
CHARACTERISING PER-PIXEL RENDERING DIFFICULTY BY TRANSPORT MECHANISM

Po-Ting Lin
Independent Researcher
Tainan, Taiwan
botimlin@gmail.com

ABSTRACT

Per-pixel rendering difficulty is conventionally characterised by one noisy scalar: the sample variance of a Monte Carlo estimator. We argue that it should instead be characterised through *transport structure* — a discrete description of how each contribution’s energy reaches the sensor, deterministic under stated renderer conventions — with variance treated as a measurement whose reliability that structure helps predict. We make this concrete with a seven-class transport-mechanism descriptor assigned per contribution event, evaluate it on eleven scenes, and measure variance reliability on the seven first-party ones. The dominant label agrees 87–99.6% between 64 and 4096 samples per pixel, where quantile-binned variance agrees as little as 21%; its stability on unseen scenes is predicted from their pilots. Conditioning a pilot variance on the label improves equal-budget sample allocation wherever heavy-tailed buckets carry appreciable population, reduces to the incumbent where they are absent or negligible, and is neither reproduced by a random partition nor absorbed by a median-of-means estimator; where the pilot fails for other reasons, as on the classical ajar-door scene (6.8 dB below uniform), the label says so from the pilot alone.

Keywords Monte Carlo rendering · light transport · transport mechanism · per-pixel rendering difficulty · variance reliability · adaptive sample allocation

CCS Concepts: • *Computing methodologies* → *Ray tracing; Rendering.*

1 Introduction

For thirty-five years the operational signal of adaptive Monte Carlo rendering has been one scalar: the per-pixel sample variance $\hat{\sigma}^2(p)$ of an unbiased pixel estimator. Adaptive samplers refine on it, denoisers steer their filter footprint by it, efficiency-aware Russian roulette branches off it, and inverse-rendering pipelines differentiate through it. It is also, implicitly, the *reference*: the quantity difficulty predictors are trained against and evaluated on. The assumption underneath is that a variance estimated from a pilot render is a reliable measurement of how hard each pixel is.

This paper makes a methodological proposal and supports it with measurements. The proposal is that per-pixel difficulty should be characterised through *transport structure* — the discrete mechanism by which each contribution’s energy reaches the sensor, deterministic under stated renderer conventions — and not only through a noisy continuous statistic. Variance does not disappear from this picture; it becomes a measurement whose reliability is itself to be measured, and which the structural description helps predict. The measurements show three things. First, the reliability of variance-derived difficulty references is low exactly where difficulty concentrates, and the failure is associated with specific transport mechanisms. Second, a mechanism descriptor cheap enough to read from any path tracer is stable where variance is not (Fig. 1), and its own reliability on three third-party scenes was predicted from their pilots. Third, conditioning the variance estimate on the descriptor repairs it where it is least reliable, with an effect a random partition does not reproduce and a median-of-means estimator does not absorb.

Why variance is an unreliable reference under heavy tails. Three distinct quantities are routinely conflated: the variance σ^2 of individual per-sample contributions, the variance σ^2/N of the N -sample pixel mean, and the

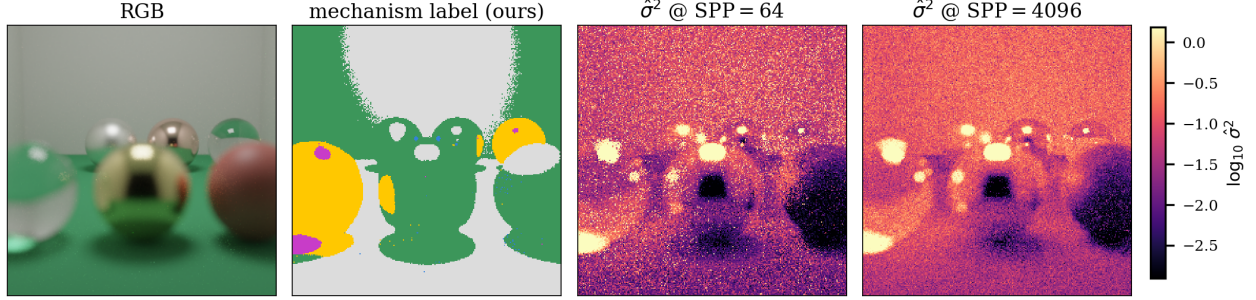


Figure 1: The discrete mechanism label (second panel) is determined at SPP= 64 and barely changes at SPP= 4096; the residual disagreement sits on the boundary between two light-tailed buckets (§4). The per-pixel sample variance (right two panels) drifts substantially across the same $64\times$ budget change. *Under heavy-tailed transport, finite-sample variance is an unstable measurement; the transport mechanism of each contribution is a stable structure — deterministic under stated renderer conventions — to characterise difficulty by, and to condition on.*

sampling uncertainty of the estimator $\hat{\sigma}_N^2$. It is the third that matters for a reference signal. Under finite-fourth-moment assumptions, the relative standard deviation of the sample-variance estimator satisfies $\sqrt{\text{Var}(\hat{\sigma}_N^2)}/\sigma^2 \approx \sqrt{(\kappa - 1)/N}$, where κ is the integrand’s fourth standardised moment [1, 2]. Three regimes follow. When κ is finite but large — caustic and near-specular integrands — pilot-budget estimates are useless in practice. When $\kappa = \infty$ with σ^2 finite, $\hat{\sigma}_N^2$ remains consistent but admits no finite-rate Gaussian limit [3], so no error bar of the usual form exists. When σ^2 itself is infinite — unbounded contributions near delta paths [4] — the target of the estimate does not exist. Estimating κ from a pilot to budget accordingly fails by regress: it needs the eighth moment. The rendering literature knows this pathology under other names (weak singularities [5], infinite Hardy–Krause variation [6]) and suppresses it at the *mean* estimator [7, 8]; its implication for $\hat{\sigma}^2$ as a reference has not been examined.

The implication is measurable. Rank-correlating per-pixel $\hat{\sigma}^2$ between a 64-SPP pilot and a 4096-SPP render gives Spearman $\rho = 0.26$ on a caustic-dominated scene (§5.1). Splitting ten independent seeds into two halves and fitting a per-pixel convergence exponent on each — $\sim 40,000$ samples per pixel in total — gives a split-half correlation of only 0.23–0.29 on the two geometrically simplest scenes of our test set. In the language of classical measurement theory [9, 10] this is a *reliability* measurement: two independent measurements of the same target barely agree, and any correlation between that target and a predictor is attenuated accordingly. We report it as an empirical reliability figure; turning it into a formal ceiling on predictor performance would require measurement-error assumptions we do not need, since low reliability of the evaluation target is damaging under any of them. Robust estimators of mean and variance under heavy tails exist [11, 12, 13] and qualify this critique; §2 discusses them and §7 tests one.

One objection deserves a direct answer. Where σ_∞^2 exists, the σ^2 -field is a deterministic functional of the integrand; the stored buffer $\hat{\sigma}_N^2/N$ is unbiased for σ_∞^2/N , rank correlations are invariant to the global $1/N$ factor, and cross-SPP experiments reporting near-unity *shape* stability in this regime are not in error [14, 15]. What is not stable is the per-pixel *rank* on a scene that contains heavy-tailed pixels, because those pixels do not contract uniformly with N . A difficulty reference must rank pixels consistently across the entire scene, including exactly the pixels where σ_∞^2 fails to exist or is recoverable only through fourth-moment-limited estimators.

What we propose instead — and what we do not claim. Three notions must be kept apart. *Rendering difficulty* is operational and admits several non-equivalent definitions — variance under a fixed estimator, samples to an error threshold, time to quality; we measure the first and third and claim no single scalar ordering. *Transport mechanism* is structural: how a contribution’s energy reaches the sensor. A mechanism class does not by itself order pixels by cost — our own results show large cost heterogeneity within classes (§5.2) — so it is not a difficulty measure; it is an explanatory and *conditioning* variable for one. *Stability* is what makes a signal usable as a reference, and proves neither of the other two. Our claim, in this vocabulary, is that the community conditions on an unstable measurement when a stable structural descriptor of the same pixels is available for free. The descriptor we use is the transport mechanism of each contribution event — the BSDF lobe at its end-vertex, whether the path to that vertex passes through a delta interface, and a single- versus multi-bounce distinction — which yields seven mutually exclusive labels assigned from quantities the path tracer already maintains (§3). The per-event label is deterministic; the per-pixel label is an estimator of a well-defined population quantity, and we measure how well it recovers it.

Contributions. The paper offers three things that travel beyond its test scenes, and the findings they produce here.

1. **A reliability test for difficulty references.** Split-half reliability and cross-budget rank agreement, applied per mechanism bucket, audit any per-pixel difficulty signal. Applied to variance, they show reliability that is tail-dependent and localises to specific mechanisms (§5.1).
2. **A stable coordinate system.** The seven-class descriptor with continuous sidefields, formally separated into per-event label, population per-pixel label, and finite-sample estimator. On eleven scenes the dominant label agrees at 87–99.6% across a $64\times$ budget change against 21–34% for 7-bin quantile-discretised variance; it survives a controlled estimator perturbation; and on three third-party scenes its agreement was predicted from each scene’s pilot to within 0.018 (§4).
3. **A recommendation with evidence.** Conditioning a pilot variance on the label improves equal-budget sample allocation wherever heavy-tailed buckets carry appreciable population, reduces to the incumbent where they are absent (exactly) or negligible (to within 0.004 dB), is not reproduced by a random partition, and is not absorbed by a median-of-means variance estimator (§7).

Read per mechanism, difficulty shows structure a scalar hides: in our scenes caustic transport is not the most expensive — narrow-lobe glossy and deep specular chains are, and time-to-quality is right-censored there (§5.2); lobe width monotonically governs the glossy firefly tail (§5.3); the coupling between delta-mediated and glossy energy follows mutual visibility between object families, and held as a pre-registered prediction on one third-party scene (§6.1); and a scene can be heavy-tailed in variance yet light-tailed in mechanism, which the label indicates from the pilot alone (§7.5).

2 Related Work

Three bodies of work produce or rely on a per-pixel quantity treated as a difficulty signal: sample-statistic estimates of variance, analytically traced continuous descriptors, and path mechanism used inside the estimator or for compositing.

2.1 Variance as the de facto difficulty signal

Image-space adaptive sampling refines where a perceptually motivated contrast test fires [16] or greedily minimises per-pixel relative MSE [17]; variance-guided denoising builds its filter footprint from a variance buffer [18]; efficiency-aware path tracing iterates over second-moment estimates [19]; Bayesian direct illumination regresses over per-light statistics [20]; the Eurographics STAR [21] surveys the family as the dominant paradigm. Inverse rendering derives unbiased estimators for the derivatives of variance [14], and denoising-aware adaptive sampling drives its sampling map from a variance estimate of the denoiser’s output [15]. The community has noticed that $\hat{\sigma}^2$ behaves badly under firefly-prone transport — density-based outlier rejection [7], firefly re-weighting [8], and the observation that variance and contrast cannot be reliably computed at low sample counts [22] — but in each case the instability is suppressed at the mean estimator rather than examined as a property of the reference.

Robust estimation qualifies the critique. Estimators of mean and variance with sub-Gaussian deviation guarantees under heavy tails — Catoni’s deviation study [11], sub-Gaussian mean estimators [12], the median-of-means family [13] — need only the moments their bounds assume, not the finite fourth moment that error bars on $\hat{\sigma}^2$ require; within graphics, recent work corrects estimated radiance and variance from statistical models of the sample population [23]. These methods weaken the practical force of the instability wherever their moment assumptions hold. They do not remove the regime structure — when the fourth moment is unbounded, variance estimation degrades regardless of estimator — nor supply mechanism-level structure. We treat robust estimation and mechanism conditioning as complementary and test the relationship for one such estimator, median-of-means (§7.4). Per-pixel guidance maps produced by denoising pipelines [24, 25] inherit the reliability of their variance inputs at low sample counts; the split-half methodology of §5.1 applies to them directly, though quantifying it is beyond our scope.

2.2 A-priori prediction via traced continuous descriptors

A second line forgoes sample variance and predicts difficulty analytically by tracing a continuous descriptor of the radiance signal through the transport operators: frequency analysis [26], its specialisation to depth of field [27], the 5×5 covariance matrix propagated through transport, BRDF, occlusion, lens and motion [28], and first-order sensitivities of lighting, shading and shadow [29]. These are the closest prior work by goal. Their object — a covariance, spectrum or gradient field — is itself continuous and must be traced and approximated under Gaussian-family assumptions; our label is a deterministic function of path structure read from a single contribution event.

2.3 Path mechanism inside the estimator

Path mechanism has been load-bearing for decades, but inside the estimator or as a compositing convenience. Heckbert’s $L(D|S)*E$ notation [30] described paths and motivated hybrid algorithms; Veach [4] formalised it as a classifier of which algorithms can sample which configurations. Photon mapping separates a caustic from a global map [31]; path guiding learns incident-radiance proposals as parametric mixtures [32], SD-trees [33] or normalising flows [34]; manifold next-event estimation [35] and specular manifold sampling [36] target specific mechanisms. Path-space regularization makes the mechanism–difficulty connection explicit from the estimator side, mollifying unsampleable or high-variance path classes by path-space criteria [37] or by widening microfacet lobes [38] — direct evidence that particular mechanisms are trouble spots for particular estimators, though neither produces a per-pixel label, and estimator-specific trouble does not by itself imply an estimator-independent ordering of difficulty. Variance-aware MIS [39] combines mechanism and variance at sampler-family granularity. Production light-path expressions [40] are the closest analogue to our taxonomy — a regex over $L/D/G/S/E$ selecting events into AOVs — but are user-defined, overlapping, and framed as compositing conveniences. The mechanism abstraction has appeared in every role except as a per-pixel descriptor of, and conditioning signal for, rendering difficulty. That is the role proposed here.

3 The Transport-Mechanism Descriptor

We describe each pixel by a discrete label assigned per contribution event. The *per-event* label is a deterministic function of the lobe sampled at the event’s end-vertex and the lobe sequence leading to it, under the estimator conventions of §3.5; nothing at the event level is estimated or thresholded. The *per-pixel* label aggregates events by an energy-weighted argmax and is therefore a Monte Carlo estimate at finite sample counts.

3.1 The contribution event

Each contribution to a pixel value reaches the sensor through a specific sequence of vertices and BSDF events; we call each accumulation a *contribution event*. In a path tracer with next-event estimation they arise in three forms: a successful NEE shadow ray, a BSDF-sampled bounce that lands on an emitter, and the trivial primary-ray-into-emitter case. For each event we record (a) the BSDF lobe at the *end-vertex* V_n , the last surface vertex before the emitter; (b) whether the light-to-end-vertex sequence contains any delta-specular interaction; and (c) the event’s surface-bounce count. All three are read from lobe types and counters the path tracer already maintains; no additional rays are cast.

End-vertex lobe under NEE. For a BSDF-sampled bounce the lobe at V_n is the BSDF’s reported sampled lobe. For an NEE connection the BSDF is only evaluated, so V_n takes a lobe from the BSDF’s flags: **D** if it advertises a diffuse component, **G** otherwise; multi-lobe BSDFs advertising both (e.g. `roughplastic`) are routed to **D**, the light-tailed side. Delta surfaces have zero NEE connection probability, so an NEE end-vertex is never **S** and specular-direct receives only BSDF-sampled contributions.

Population label versus finite-sample label. Let $L_b(p)$ be the population path-space contribution to pixel p restricted to bucket b — the integral of the measurement contribution function over paths whose events classify to b . The *ideal dominant label* $b^*(p) = \arg \max_b L_b(p)$ is a deterministic property of the integrand and the stated conventions. A renderer at budget N produces estimates $\hat{L}_b^{(N)}(p)$, and the reported label is $\hat{b}_N(p) = \arg \max_b \hat{L}_b^{(N)}(p)$: an estimator of $b^*(p)$, and at finite SPP sample-, estimator- and implementation-dependent. The stability results of §4 are evidence that this estimator recovers $b^*(p)$ reliably at small budgets in the tested settings; the practical cost of estimating it from a 4-SPP pilot is measured in §7 and found to be negligible.

3.2 The partition

Each event maps to one of seven buckets (Table 1) along two axes — end-vertex lobe $\ell \in \{D, G, S\}$ and presence of a δ -S event in the sequence — refined within the no- δ branch by a single- versus multi-bounce predicate, with a seventh row for the trivial primary-ray case. Operationally it is a first-match cascade, with end-vertex lobe ℓ , delta flag δ and bounce count n : (1) $n = 0$: emitter-direct; (2) $\ell = S$: specular-direct; (3) $\ell \notin \{D, G\}$: other; (4) δ : delta-mediated; (5) $n = 1$: direct; (6) $\ell = D$: diffuse-indirect; (7) $\ell = G$: glossy. The cascade matches the integrator implementation line for line.

The partition is mutually exclusive and formally complete by construction — the `other` row catches anything outside the named cells. What construction does not guarantee is that the six *named* mechanisms cover the transport that actually occurs; §4.2 measures this. Every cell boundary is a discrete predicate on a lobe type, a sequence property or the bounce count; no continuous threshold enters the main axis.

Table 1: The seven contribution-event buckets. *delta-mediated* marks a *transport class* — energy that passed through a delta interface — of which the bright image-space caustic footprint is a subset, resolved by the continuous δ -share sidefield (§3.3).

end-vertex lobe	δ -S in seq.	bounces	bucket
(no surface bounce)	—	= 0	emitter-direct
D or G	no	= 1	direct
D	no	> 1	diffuse-indirect
G	no	> 1	glossy
D or G	yes	any	delta-mediated
S	—	—	specular-direct
other	—	—	other

3.3 Continuous sidefields

The dominant label discards the per-pixel mixture. Four continuous sidefields retain it: *purity* $\pi(p) \in [0, 1]$, the energy fraction of the dominant bucket; the *glossy energy fraction* $\gamma(p)$, the share of events whose sequence contains any glossy bounce; the *energy-weighted bounce depth* $d(p)$; and the *per-bucket energy fractions* $\hat{L}_b / \sum_{b'} \hat{L}_{b'}$, of which the δ -share (the delta-mediated fraction) resolves the bright visual caustic footprint from pixels that merely receive some delta-mediated energy. Any threshold in the framework is confined to sidefields, never to the main partition.

3.4 Design choices

Why discrete. A discrete label absorbs small perturbations — a design property, not by itself evidence that the label measures difficulty, since any coarse quantisation is more stable than the continuous signal it coarsens. Two things elevate the property beyond quantisation: the fair comparison of §4.3, in which $\hat{\sigma}^2$ discretised onto quantile bins of matching cardinality is still $1.4\text{--}4.5\times$ less stable; and the placebo test of §7, in which a random partition with identical class sizes does not reproduce the label’s operational value. What matters is *which* pixels share a class, not that classes exist.

Why γ is continuous. A one-bit “any glossy interaction” flag saturates (0.0, 0.0, 1.0 on `glass_caustic`, `cubes_dof`, `snooker` at SPP=4096) and would saturate at 1.0 on any glossy-rich scene; the continuous fraction recovers the distinction between “occasionally glossy” and “glossy-dominated” at no cost, since the path tracer accumulates the glossy-touching stream anyway.

Why glossy is its own bucket. Folding glossy focusing into delta-mediated would require a roughness threshold on the main axis; and the empirical relationship between the two has no fixed sign — it reverses with scene geometry (§6.1), so any rule pre-committing glossy to either neighbour mislabels at least one observed regime.

3.5 Scope and conventions

The descriptor is a property of the integrand *together with* stated conventions, not of the integrand alone. The NEE lobe rule and the D/G/S trichotomy depend on the renderer’s material model and flag semantics; a different BSDF decomposition, roughness parameterisation or layered-material factoring may label the same physical scattering function differently. Appendix A gives the complete mapping for every material used; rather than assert insensitivity, we perturb the estimator and report which pixels move (§4.4). All claims concern surface path tracing on scenes of diffuse, rough- and smooth-conductor, rough- and smooth-dielectric, and plastic-family materials. We make no claims for participating media, subsurface transport, hair and foliage visibility complexity, glinty or measured materials, environment emitters, or null scattering — several of which carry substantial difficulty regimes of their own (§8).

4 Coverage, Stability and Robustness of the Descriptor

Before using the descriptor to characterise anything, we establish its own reliability: the named mechanisms cover the transport that occurs on the tested scenes; the per-pixel label is stable across an order-of-magnitude budget change in the regime where variance-derived references are not; it survives a controlled perturbation of the estimator; and on three third-party scenes its stability was predicted from each scene’s own pilot.

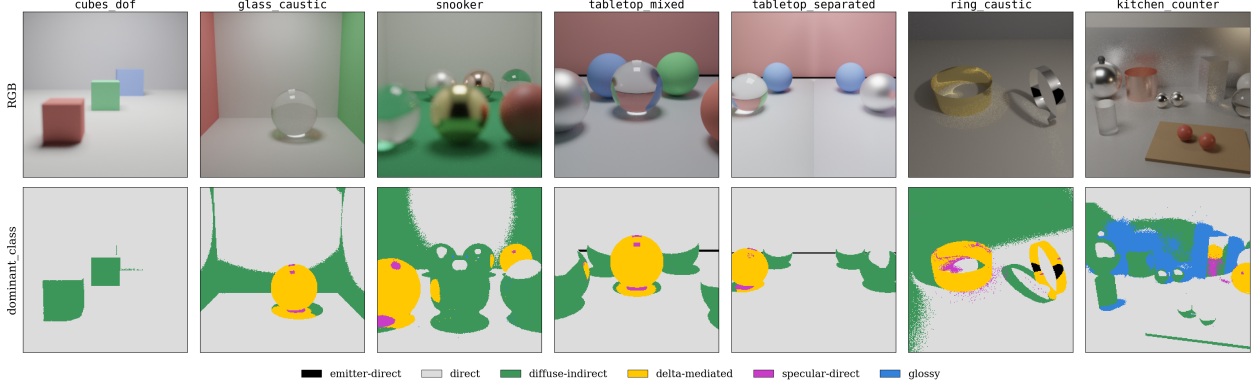


Figure 2: RGB renders (top) and dominant-label maps (bottom) for the seven first-party scenes. The tabletop pair differs only by the blocking partition; `ring_caustic`’s lying-ring cardioid is directly visible in the RGB render; `kitchen_counter`’s metal cluster and brushed backsplash form the 10,137-pixel glossy population used for the population-scale analyses of §4.4 and §5.2.

4.1 Test matrix

Seven first-party scenes span the regime of interest: `cubes_dof` (three cubes, thin lens, Lambertian only); `glass_caustic` (Cornell variant, glass sphere, L-S-D caustic on the floor); `snooker` (six balls — matte plastic, polished gold/silver/copper, glass — on felt); `tabletop_mixed`/`tabletop_separated` (a controlled pair sharing geometry and materials, the second adding an opaque partition that blocks delta-S $\leftrightarrow$ glossy sight lines); `ring_caustic` (two polished metal rings on a diffuse table under a small bright light — a lying ring focusing the classic cardioid, a standing ring projecting a caustic streak; every caustic path is L S^+ DE, giving an unambiguous, visually verifiable delta-mediated population of 6,949 pixels); and `kitchen_counter` (~ 15 objects spanning diffuse, rough plastic, brushed and polished conductors at several roughnesses, smooth glass and a frosted tumbler, two lights; metal objects before a brushed-steel backsplash give the glossy bucket 10,137 pixels). Four third-party scenes from Bitterli’s collection [41] are used in their native Mitsuba 3 form — `veach-ajar`, `veach-mis`, `bathroom2`, `matpreview` — each chosen to test one claim on geometry the authors did not design (§4.5, §7.5, §8).

Each first-party scene is rendered at $SPP \in \{64, 512, 4096\}$; the $SPP=4096$ render is the reference and the partition is computed at each SPP independently. The implementation is a Mitsuba 3 custom `SamplingIntegrator` recording, per contribution event, the end-vertex lobe, a delta-in-sequence bit and the bounce count. All tabulated matrices use a single renderer version, and re-rendering the original test matrix from scratch reproduced every statistic to within seed noise; four auxiliary measurements — the cross-budget convergence study behind $\rho = 0.26$, the $\hat{\alpha}$ reliability figure, the `snooker` glossy sub-population seed statistics, and the light-count control — date from the original runs and are used unchanged (Appendix A).

4.2 Coverage

The other catch-all receives 0% of contribution-event energy on all 21 first-party runs and on every third-party render (both budgets, both estimator modes, all material variants): on the scenes tested, the six named mechanisms account for all observed energy. This establishes coverage for the tested scene class, not universal completeness — materials outside the D/G/S trichotomy would land in `other` by design. Per-scene bucket maps agree with physical expectation (Fig. 2): `cubes_dof` is 90.9% direct; `glass_caustic` places 7.4% of pixels (10.0% of energy) into delta-mediated on the floor below the sphere; `snooker` sorts glass and chrome balls into delta-mediated and the gold ball into glossy; `ring_caustic` concentrates delta-mediated on and inside the rings; `kitchen_counter` assigns its metal cluster and backsplash reflections to glossy. Such agreement is a sanity check, not proof of a uniquely correct partition.

4.3 Cross-budget stability

Table 2 reports per-pixel agreement of the dominant label between budgets and the largest change in any bucket’s pixel share. Raw agreement could be inflated by a majority bucket; Cohen’s κ [42], which removes chance agreement, stays 0.79–0.98 across all seven scenes (Table 3), “substantial” to “almost perfect” on the Landis–Koch scale [43], used informally.

Table 2: Cross-budget stability of the dominant label: per-pixel agreement between budgets, and the largest change in any bucket’s pixel share between SPP= 64 and 4096.

scene	64 vs 4096	512 vs 4096	max drift
cubes_dof	99.58%	99.82%	0.12%
tabletop_separated	99.43%	99.77%	0.12%
tabletop_mixed	98.65%	99.45%	0.37%
glass_caustic	95.64%	98.70%	1.19%
snooker	89.55%	95.77%	0.58%
kitchen_counter	88.69%	94.12%	2.88%
ring_caustic	87.07%	92.81%	1.92%

 Table 3: Same-scale comparison and chance correction (64 vs 4096). $\hat{\sigma}^2$ is binned into per-scene quantile bins computed independently at each SPP — which removes the global $1/N$ rescaling and scores only the stability of the field’s shape, the reading most favourable to variance — and scored by the same agreement metric as the label. Weighted κ [44] uses penalty 0.2 for swaps within {emitter-direct, direct, diffuse-indirect} and 1.0 for any swap touching {delta-mediated, specular-direct, glossy}; the classes are not ordinal, and this weighting is declared as our reading of boundary flicker versus structural change. “struct%” is the share of disagreeing pixels involved in a swap touching the second family.

scene	$\hat{\sigma}^2$ 3-bin	7-bin	label	κ	wtd. κ	struct%
cubes_dof	0.484	0.264	0.996	0.975	0.975	0.0%
tabletop_sep.	0.652	0.335	0.994	0.974	0.972	47.2%
tabletop_mixed	0.596	0.286	0.987	0.964	0.953	70.3%
glass_caustic	0.424	0.212	0.956	0.890	0.900	32.9%
snooker	0.560	0.307	0.896	0.828	0.885	17.0%
kitchen_counter	0.611	0.344	0.887	0.789	0.718	81.3%
ring_caustic	0.524	0.265	0.871	0.786	0.859	14.3%

The fair comparison. A coarse categorical variable is naturally more stable than a continuous one, so the raw juxtaposition “95.6% versus $\rho = 0.26$ ” (both *glass_caustic*) proves little. Discretising $\hat{\sigma}^2$ onto quantile bins of matching cardinality and scoring it with the identical metric (Table 3, first columns) gives 0.42–0.65 (3-bin) and 0.21–0.34 (7-bin) against the label’s 0.87–0.996: on every scene the label is 1.4–2.3 $\times$ more stable than coarse-binned variance and 2.5–4.5 $\times$ more than fine-binned. Class imbalance is handled by κ , not raw agreement.

Where the residual disagreement lives. Two kinds of label boundary must be distinguished. *Predicate boundaries* are decided by a discrete test — membership of *delta-mediated* and *specular-direct* turns on a δ -S event. *Energy-argmax boundaries* are decided by which bucket carries more energy: *direct* versus *diffuse-indirect* (bounce-count split) and *diffuse-indirect* versus *glossy* (lobe split). Across all seven scenes disagreement concentrates on the energy-argmax boundaries and spares the predicate buckets. On the delta-heavy scenes the dominant flicker is *direct* $\leftrightarrow$ *diffuse-indirect* (Fig. 3), and weighted κ , which forgives that swap, rises above Cohen’s (0.828 $\rightarrow$ 0.885 on *snooker*, 0.786 $\rightarrow$ 0.859 on *ring_caustic*). On the glossy-rich *kitchen_counter* the dominant flicker is instead *diffuse-indirect* $\leftrightarrow$ *glossy* (2,575 of 7,412 disagreeing pixels at SPP=64), which our weighting scores as structural — hence the one weighted κ that drops and the 81% struct share. That is the glossy *energy* boundary, not a predicate failure: *kitchen_counter*’s *delta-mediated* share drifts by 0.01% across the 64 $\times$ change while its *diffuse-indirect/glossy* shares exchange up to 2.9% of pixels. Structural-bucket shares drift by at most 2.9% anywhere, and by $\leq 1.2\%$ outside that scene. The continuous sidefields drift as expected for continuous quantities: purity median by 0.94–2.4%, glossy fraction by $< 0.5\%$ where non-zero.

4.4 Robustness to a controlled estimator perturbation

The pilot integrator resolves glossy and *delta-mediated* through NEE only, suppressing the BSDF-sampled emitter hit by an acceptance rule that avoids double counting without MIS (Appendix A); the energy accounting on those buckets is by construction half of a complete MIS estimator. If the argmax were sensitive to this, the label would inherit exactly the estimator dependence we fault $\hat{\sigma}^2$ for. We therefore restore the missing half — a power-heuristic MIS on the emitter-hit branch of those two buckets, the two on which the suppressed half carries substantial energy on narrow lobes (*emitter-direct* and *specular-direct* admit no competition at all: primary rays have no NEE branch, delta vertices zero NEE probability), leaving *direct* and *diffuse-indirect* untouched so that the perturbation is not conflated with the boundary flicker already characterised — and recompute the label at SPP= 4096. The scope is deliberately narrow: this is robustness to *one* controlled perturbation, the largest our own pilot construction introduces,

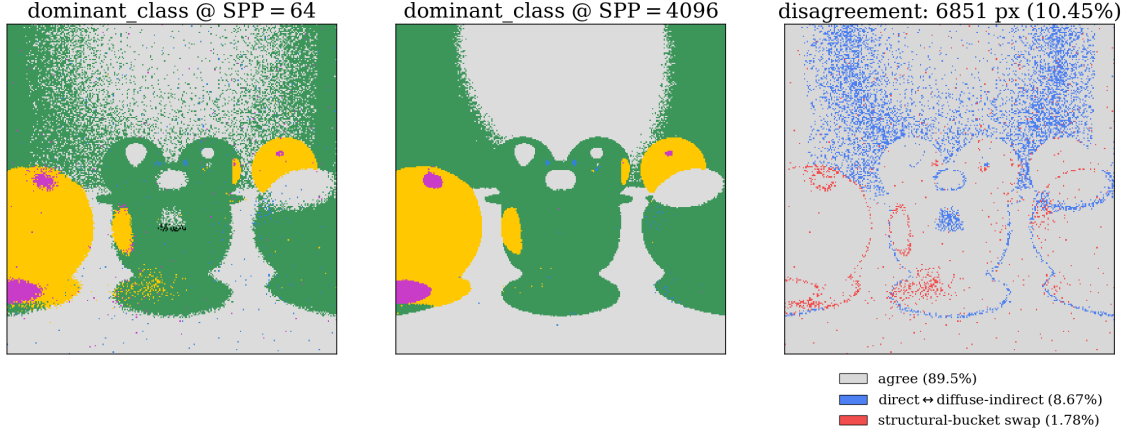


Figure 3: Cross-budget stability on snooker. Left/middle: dominant label at SPP= 64 and 4096. Right: agreement map (grey = agree, blue = direct ↔ diffuse-indirect swap, red = swap touching a structural bucket). 89.55% agree; the 10.45% disagreement is dominated by the light-tailed boundary (8.67%), with 1.78% of pixels in any structural swap; no bucket’s share moves by more than 0.58%.

not a demonstration of general estimator independence; the argmax can in principle still depend on sampling strategy, MIS heuristic, lobe probabilities and omitted path families.

Table 4: Robustness to estimator completion: per-pixel argmax agreement between the NEE-only and MIS-completed estimators at SPP= 4096, and the 64-vs-4096 agreement recomputed under MIS.

scene	MIS agree (4096)	64 vs 4096 (MIS)
cubes_dof	100.00%	99.58%
tabletop_separated	100.00%	99.42%
tabletop_mixed	100.00%	98.64%
glass_caustic	100.00%	95.64%
ring_caustic	100.00%	87.07%
snooker	99.86%	89.58%
kitchen_counter	98.52%	89.87%

Within that scope the argmax is nearly insensitive (Table 4): five of seven scenes are unchanged to the pixel — including ring_caustic’s 6,949-pixel delta-mediated population — snooker moves by 0.14% and kitchen_counter by 1.48%; cross-budget agreement recomputed under MIS lands within a whisker of the NEE-only values. Reweighting a bucket whose membership is a δ -predicate cannot flip it, and every flip that occurs lies on the energy boundaries the MIS actually reweights (401 of 403 diffuse-indirect departures on kitchen_counter go to glossy).

The flips are a correction, and they are lobe-width dependent. On snooker, of the 83 pixels the NEE-only pilot calls glossy, 44 re-label as diffuse-indirect under MIS and 8 move elsewhere, leaving 31 (the MIS render’s own glossy bucket numbers 37). The 44 are the narrowest-lobe pixels: on a near-specular lobe the BSDF PDF dominates the NEE PDF, the power heuristic correctly drives the NEE weight toward zero, and the NEE-only estimate — which kept that half at full weight — had overstated glossy energy enough to push the argmax across the boundary (Fig. 4). The lobe-width account predicts that a glossy population not concentrated at the narrowest lobes should be far more robust, and kitchen_counter tests this at scale: its 10,137-pixel bucket, spanning $\alpha \in [0.03, 0.16]$, is 95.8% stable under the same completion, against 37% for snooker’s 83 pixels at $\alpha \approx 0.03$. The snooker fragility is the narrow-lobe boundary behaviour the analysis says it is, not a property of the glossy class; §5.3 isolates the variable.

4.5 Stability on third-party scenes, predicted from the pilot

Every scene above is author-built. The third-party scenes address the objection that the test distribution’s generating process was under our control. Predictions for each were fixed in writing before any bucket-instrumented render of a third-party scene existed (one amendment, registered after a first-batch outcome, is marked as such in the supplementary material), and all are reported against their outcomes, pass or fail, in the supplementary material. Here we report the two whole-image results. Coverage transfers: other receives exactly 0% in every third-party render. Stability

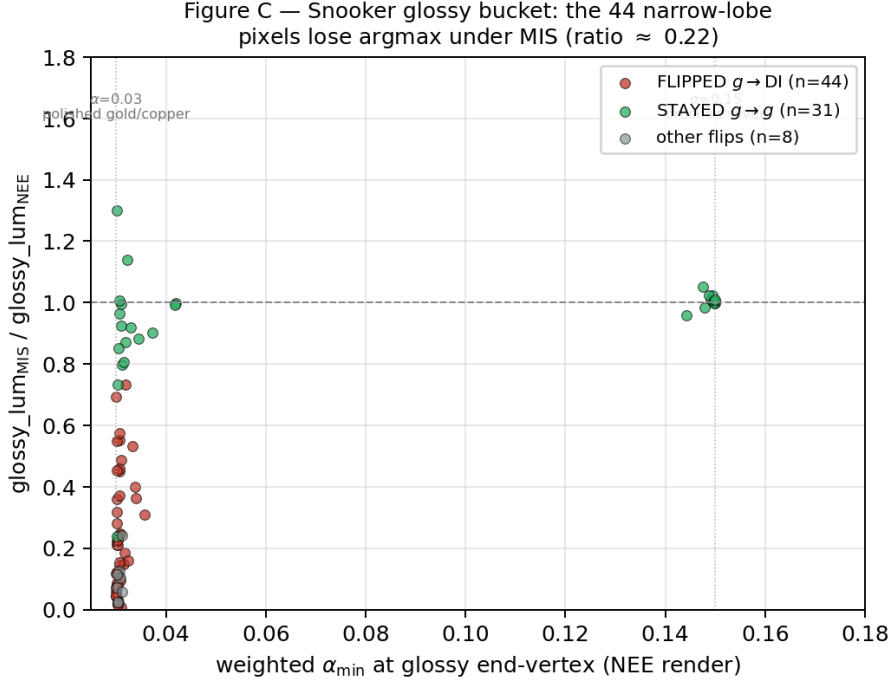


Figure 4: Mechanism of the glossy re-labelling on snooker. Each point is a pixel of the NEE-only glossy bucket, by end-vertex lobe width against the ratio of glossy energy under MIS to NEE-only. The 44 flipped pixels (red) cluster at $\alpha \approx 0.03$ (polished gold/copper), where MIS down-weights the NEE half and glossy energy contracts to ≈ 0.22 of its NEE-only value; the 31 that stay (green) span wider lobes and keep a ratio near unity.

transfers: 64-vs-4096 agreement is 0.978 / 0.990 / 0.951 on *veach-ajar* / *veach-mis* / *bathroom2* (Table 5), inside the first-party range.

These values were predicted before the reference renders existed. On the seven first-party scenes, agreement rises along pilot argmax-margin (and purity) bins in every scene — monotonically up to a single late-bin fluctuation of ≤ 0.04 on two — and a pooled 20-bin monotone curve transfers across scenes with a leave-one-scene-out scene-level error of at most 0.061 (purity; 0.079 for margin), the worst case being the most mechanism-mixed scene. The curve is fitted from first-party data alone and was registered before any third-party render; each scene’s predicted agreement was issued from its own SPP-64 pilot before its reference render existed, and the outcomes lie within 0.018 on *veach-ajar* and 0.005 on the other two (Fig. 5). This suggests that the label’s stability on a new scene can be estimated from the scene’s own pilot before a reference exists. The error range is an empirical one — the maximum leave-one-scene-out error on author-generated scenes — and the three third-party outcomes are consistent with it but do not constitute an independent re-estimate.

Table 5: Third-party stability and frozen-curve predictions: 64-vs-4096 agreement predicted from each scene’s own SPP-64 pilot via the pooled margin/purity curves fitted on the first-party scenes.

scene	agree	pred. (margin)	pred. (purity)	other
<i>veach-ajar</i>	0.978	0.995	0.996	0%
<i>veach-mis</i>	0.990	0.990	0.991	0%
<i>bathroom2</i>	0.951	0.950	0.946	0%

5 Variance-Based Difficulty Read per Mechanism

We now use the descriptor to resolve the variance-based signals per mechanism. Each subsection reports one observation that a scalar per-pixel statistic does not expose.

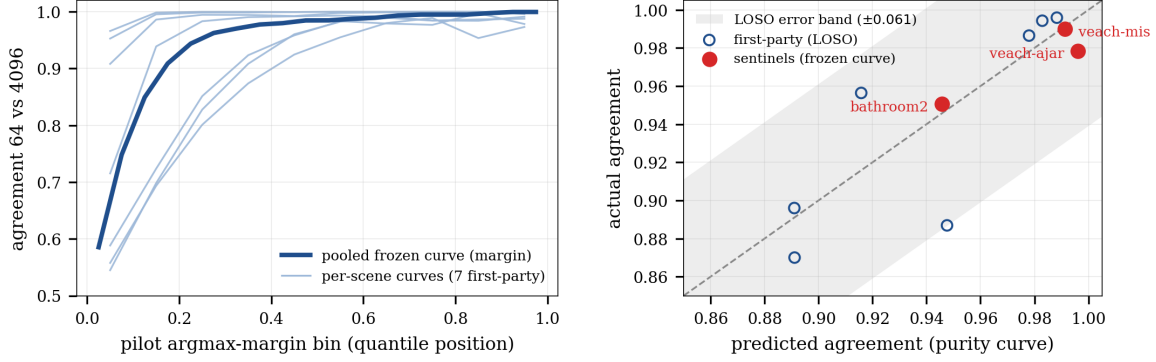


Figure 5: Pilot-based prediction of label stability. *Left*: 64-vs-4096 agreement as a function of the pixel’s pilot argmax-margin bin, rising in every first-party scene (thin), and the pooled 20-bin monotone curve (bold), fitted on first-party data and registered before any third-party render. *Right*: predicted versus actual scene-level agreement. Open circles: leave-one-scene-out predictions on the first-party scenes; the grey band is their observed maximum error (± 0.061 ; the off-diagonal point is `kitchen_counter`). Filled points: the three third-party scenes, each predicted from its own pilot — errors 0.001–0.018.

5.1 Variance reliability is tail-dependent and localises to mechanism

We measure two split-half quantities. First, a per-pixel convergence exponent $\hat{\alpha}(p)$ is fit from $\hat{\sigma}^2$ at SPP $\in \{64, 256, 1024, 4096\}$ with ten independent seeds per level ($\sim 40,000$ samples per pixel); the seeds are partitioned into two halves, $\hat{\alpha}$ refit on each, and the two fields Spearman-correlated, averaged over 20 random 5+5 partitions (partition-to-partition standard deviation ≤ 0.003). Second, the same procedure is applied to $\hat{\sigma}^2$ itself at each fixed SPP. A noiseless reference would score 1.0.

Table 6: Split-half self-consistency of $\hat{\sigma}^2$: Spearman ρ between the fields of two disjoint five-seed halves, per fixed SPP, mean over 20 partitions. Reliability is *tail-dependent*: it collapses on the caustic-dominated scene and is stable on light-tailed ones.

scene	SPP=64	256	1024	4096
glass_caustic	0.181	0.143	0.206	0.257
cubes_dof	0.311	0.310	0.308	0.309
ring_caustic	0.385	0.426	0.498	0.543
snooker	0.483	0.517	0.549	0.577
tabletop_mixed	0.512	0.546	0.599	0.669
kitchen_counter	0.553	0.601	0.619	0.630
tabletop_separated	0.689	0.698	0.707	0.712

The $\hat{\alpha}$ -derived ranking on `cubes_dof` and `glass_caustic` has split-half $\rho = 0.23$ – 0.29 — the reliability figure of §1. The $\hat{\sigma}^2$ measurement (Table 6) is sharper: its self-consistency is a function of the scene’s tail. On `glass_caustic` it sits at 0.18–0.26, non-monotone in SPP — more samples do not buy stability, itself a heavy-tail signature — and the complementary cross-budget measurement on the same scene, rank-correlating $\hat{\sigma}^2$ between the SPP-64 and SPP-4096 renders, gives $\rho = 0.26$. On light-tailed `tabletop_separated` it sits at 0.69–0.71; on Lambertian `cubes_dof` a flat ~ 0.31 . Where the integrand is light-tailed, $\hat{\sigma}^2$ is serviceable and nobody should stop using it; where it is heavy-tailed — the regime holding the hardest pixels — it fails to agree with itself across a seed split, and evaluations against a $\hat{\sigma}^2$ - or $\hat{\alpha}$ -derived target are attenuated accordingly. On the same scenes the mechanism label agrees with its own reference at 87–99.6% from a single 64-SPP pilot.

Resolving it per bucket. The scene-level statement localises to the pixel level when the same measurement is resolved per mechanism (Table 7). The clearest case is the delta-mediated bucket of `glass_caustic`: split-half $\rho = 0.13$ at SPP= 4096 and 0.03 at SPP= 64 — on the pixels the label marks as delta-mediated in a caustic-dominated scene, the variance estimator does not agree with itself across a seed split at all, and the pixels were isolated into that bucket without ever looking at $\hat{\sigma}^2$. The bucket is not uniformly the least reliable: in the tabletops `specular-direct` is comparable, and on `ring_caustic` the bright focused cardioid makes it the scene’s *most* reliable bucket at high SPP (0.72, from 0.32 at SPP= 64). The pattern is that within-bucket reliability tracks how heavy-tailed the delta-mediated energy is at its pixels — collapsing where it is dim and firefly-borne, recovering where it is bright and focused, and

Table 7: Split-half self-consistency of $\hat{\sigma}^2$ at SPP= 4096, resolved *per mechanism bucket* (mean over 20 partitions; “—” marks buckets with fewer than 50 pixels).

scene	dir.	diff.	d-m	spec.	gloss.
snooker	0.50	0.65	0.41	0.59	0.83
glass_caustic	0.18	0.42	0.13	0.46	—
tabletop_mixed	0.67	0.59	0.54	0.50	—
tabletop_separated	0.71	0.57	0.50	0.46	—
cubes_dof	0.20	0.72	—	—	—
ring_caustic	0.46	0.24	0.72	0.39	—
kitchen_counter	0.59	0.57	0.38	0.40	0.44

in between (0.38–0.41) where it is prominent but mixed — with the residual variation carried by the δ -share and luminance sidefields. A variance-based reference has no representation to flag, in advance, which pixels carry this risk; the delta-mediated bucket is a pre-computed map of where the variance baseline is at risk, and the sidefields resolve where within it the risk is realised.

One entry appears to contradict the cost analysis to follow: snooker’s glossy bucket has the *highest* self-consistency (0.83) yet is slowest-converging (§5.2). There is no contradiction: self-consistency asks whether the *estimator* agrees with itself, time-to-quality whether the *pixel* converges. A glossy pixel can carry a stably-estimated, consistently high variance — $\hat{\sigma}^2$ reliably reports “hard” — whereas the delta-mediated failure is that $\hat{\sigma}^2$ cannot report a consistent value at all. The partition separates these two failure modes into two buckets where one continuous number conflates them. (The 0.83 is itself an average over sub-populations: §5.3 shows the narrowest-lobe glossy pixels have *negative* ρ .)

5.2 Mechanism is not difficulty: cost heterogeneity and censoring

A mechanism class describes how energy arrives, not how expensive it is to resolve, and per-bucket resolution makes the distinction measurable. For each pixel we compute a time-to-quality cost $c(p)$: the smallest SPP at which $|\text{lum}(I_N) - \text{lum}(I_{\text{ref}})| / \max(\text{lum}(I_{\text{ref}}), \epsilon) < \tau$ ($\tau = 0.10$, $\epsilon = 10^{-4}$, doubling SPP grid to 2048, independent renders per level, SPP-4096 reference). Pixels that fail to cross by 2048 are right-censored: their cost is unknown but above the cap. Table 8 gives the median per bucket.

 Table 8: Median time-to-quality cost (SPP) per dominant-label bucket. “—”: no pixels. $\dagger$: bucket heavily right-censored at $\text{SPP}_{\text{max}} = 2048$ (rate below); a censored median is a lower bound, not a measurement. The glossy population is NEE-defined (§4.4).

bucket	sn	gc	tm	ts	cb	rc	kc
direct	16	16	16	16	16	32	32
diffuse-indirect	32	16	32	32	64	32	64
delta-mediated	32	32	32	32	—	32	64
specular-direct	128	256	256	128	—	2048 $\dagger$	2048 $\dagger$
glossy	2048 $\dagger$	—	—	—	—	—	128

$\dagger$ censored: glossy on sn 51.8%; specular-direct on rc 53.0%, on kc 48.8%.

Three observations follow that a scalar signal does not expose. First, **the caustic is not the worst case**: on every scene where it exists, delta-mediated converges at a median of 32–64 SPP — including ring_caustic, whose bright cardioid it contains. The right-censored long tails live elsewhere: in snooker’s narrow-lobe glossy bucket (51.8% unconverged at 2048) and in the deep delta chains of specular-direct on ring_caustic (53.0%, ring interiors) and kitchen_counter (48.8%, mirror paths among polished metals). Second, **the cost measurement itself saturates** on exactly those buckets — censoring means the experiment has ceased to resolve costs above the cap, not that the pixels have become constant in cost — while the discrete label remains discriminating on the same pixels. Third, **buckets are internally heterogeneous**: kitchen_counter’s population-scale glossy bucket, spread across lobe widths, converges at a median of 128 SPP with only 5.9% censoring, whereas snooker’s 83-pixel bucket, concentrated at $\alpha \approx 0.03$ and NEE-defined, is half censored; part of that apparent hardness is the high variance of the NEE estimator on narrow lobes rather than an intrinsic property of the transport — itself an instance of estimator dependence. A mechanism class is therefore not an ordering of difficulty, and we do not claim one; it is a coordinate system in which the heterogeneity becomes visible and attributable. Per-dimension correlations between cost and the continuous sidefields, reported in the supplementary material, confirm the point: no single component is a stable cost proxy, and the weak correlations are themselves a consequence of censoring on the hardest pixels.

5.3 Lobe width governs the glossy tail

Both the estimator-perturbation flips (§4.4) and the cost censoring point at end-vertex roughness α as the controlling variable of the glossy tail. A controlled single-sphere experiment isolates it (Fig. 6): a single isotropic rough conductor sphere in an otherwise diffuse box — no delta surfaces, so no δ -S contaminates the bucket — with α swept over eight values (0.01 to 0.30; 30 seeds $\times$ 128 SPP per level) and geometry, illumination and camera fixed. The prediction, fixed before rendering, was that the per-pixel firefly ratio (seed-wise maximum over mean glossy contribution; statistics over glossy-dominant pixels, those whose dominant label is glossy) and the split-half ρ of $\hat{\sigma}^2$ would both co-vary with α .

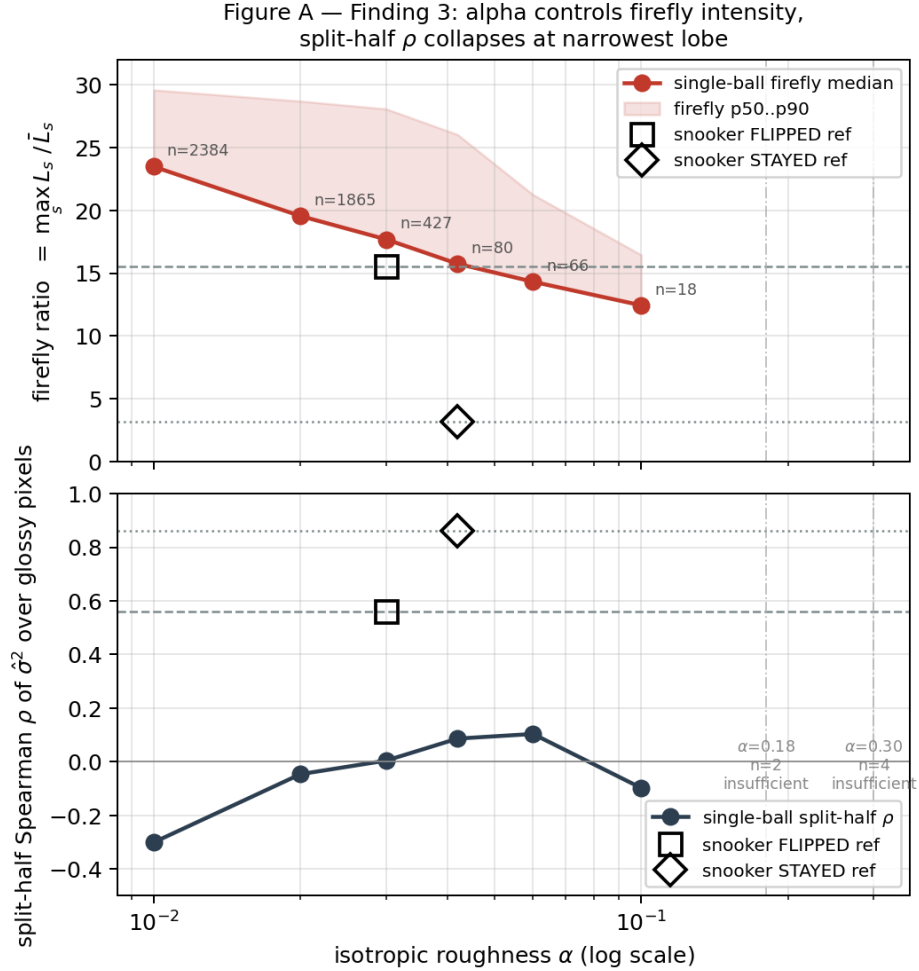


Figure 6: Controlled roughness sweep. *Top*: median firefly ratio falls monotonically as the lobe widens, 23.5 at $\alpha=0.01$ to 12.4 at $\alpha=0.10$ (p50–p90 band); the snooker flipped pixels ($\square$, narrow lobe) land on the curve, the surviving pixels ($\diamond$, wide lobe, multi-path) sit far below it. *Bottom*: split-half ρ of $\hat{\sigma}^2$ is not monotone in α : it collapses to -0.30 at the narrowest lobe and otherwise sits in a noise band, while the surviving pixels’ $\rho \approx 0.86$ floats above the whole curve. The two widest levels yield too few glossy-dominant pixels to measure.

The firefly prediction holds cleanly: the median firefly ratio falls monotonically from 23.5 at $\alpha=0.01$ to 12.4 at $\alpha=0.10$ across the six populated levels. Narrow lobes carry a heavier per-pixel tail — the signature the kurtosis bound predicts, the narrowest lobe being the highest-kurtosis integrand. The reliability prediction holds only in its weaker form: the split-half ρ of $\hat{\sigma}^2$ does not rise smoothly with α , but at $\alpha=0.01$ it is -0.30 — the two seed halves rank pixels in *anti*-correlation, the fingerprint of an estimator dominated by rare large samples — and elsewhere sits in a ± 0.10 noise band. We claim no monotone $\rho(\alpha)$; what the controlled bucket shows is that $\hat{\sigma}^2$ fails to agree with itself at every α we can measure and agrees in reverse on the narrowest lobe. This is the glossy counterpart of the delta-mediated failure above. One regime resists the lobe-width account: snooker’s 31 surviving glossy pixels combine a low firefly ratio (≈ 3)

with high ρ (≈ 0.86), which neither the sphere sweep nor a light-count sweep at fixed α reproduces (supplementary); we record it as open rather than claim a mechanism we have not isolated.

6 Cross-Scene Structure of the Descriptor

The continuous sidefields turn each pixel into a small vector — δ -share C , glossy fraction G , depth D and inverse purity $P = 1 - \pi$ — and the correlations among these components, computed per scene and compared across scenes, expose geometric variables a scalar variance has no representation for. Table 9 reports all six pairs on all seven scenes; two of the sign reversals are controlled by an identifiable variable and are examined here, the rest are recorded in the supplementary material.

Table 9: Per-pixel Spearman ρ among the four continuous components, all seven scenes (cb, gc, sn, tm, ts, rc, kc: cubes_dof, glass_caustic, snooker, tabletop_mixed, tabletop_separated, ring_caustic, kitchen_counter). “deg.”: one component identically zero. ‡: sign reversal controlled by an identified variable (§6.1, §6.2); †: sign-stable; ○: recorded observation (supplementary). Every cell is reported. Three depth–purity and three depth–glossy entries differ from the PG-submitted version, which cyclically mis-assigned them across scenes — a transcription error, corrected here from the regenerated pipeline (revision report, Corrections 1).

pair	cb	gc	sn	tm	ts	rc	kc	note
$C \leftrightarrow P$	deg.	+0.22	−0.09	−0.30	−0.35	−0.18	+0.18	○
$P \leftrightarrow G$	deg.	deg.	−0.19	−0.25	−0.19	deg.	+0.55	○
$C \leftrightarrow G$	deg.	deg.	−0.12	+0.44	−0.27	deg.	−0.02	‡
$D \leftrightarrow C$	deg.	+0.73	+0.45	+0.43	−0.21	+0.29	+0.13	○
$D \leftrightarrow G$	deg.	deg.	+0.42	+0.23	+0.20	deg.	+0.87	†
$D \leftrightarrow P$	+0.66	+0.52	−0.37	+0.24	+0.50	−0.40	+0.60	‡

6.1 Delta-mediated/glossy coupling follows mutual visibility

A contribution event lands in delta-mediated if its sequence contains a δ -S event; the same path may also pass through a glossy interaction, making γ non-zero on the same pixel. The two fractions therefore co-occur when scene topology makes combined paths common and compete when geometry forces a path to pick one, so the sign of their per-pixel correlation should depend on the object-space separation of the two material families and cross zero along it. “Separation” must be a measured quantity; we declare two: the minimum surface-to-surface distance between any delta-S and any glossy object, and the *mutual visibility fraction* — the fraction of unoccluded segments between surface points sampled on the two families (64 points per surface). The second is the mechanism-relevant one, since co-occurrence requires a path segment connecting the families.

Table 10: Per-pixel Spearman correlation between the δ -share and glossy fraction against the two declared separation measures. The ordering follows mutual visibility on the tabletop-family scenes; metric distance alone does not order them (snooker’s closest pair is nearer than tabletop_mixed’s). kitchen_counter is an out-of-family point: high visibility but large metric separation, $\rho \approx 0$. Last row: third-party pre-registered test (measures computed and sign interval fixed before rendering).

scene	min dist. (m)	mut. vis.	ρ_{C-G}
tabletop_mixed	0.15	0.094	+0.441
snooker	0.06	0.078	−0.121
tabletop_separated	0.78	0.000	−0.270
kitchen_counter	0.53	0.125	−0.017
bathroom2 (3rd-party, pre-reg.)	0.34	0.327	+0.251

The controlled pair carries the test: tabletop_separated shares tabletop_mixed’s geometry and materials exactly and adds an opaque partition that drops mutual visibility from 0.094 to 0.000; the correlation reverses from +0.441 to −0.270, and since the partition is the only change, the reversal on this pair is causal (Table 10, Fig. 7). Across the tested configurations the ordering follows mutual visibility and crosses zero between tabletop_mixed and snooker. We state the claim at the strength the data supports: an observed ordering with one controlled two-point contrast — not a continuous monotonic law, which would require a parametric separation sweep. kitchen_counter bounds the account: with sight lines present but the families metrically distant, $\rho \approx 0$ — visibility is necessary for co-occurrence and coupling decays with distance.

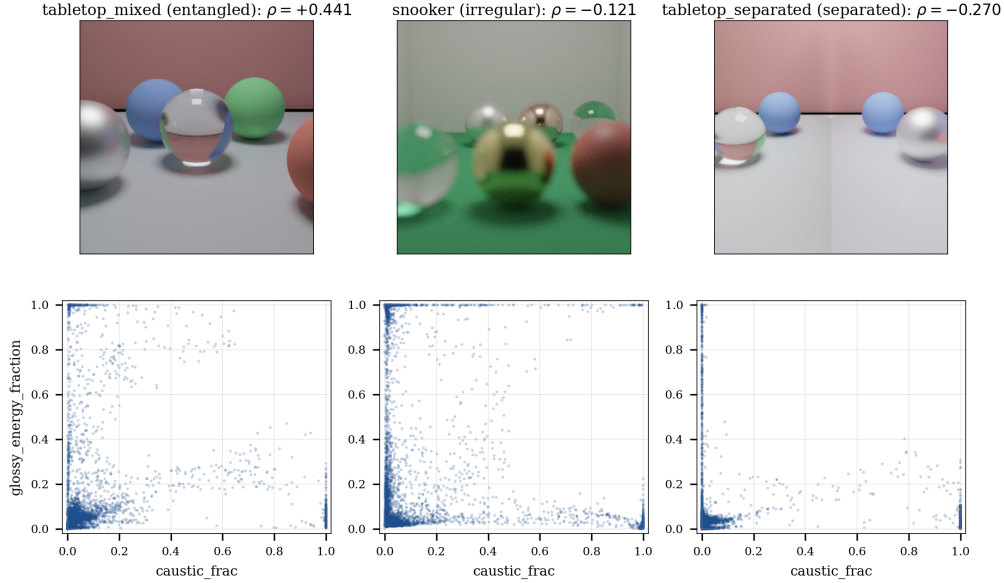


Figure 7: Sign reversal of the delta-mediated/glossy correlation with object-space separation. Top: the three scenes ordered by mutual visibility between the delta-S and glossy object families. Bottom: per-pixel scatter of δ -share against glossy fraction; ρ falls from $+0.441$ (entangled) through -0.121 (snooker) to -0.270 (separated). `tabletop_separated` is the controlled test: identical geometry and materials except for the sight-line-blocking partition.

A pre-registered test on third-party geometry. For the bathroom2 interior — mirror and chrome fixtures beside glossy tiles, an uncontrolled sight-line structure — the two separation measures were computed first (mutual visibility 0.327, the highest of any tested scene, at 0.34 m), and the prediction $0 < \rho_{C-G} < +0.441$ — positive by the visibility rule, attenuated below the entangled anchor by the distance qualification — was registered before any bucket-instrumented render existed. The outcome is $+0.251$: sign and interval as registered. An uninformed sign guess succeeds half the time and the interval’s upper end is the largest value previously observed, so this is one successful out-of-sample test, not a validated law. For the design question of §3.4 it suffices: the sign of the relationship is a function of geometry that passes through zero, so glossy stays its own bucket.

6.2 Depth and purity: delta chains versus lobe mixing

If “deeper paths are messier paths”, depth would predict mechanism mixing and $\rho(D, P)$ would be positive on every scene. It is not: $+0.66$, $+0.52$, $+0.24$, $+0.50$, $+0.60$ on `cubes_dof`, `glass_caustic`, `tabletop_mixed`, `tabletop_separated`, `kitchen_counter`, but -0.37 on `snooker` and -0.40 on `ring_caustic`. A single mechanism accounts for the sign. Where deep paths are dominated by δ -S chains — `snooker`’s inter-reflections among chrome and glass balls, `ring_caustic`’s multi-bounce reflections inside the rings — every added vertex is the same delta event, so paths gain length without gaining lobe diversity: the deepest pixels are the purest and $\rho(D, P) < 0$. Where deep paths accumulate through lobe-mixing interreflection — diffuse bounce chains, glass transmission into diffuse interiors, glossy-diffuse mixing — every added vertex can change lobe class, depth accumulates mixture and $\rho(D, P) > 0$. `ring_caustic`, whose only non-diffuse transport is delta chains, lands on the negative side as the account requires. Depth and purity are independent dimensions by data, not merely by construction; the account is consistent with seven scenes but, except for `ring_caustic`, derived from uncontrolled comparisons (§8).

7 Mechanism-Conditioned Pilot Correction for Sample Allocation

The preceding sections establish that the descriptor is stable where variance estimates are not, and that it localises where they fail. This section asks whether that buys anything operational, with a controlled equal-budget sample-allocation experiment whose hypotheses, metrics and decision rules were fixed before the runs (protocol summary, the full control set, and the companion budgets in the supplementary material; the complete per-cell record, including tail quantiles and allocation-stability diagnostics, in the research artifact).

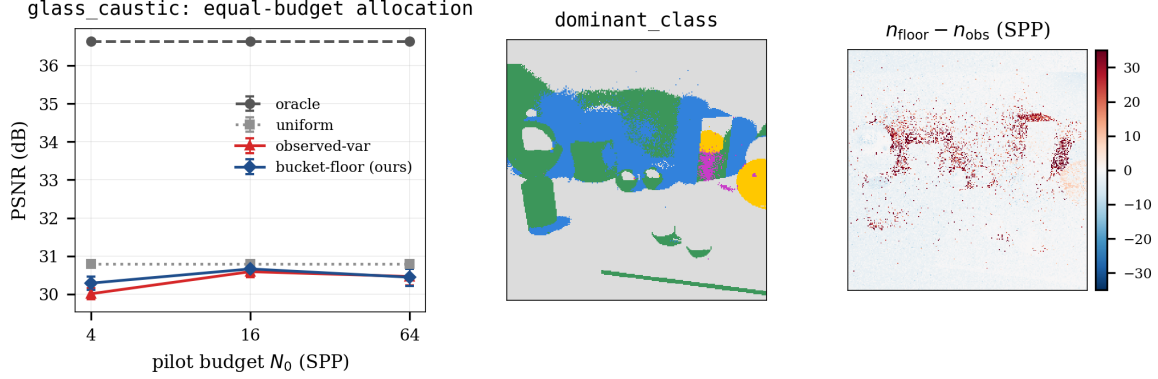


Figure 8: *Left*: equal-budget PSNR versus pilot budget on `glass_caustic` (mean $\pm$ std over 10 pilots). The pilot-variance allocator (red) sits below uniform (dotted) at every pilot budget, while the oracle (dashed) shows +5.9 dB of headroom; bucket-floor (blue) recovers part of the gap at small pilots and converges to the baseline as the pilot improves. *Middle/right*: `kitchen_counter` at $N_0=4$: the allocation difference $n_{\text{floor}} - n_{\text{obs}}$ (right; red = more samples under bucket-floor) concentrates on the pixels the label marks glossy and delta-mediated (middle) — the pixels whose pilot variance is systematically underestimated.

7.1 Protocol and strategies

Two-stage adaptive rendering at 256^2 on all seven first-party scenes. A pilot of $N_0 \in \{4, 16, 64\}$ SPP yields both a pilot variance $\hat{\sigma}^2(p)$ (luminance, unbiased) and pilot bucket energies from which the pilot label $\hat{b}(p)$ is read — label estimation shares the pilot budget. Each strategy then allocates a fixed extra budget of 32 SPP average per pixel, Neyman-proportionally ($n(p) \propto \sqrt{D(p)}$, per-pixel baseline of 1, identical totals) from a strategy-specific score $D(p)$; the final image is the per-pixel mean of the first $n(p)$ frames of a shared 256-frame pool disjoint from every pilot, and error is measured against an independent 4096-SPP reference. Every number is the paired outcome over $K=10$ disjoint pilot draws: single-pilot comparisons can reverse sign under re-piloting, so we treat multi-pilot pairing as mandatory.

Baselines are *uniform*, *observed-var* ($D = \hat{\sigma}^2$, the standard pilot-variance allocator) and an *oracle* using the pool’s own converged variance. The proposed treatment is **bucket-floor**: for pixels whose pilot label lies in the heavy-tailed family {delta-mediated, specular-direct, glossy}, replace $\hat{\sigma}^2(p)$ by $\max(\hat{\sigma}^2(p), \text{med}_b)$, the median pilot variance within the pixel’s own bucket; light-tailed pixels are untouched. The rationale is §5.1’s: on heavy-tailed pixels a small pilot usually *misses* the tail, so $\hat{\sigma}^2$ is biased low exactly where the label says the tail lives. Controls: a *random-partition placebo* (the same floor on a label map with identical class sizes but shuffled assignment, fresh per draw) tests whether coarse partitioning alone explains any gain; a *label-oracle* variant (labels from the reference) prices pilot-label error. A *robust-variance* incumbent, $D = \hat{\sigma}_{\text{MoM}}^2$ — median-of-means variance over $k = \lceil \sqrt{N_0} \rceil$ contiguous blocks of the pilot frames — and its floored variant *robust-floor*, with its own matched placebo, test whether robust estimation absorbs the effect; MoM is not meaningfully defined at $N_0=4$ (two-sample blocks), a limitation with content of its own, since at the smallest pilots mechanism conditioning is the only correction that still applies. The label’s overhead is negligible: bucket-energy capture shares the pilot’s path tracing, the floor is a per-bucket median and a maximum (< 1 ms per image).

Table 11: Equal-budget allocation at pilot $N_0=4$: PSNR (dB, mean over 10 disjoint pilot draws; seed-to-seed std 0.08–3.0) and the per-draw win rate of bucket-floor against observed-var on paired MSE and paired p99 tail error. Bold: scenes on which the pilot-variance allocator is at or below *uniform*. On `cubes_dof` (no heavy-tailed buckets) bucket-floor reduces to observed-var exactly.

scene	uniform	obs-var	b-floor	oracle	MSE	p99
glass_c.	30.79	30.01	30.29	36.64	10/10	8/10
kitchen	33.71	32.75	33.03	41.18	9/10	10/10
snooker	34.97	37.83	38.17	42.46	10/10	10/10
ring_c.	58.30	57.17	57.50	67.09	7/10	8/10
t._mixed	37.57	37.67	38.08	44.39	10/10	2/10
t._sep.	38.04	38.00	38.35	40.85	10/10	7/10
cubes	29.51	27.29	27.29	29.71	—	—

7.2 The pilot fails for two distinguishable reasons

At $N_0=4$ pilot-variance allocation sits at or below uniform on five of seven scenes (Table 11), and the failures decompose into two components. The first is *generic weight-estimation noise*: Neyman weights from four samples are noisy on any scene, and where the true difficulty field is nearly homogeneous there is little to gain from concentration and much to lose from misplacing it. Lambertian `cubes_dof` is the pure case — observed-var loses 2.22 dB to uniform with no heavy-tailed transport in the scene — and `tabletop_separated`’s -0.04 dB is within seed noise. This component shrinks as the pilot grows (`cubes_dof` is within 0.15 dB of uniform by $N_0=64$), is repaired almost entirely by a per-pixel sample floor of 4 (`ring_caustic@4`: $-1.13 \rightarrow +0.02$ dB; `cubes_dof@4`: $-2.22 \rightarrow -0.94$ dB; supplementary), and is not what the descriptor addresses. The second is the *heavy-tail pathology* of §1: the pilot systematically underestimates variance on the tail pixels. Its budget-persistent form — a loss to uniform that does not vanish as the pilot grows — is exhibited by `glass_caustic` (-0.20 dB at $N_0=16$, -0.32 at 64; narrowing to -0.04 at a $4\times$ allocation budget, so the persistence is strongest at the pilot and budget scales adaptive sampling typically uses), the $n=1$ demonstration it is; the other two heavy-transport deficits flip positive by $N_0=16$. The tail-underestimation component is evidenced more directly by bucket-floor’s paired gains over observed-var (next subsection) and by the oracle’s $+5.9$ dB of unrealised headroom on `glass_caustic`. The deficits *amplify* with the budget being allocated ($-0.97 \rightarrow -4.39$ dB on `kitchen_counter@4` as the extra budget grows from 32 to 128 SPP): a noisy pilot does more damage the more samples it is trusted with. We measure no-harm against observed-var rather than uniform because the experiment isolates the label’s *marginal* contribution to the incumbent; whether a system should fall back toward uniform when its pilot is weak is an orthogonal control, and not a trivial one — a bucket-free global-shrinkage placebo, which is exactly such a fallback, underperforms observed-var on all seven scenes at $N_0=4$.

7.3 Conditioning repairs the tail component, with specificity

On paired MSE, bucket-floor improves on observed-var in 7–10 of 10 pilot draws on every first-party scene containing heavy-tailed buckets ($+0.28$ to $+0.41$ dB at $N_0=4$; delta-mediated-bucket MSE halved on `snooker`), and reduces to observed-var exactly on `cubes_dof`. Win counts are paired descriptive statistics, not significance tests; under the stricter pre-fixed criterion ($\geq 8/10$ wins *and* mean gain above one seed standard deviation, taking the larger of the two strategies’) the $N_0=4$ MSE wins on `glass_caustic` and `tabletop_separated` qualify, `kitchen_counter` and `snooker` qualify on the tail metric (10/10 p99 each), and `ring_caustic`’s 7/10 stays within seed noise at $N_0=4$ and reaches 10/10 at $N_0=16$. The two metrics can disagree informatively: on `tabletop_mixed@4` bucket-floor lowers delta-mediated-bucket MSE by 33% yet slightly raises image-wide p99 in 8/10 draws — the floor spreads budget across the bucket while observed-var concentrates on the few extreme pixels that set the global quantile whenever its pilot happens to catch them.

The spatial structure of the correction is the predicted one (Fig. 8, right): extra samples flow to the pixels labelled glossy and delta-mediated. The random-partition placebo reproduces none of the gains — its paired deltas are several times smaller or negative, and in every qualifying cell below half the treatment’s — so the gain is attributable to *which* pixels share a bucket, not to quantisation. The label-oracle variant performs within noise of the pilot-labelled one (38.15 vs 38.17 dB on `snooker@4`): estimating the label from the same 4-SPP pilot costs essentially nothing, which prices the finite-sample-label concern of §3.1 operationally. Gains shrink as the pilot grows — by $N_0=64$ every delta lies within ± 0.06 dB except `ring_caustic`’s $+0.24$ — the repair targets pilot noise and largely vanishes when the pilot no longer needs repairing.

7.4 Interaction with a robust variance estimator

The qualification of §2.1 is here tested for one robust estimator, median-of-means, and the result cuts both ways. As an incumbent, $\hat{\sigma}_{\text{MoM}}^2$ helps only where its assumptions hold: it beats observed-var under the strict criterion in exactly one cell (`glass_caustic@64`, 9/10), and on the most delta-mediated scene it is harmful (-2.35 dB vs observed-var on `ring_caustic@16`): median-of-means gains stability by discarding rare-tail evidence, and on single-firefly pixels that evidence *is* the signal. The mechanism floor is not absorbed by the robust incumbent: robust-floor improves on robust-var on paired MSE in 9–10 of 10 draws in eleven of twelve cells on the six scenes with heavy-tailed buckets (Table 12), and its matched placebo reproduces none of it. A block-count sweep localises the interaction (Fig. 9): within-bucket rank fidelity against the converged variance is flat in k for delta-mediated but collapses for specular-direct (Spearman $0.52 \rightarrow 0.20$ on `ring_caustic@16` as $k = 2 \rightarrow 8$), the error explosion localises in that bucket ($0.99 \rightarrow 13.1$ bucket MSE), and the mechanism-floored level is essentially estimator-invariant (0.66 – 0.70 across the naive pilot and $k \leq 4$). Choosing k safely would require knowing in advance which pixels are single-firefly dominated — precisely what $\hat{\sigma}^2$ is least able to report about itself and what the label supplies for free — and no fixed k serves both regimes, since mid-tail buckets *improve* under aggressive blocking (glossy rank fidelity $0.33 \rightarrow 0.74$ on `snooker@64`) while rare-tail buckets degrade. In the tested configuration, median-of-means estimation and mechanism

conditioning repair different failures, and the former does not substitute for the latter on rare-tail pixels; other robust constructions remain untested (§8).

Table 12: Pilot $N_0=16$: PSNR (dB, mean over 10 draws) for the naive incumbent, the robust incumbent (MoM, $k=4$), and their mechanism-floored variants; last columns: robust-floor’s paired-MSE win rate against robust-var at $N_0=16$ and 64. Bold: the robust incumbent falls below the naive one — on heavy-tailed scenes because blocking discards tail evidence, on cubes_dof because small-sample blocking is itself costly. On cubes_dof both floors are exact no-ops.

scene	obs-var	MoM	b-floor	r-floor	@16	@64
glass_c.	30.59	30.69	30.66	30.83	10/10	6/10
kitchen	34.14	34.32	34.34	34.69	9/10	10/10
snooker	39.35	39.31	39.57	39.54	10/10	9/10
ring_c.	62.24	59.89	62.87	62.66	10/10	9/10
t._mixed	39.14	38.72	39.17	39.00	10/10	9/10
t._sep.	39.60	39.29	39.67	39.59	10/10	10/10
cubes	29.37	28.79	29.37	28.79	—	—

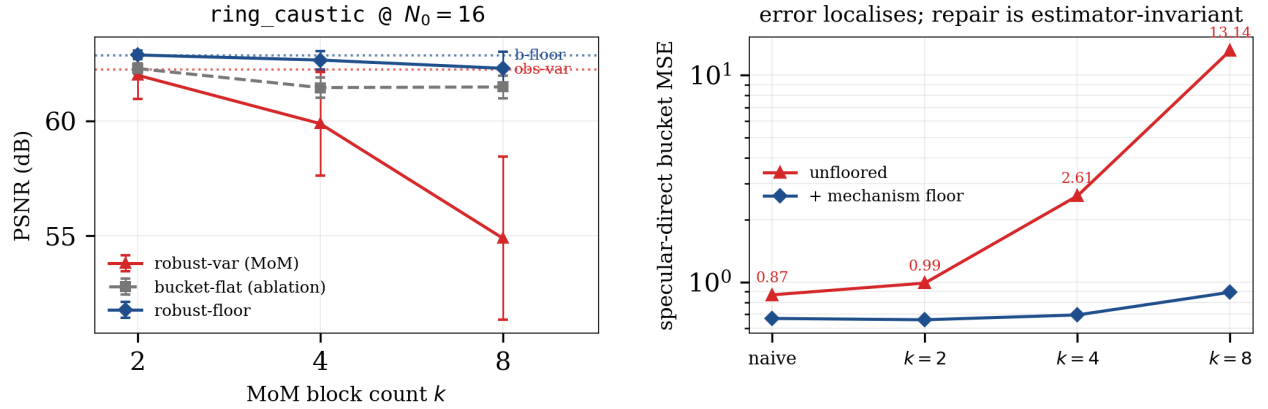


Figure 9: Robust estimation versus mechanism conditioning on *ring_caustic* at $N_0=16$. *Left*: PSNR versus median-of-means block count k : the robust incumbent (red) degrades monotonically as blocking becomes more aggressive; a per-bucket-uniform ablation (grey) recovers most of the loss — the partition component — and robust-floor (blue) holds a roughly constant level above it — the per-pixel component. Dotted: the naive incumbent and its floored variant. *Right*: the damage localises in specular-direct, whose within-bucket ranking collapses (bucket MSE 0.99 $\rightarrow$ 13.1 as $k = 2 \rightarrow 8$), while the floored level holds at 0.66–0.70 for $k \leq 4$.

7.5 Third-party geometry: veach-ajar

The same protocol on the classical *veach-ajar* scene (Table 13) gives the largest pilot failure measured in this work: observed-var loses to uniform by 6.8 dB at $N_0=4$ — roughly three times the largest first-party deficit — with the oracle a further 10.5 dB above observed-var. The fragility of pilot-variance allocation is therefore not specific to our scene construction. Bucket-floor ties observed-var (-0.004 dB, 5/10), for a reason the label itself reports: *veach-ajar* is heavy-tailed in variance but light-tailed in mechanism — its difficulty lies in dim diffuse-indirect transport, and its heavy-bucket population is $\approx 1\%$ of pixels. The floor acts on almost nothing, as designed. The label thus indicates, from the pilot alone, that this failure belongs to the generic-starvation component of §7.2 rather than the tail component it corrects; a shared per-pixel sample floor, which targets starvation, recovers 3.2 of the 6.8 dB (supplementary). The no-op behaviour observed on *cubes_dof* thus carries over to third-party geometry, and so does the distinction the paper rests on: mechanism is not variance.

7.6 What this demonstrates, and what it does not

No pilot-based strategy in Table 11 — ours included — beats uniform sampling on all seven scenes at $N_0=4$, and on the heavy-transport scenes where the pilot fails outright, bucket-floor narrows but does not close the gap to uniform. The demonstrated claim is improvement over the pilot-variance allocator: bucket-floor improves on observed-var wherever heavy-tailed buckets carry appreciable population and reduces to it where they are absent, so a system already running pilot-variance allocation loses nothing in our tests and gains on the heavy-transport scenes. The

Table 13: Equal-budget allocation on `veach-ajar` at $N_0=4$ (10 disjoint pilots). The pilot-variance allocator loses to uniform by 6.8 dB; bucket-floor ties observed-var because the scene’s heavy-bucket population is $\approx 1\%$.

strategy	PSNR (dB)	vs. obs-var
uniform	17.23 ± 0.00	+6.79
observed-var	10.45 ± 0.46	—
bucket-floor	10.44 ± 0.46	−0.004 (5/10)
oracle	20.93 ± 0.00	+10.48

oracle’s +5.9 to +8.8 dB headroom over uniform on those scenes locates the remaining loss in pilot *quality*, not in the variance-proportional principle. The experiment shows the descriptor in the role argued for: a stable conditioning variable that identifies a priori the pixels on which a continuous estimate needs repair — not a replacement difficulty scalar and not a complete allocator. It is one demonstration on one task with deliberately simple estimators; richer uses (denoising auxiliaries, population-model corrections in the spirit of [23], Bayesian shrinkage with mechanism priors, per-bucket adaptive block counts) remain open.

8 Scope and Limitations

Taxonomy scope. The descriptor covers surface path tracing under a stated D/G/S decomposition. It does not address participating media, subsurface transport, hair and foliage visibility complexity, glints, environment emitters or measured materials with non-unique lobe decompositions — several of which are major difficulty sources in production; extending the taxonomy there, e.g. with phase-function lobes as an additional axis, is future work. Coverage of the named mechanisms is established on the tested scene class only.

What the label sees, and what the sidefields carry. Two third-party scenes test the taxonomy’s boundary. On `veach-mis` — four metal plates of graded roughness under a graded light row, the scene designed to stress NEE/BSDF balance — the plate faces are dominated by *diffuse-indirect* (13,825 of 13,863 pixels), not *glossy*, contrary to our prediction: what a rough plate mostly shows the camera is a glossy-mediated *reflection of the diffuse surround*, camera $\rightarrow$ plate (G) $\rightarrow$ diffuse end-vertex, which the end-vertex rule routes to the diffuse end-vertex by design. Camera-side glossy mediation is invisible to the dominant label by construction and carried entirely by γ (plate-face glossy fraction 0.94–1.00 against ≈ 0.03 elsewhere). On `matpreview` with the principled BSDF’s clearcoat swept over $\{0, 0.25, 0.5, 1.0\}$, the NEE rule routes the coated material to **D** at every coat value, agreement is 0.989 throughout and the sweep moves 0.2% of pixels, while γ responds monotonically (mean 0.040 $\rightarrow$ 0.047; 90th percentile 0.166 $\rightarrow$ 0.196). Both are design consequences, not label instability: the discrete label holds the main axis still and the continuous mixture carries the rest. Two of our sharpened per-plate predictions on `veach-mis` failed for the same reason — they misread the taxonomy’s scope — and are reported as registered in the supplementary material. Measured BRDFs remain untested.

Convention and estimator dependence. The per-event label depends on renderer conventions; the per-pixel label is a finite-sample estimator. We measured robustness to budget changes and to one controlled estimator perturbation; bidirectional, guided and regularised estimators are untested.

Findings. The visibility account (§6.1) rests on a controlled pair, an ordering, and one blind prediction — not a parametric separation sweep. The depth/purity account is consistent with seven scenes but, except for `ring-caustic`, derived from uncontrolled comparisons. Two further sign-varying pairs and the stable-glossy sub-population of `snooker` remain open (supplementary). By construction the *direct* bucket admits both Lambertian single-bounce transport and single-bounce glossy highlights; on `kitchen_counter` a γ -split isolates a 4.7% glossy-lit sub-population with twice the median cost and half the $\hat{\sigma}^2$ reliability of its Lambertian complement (supplementary). An explicit lobe sub-split of *direct* would resolve this without disturbing the main partition.

Downstream. The allocation experiment uses a simple NEE-only pilot, a two-stage budget split and one robust incumbent (median-of-means, chosen for having no free parameter beyond the block count); Catoni-type M -estimators and production adaptive pipelines are untested, and we deliberately do not evaluate the label as a denoiser auxiliary channel, where any conclusion would be dominated by the choice of denoiser family. The split-half methodology applies to denoiser guidance maps directly and is the form in which we expect the descriptor to be useful there first.

9 Conclusion

Per-pixel rendering difficulty has had thirty-five years of operational consensus around a single noisy scalar. We have argued that difficulty should be characterised through transport structure, with variance treated as a measurement whose reliability that structure helps predict, and we have supported the argument with measurements. The finite-sample variance is governed by the integrand’s fourth moment, of low split-half reliability as an evaluation target, and least reliable on the pixels a mechanism descriptor isolates in advance. The descriptor itself is stable across a $64\times$ budget change where quantile-binned variance is not, survives a controlled estimator perturbation, and on three third-party scenes its stability was predicted from each scene’s pilot. Read per mechanism, difficulty shows structure a scalar hides: in our scenes the caustic is not the most expensive transport, lobe width governs the glossy firefly tail, and coupling between mechanisms follows scene geometry. Conditioning on the descriptor has measurable value: a pilot variance corrected by the label improves equal-budget allocation wherever heavy-tailed buckets carry appreciable population, reduces to the incumbent where they are absent (exactly) or negligible (to within 0.004 dB), and repairs a failure that the tested median-of-means estimator does not. Under heavy-tailed transport, finite-sample variance is an unstable measurement; the transport mechanism of each contribution — deterministic under stated renderer conventions — is a stable structure to characterise difficulty by, and conditioning on it yields a measurable gain.

Several extensions follow. A perceptual variant of the cost analysis (FLIP, FovVideoVDP) may censor differently; a Bayesian shrinkage formulation with mechanism as prior and sample statistics as likelihood is the natural composition; a per-bucket adaptive median-of-means block count is motivated directly by §7.4; and a richer explicit path-length axis would complement the lobe-based labels along a dimension orthogonal to mechanism type.

A Reproducibility

All first-party results were produced with Mitsuba 3.8.0 (cuda_ad_rgb) on a single NVIDIA RTX 5090; the third-party suite ran under the same renderer version on an RTX 4090, recorded per render. Four auxiliary measurements predate the uniform re-render and are used unchanged from the original runs: the cross-budget convergence study behind $\rho = 0.26$, the $\hat{\alpha}$ reliability estimate, the snooker glossy sub-population seed renders, and the light-count control. The complete first-party pipeline re-runs in under one hour. This appendix and the supplementary material give the implementation and protocol specifications in detail; the corresponding code — the custom integrator, all procedural scene builders, analysis scripts, seed layouts, and the allocation protocol, from whose entry points every table and figure regenerates — is being consolidated into a public research artifact and will be released separately.

Classifier. Per-event classification follows Table 1 with the sampled-lobe cascade $\Delta > \text{Diffuse} > \text{Glossy}$ for BSDF-sampled bounces and the flag rule of §3.1 for NEE events. BSDF-to-lobe mapping: `diffuse` $\rightarrow$ D; `roughplastic` $\rightarrow$ D under NEE, sampled lobe otherwise; `roughconductor`, `roughdielectric` $\rightarrow$ G; `conductor`, `dielectric` (smooth) $\rightarrow$ S; `twosided` transparent to flags. Contribution energy is Rec. 709 luminance of the full-throughput contribution (post Russian-roulette compensation, post MIS weight where applicable); zero-energy pixels carry no label; argmax ties break to the lower bucket index (measure zero). Contributions are non-negative by construction; non-finite values do not arise in the bucket accumulators, and rare non-finite pixels in raw float32 RGB frames are sanitised to zero by the analysis loaders and counted as background. The `delta`-mediated bucket and the `$\delta$ -share` sidefield carry the keys `caustic` and `caustic_frac` in the implementation.

Estimator. NEE at every Smooth-flagged vertex (one NEE and one BSDF sample per vertex); BSDF-into-emitter accepted iff the primary ray or the previous bounce was delta; Russian roulette from depth 5 at $p = \min(\max \beta, 0.95)$; max depth 12; 256^2 , box filter, independent sampler. Restricted MIS: power heuristic ($\beta=2$, solid-angle measure) on the NEE and emitter-hit branches of the glossy and `delta`-mediated buckets only.

Statistics. $\hat{\sigma}^2$ at budget K is $K \cdot \text{Var}$ across 10 seeds (ddof 1) of per-seed luminance; background dropped at mean luminance $\leq 10^{-4}$; split-half values are means over 20 random 5+5 partitions (base seed 0). Quantile bins: per scene, independently at each SPP. Weighted κ as in Table 3. Time-to-quality: $\tau = 0.10$, $\epsilon = 10^{-4}$, SPP grid $\{16, \dots, 2048\}$ doubling, independent renders per level, first-crossing rule, reference the seed-disjoint SPP-4096 stability render. Convergence exponent $\hat{\alpha}(p)$: per-pixel unweighted OLS of $\log_{10}$ luminance variance on $\log_{10}$ SPP over four levels, $\hat{\alpha} = -\text{slope}$; pixels with non-positive or non-finite variance at any level excluded; per-pixel R^2 is stored alongside. Separation measures: minimum surface distance and mutual visibility over 64 surface points per object pair (ray-tested, offset 10^{-3} , seed 0); on `bathroom2` a 20k-pair unbiased subsample, family assignment by BSDF flags. Roughness sweep: $\alpha \in \{0.01, 0.02, 0.03, 0.042, 0.06, 0.10, 0.18, 0.30\}$, 30 seeds $\times$ 128 SPP per level plus a 4096-SPP reference at 384^2 . Allocation protocol: seed layout pool = $1000 + i$, pilot draw k frame $j = 100000 + 1000k + j$, reference chunks = $900000 + c$, disjoint by construction; median-of-means over $k = \lceil \sqrt{N_0} \rceil$ contiguous render-order blocks (per-block

variance $\text{ddof } 1$), sensitivity $k \in \{2, 4, 8\}$ at $N_0=16$ and $\{4, 8, 16\}$ at $N_0=64$; pilot frames re-rendered with per-frame luminance storage under identical seeds, SHA-256 manifests verifying every pool and pilot statistic bit-identical. Third-party scenes: Bitterli’s collection in native Mitsuba-3 XML, 448×256 , paper estimator configuration overriding native defaults; `matpreview`’s environment map replaced by an overhead 3×3 area panel at radiance 7.2 (mean image luminance 0.38, matching the original). All predictions for the third-party scenes, the frozen reliability curves, and every prediction/outcome file are retained for inclusion in the public artifact and summarised in the supplementary material.

References

- [1] George Casella and Roger L. Berger. *Statistical Inference*. Duxbury / Thomson Learning, Pacific Grove, CA, 2nd edition, 2002.
- [2] A. W. van der Vaart. *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.
- [3] Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling Extremal Events for Insurance and Finance*, volume 33 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, Berlin, 1997.
- [4] Eric Veach. *Robust Monte Carlo Methods for Light Transport Simulation*. Ph.D. dissertation, Stanford University, December 1997. Technical Report STAN-CS-TR-98-1610.
- [5] Thomas Kollig and Alexander Keller. Illumination in the presence of weak singularities. In Harald Niederreiter and Denis Talay, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2004*, pages 245–257. Springer, Berlin, Heidelberg, 2006.
- [6] Art B. Owen. Quasi-Monte Carlo for integrands with point singularities at unknown locations. In Harald Niederreiter and Denis Talay, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2004*, pages 403–417. Springer, Berlin, Heidelberg, 2006.
- [7] Christopher DeCoro, Tim Weyrich, and Szymon Rusinkiewicz. Density-based outlier rejection in Monte Carlo rendering. *Computer Graphics Forum*, 29(7):2119–2125, 2010. Proceedings of Pacific Graphics 2010.
- [8] Tobias Zirr, Johannes Hanika, and Carsten Dachsbacher. Re-weighting firefly samples for improved finite-sample Monte Carlo estimates. *Computer Graphics Forum*, 37(6):410–421, 2018.
- [9] Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.
- [10] Charles Spearman. Correlation calculated from faulty data. *British Journal of Psychology*, 3(3):271–295, 1910.
- [11] Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, 2012.
- [12] Luc Devroye, Matthieu Lerasle, Gábor Lugosi, and Roberto I. Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.
- [13] Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.
- [14] Kai Yan, Vincent Pegoraro, Marc Droske, Jiří Vorba, and Shuang Zhao. Differentiating variance for variance-aware inverse rendering. In *SIGGRAPH Asia 2024 Conference Papers (SA ’24)*, pages 1–10, New York, NY, USA, 2024. ACM.
- [15] Arthur Firmino, Jeppe Revall Frisvad, and Henrik Wann Jensen. Denoising-aware adaptive sampling for Monte Carlo ray tracing. In *ACM SIGGRAPH 2023 Conference Proceedings (SIGGRAPH ’23)*, pages 32:1–32:11, New York, NY, USA, 2023. ACM.
- [16] Don P. Mitchell. Generating antialiased images at low sampling densities. In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH ’87)*, volume 21 of *Computer Graphics*, pages 65–72. ACM, 1987.
- [17] Fabrice Rousselle, Claude Knaus, and Matthias Zwicker. Adaptive sampling and reconstruction using greedy error minimization. *ACM Transactions on Graphics*, 30(6):159:1–159:12, 2011. Proceedings of SIGGRAPH Asia 2011.
- [18] Christoph Schied, Anton Kaplanyan, Chris Wyman, Anjul Patney, Chakravarty R. Alla Chaitanya, John Burgess, Shiqiu Liu, Carsten Dachsbacher, Aaron Lefohn, and Marco Salvi. Spatiotemporal variance-guided filtering: Real-time reconstruction for path-traced global illumination. In *Proceedings of High Performance Graphics (HPG ’17)*, pages 1–12. ACM, 2017.

- [19] Alexander Rath, Pascal Grittmann, Sebastian Herholz, Philippe Weier, and Philipp Slusallek. EARS: Efficiency-aware Russian roulette and splitting. *ACM Transactions on Graphics*, 41(4), 2022. Proceedings of SIGGRAPH 2022.
- [20] Petr Vévoda, Ivo Kondapaneni, and Jaroslav Křivánek. Bayesian online regression for adaptive direct illumination sampling. *ACM Transactions on Graphics*, 37(4), 2018. Proceedings of SIGGRAPH 2018.
- [21] Matthias Zwicker, Wojciech Jarosz, Jaakko Lehtinen, Bochang Moon, Ravi Ramamoorthi, Fabrice Rousselle, Pradeep Sen, Cyril Soler, and Sung-Eui Yoon. Recent advances in adaptive sampling and reconstruction for Monte Carlo rendering. *Computer Graphics Forum*, 34(2):667–681, 2015. Eurographics State of the Art Reports.
- [22] Alexandr Kuznetsov, Nima Khademi Kalantari, and Ravi Ramamoorthi. Deep adaptive sampling for low sample count rendering. *Computer Graphics Forum*, 37(4):35–44, 2018. Proceedings of EGSR 2018.
- [23] Hiroyuki Sakai, Christian Freude, Michael Wimmer, and David Hahn. Statistical error reduction for Monte Carlo rendering. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, New York, NY, USA, 2025. ACM.
- [24] Steve Bako, Thijs Vogels, Brian McWilliams, Mark Meyer, Jan Novák, Alex Harvill, Pradeep Sen, Tony DeRose, and Fabrice Rousselle. Kernel-predicting convolutional networks for denoising Monte Carlo renderings. *ACM Transactions on Graphics*, 36(4):97:1–97:14, 2017. Proceedings of SIGGRAPH 2017.
- [25] Thijs Vogels, Fabrice Rousselle, Brian McWilliams, Gerhard Röhlin, Alex Harvill, David Adler, Mark Meyer, and Jan Novák. Denoising with kernel prediction and asymmetric loss functions. *ACM Transactions on Graphics*, 37(4):124:1–124:15, 2018. Proceedings of SIGGRAPH 2018.
- [26] Frédo Durand, Nicolas Holzschuch, Cyril Soler, Eric Chan, and François X. Sillion. A frequency analysis of light transport. *ACM Transactions on Graphics*, 24(3):1115–1126, July 2005. Proceedings of SIGGRAPH 2005.
- [27] Cyril Soler, Kartic Subr, Frédo Durand, Nicolas Holzschuch, and François Sillion. Fourier depth of field. *ACM Transactions on Graphics*, 28(2):18:1–18:12, May 2009.
- [28] Laurent Belcour, Cyril Soler, Kartic Subr, Nicolas Holzschuch, and Frédo Durand. 5d covariance tracing for efficient defocus and motion blur. *ACM Transactions on Graphics*, 32(3):31:1–31:18, July 2013.
- [29] Ravi Ramamoorthi, Dhruv Mahajan, and Peter Belhumeur. A first-order analysis of lighting, shading, and shadows. *ACM Transactions on Graphics*, 26(1):2:1–2:21, January 2007.
- [30] Paul S. Heckbert. Adaptive radiosity textures for bidirectional ray tracing. *ACM SIGGRAPH Computer Graphics*, 24(4):145–154, September 1990.
- [31] Henrik Wann Jensen. Global illumination using photon maps. In Xavier Pueyo and Peter Schröder, editors, *Rendering Techniques '96 (Proceedings of the 7th Eurographics Workshop on Rendering)*, pages 21–30, Porto, Portugal, 1996. Springer-Verlag.
- [32] Jiří Vorba, Ondřej Karlík, Martin Šik, Tobias Ritschel, and Jaroslav Křivánek. On-line learning of parametric mixture models for light transport simulation. *ACM Transactions on Graphics*, 33(4):101:1–101:11, July 2014. Proceedings of SIGGRAPH 2014.
- [33] Thomas Müller, Markus Gross, and Jan Novák. Practical path guiding for efficient light-transport simulation. *Computer Graphics Forum*, 36(4):91–100, 2017. Proceedings of EGSR 2017.
- [34] Thomas Müller, Brian McWilliams, Fabrice Rousselle, Markus Gross, and Jan Novák. Neural importance sampling. *ACM Transactions on Graphics*, 38(5):145:1–145:19, October 2019.
- [35] Johannes Hanika, Marc Droske, and Luca Fascione. Manifold next event estimation. *Computer Graphics Forum*, 34(4):87–97, 2015. Proceedings of EGSR 2015.
- [36] Tizian Zeltner, Iliyan Georgiev, and Wenzel Jakob. Specular manifold sampling for rendering high-frequency caustics and glints. *ACM Transactions on Graphics*, 39(4):149:1–149:15, July 2020. Proceedings of SIGGRAPH 2020.
- [37] Anton S. Kaplanyan and Carsten Dachsbacher. Path space regularization for holistic and robust light transport. *Computer Graphics Forum*, 32(2):63–72, 2013. Proceedings of Eurographics 2013.
- [38] Johannes Jendersie and Thorsten Grosch. Microfacet model regularization for robust light transport. *Computer Graphics Forum*, 38(4):39–47, 2019. Proceedings of the 30th Eurographics Symposium on Rendering (EGSR 2019).
- [39] Pascal Grittmann, Iliyan Georgiev, Philipp Slusallek, and Jaroslav Křivánek. Variance-aware multiple importance sampling. *ACM Transactions on Graphics*, 38(6):152:1–152:9, November 2019. Proceedings of SIGGRAPH Asia 2019.

- [40] Pixar Animation Studios. Light path expressions — RenderMan 26 documentation. <https://rmanwiki-26.pixar.com/space/REN26/19661883/Light+Path+Expressions>, 2024. Accessed: 2026-05-27.
- [41] Benedikt Bitterli. Rendering resources. <https://benedikt-bitterli.me/resources/>, 2016. Scene collection; individual scene licenses as listed (CC0 / CC-BY 3.0).
- [42] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.
- [43] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.
- [44] Jacob Cohen. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968.

SUPPLEMENTARY MATERIAL

FOR CHARACTERISING PER-PIXEL RENDERING DIFFICULTY BY TRANSPORT MECHANISM

This document collects material referenced from the main paper that supports, but is not required for, its argument: additional correlation observations and controlled sweeps (§S1), the per-dimension cost analysis (§S2), within-bucket heterogeneity of the `direct` bucket (§S3), per-bucket estimator-perturbation robustness (§S4), the full control set of the allocation experiment (§S5), and the complete record of the third-party scene suite, including every registered prediction and its outcome (§S6). Section numbers of the form § n refer to the main paper; § Sn to this document.

S1 Additional Correlation Observations and Sweeps

Two further sign-varying pairs. Main-paper Table 9 records two pairs whose sign varies across scenes without, so far, an identified controlling variable. *Depth versus δ -share* runs $+0.73, +0.45, +0.43, +0.29, +0.13$ across five non-degenerate scenes but -0.21 on `tabletop_separated`. The positive values are unsurprising — delta-mediated paths include a specular bounce and so tend to be at least one vertex longer than direct paths. The negative outlier is harder to interpret with confidence: the partition geometry constrains both depth (paths must detour around the wall) and the δ -share (which surfaces the partition occludes) simultaneously, and a single-scene anomaly does not distinguish a genuine mechanism from a partition-induced artefact. *Inverse purity versus glossy fraction* is mildly negative on `snooker` and the tabletops (-0.19 to -0.25) but $+0.55$ on `kitchen_counter`, where glossy energy arrives layered over direct and diffuse transport on the same pixels, so glossy-heavy pixels are also mixed pixels. Both observations would need controlled sweeps analogous to the tabletop pair to be promoted to findings. The *δ -share versus inverse purity* pair is weak ($|\rho| \leq 0.35$) on every non-degenerate scene.

What lobe width does not explain. The `snooker` glossy bucket contains a sub-population — the 31 pixels that survive MIS completion as glossy (main paper §4.4) — with both a low firefly ratio (≈ 3) and a high split-half ρ (≈ 0.86), a “stable” glossy regime that the single-sphere roughness sweep (main paper §5.3) never reproduces at any α . We tested whether per-pixel path multiplicity was the missing variable by holding $\alpha=0.03$ fixed and sweeping the number of energy-conserving light sources from one to sixteen. It is not: the firefly ratio stayed flat (as predicted, since α was fixed), but ρ did not climb out of the noise band, because adding light *sources* did not increase per-pixel supplying-path *multiplicity* — each pixel still saw only one or two caustic hits. The stable-glossy regime of `snooker` is thus governed by some scene-topology variable — multi-object inter-reflection is our leading candidate — that neither controlled scene reproduces; isolating it requires a controlled two-object experiment.

S2 Cost Versus the Descriptor Dimensions

Main-paper §5.2 reports per-bucket time-to-quality cost (per-bucket medians in Fig. S1). Table S1 asks the complementary question: whether the continuous sidefields correlate with cost in a stable way — whether depth, δ -share, inverse purity or glossy fraction would each serve, alone, as a cheap cost proxy. They would not, and the reasons are informative about both the cost measurement and the vector.

Glossy fraction is the only sign-stable dimension, and its magnitude is structurally capped. $\rho(\text{glossy frac.}, c)$ is positive on all four non-degenerate scenes, in the narrow range $+0.20$ to $+0.29$; no other component is as stable. But $\rho \approx 0.2$ explains only $\sim 4\text{--}8\%$ of the variance in c , and on `snooker` this is not the intrinsic strength of the relationship: with 51.8% of the NEE-defined glossy bucket right-censored, Spearman ρ assigns those pixels the same maximal rank, compressing the top of the distribution into a plateau that pulls correlations toward zero. The weak ρ does not mean the descriptor fails to track cost; it means cost itself has gone flat on the very pixels the vector flags as glossy-dominant.

Table S1: Per-pixel Spearman correlation between time-to-quality $c(p)$ and the four continuous components, all seven scenes (column abbreviations as in the main paper). “n/a”: no energy in that scene. Every cell is reported.

dim	sn	gc	tm	ts	cb	rc	kc
depth	+0.27	+0.22	+0.44	+0.04	+0.35	+0.22	+0.34
δ -share	+0.07	+0.17	+0.21	+0.01	n/a	+0.15	+0.05
inv. purity	-0.16	+0.06	+0.07	-0.10	+0.09	+0.05	+0.26
glossy frac.	+0.23	n/a	+0.22	+0.20	n/a	n/a	+0.29

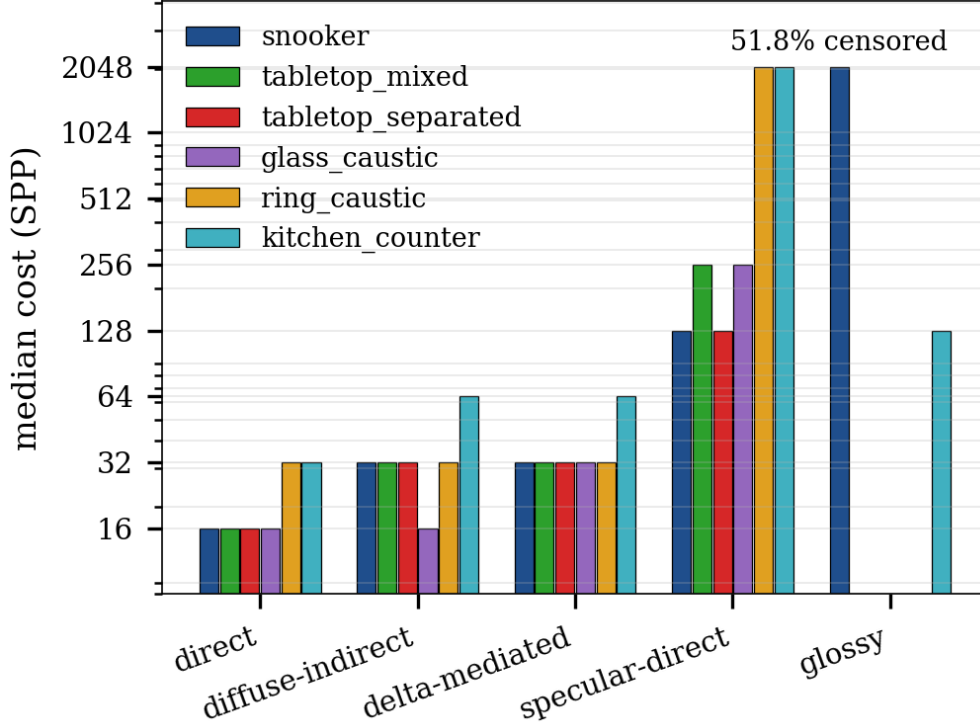


Figure S1: Median time-to-quality cost per dominant-label bucket across the six scenes with non-trivial transport (main-paper Table 8 as a figure). Three censored medians sit at the SPP= 2048 cap: the NEE-defined glossy bucket on snooker (51.8% right-censored) and the deep-chain specular-direct buckets on ring_caustic (53.0%) and kitchen_counter (48.8%), while kitchen_counter’s population-scale glossy bucket converges at a median of 128 SPP.

Inverse purity versus cost is sign-unstable. $\rho(\text{inv. purity}, c)$ runs -0.16 to $+0.26$ across the seven scenes — mostly small, signs disagreeing. The instability is consistent within each scene but needs care: the link between mechanism mixing and cost is mediated by which bucket the mixed pixels fall into and which buckets are expensive there. The component flags mixture; it is not, alone, a stable cost proxy.

Depth versus cost is geometry-driven. $\rho(\text{depth}, c)$ sits at $+0.22$ to $+0.44$ on six scenes but collapses to $+0.04$ on tabletop_separated — the same scene where depth versus δ -share inverts sign (§S1), and plausibly for the same reason: the partition forces a family of long detour paths whose depth does not coincide with the transport that drives cost elsewhere. (This is not an instance of the depth/purity sign structure of main-paper §6.2, whose negative- ρ scenes are snooker and ring_caustic, where depth–cost remains positive.) Depth is the component whose relationship with cost is most geometry-sensitive; any predictor using it as a cost proxy must do so per scene.

Reading the table. A naive reading is that the descriptor vector tracks cost weakly. The closer reading is that the continuous-cost signal itself is locally degenerate on the hardest pixels, sign-unstable on the mixture component and geometry-varying on the depth component. The mechanism partition sits ahead of all three: it separates the saturating pixels into their own buckets before averaging, and need not correlate with cost to be informative — the bucket identity is the signal.

S3 Within-Bucket Heterogeneity of direct

By construction the `direct` bucket admits both Lambertian single-bounce transport and single-bounce glossy highlights, and the two differ where they coexist. Splitting `direct` by the γ sidefield (threshold 0.5; a *proxy* for the end-vertex lobe, not a definition — fixed in the floor-experiment registration before the analysis ran) on `kitchen_counter` isolates a glossy-lit sub-population of 1,960 pixels (4.7% of the bucket) with twice the median time-to-quality (64 vs 32 SPP), a 12 $\times$ censored fraction (3.7% vs 0.3%), and half the $\hat{\sigma}^2$ split-half reliability (0.29 vs 0.60) of its Lambertian complement — quantified heterogeneity inside the cheapest bucket. The other tested scenes carry essentially no such sub-population (13 pixels on `snooker`, 8 on `veach-mis`); on `veach-mis` the glossy-mediated energy instead lands in `diffuse-indirect` and is carried by γ (§S6.3). An explicit lobe sub-split of `direct` would resolve this axis without disturbing the main partition; we do not re-partition, since every downstream number would move for a $\leq 5\%$ sub-population.

S4 Per-Bucket Estimator-Perturbation Robustness

Figure S2 resolves main-paper Table 4 per bucket on `snooker`. Two implementation checks accompany the experiment. First, on the five scenes without a glossy population — four of which carry substantial delta-mediated content, up to `ring_caustic`’s 6,949 pixels — the labelling is unchanged *to the pixel*: reweighting a bucket whose membership is a δ -event predicate cannot flip it. Second, every flip that does occur on the remaining two scenes lies on the energy boundaries the restricted MIS actually reweights — overwhelmingly `diffuse-indirect` $\leftrightarrow$ `glossy` exchanges (401 of the 403 `diffuse-indirect` departures on `kitchen_counter` go to `glossy`; the residue trades with `delta-mediated`) — and the single `specular-direct` $\rightarrow$ `delta-mediated` transition per scene is a known artefact of an upper-bound test rather than a classification ambiguity. `delta-mediated` agrees at 99.2% or better on every scene.

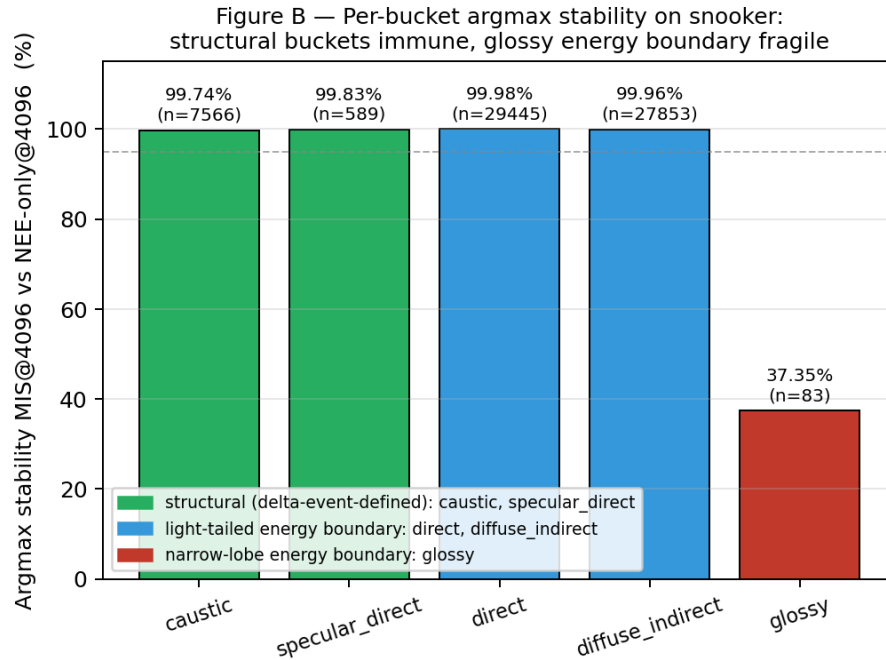


Figure S2: Per-bucket argmax stability on snooker under estimator completion (MIS@4096 vs NEE-only@4096). The predicate buckets (delta-mediated, specular-direct) and the light-tailed energy boundaries (direct, diffuse-indirect) are all stable above 99.7%; only the narrow-lobe glossy energy boundary moves. Dashed line: 95%.

S5 Allocation Experiment: Full Control Set

The allocation experiment of main-paper §7 was executed under a written protocol with hypotheses, metrics and decision rules fixed before the runs. The first execution required three disclosed corrections, none of which altered thresholds,

metrics, scenes or treatment definitions: the allocation exponent was fixed to Neyman $n \propto \sqrt{D}$ for every strategy alike; strict equal totals were enforced by construction; and a shrink-placebo control absent from the original design was added. The robust-variance incumbent and its hypotheses were added by a protocol amendment with estimator definition, block-count rule and decision criteria fixed before any robust-variance run was executed; the amendment and its timestamps are retained for inclusion in the public artifact, kept separate from the original registration.

The per-pixel sample floor ($f=4$). One configuration choice is disclosed rather than pre-registered: a per-pixel sample floor of $f=4$, applied to every strategy with the floor’s budget deducted from the Neyman share, first appeared as a control in the budget sweeps, and we considered promoting it to the shared base configuration. The selection rule was fixed before inspecting its inputs and does not depend on the vs-uniform outcomes: $f=4$ becomes primary only if it reduces bucket-floor’s paired win rate on no scene and preserves the cubes_dof no-op. The verdict: the no-op holds exactly, but the win rate drops in seven scene–budget cells (e.g. kitchen_counter@4: 9/10 $\rightarrow$ 6/10) — the two floors partially overlap in what they repair, so the mechanism floor’s marginal gain shrinks once a sample floor absorbs the pilot’s worst starvation. $f=1$ therefore remains primary and $f=4$ is reported as the control (Table S2 at $N_0=4$; the remaining budgets in the artifact). The control table also answers the question it invites — is a one-line sample floor simply enough on its own? No: on top of the shared floor, bucket-floor retains a positive mean gain over observed-var on every one of the six scenes with heavy-tailed buckets at $N_0=4$ (paired MSE wins 6–10/10), and still degenerates to the incumbent exactly where those buckets are absent. The sample floor repairs starvation; the mechanism floor adds information the sample floor does not have.

Table S2: The $f=4$ shared-sample-floor control configuration at pilot $N_0=4$ (same protocol and pool as main-paper Table 11; the floor’s 4 SPP per pixel is deducted from the Neyman share, totals unchanged). The floor repairs much of the generic-starvation component — and only it: ring_caustic’s deficit closes to within seed noise (+0.02 dB against a ± 2.2 seed std), glass_caustic and kitchen_counter remain below uniform, and cubes_dof’s no-op is exact. Bold: observed-var at or below uniform.

scene	uniform	obs-var	b-floor	oracle	MSE	p99
glass_c.	30.79	30.29	30.55	36.69	10/10	5/10
kitchen	33.71	33.00	33.18	41.16	6/10	10/10
snooker	34.97	37.95	38.28	42.46	9/10	10/10
ring_c.	58.30	58.32	58.56	67.09	8/10	9/10
t._mixed	37.57	37.90	38.25	44.39	10/10	2/10
t._sep.	38.04	38.53	38.76	40.89	10/10	5/10
cubes	29.51	28.57	28.57	29.72	—	—

Companion budgets. Table S3 completes the primary ($f=1$) configuration at the two larger pilot budgets; gains shrink as the pilot improves, as main-paper §7.3 reports.

Table S3: Primary configuration at pilot $N_0=16$ (top) and $N_0=64$ (bottom): PSNR (dB, mean over 10 disjoint draws) and bucket-floor’s per-draw win rates against observed-var on paired MSE and p99. Bold: observed-var at or below uniform.

scene	uniform	obs-var	b-floor	oracle	MSE	p99
$N_0=16$						
glass_c.	30.79	30.59	30.66	36.64	8/10	9/10
kitchen	33.71	34.14	34.34	41.18	8/10	10/10
snooker	34.97	39.35	39.57	42.46	8/10	10/10
ring_c.	58.30	62.24	62.87	67.09	10/10	5/10
t._mixed	37.57	39.14	39.17	44.39	6/10	1/10
t._sep.	38.04	39.60	39.67	40.85	9/10	4/10
cubes	29.51	29.37	29.37	29.71	—	—
$N_0=64$						
glass_c.	30.79	30.47	30.45	36.64	3/10	3/10
kitchen	33.71	35.26	35.24	41.18	4/10	10/10
snooker	34.97	39.88	39.93	42.46	4/10	6/10
ring_c.	58.30	62.98	63.22	67.09	8/10	5/10
t._mixed	37.57	39.46	39.46	44.39	4/10	3/10
t._sep.	38.04	39.90	39.90	40.85	4/10	1/10
cubes	29.51	29.66	29.66	29.71	—	—

Separating the two failure components experimentally. Three single-variable sweeps separate the generic-noise and heavy-tail components of the pilot’s failure. On the three diagnostic scenes chosen to represent the pure small-sample case, the pure heavy-tail case and the boundary case (*cubes_dof*, *glass_caustic*, *tabletop_separated*), every gap is unchanged when the test is repeated at 512^2 , ruling out a resolution artefact. The small-pilot deficits *amplify* with the budget being allocated ($-0.97 \rightarrow -4.39$ dB on *kitchen_counter* at $N_0=4$ as the extra budget grows from 32 to 128 SPP, while the positive gaps compress — *snooker@64* $+4.90 \rightarrow +1.77$ dB; one near-zero case, *tabletop_mixed@4* at $+0.10$, slips just below zero): a noisy pilot does more damage the more samples it is trusted with, so the deficit is a liability that scales with commitment rather than a fixed cost of small pilots. And the per-pixel sample floor of 4 repairs the small-sample component almost entirely (*ring_caustic@4*: $-1.13 \rightarrow +0.02$ dB; *cubes_dof@4*: $-2.22 \rightarrow -0.94$ dB) yet leaves *glass_caustic* below uniform at every pilot budget, without touching bucket-floor’s gain ($+0.26$ vs $+0.28$ dB). The first component is a starvation pathology a dumb floor can fix; the second is a mis-ranking pathology it cannot. The *glass_caustic* deficit persists at $N_0=16$ (-0.20 dB) and 64 (-0.32 dB); at a $4\times$ allocation budget the $N_0=64$ deficit narrows to -0.04 dB. *kitchen_counter* and *ring_caustic* flip positive by $N_0=16$ ($+0.43$ and $+3.94$ dB over uniform).

Ablations disqualified by our own controls. Two further treatments were run and are reported as ablations, not methods: cross-scene transfer of per-bucket scale factors fails the no-harm criterion under leave-one-scene-out transfer, and reliability-weighted within-bucket shrinkage has gains largely reproduced by a bucket-free global-shrinkage placebo. That global-shrinkage placebo — a fallback-by-regularisation toward uniform — underperforms observed-var on all seven scenes at $N_0=4$. Under the shared sample floor the shrink of bucket-floor’s gain with pilot size flattens (configuration-dependent).

Robust-variance incumbent. $\hat{\sigma}_{\text{MoM}}^2$ is the median-of-means variance over $k = \lceil \sqrt{N_0} \rceil$ contiguous equal blocks of the pilot frames in render order (per-block unbiased variance, median across blocks; trailing remainder discarded and counted). Block-count sensitivity ($k \in \{2, 4, 8\}$ at $N_0=16$, $\{4, 8, 16\}$ at $N_0=64$) changes no conclusion. Any global multiplicative bias of the MoM variance cancels under Neyman normalisation; only cross-pixel ordering acts. The median-of-means estimate costs on the order of 0.1 s per 256^2 pilot against ~ 1 ms for the naive sample variance — a large relative factor, negligible against rendering. Mean PSNR can favour MoM without meeting the paired criterion (*glass_caustic@16* and *kitchen_counter@16*: 7/10 paired wins with gains below one seed standard deviation — throughout, the larger of the two strategies’ seed stds). Robust-floor qualifies under the strict criterion on *tabletop_mixed@16* and *tabletop_separated@16/64*; its matched placebo deltas are ≈ 0 or negative in every qualifying cell. The block-count mechanism sweep on *ring_caustic@16* reads, for the specular-direct bucket at the registered k : naive $0.87 \rightarrow$ MoM $2.6 \rightarrow$ MoM+floor 0.70 bucket MSE; the floor already improves the naive pilot ($0.87 \rightarrow 0.67$), and the repaired level is $0.66\text{--}0.70$ across the naive pilot and $k \leq 4$, with residual degradation to 0.895 only at $k=8$, whose two-sample blocks carry a single degree of freedom — the same minimum-sample-size limit that rules out $N_0=4$ entirely.

S6 Third-Party Scenes: Complete Record

Every first-party scene is procedurally generated by the authors. The third-party scenes address the objection that the test distribution’s generating process was under our control. Rather than adding scenes indiscriminately, we selected one third-party scene per claim that author-built scenes cannot test. The four scenes were run under a written registration (assets, protocol constants and every prediction fixed before any bucket-instrumented render of a third-party scene); one amendment was registered after a first-batch outcome and is marked as such below. All predictions are reported against their outcomes, pass or fail.

S6.1 Assets and protocol

Scenes are from Bitterli’s *Rendering Resources* collection in native Mitsuba 3 form (no format conversion): *veach-ajar* (CC0; the classical ajar-door indirect-transport stress case), *veach-mis* (CC0; four metal plates of graded roughness under a graded light row, designed to stress NEE/BSDF balance), *bathroom2* (CC-BY 3.0; an uncontrolled interior with mirror and chrome fixtures beside glossy tiles), and *matpreview* (the material test ball distributed with the collection, originally from the Mitsuba repository). A pre-render audit checks each against the scope of main-paper §3.5: all emitters are area lights except *matpreview*’s environment map, which is replaced by an overhead area panel calibrated to the original’s mean image luminance (geometry and camera untouched). Renders use the paper’s estimator configuration (max depth 12, NEE, seed layout as in the main-paper appendix) at 448×256 — native aspect at the paper’s pixel scale — with the identical 64-vs-4096-SPP stability protocol. One deviation: the RGL measured-BRDF half of the taxonomy scene could not be executed (the database’s downloads are not fetchable non-interactively); the

layered-material half ran in full, and the measured half is deferred with the pipeline in place. Total compute for the entire suite, including the allocation experiment, is under one GPU-hour.

S6.2 Sentinel 1: veach-ajar

Six bucket-structure predictions were registered from XML inspection alone (one area light in the adjacent room; one smooth dielectric; four rough conductors): diffuse-indirect plurality; direct confined to the door-gap wedge; a small nonzero delta-mediated population; glossy present; emitter-direct ≈ 0 ; other = 0. All six hold (measured dominant populations: diffuse-indirect 0.988, delta-mediated 0.010, glossy 0.002, direct and emitter-direct below 10^{-3}). The registered operational prediction was that pilot-variance allocation would sit at or below uniform at $N_0=4$; the outcome far exceeds it (observed-var loses 6.8 dB; main-paper Table 13). The second half of the prediction — bucket-floor $\geq$ observed-var with the win-rate pattern of main-paper §7.3 — holds only as an exact tie (-0.004 dB, 5/10; the registered win-rate pattern did not materialise), because the scene’s heavy-bucket population is $\approx 1\%$. Under the $f=4$ control configuration, reported per a commitment registered before the number was computed, the shared sample floor recovers 3.2 of the 6.8 dB (observed-var 10.45 $\rightarrow$ 13.61; deficit $-6.79 \rightarrow -3.62$); about half the failure is starvation the floor repairs, while the residual -3.6 dB — still larger than any first-party deficit — persists in the dim-indirect Neyman-noise regime neither floor addresses. The label’s abstention diagnosis is unchanged in both configurations. The veach-ajar time-to-quality censoring measurement was registered but not executed and is recorded as deferred.

S6.3 Sentinel 2: veach-mis

Whole-image results are unremarkable in the best sense: agreement 0.990, predicted 0.990; other = 0%. The per-plate analysis is where the failures live. The original prediction that glossy would dominate the plates failed: the plate faces are dominated by diffuse-indirect (13,825 of 13,863 plate-face pixels; the image-wide direct = 0.83 is the diffuse surround, not the plates). What a rough plate mostly shows the camera is a glossy-mediated reflection of the diffuse surround — camera $\rightarrow$ plate (G) $\rightarrow$ diffuse end-vertex — which the end-vertex taxonomy routes to diffuse-indirect by design (Fig. S3); the plate’s own source highlights (single-bounce G-end-vertex NEE events, direct) are small ellipses; and the near-mirror plate’s accepted energy is further suppressed by the NEE-only acceptance rule. This reading corrects an earlier one of our own, caught by a cross-check during the follow-up direct-bucket analysis. The amendment registered three sharpened predictions: (i) plate-face glossy energy fraction 0.94–1.00 against ≈ 0.03 elsewhere — **pass**; (ii) per-plate agreement monotone in lobe width — **fail** (0.989/0.988/0.980/0.994 from narrowest to widest, $\hat{\alpha} \in [0.0007, 0.08]$; the narrow-lobe instability of the glossy *bucket* does not transfer to these diffuse-indirect-labelled pixels); (iii) restricted-MIS label changes concentrate on the narrowest plate — **fail** (4/1/8/4 flips across plate faces; 17 of the image’s 150 flips lie on them at all; consistent in retrospect, since the restricted completion touches only glossy and delta-mediated and the plate-face pixels are diffuse-indirect). Both failures are the predictor’s misreading of the taxonomy’s scope, not label instability — the labels moved nowhere unexpected — and are retained in the record as registered.

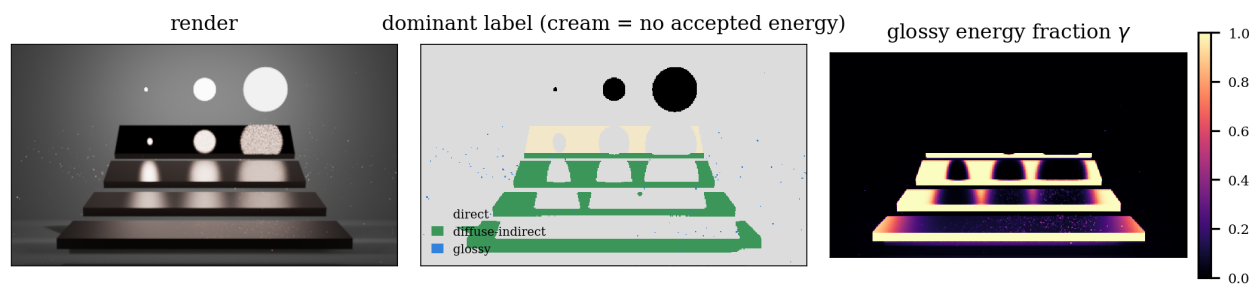


Figure S3: What the veach-mis plates are, per the corrected reading. The plate faces are diffuse-indirect (green) with glossy energy fraction $\gamma \approx 1$ (right): each face shows the camera a glossy-mediated reflection of the diffuse surround, which the end-vertex rule routes to the diffuse end vertex. The source highlights are small direct ellipses; the near-mirror top plate has almost no accepted energy under the NEE-only acceptance rule (cream). Camera-side glossy mediation is invisible to the dominant label by construction and carried by γ .

S6.4 Sentinel 3: bathroom2

The declared separation measures were computed on the scene first (mutual visibility 0.327 — higher than any tested scene — at 0.34 m, between the tabletop_mixed and kitchen_counter anchors), the sign prediction

$0 < \rho_{C-G} < +0.441$ was registered, and only then was the scene rendered: $\rho_{C-G} = +0.251$ — **pass**, sign and interval as registered (main-paper §6.1).

S6.5 Sentinel 4: matpreview

Three registered questions on the principled BSDF with clearcoat swept over $\{0, 0.25, 0.5, 1.0\}$ (a smooth coating layered over a rough base). Does the NEE end-vertex rule remain well-defined at every coat value? Yes: the BSDF advertises Diffuse and Glossy components throughout, so the D-if-any-diffuse-flag rule routes NEE events to **D** deterministically. Do clearcoat-bearing pixels land under D-routed labels and stay there? Yes: **direct** 0.88, **diffuse-indirect** 0.12; agreement 0.989 at every coat value; sweeping the coat from 0 to 1 moves 0.2% of pixels. Does **other** stay at zero? Yes, exactly. The glossy energy fraction γ responds monotonically to the sweep (mean $0.040 \rightarrow 0.047$, +16%; 90th percentile $0.166 \rightarrow 0.196$) while the dominant label holds still — the division of labour between discrete label and continuous mixture as designed.

S6.6 Scoreboard

Of the nineteen registered predictions scored, fifteen passed, one held as a boundary case (the **veach-ajar** tie, consistent with the descriptor’s intended scope), and three failed — the original plate-dominance prediction and two of the amendment’s sharpened successors, all three the predictor’s misuse of the taxonomy’s scope. Two further registered items are disclosed rather than scored: the superseded original wording of one plate prediction (its sharpened successor is scored in its place), and the **veach-ajar** time-to-quality censoring measurement, registered but not executed. The per-scene stability predictions (main-paper Table 5) were issued from each scene’s own pilot via the frozen margin/purity curves before the reference render existed, and land within 0.018 on **veach-ajar** and within 0.005 on the other two; the curves’ leave-one-scene-out error on the first-party scenes is at most 0.061 (purity) and 0.079 (margin), the worst case being **kitchen_counter**, the most mechanism-mixed scene. The registration document, its amendment, the frozen curves, every per-scene prediction and outcome file, and the variant-generation script are retained for inclusion in the public artifact (main-paper Appendix A).