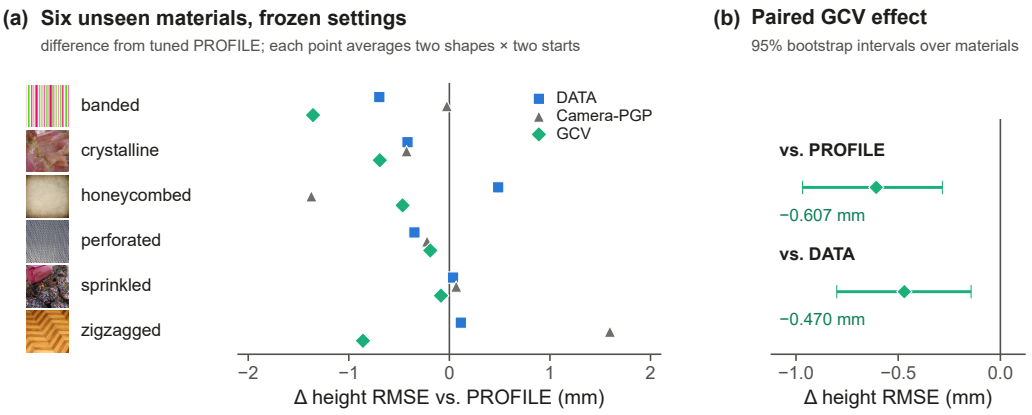

Figure A3: **The score effect transfers to unseen materials.** (a) The six DTD source images and each material's error difference from tuned PROFILE (two shapes × two mirrored starts); GCV improves every material. (b) Paired 95% bootstrap intervals over materials; GCV−DATA shares the estimator fitted on all training rows at matched geometry and $\lambda = 10^{-6}$, isolating the complexity correction.

weighted projection of the observed colors of all training cameras with the residual scheme of the original method, the closed form pixels are composited with the native weights and exterior term, and L_{cf} uses the normalization of the data term. The method sees the same sampled observations, estimator, starts, optimizer, and budget as every other row, with gradients through the full closed form estimate. The grid $\lambda_{\mathrm{cf}} \in \{0.1, 1, 10, 100\}$ times SH degree $\{0, 2\}$ is searched on the development materials and frozen, and the published setting ($\lambda_{\mathrm{cf}} = 10$, degree 2) gives 6.42 mm. With the shared sparse observations, L_{cf} attains its minimum 4.4 mm toward the cameras on average, whereas a dense observation oracle moves the minimum to within 1 mm of the truth, so the observation budget shared by all methods limits this photo consistency term.

Multi view NCC baseline. We port the multi view photometric term of PGSR (Chen et al., 2025) as $S_{\mathrm{NCC}} = F + \lambda_{\mathrm{ncc}} \kappa L_{\mathrm{ncc}}$, with per pixel planes from the native depth and normal, homography warps of gray patches into neighbor views selected by the PGSR rule, its NCC with occlusion aware weights, and $\kappa = F(g_0)/L_1(g_0)$ from the start fit, which expresses the PGSR weight relative to the image loss in the units of F. The sampled observations cannot form a 7×7 pixel patch, so patches are 3×3 or 5×5 grid neighborhoods, and the grid $\lambda_{\mathrm{ncc}} \in \{0.01, 0.1, 1, 10\}$ times both patches is searched on the development materials and frozen. PGSR's own weight (0.15, 3×3 patch) gives 5.59 mm on the default block and 4.66 mm on unseen materials. With a dense 7×7 pixel patch, L_{ncc} attains its minimum at the truth in all 8 diagnostic cases, whereas on the sampled grid its profile is nearly flat, so the protection that NCC gives PGSR on dense images (Appendix D) depends on dense observations. Several NCC and closed form runs from the truth stop early, which contributes to their small drift.

Controlled preservation. From a correct start (18 conditions, 40 opportunities), final height error is 4.74 mm for PROFILE, 2.63 for DATA, 3.03 for Camera-PGP, and 2.66 for GCV, so separating the score removes 44% of the joint score's drift. Under weak parallax (18 conditions), GCV reaches 6.60 mm against PROFILE's 8.24 mm and improves all nine material means; with image noise ($\sigma = 0.01$) GCV reaches 4.39 mm against PROFILE's 9.25 mm. In the predictive branch, retaining the estimator response lowers Camera-PGP error from 6.43 to 5.13 mm.

Response ablation and replication. FULL and STOP-COEF share the regularized solve, forward scores, and every geometry derivative except the response of the fitted coefficients; every trial refits appearance and is accepted by the actual GCV score. A discovery on six sources suggested that FULL improves recovery but increases drift from a correct start. We then froze the contrast between the two arms and both endpoints before selecting twelve new DTD photographs (chequered, cobwebbed, cracked, frilly, gauzy, lined, marbled, polka-dotted, porous, stratified, striped, swirly) and completed all 144 prescribed trajectories.

Table A7: Independent replication (twelve materials). Differences are STOP-COEF minus FULL (mm) with paired bootstrap intervals over materials; exact sign flip tests over materials are adjusted by Holm's method jointly across both endpoints.

Endpoint	FULL	STOP-COEF	Difference [95% CI]	Holm p
Recovery RMSE	**3.167**	3.798	$+0.632\ [0.370, 0.881]$	.00244
Truth drift	1.727	**0.762**	$-0.965\ [-1.216, -0.727]$	.000977

FULL moves in all 24 runs from a correct start, while STOP-COEF takes no accepted step in six; mean accepted steps are 26.6 versus 4.4. Recomputation from the raw logs reproduces all 144 chains, and cached directional derivatives agree with the FULL response to within relative 2×10^{-5}.

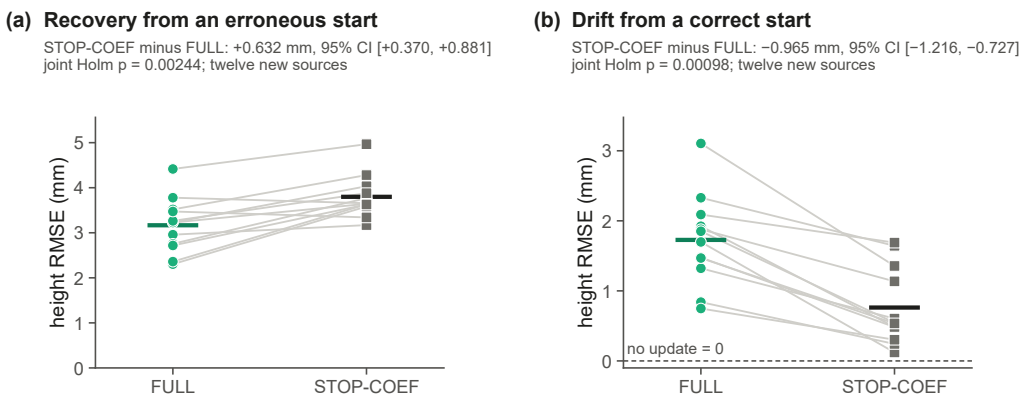

Figure A4: **Recovery and preservation respond differently.** Each connected pair is one of twelve replication materials averaged over its repeats; bars mark means. FULL improves recovery on 10/12 materials, whereas STOP-COEF drifts less from a correct start on all twelve.

Computational cost. We time the unmodified benchmark code on the default setting of Table 1 for three materials in two contexts with three repeats (one RTX 5090 Laptop GPU, one CPU thread, no other GPU process). Timers wrap each component along the actual trajectories, and the complete response is isolated as the difference between backward passes with and without the dependence of the fitted coefficients. The timed runs start from the stored initial scores and reproduce the stored endpoints to within 0.01 mm in 28 of 30 conditions, with the remainder inside the run to run spread of GPU atomics. DATA costs the same per step as PROFILE, because the benchmark differentiates PROFILE through the coefficient solve as well, where its response term vanishes by stationarity. GCV adds the trace of its degrees of freedom, and the fold refits of Camera-PGP and Pixel-CV raise refit and response costs about fivefold (Table A8). Over the 508 refinements on reconstructed scenes, the paired median time per objective call relative to PROFILE is 1.00 for DATA, 0.95 to 1.02 for GCV, 1.10 to 1.23 for the predictive scores, and 1.58 to 2.20 for SH refitting across the four cohorts, measured on a shared GPU.

G Refinement cohorts on reconstructed scenes

Cohorts. The first cohort refines four regions in each of eight DTU scenes (24, 37, 40, 55, 63, 65, 69, 83) and two Replica scenes (office0, room0; 150 mm regions) with PROFILE, DATA, Camera-PGP, GCV, Pixel-CV, and native SH (fixed, or refitted with 60 Adam steps). The confirmation cohort trains new common hosts for the seven further DTU scenes (27 qualified regions; scan114 qualifies three) and all eight Replica scenes (32 regions of 300 mm radius) and runs PROFILE, DATA, GCV, and native SH refitting with 40 opportunities each; its protocol and primary contrasts were fixed before any complete aggregate.

Table A8: **Computational cost of the evaluated geometry scores.** Milliseconds per geometry step on the default controlled setting, as the mean over six conditions of per condition medians, for rendering with the native operator and its backward pass, the appearance refit, the score including degrees of freedom, and the complete response term, together with the measured step, seconds per run, and changes relative to PROFILE. Rendering runs identical code for every score, so its differences reflect GPU latency, and runs of GCV and the predictive scores use more of the 60 opportunities, which enlarges their cost per run.

	Component (ms per step)				Step		Run	
Score	Render	Refit	Score	Response	ms	Δ	s	Δ
PROFILE	60.7	8.5	2.2	5.1	80.6		4.19	
DATA (ours)	60.2	8.4	2.1	5.1	79.9	−0.9%	4.34	+3.6%
GCV (ours)	53.0	7.8	12.1	0.9	83.8	+3.9%	4.92	+17.4%
Camera-PGP	57.4	43.8	5.3	28.6	129.8	+61.0%	7.51	+79.2%
Pixel-CV	57.7	44.8	5.4	28.6	130.9	+62.4%	7.39	+76.4%

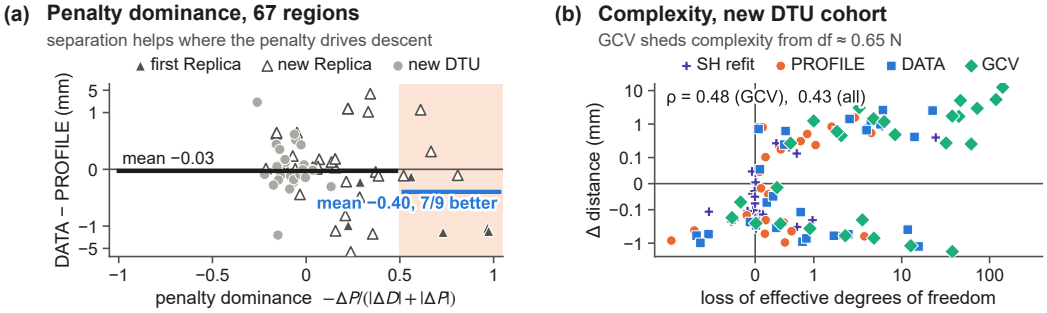

Figure A5: **Where each score lever acts on scene reconstructions.** (a) Penalty dominance of PROFILE's descent against the final distance of DATA minus PROFILE in the 67 regions with a recorded decomposition, with means as bars and filled triangles for the first Replica cohort. The nine penalty dominated regions come from six Replica scenes, and a sign flip test over them gives $p = 0.195$. (b) Loss of effective degrees of freedom against change in distance on the new DTU cohort.

Primary contrasts of the confirmation cohort. With scenes as units (20,000 scene bootstrap resamples; exact sign flip tests, adjusted by Holm's method over the two score contrasts within each dataset), DATA−PROFILE is $+0.044$ mm on DTU ($[-0.148, 0.252]$, DATA closer in 5/7 scenes, $p = 0.77$) and $+0.001$ mm on Replica ($[-0.549, 0.469]$, $p = 1.0$). GCV−PROFILE is $+1.161$ mm on DTU ($[0.592, 1.796]$, 0/7, $p = 0.031$) and $+0.483$ mm on Replica ($[-0.129, 1.264]$, $p = 0.55$). One DTU GCV run reaches the boundary of its training support; excluding that region for every method still leaves GCV farther from the reference in all seven scenes.

Score mechanics. On the common kernel head, the penalty is 65% of the initial objective in the first Replica cohort, against 6% (DTU) and 33% (Replica) in the confirmation cohort. PROFILE's descent improves the data term in all 27 confirmation DTU regions and in 31 of 32 Replica regions, and the penalty change is a median 10% (DTU) and 23% (Replica) of the data change; in the first Replica cohort it is 125%. The effective degrees of freedom start at $0.65N$ (DTU) and $0.61N$ (Replica). Along the DTU refinements, GCV endpoints lose 23.9 effective degrees of freedom on average against 2.4 (PROFILE) and 3.3 (DATA); across all DTU endpoints the loss correlates with added distance (Spearman $\rho = 0.43$, $p = 3 \times 10^{-6}$). The first DTU cohort did not record the decomposition, so the regime analysis of Section 4.5 uses the other 67 regions.

Contribution depth. CDE uses full native rays on fixed queries of the projected reference; empty native support keeps its reference ray with a point mass at zero depth, missing references are never counted as zero error, and squared errors are aggregated before roots. In the confirmation DTU cohort, 26 of 27 regions have support in the projected reference (17,948 valid queries); among 48 endpoints

Table A9: Final distance to the reference (mm), averaged over regions within scenes and then scenes. Change is endpoint minus initial state. The two datasets use different reference domains and are not pooled.

	First cohort		Confirmation cohort			
Method	DTU (8)	Replica (2)	DTU (7)	change	Replica (8)	change
Initial reconstruction	12.484	4.291	28.088	0	13.993	0
PROFILE	12.434	5.189	28.183	$+0.095$	**12.697**	-1.295
DATA	12.498	4.425	28.227	$+0.139$	12.698	-1.295
GCV	12.590	4.675	29.344	$+1.256$	13.180	-0.812
Camera-PGP / Pixel-CV	12.516 / 12.548	4.460 / 4.559				
Native SH, fixed	**12.145**					
Native SH, refit	12.206	**4.285**	**28.074**	-0.014	13.896	-0.097

that improve both training and test RGB by more than 10^{-8}, CDE flags 15 deteriorations and mean depth 12, with three visible to CDE alone. The first cohort yields the 99 endpoints of Figure 5.

Table A10: **CDE flags deterioration that mean depth misses.** DTU endpoints that improve training and test RGB. A flag marks worsening above 0.05 mm, and better means an improvement above 0.05 mm.

DTU cohort	First	Confirm.
Endpoints	99	48
CDE flags	**43**	**15**
Mean depth flags	27	12
CDE only	17	3
mean depth better	14	2
Mean depth only	1	0

H NATIVE VOLUMETRIC SH REGIMES

Setup. The installed volumetric 3DGS CUDA rasterizer renders 25 anisotropic Gaussians (scales $(0.09, 0.075, 0.045)$ m, opacity 0.82, fixed rotations) whose centers sample a saddle or a rippled surface; two height modes, orthonormal in RMS, move the centers, and error is the RMS center displacement. Eight cameras render 48×48 images (four train, four held out; two train in the sparse regime); wide and weak camera rings have radii 0.48 and 0.035 m at 1.2 m depth. The native shader computes $\max(0.5 + \mathrm{SH}_3, 0)$. The tied head holds a shared 3×3 spatial field of SH coefficients of degree 3; the free head gives every primitive its own coefficients. Appearance minimizes half mean squared error plus $10^{-4}/2$ times the mean squared coefficients (L-BFGS, up to 60 iterations when the clamp is active); JOINT and DATA use the same fitted appearance at matched geometry, and FIXED freezes the initial fit. Twelve photo sources, three starts, three arms, and eight regimes give 864 trajectories of local coordinate descent (± 2 mm moves, at most 16 opportunities).

Outcomes. Table A11 lists every regime. The 12 photo sources are reused across regimes, so intervals are descriptive paired bootstraps over sources. With the tied head, DATA minus JOINT in recovery is -2.63 mm on the ripple with a directional lobe ($[-4.62, -0.92]$; DATA better for 8 sources, worse for 1) and $+0.81$ mm with weak cameras ($[0.11, 1.67]$), the one tied regime where JOINT recovers better.

Sensitivity. With 180 instead of 60 inner iterations (81 repeats), saddle terminals are unchanged and runs with free SH still recover the surface, while tied ripple/lobe runs move by up to 14.4 mm and keep DATA ahead (9.09 versus 12.85 mm on the subset of three sources). Under a matched native SH coefficient prior (108 trajectories), the gap between tied and free heads persists (13.2 to 13.7 versus 0.26 mm) and DATA's small recovery advantage reverses (13.69 versus 13.16 mm). The capacity of the head therefore dominates, and the score adds a smaller effect that depends on the prior.

Table A11: **Capacity dominates, and with free or realizable SH both scores recover the surface.** Mean center RMSE (mm) of 24 trajectories per native volumetric regime. No update gives 8.94 mm from signed starts and 0 from the truth, and orange marks recovery worse than no update. FIXED keeps the initial appearance, and heads are tied unless noted.

	Recovery			Drift from truth		
Regime	JOINT	DATA	FIXED	JOINT	DATA	FIXED
Saddle, diffuse	8.63	7.64	6.48	8.68	7.79	3.52
Saddle, lobe	10.76	9.14	9.50	7.16	6.04	9.08
Ripple, lobe	14.04	11.41	10.89	10.71	8.59	9.66
Ripple, noisy lobe	14.33	13.13	11.08	11.28	8.57	10.09
Ripple, weak cameras	6.77	7.57	7.93	1.54	1.71	5.17
Ripple, two cameras	6.96	6.84	6.98	2.83	2.65	1.37
Ripple, free SH	**0.26**	**0.26**	5.79	**0.00**	**0.00**	**0.00**
Ripple, realizable	**0.00**	**0.00**	5.21	**0.00**	**0.00**	**0.00**