A Stable Transport-Mechanism Descriptor for Per-Pixel Rendering Difficulty

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

Abstract

Per-pixel rendering difficulty is conventionally measured by the sample variance $\hat{\sigma}^2(p)$ of a Monte Carlo estimator, yet this signal is least reliable exactly where difficulty concentrates: under heavy-tailed transport its relative error is governed by the integrand's kurtosis, and the split-half reliability of variance-derived evaluation targets reaches only $\rho = 0.23$–0.29 even at $40{,}000$ samples per pixel. We propose a complementary *discrete transport-mechanism descriptor*: every contribution event is classified by its end-vertex BSDF lobe and the presence of a delta-specular event, yielding seven mutually exclusive labels whose six named mechanisms receive all observed energy on the tested scenes, with continuous side-channels retaining the per-pixel mechanism mixture. Across seven scenes, the dominant label agrees 87–99.6% between 64 and 4096 samples per pixel — where quantile-binned $\hat{\sigma}^2$ agrees as little as 21% — and is robust to restoring the estimator's MIS half. The descriptor exposes cross-scene structure that a scalar variance cannot represent, including a geometry-controlled sign reversal of the caustic–glossy correlation. Finally, we demonstrate operational value: using the label to correct a noisy pilot $\hat{\sigma}^2$ consistently improves on pilot-variance sample allocation at equal budget, exactly on the scenes where the pilot is least reliable — narrowing, though not yet closing, its gap to uniform sampling on the scenes where the raw pilot loses even to uniform — while acting only where heavy-tailed buckets exist, with gains that survive a random-partition placebo control and persist on top of a robust (median-of-means) pilot baseline.

Keywords Ray tracing · Rendering · Monte Carlo · Adaptive sampling · Light transport

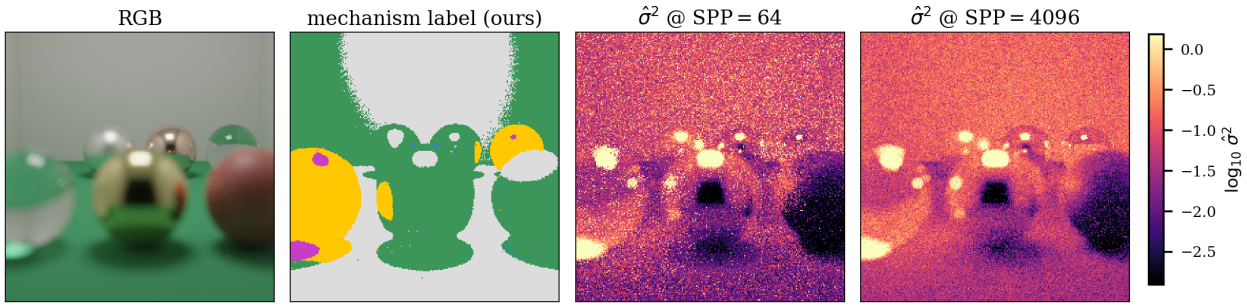

Figure 1: Our discrete mechanism label (second panel) is determined at SPP$= 64$ and barely changes at SPP$= 4096$, with the residual disagreement concentrated at the boundary between two light-tailed buckets (§4). The classical variance signal (right two panels) drifts substantially across the same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

1 Introduction

Per-pixel rendering difficulty has no agreed-upon reference signal. For thirty-five years the operational signal in adaptive Monte Carlo rendering has been one scalar — the per-pixel sample variance $\hat{\sigma}^2(p)$ of an unbiased pixel estimator — which adaptive samplers refine on, denoisers steer their filter footprint by, efficiency-aware Russian roulette branches off, and inverse-rendering pipelines now differentiate through. The default has hardened into a methodological assumption: *the per-pixel sample variance, estimated from a pilot render, is a reliable measurement of how hard each pixel is to render*. We argue this assumption fails under heavy-tailed light transport. The defect is not that $\hat{\sigma}^2$ is a little noisy; it is that at the budgets the community actually uses, finite-sample variance estimates are least reliable exactly on the pixels where difficulty concentrates. Our response is not a replacement scalar but a change of object: a discrete, deterministic *transport-mechanism descriptor* of each contribution event, which remains stable in the regime where the variance estimate fails, and which we show can be used to condition and repair variance-based signals there (§7).

The variance-based difficulty reference and its inheritors. Adaptive samplers since Mitchell [1] refine where a per-pixel, perceptually motivated contrast test fires, or in the modern formulation distribute samples to greedily minimise per-pixel relative mean-squared error [2]; real-time denoising steers a spatio-temporal filter footprint by a per-pixel variance buffer [3]; efficiency-aware Russian roulette iteratively estimates per-path second moments [4]; Bayesian direct illumination drives light-selection probabilities from online regression over per-light statistics [5]; and the Eurographics State-of-the-Art Report [6] surveys variance-driven allocation as the dominant family. We do not dispute these methods' utility; we dispute that the finite-sample quantity they all rely on can serve as a difficulty *reference* — a signal other methods are trained against, evaluated against, or steered by — in the heavy-tailed regime.

Problem 1: heavy-tail instability. $\hat{\sigma}^2$ is a sample-variance estimator, and such estimators are unstable on heavy-tailed integrands. The instability shows in the simplest measurement: rank-correlate per-pixel $\hat{\sigma}^2$ at a low pilot SPP against the same quantity at a much higher SPP. On caustic-dominated scenes we measure Spearman $\rho(\hat{\sigma}^2_{64},\, \hat{\sigma}^2_{4096})$ as low as 0.26 (§4). Prior work documents the same instability under the proxy of firefly outliers — density-based rejection [7], sample re-weighting [8] — but treats it as a removable nuisance at the *mean* estimator rather than a foundational defect of finite-sample variance as a reference signal.

Problem 2: kurtosis regress. The instability is a mathematical necessity, not an implementation defect. Three distinct quantities must be kept separate: the variance σ^2 of individual per-sample contributions (the population quantity we wish to know), the variance σ^2/N of the N-sample pixel mean (which is what adaptive samplers ultimately care about), and the sampling uncertainty of the *estimator* $\hat{\sigma}^2_N$ of σ^2. It is the third that concerns us here: under finite-fourth-moment assumptions, the relative standard deviation of the sample-variance estimator satisfies $\sqrt{\mathrm{Var}(\hat{\sigma}^2_N)}\,/\,\sigma^2 \approx \sqrt{(\kappa-1)/N}$, where κ is the integrand's fourth standardised moment [9, 10]. Three regimes follow. When κ is finite but large — caustic and near-specular integrands — $\hat{\sigma}^2_N$ converges at a rate that makes pilot-budget estimates useless in practice. When $\kappa = \infty$ with σ^2 finite, $\hat{\sigma}^2_N$ remains consistent but admits no finite-rate Gaussian limit [11], so no error bar of the usual form exists. When σ^2 itself is infinite — unbounded contributions near delta paths [12] — the target of the estimate does not exist. The natural workaround — estimate κ from a pilot, compute an error bar, budget accordingly — fails by infinite regress: estimating κ requires the eighth moment, which converges more slowly still. The rendering literature has long known this pathology under other names ("weak singularities" [13], infinite Hardy–Krause variation [14]), but to our knowledge the implications for $\hat{\sigma}^2$ *as a reference signal* — rather than as a nuisance at the mean estimator — have not been examined.

Problem 3: low reliability of the reference itself. Any reference derived from a finite-sample MC estimator inherits an unmeasured reliability limit. Concretely: we fit a per-pixel convergence exponent $\hat{\alpha}(p)$ from a four-level SPP sweep with ten seeds per level ($\sim$40,000 samples per pixel), then split the seeds in half and re-fit on each. The Spearman correlation of the two halves — 1.0 for a noiseless reference — is only 0.23–0.29 even on the two geometrically simplest scenes of our test set, a Lambertian arrangement and a single-sphere Cornell variant (§4). In the language of classical measurement theory this is a split-half *reliability* measurement [15, 16]: it demonstrates that two independent measurements of the same target barely agree, and it attenuates any observed correlation between the target and a predictor. We report it as an empirical reliability figure; translating reliability into a formal ceiling on predictor performance requires additional measurement-error assumptions (classically, a square-root attenuation relation [15]) that we do not need for our argument — low reliability of the evaluation target is damaging under any of them. Recent deep adaptive-sampling work acknowledges that variance and contrast cannot be reliably computed at low sample counts [17], but treats this as a problem of better estimation rather than a defect of the signal itself; robust estimators of mean and variance under heavy tails exist [18, 19, 20] and qualify our critique — we return to them in §2.1.

Problem 4: structural property versus measured signal. One might object that $\sigma^2(x, y)$ is, after all, a stable structural property of the scene — recent work differentiates it with respect to scene parameters [21] and builds sampling maps on refined variance estimates [22]. The objection is correct in one sense and not in another. Where σ^2_∞ exists, the σ^2-field is a deterministic functional of the integrand: the stored buffer $\hat{v}_N(p) = \hat{\sigma}^2_N(p)/N$ (Problem-2 notation) 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. What is *not* stable is the per-pixel *rank* in a regime that includes heavy-tailed pixels: Problem 1's $\rho = 0.26$ measures a rank instability that survives the $1/N$ rescaling, because heavy-tailed pixels do not contract uniformly with N. Hence the narrow form of our objection: a difficulty reference must rank pixels consistently across the *entire* scene, including exactly the heavy-tailed pixels where σ^2_∞ fails to exist or is recoverable only through fourth-moment-limited estimators — precisely where the hardest pixels live.

Our position. Three notions must be distinguished, because the argument of this paper connects them without equating them. *Rendering difficulty* is an operational quantity and admits several non-equivalent definitions — variance under a fixed estimator, samples-to-error-threshold, wall-clock time to quality; we measure the first and third explicitly (§4, §6) and claim no single scalar ordering. *Transport mechanism* is a structural property: how a contribution's energy reaches the sensor. A mechanism class does not by itself order pixels by cost — our own results show substantial cost heterogeneity within classes (§6) — so it is not a difficulty measure; it is an *explanatory and conditioning variable* for one. *Stability* is the property that makes a signal usable as a reference; stability alone proves neither of the other two. Our claim, in this vocabulary: the community conditions on an unstable measurement ($\hat{\sigma}^2$) when a stable structural descriptor of the same pixels is available for free. We therefore propose a descriptor that is (i) a deterministic function of the light-transport integrand and the stated estimator conventions (§3.5), (ii) discrete — so the sampling noise of continuous estimators does not propagate into it — and (iii) defined identically across scenes. The candidate is the *transport mechanism of each contribution event*: a small label set defined by the BSDF lobe at the contribution's end-vertex and whether the path sequence to that vertex passes through a delta (specular) interface. The seven labels that result from this two-axis partition (Section 3) are mutually exclusive and formally complete by construction — a catch-all row absorbs anything outside the named cells, and receives no energy on any tested scene — are assigned per contribution event from quantities the path tracer already maintains, and are — empirically — stable across sample budgets where $\hat{\sigma}^2$ is not (Fig. 1). Its operational value as a conditioning signal is demonstrated directly in §7.

Contributions. This paper makes five contributions.

1. **A methodological argument** that finite-sample variance-based per-pixel references are unreliable under heavy-tailed transport, supported by the kurtosis bound on the variance estimator (Problem 2), the empirical heavy-tail instability (Problem 1), and the split-half reliability measurement (Problem 3) above (§1).
2. **A mechanism descriptor** consisting of seven mutually exclusive labels assigned per contribution event, derived from a two-axis partition (end-vertex lobe, presence of delta-S) with no continuous threshold parameter on the main axis, together with a formal separation between the deterministic per-event label, the ideal population-level per-pixel label, and its finite-sample estimator (§3).
3. **Validation on seven scenes:** (a) coverage — 0% of contribution-event energy falls outside the six named mechanisms on every scene and budget tested; (b) cross-SPP stability — the dominant label agrees at 87.1–99.6% between SPP=64 and SPP=4096, against 21–34% for 7-bin quantile-discretised $\hat{\sigma}^2$ on the same seven scenes; disagreements are dominated by two energy-argmax boundaries whose identity the analysis pinpoints, with structural-bucket pixel shares drifting by at most 2.9%; and (c) robustness to a controlled estimator perturbation — completing the estimator with MIS leaves the argmax unchanged on 98.5–100% of pixels, and a population-scale glossy bucket (10,137 pixels) is 95.8% stable where the 83-pixel narrow-lobe glossy population of the original test scene was not (§4).
4. **A cross-scene correlation analysis** of the mechanism-derived descriptor vector on seven scenes, yielding falsifiable structural findings: (a) the correlation between delta-mediated and glossy contributions reverses sign with object-space separation, tested on a controlled scene pair that differs only by a blocking partition and quantified by a declared mutual-visibility measure (§5.1); (b) the correlation between bounce depth and mechanism purity separates specular-chain-dominated scenes from lobe-mixing-dominated ones by sign (§5.2); and (c) a controlled roughness sweep shows end-vertex lobe width monotonically controls the firefly intensity of the glossy bucket (§5.4) — the regime in which the time-to-quality analysis (§6) finds continuous cost right-censored at the budget cap while the discrete labels still discriminate.
5. **A downstream demonstration** (§7): conditioning a noisy pilot $\hat{\sigma}^2$ on the mechanism label improves on pilot-variance sample allocation at equal budget on every scene with heavy-tailed buckets, with 7–$10/10$ per-pilot consistency — narrowing, though not closing, the gap to uniform sampling on the scenes where we

show the raw pilot losing even to *uniform* — while reducing exactly to observed-var on the light-tailed control, and with gains a random-partition placebo cannot reproduce.

Roadmap. Section 2 surveys the variance-as-reference camp, the covariance / frequency a-priori prediction camp, and prior uses of path mechanism inside (rather than as the target of) Monte Carlo estimators. Section 3 develops the mechanism descriptor and states its scope. Section 4 reports coverage, cross-SPP stability, and estimator-perturbation robustness. Section 5 reports the cross-scene correlation structure of the descriptor vector and its sign-reversal findings. Section 6 presents the time-to-quality analysis. Section 7 demonstrates the descriptor's operational value for pilot-variance correction in equal-budget sample allocation. Section 8 states limitations and scope; reproducibility details are collected in the appendix.

2 Related Work

The methodological context for our claim falls into three camps, each producing or relying on a per-pixel quantity that has been treated — implicitly or explicitly — as a difficulty signal. The first uses a sample-statistic-based estimate of variance; the second traces an analytically propagated continuous descriptor of the radiance signal; the third already uses path-transport mechanism, but exclusively inside the estimator or for compositing. We discuss each in turn.

2.1 Variance as the de facto difficulty signal

Three and a half decades of adaptive Monte Carlo rendering use sample statistics as the operational signal: image-space adaptive sampling refines where a perceptually motivated contrast test fires [1] or greedily minimises per-pixel relative MSE [2]; variance-guided denoising builds its filter footprint from a variance buffer [3]; efficiency-aware path tracing iterates over second-moment estimates [4]; Bayesian direct illumination regresses over per-light statistics [5]; and the Eurographics STAR [6] surveys this family as the dominant paradigm. Recent inverse-rendering work derives unbiased estimators for the *derivatives* of variance with respect to scene parameters [21], and denoising-aware adaptive sampling drives its sampling map from a variance estimate of the denoiser's output [22] — both extending the variance-family program rather than abandoning it.

Heavy-tailed instability is known but not foundationalised. The community has noticed that $\hat{\sigma}^2$ behaves badly under firefly-prone transport: density-based outlier rejection [7] and firefly re-weighting [8] both target the fingerprint of an unstable estimator, and deep adaptive sampling states that variance and contrast cannot be reliably computed at low sample counts [17]. In each case the instability is suppressed at the *mean* estimator, and its implications for $\hat{\sigma}^2$ *as a reference signal* remain unexamined.

Robust estimation qualifies, but does not void, the critique. The statistics literature provides estimators of mean and variance with sub-Gaussian deviation guarantees under heavy tails — Catoni's deviation study for the empirical mean and variance [18], sub-Gaussian mean estimators [19], and the median-of-means family surveyed by Lugosi and Mendelson [20] — and convergence diagnostics or stopping rules built on such estimators need only the moments their deviation bounds assume (finite variance for the mean), rather than the finite fourth moment that error bars on $\hat{\sigma}^2$ require. Within graphics, recent work corrects estimated radiance and variance using statistical models of the sample population [23]. These methods weaken the practical force of Problem 1 wherever their moment assumptions hold: a variance-derived target need not be built from the naive sample variance. They do not, however, remove Problem 2's regime structure — when the fourth moment is unbounded, guarantees for variance estimation degrade regardless of the estimator — nor do they supply the mechanism-level structure our descriptor exposes (§5). We view robust variance estimation and mechanism conditioning as complementary, and §7 (Result 3) now tests that relationship directly: a median-of-means pilot variance enters the allocation experiment as a second incumbent. It repairs the light-tailed part of pilot noise at the larger budgets, but it purchases that stability by discarding exactly the rare-tail evidence that delta-mediated pixels carry — and the mechanism label identifies those pixels a priori and restores them, with gains a matched placebo does not reproduce.

Per-pixel guidance maps produced by denoising pipelines are a related practical reference: kernel-predicting networks and their successors consume noisy auxiliary buffers — including variance — and produce per-pixel filtering decisions [24, 25]. These maps inherit the reliability of their variance inputs at low sample counts; our split-half methodology (§4) applies to them directly, though quantifying their consistency is beyond our scope.

Our distinction. All these methods build on $\hat{\sigma}^2$ as an estimated continuous quantity and inherit its sampling behaviour. We complement it with a deterministic discrete label, read from path structure rather than estimated from samples.

2.2 A-priori prediction via traced continuous descriptors

A second line — closer in spirit, forgoing sample variance to predict difficulty analytically — traces a continuous descriptor of the radiance signal through the transport operators. The frequency-analysis framework [26] treats radiance as a band-limited signal in space–angle frequency, occlusion as Fourier convolution by the blocker spectrum, free-space propagation as a shear. Later work specialised it to depth-of-field sampling via local lens bandwidth [27] and generalised it to a 5×5 covariance matrix propagated through transport, BRDF, occlusion, lens, and motion [28] — the closest prior work to ours by goal, both targeting an a-priori per-pixel prediction without sample variance. A non-Fourier antecedent expresses the same philosophy through first-order sensitivities of lighting, shading, and shadow [29].

Our distinction. The object of prediction here — a covariance matrix, spectrum, or gradient field — is itself continuous: it must be traced, propagated under Gaussian-family assumptions, and approximated as it accumulates. Our mechanism label is a deterministic function of path structure, read from a single contribution event, not traced or estimated.

2.3 Path mechanism as an estimator-internal construct

Path-transport mechanism is not new in light-transport research; it has been load-bearing for decades, but only inside the estimator or as a compositing convenience. Heckbert's `L(D|S)*E` regex [30] described paths and motivated a hybrid algorithm tied to particular path classes (the glossy state `G` was a later production addition); Veach [12] formalised the regex as a classifier of which algorithms can sample which configurations. Photon mapping separates a caustic from a global map for density estimation [31] — an estimator structure, not a per-pixel label. Path guiding learns incident-radiance proposals from a particle stream, as a parametric mixture [32] or an SD-tree [33]; the neural variant trains a normalising-flow sampler on KL or χ^2 divergence [34] — a loss internal to sampler training, driving a sampler toward the target density, the opposite direction from our use of a mechanism label as the predicted target. Manifold-based samplers target specific mechanisms: manifold next-event estimation connects shading points to lights across refractive interfaces to render refractive caustics [35], and specular manifold sampling generalises the approach to high-frequency caustics and glints [36]; there the mechanism is what the sampler targets, not how the pixel is labelled. Path-space regularization makes the connection between mechanism and difficulty explicit from the estimator side: unsampleable or high-variance path classes are selectively mollified, either by path-space criteria [37] or by widening microfacet lobes [38] — direct evidence that particular mechanisms are intrinsic trouble spots for particular estimators, though neither work produces a per-pixel label, and estimator-specific trouble does not by itself imply an estimator-independent ordering of pixel difficulty. Variance-aware MIS injects variance into the balance heuristic [39] — the closest prior work combining mechanism and variance, though at sampler-family granularity and for sampler weighting, not per-pixel difficulty.

Production light-path expressions. The closest analog to our seven-bucket taxonomy is the light-path-expression facility of production renderers [40]: a regex over `L/D/G/S/E` selecting events into named AOVs. But the expressions are user-defined, not a closed exhaustive taxonomy; framed as compositing conveniences, not a difficulty signal; and overlap is permitted, so the exhaustive-and-mutually-exclusive property our taxonomy rests on (§3) is not part of their design.

Our distinction. The path-mechanism abstraction has appeared in every role — Heckbert's notation, Veach's classifier, path-guiding proposals, manifold samplers, regularization criteria, production light-path expressions — except as a per-pixel descriptor and conditioning signal for rendering difficulty. That is the role we propose.

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 stated in §3.5; nothing at the event level is estimated, and nothing on the main axis is thresholded. The *per-pixel* label aggregates events by an energy-weighted argmax and is therefore itself a Monte Carlo estimate at finite sample counts — a distinction we make precise in §3.1 and quantify throughout §4. We give the construction in three pieces — the unit of analysis (§3.1), the two-axis partition into seven buckets (§3.2), and the continuous sidefields that retain the per-pixel mechanism mixture (§3.3) — then motivate three design choices (§3.4) and state the descriptor's scope and conventions (§3.5).

3.1 The contribution event as the unit of analysis

A path-traced estimator at pixel p accumulates radiance from many sources, but each contribution to the pixel value reaches the sensor through a specific sequence of vertices and BSDF events. We call each such accumulation event a *contribution event*. In a standard path tracer with next-event estimation, contribution events arise in three forms:

- a successful NEE shadow ray connecting the camera-traced subpath to an emitter,
- a BSDF-sampled bounce that lands directly on an emitter, or
- the trivial case in which the camera primary ray intersects an emitter with no intervening surface bounce.

For each contribution event we observe two quantities: (a) the BSDF lobe at the *end-vertex* V_n, the last surface vertex before the emitter, and (b) whether the light-to-end-vertex sequence contains any delta-specular interaction. Both are computed from lobe types already sampled during path construction; no additional rays are cast.

End-vertex lobe under NEE. The two NEE-path-tracer branches need different end-vertex rules, one requiring a convention. For a BSDF-sampled bounce landing on an emitter, V_n is the last surface vertex before the emitter, and its lobe is read from the BSDF's reported sampled lobe. For an NEE shadow-ray connection, V_n is the same vertex but the BSDF is only *evaluated*, not sampled, so we assign V_n 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 labelled **D**, conservative in the sense of §3.2: it routes the event to the light-tailed side rather than the long-tailed glossy bucket. Delta-specular surfaces do not participate in NEE (zero connection probability), so an NEE end-vertex is never **S**; consequently the `specular-direct` bucket (Table 1) receives only BSDF-sampling contributions.

Population label versus finite-sample label. Let $L_b(p)$ denote the population path-space contribution to pixel p restricted to bucket b — the integral of the measurement contribution function over the paths whose contribution events classify to b under the rules above. The *ideal dominant label* is

$$b^*(p) \;=\; \arg\max_b \; L_b(p),$$

a deterministic property of the integrand and the stated conventions. A renderer at budget N produces Monte Carlo estimates $\hat{L}_b^{(N)}(p)$ of the bucket energies, and the reported per-pixel label is the finite-sample argmax $\hat{b}_N(p) = \arg\max_b \hat{L}_b^{(N)}(p)$. Thus: the classification of an individual event is deterministic; the per-pixel label is an *estimator* of $b^*(p)$ and is, at finite SPP, sample-, estimator-, and implementation-dependent. The stability experiments of §4 are properly read as evidence that this estimator recovers $b^*(p)$ reliably at small budgets in the tested settings — not that the finite-sample label is free of estimation. The practical cost of estimating $\hat{b}_N$ from a small pilot is measured directly in §7, where it is found to be negligible.

This construction differs from prior uses of path-mechanism information (§2.3) in two ways. First, the unit is a single contribution event, not a whole path: one path-tracer step may contribute through both NEE and BSDF branches, which we label separately. Second, the label is itself the per-pixel descriptor — not consumed inside an estimator (photon-map split, path-guiding proposal, manifold-sampler target) nor user-defined for compositing (light-path expression).

3.2 The two-axis partition

Each contribution event maps into one of seven buckets through a partition along two axes (Table 1). The first axis is the end-vertex BSDF lobe, which takes one of three values: **D** (diffuse), **G** (glossy, non-delta), or **S** (delta specular). The second axis is binary — whether the light-to-end-vertex sequence contains any delta-S event — and a seventh row covers the trivial primary-ray-into-emitter case, disjoint from all surface-bounce cases.

The partition is formally exhaustive by construction — the `other` row catches any event outside the enumerated cells — and mutually exclusive: every contribution event has a unique (end-vertex lobe, δ-S presence) pair, and the trivial primary-ray case is disjoint from all surface-bounce cases. What formal exhaustiveness does not guarantee is that the six *named* mechanisms cover the transport that actually occurs; §4 measures this, finding that `other` receives 0% of contribution-event energy on every scene and budget of our seven-scene test matrix. This establishes coverage for the scene class tested (§3.5), not universal completeness.

The seven labels are not arbitrary categories: they enumerate a structural partition along two well-defined axes, the structure preceding the labels. The partition admits no continuous threshold on the main axis — every cell boundary is a discrete predicate on a lobe type or sequence property.

Table 1: The seven contribution-event buckets, defined by end-vertex BSDF lobe and presence of a delta-specular event in the light-to-end-vertex sequence. The `delta-mediated` bucket carries the key `caustic` in the released code; we avoid that name in the text because the bucket marks a *transport class* — energy that passed through a delta interface — whose bright image-space caustic footprint is only a subset, resolved by the continuous `caustic_frac` sidefield (§3.3, §4.1).

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`

Classification order (Algorithm 1). Operationally, Table 1 is a first-match cascade. Per contribution event, with end-vertex lobe ℓ, delta flag δ (any delta-specular interaction in the light-to-end-vertex sequence), and surface-bounce count n:

1. if the camera primary ray reaches the emitter with no surface bounce ($n = 0$): `emitter-direct`;
2. else if $\ell = \mathrm{S}$: `specular-direct`;
3. else if $\ell \notin \{\mathrm{D}, \mathrm{G}\}$: `other`;
4. else if δ: `delta-mediated`;
5. else if $n = 1$: `direct`;
6. else if $\ell = \mathrm{D}$: `diffuse-indirect`;
7. else ($\ell = \mathrm{G}$): `glossy`.

Here ℓ is read by the end-vertex rules of §3.1 (sampled lobe on the BSDF branch; D-if-any-diffuse-flag, else G, on the NEE branch), and the cascade matches the released integrator's classification code line for line.

3.3 Purity and energy-fraction sidefields

The dominant-class field captures only the energy-argmax bucket. To retain the per-pixel mechanism mixture — and support the cross-scene analysis of §5 — we add two continuous sidefields:

- *purity* $\pi(p) \in [0, 1]$: the energy fraction of the dominant bucket. $\pi = 1$ marks a mechanically pure pixel; lower values indicate mixing.
- *glossy energy fraction* $\gamma(p) \in [0, 1]$: the energy share of events whose light-to-end-vertex sequence contains any glossy bounce, regardless of dominant class.
- *per-bucket energy fractions* $\hat{L}_b(p) / \sum_{b'} \hat{L}_{b'}(p)$: the full mechanism mixture underlying the argmax. In particular `caustic_frac` — the `delta-mediated` share (the name records the released-code key) — resolves, within the delta-mediated class, the bright visual caustic footprint (high share) from pixels that merely receive some delta-mediated energy (§4.1).

A derived mixed-flag $m(p) = [\pi(p) < \theta]$ ($\theta = 0.7$) marks low-purity pixels for downstream consumers wanting a binary signal. The threshold θ is post-hoc and does not enter dominant class: any continuous threshold in our framework is confined to sidefields, never the main partition.

3.4 Three design choices

Why discrete. A discrete label absorbs small pipeline perturbations — and we are explicit that this absorption is a *design property*, not by itself evidence that the label measures difficulty: any coarse quantisation is more stable than the continuous signal it coarsens. Two things elevate the property beyond quantisation. First, the fair comparison: §4.5 discretises $\hat{\sigma}^2$ onto quantile bins of matching cardinality and scores it with the identical agreement metric, and the mechanism label remains 1.4–$4.5\times$ more stable, depending on scene and bin granularity. Second, the placebo test: §7 shows that a random partition with the same class sizes does not reproduce the label's operational value — what matters is *which* pixels share a class, not that classes exist. Empirically, dominant class agrees at 87.1–99.6%

between SPP$= 64$ pilot and SPP$= 4096$ reference renders across the seven scenes, with disagreement dominated by two identifiable energy-argmax boundaries (`direct`↔`diffuse-indirect`, and `diffuse-indirect`↔`glossy` on the glossy-rich scene) and structural-bucket shares stable to within 2.9% (§4.3).

Why γ is continuous, not a boolean flag. A simpler design for the soft-caustic sidefield is a single bit: "does any contribution to this pixel pass through a glossy interaction?" We tested it on the three original scenes: the boolean `has_glossy_in_path` saturates on each (0.0, 0.0, 1.0 at SPP=4096 on `glass_caustic`, `cubes_dof`, `snooker`), with no intermediate value separating "occasionally hits a glossy surface" from "glossy energy dominates" — and it would saturate at 1.0 on the glossy-rich `kitchen_counter` by construction. The continuous fraction $\gamma(p) = L_{\text{has-glossy}}(p)/L_{\text{total}}(p)$ recovers this distinction without a threshold, and the path tracer already accumulates the glossy-touching stream for free: the boolean saved no work and lost the distribution. Booleanising at any threshold would also re-introduce a continuous parameter into a sidefield — exactly what the main axis avoids (§3.2). Continuous storage, threshold-at-use-site.

Why glossy stays its own bucket. One might fold glossy specular focusing — informal "soft caustics" — into the `delta-mediated` bucket. We resist this for two reasons. First, glossy-vs-specular is a continuous roughness spectrum, not a discrete transition; a roughness-threshold rule would re-introduce a continuous parameter on the main axis, violating the principle above. Second, the empirical delta-mediated–glossy relationship has no fixed sign: it reverses with object-space separation between specular and glossy objects across our tested configurations (§5). Any rule pre-committing glossy to the delta-mediated bucket, or to diffuse-indirect, would mislabel at least one observed regime. We therefore keep glossy separate and preserve $\gamma(p)$ as the continuous handle through which the interaction is recoverable downstream.

3.5 Scope and conventions

The descriptor is *not* a property of the light-transport integrand alone; it is a property of the integrand together with a set of stated conventions, and we scope our claims accordingly.

Convention dependence. The NEE branch has no sampled terminal lobe, so the end-vertex lobe is read from advertised BSDF component flags, with multi-lobe materials advertising a diffuse component routed to **D** (§3.1). This rule — and the D/G/S trichotomy itself — depends on the renderer's material model and flag conventions; a different BSDF decomposition, roughness parameterisation, or layered-material factoring may label the same physical scattering function differently. The appendix specifies the complete mapping for every material we use. Rather than assert insensitivity, we measure it where it matters most: §4.4 perturbs the estimator itself and reports exactly which pixels move.

Domain scope. All claims concern surface path tracing under the D/G/S decomposition above, on scenes composed of diffuse, rough- and smooth-conductor, rough- and smooth-dielectric, and plastic-family materials. We make no claims for participating media or subsurface transport, stochastic geometry and visibility complexity (hair, foliage), glinty or measured materials whose diffuse/glossy decomposition is not unique, environment emitters, or null-scattering events. Several of these carry substantial difficulty regimes of their own — hair and volumes prominently — and extending the taxonomy there, e.g. with phase-function lobes as an additional axis, is future work (§8).

4 Validation: Exhaustiveness and Cross-SPP Stability

This section reports three empirical results on the same data: (i) coverage — the `other` catch-all takes zero energy on every scene, at every sample budget tested; (ii) the discrete mechanism labels are stable across an order-of-magnitude sample-budget change in a regime where finite-sample variance-derived references are not; and (iii) the labels are robust to a controlled perturbation of the estimator itself. The contrast between (ii) and the variance baseline (§4.5) is the empirical core of our claim that mechanism is a stable object on which to condition a per-pixel difficulty signal.

4.1 Test matrix

Seven scenes span the difficulty regime of interest:

- `cubes_dof` — three coloured cubes on a ground plane viewed through a thin-lens camera; Lambertian materials only, no specular or glossy surfaces. A reference scene where the partition is expected to be dominated by `direct` and `diffuse-indirect`.

- `glass_caustic` — a Cornell-box variant with a glass sphere casting an L-S-D caustic on the floor; the prototypical heavy-tailed transport scene.
- `snooker` — six balls of mixed materials (matte plastic, polished gold/silver/copper, glass) on a felt table; mixed glossy and specular surfaces, with inter-ball reflections producing a broad range of mechanism mixtures.
- `tabletop_mixed` / `tabletop_separated` — a controlled scene pair sharing geometry and materials, the second adding an opaque partition that blocks delta-S $\leftrightarrow$ glossy sight lines (§5.1).
- `ring_caustic` — two polished metal rings on a diffuse table under a small bright light: a lying ring whose inner wall focuses the classic cardioid catacaustic, and a standing tilted ring projecting a caustic streak. Every caustic path is L S^+ D E, so the scene provides an unambiguous, visually verifiable delta-mediated population ($6{,}949$ pixels at the SPP=4096 reference) — and demonstrates that the `delta-mediated` *bucket* (whole neighbourhood of the rings) and the bright visual caustic *footprint* (its high-`caustic_frac` subset) are distinct, resolvable notions.
- `kitchen_counter` — a counter-top scene with $\sim$15 objects spanning diffuse, rough-plastic, brushed and polished conductors at several roughnesses, smooth glass, and a frosted (rough-dielectric) tumbler, lit by two lights. Metal objects clustered before a brushed-steel backsplash make final-glossy transport dominant over large regions, giving the `glossy` bucket a population of $10{,}137$ pixels — two orders of magnitude beyond `snooker`'s 83 — and the scene the hardest cost profile in the set (§6).

We render each scene at three sample budgets, SPP $\in \{64, 512, 4096\}$, for 21 (scene$\times$SPP) runs total. The SPP=4096 render serves as the reference; the partition is computed at each SPP independently. Our implementation is a Mitsuba 3 custom `SamplingIntegrator` that, for every contribution event, records the lobe sampled at V_n and a single bit indicating whether any earlier vertex along the light-to-end-vertex sequence used a delta-specular lobe. These two quantities determine the bucket (Table 1) without any additional ray casting. All quantitative results in this paper were produced with a single renderer version on a single machine (appendix); re-rendering the original test matrix from scratch reproduced every previously reported statistic to within seed noise.

4.2 The named mechanisms cover the tested transport

The `other` catch-all takes 0% of contribution-event energy on all 21 runs: on the scenes tested, the six named surface mechanisms account for all observed energy. (As noted in §3.2, 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 distributions agree with physical expectation (Fig. 2): `cubes_dof` is 90.9% `direct`, with the remainder in `diffuse-indirect`; `glass_caustic` places 7.4% of pixels (10.0% of energy) into `delta-mediated`, localised on the floor below the glass sphere; `snooker` sorts its glass and chrome balls into `delta-mediated` and its gold ball into `glossy`, with the felt as `direct`; `ring_caustic` concentrates `delta-mediated` on and inside the rings; and `kitchen_counter` assigns its metal cluster and backsplash reflections to `glossy`. Such physical agreement is a sanity check on the partition's usefulness, not a proof that this is the uniquely correct 7-way grouping.

4.3 Cross-SPP stability

We measure cross-SPP stability in two ways: per-pixel dominant-class agreement across SPP levels, and per-scene bucket-fraction drift (Table 2).

Table 2: Cross-SPP stability of `dominant_class`. "agree" is the percentage of pixels whose dominant class matches between the two SPP levels. "max drift" is the largest change in any bucket's pixel share between SPP$= 64$ and SPP$= 4096$.

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%

The headline observation is that the label's cross-SPP agreement survives chance-correction on every scene. Raw agreement could in principle be inflated by a dominant majority bucket; it is not. Cohen's κ stays 0.79–0.98 across

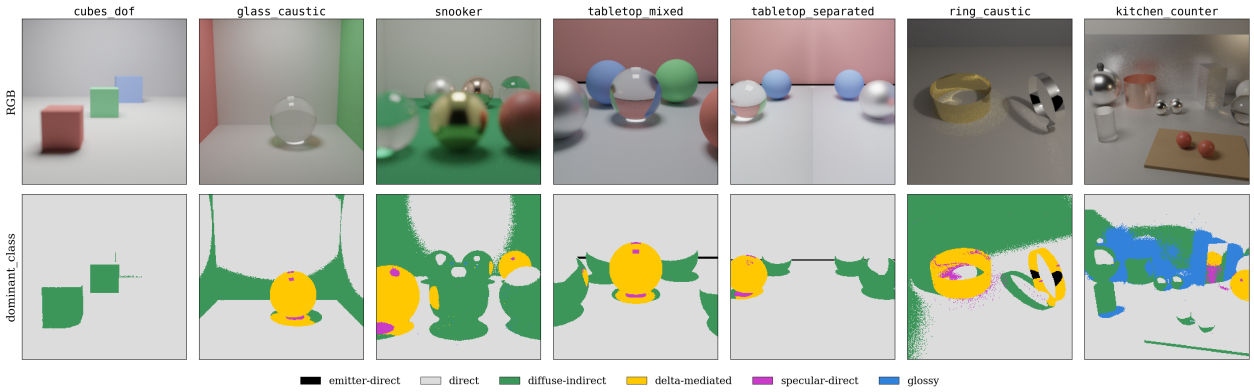

Figure 2: Per-scene RGB renders (top) and `dominant_class` maps (bottom) for the seven test scenes. `cubes_dof` is 90.9% `direct` (Lambertian only); `glass_caustic` places 7.4% of pixels into `delta-mediated` on the floor below the glass sphere; `snooker` sorts its glass and chrome balls into `delta-mediated` and its gold ball into `glossy`; the tabletop pair differs only by the blocking partition (§5.1); `ring_caustic` concentrates `delta-mediated` on and inside the polished rings, whose lying-ring cardioid is directly visible in the RGB render; and `kitchen_counter`'s metal cluster and brushed backsplash form the 10,137-pixel `glossy` population that powers the population-scale analyses of §4.4 and §6.

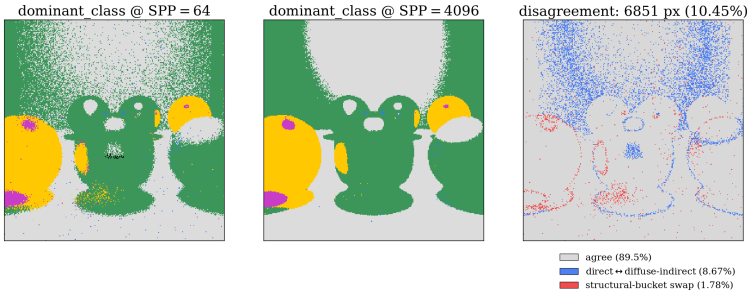

Figure 3: Cross-SPP stability of `dominant_class` on `snooker`. Left/middle: `dominant_class` at SPP$=64$ and SPP$=4096$. Right: per-pixel agreement map (grey = agree, blue = `direct`↔`diffuse-indirect` swap, red = swap to or from a structural bucket `delta-mediated/specular-direct/glossy`). 89.55% of pixels agree across the $64\times$ budget change; the 10.45% disagreement is dominated by the light-tailed boundary (blue, 8.67%), with only 1.78% of pixels participating in any swap to or from a structural bucket. On this scene, no bucket's pixel share moves by more than 0.58% (Table 2).

Table 3: Same-scale comparison: cross-SPP (64 vs 4096) agreement of the discretised variance signal versus the mechanism label. $\hat{\sigma}^2$ is binned into quantile bins (3 and 7 bins; bin edges are data-driven per-scene quantiles computed independently at each SPP level, which removes the global $1/N$ rescaling — the reading most favourable to variance), then scored by the same agreement metric used for the mechanism label. On every scene the mechanism label is 1.4–$2.3\times$ more stable than 3-bin $\hat{\sigma}^2$ and 2.5–$4.5\times$ more stable than 7-bin.

scene	3-bin $\hat{\sigma}^2$	7-bin $\hat{\sigma}^2$	mechanism
`cubes_dof`	0.484	0.264	0.996
`tabletop_separated`	0.652	0.335	0.994
`tabletop_mixed`	0.596	0.286	0.987
`glass_caustic`	0.424	0.212	0.956
`snooker`	0.560	0.307	0.896
`kitchen_counter`	0.611	0.344	0.887
`ring_caustic`	0.524	0.265	0.871

all seven scenes (Table 4) — "substantial" to "almost perfect" on the Landis–Koch verbal scale [43], which we use informally.

The structure of the residual disagreement is itself informative, and separating it correctly requires distinguishing two kinds of label boundary. *Predicate boundaries* are decided by a discrete test — membership of `delta-mediated` and

Table 4: Chance-corrected cross-SPP (64 vs 4096) agreement of the mechanism label. Cohen's κ [41] removes majority-class agreement; weighted κ [42] uses agreement weights $1-$penalty with penalty 0.2 for swaps within the light-tailed family {`emitter-direct`, `direct`, `diffuse-indirect`} and 1.0 for any swap touching {`delta-mediated`, `specular-direct`, `glossy`}. The mechanism classes are not ordinal; this weighting is declared as part of our reading of the taxonomy (boundary flicker versus structural change), and raw and unweighted κ are reported alongside. "struct%" is the fraction of disagreeing pixels involved in a swap touching the second family.

scene	raw	Cohen κ	weighted κ	struct%
`cubes_dof`	0.996	0.975	0.975	0.0%
`tabletop_separated`	0.994	0.974	0.972	47.2%
`tabletop_mixed`	0.987	0.964	0.953	70.3%
`glass_caustic`	0.956	0.890	0.900	32.9%
`snooker`	0.896	0.828	0.885	17.0%
`kitchen_counter`	0.887	0.789	0.718	81.3%
`ring_caustic`	0.871	0.786	0.859	14.3%

`specular-direct` turns on the presence of a δ-S event. *Energy-argmax boundaries* are decided by which bucket carries the larger energy share — `direct` versus `diffuse-indirect` (bounce-count energy split) and `diffuse-indirect` versus `glossy` (lobe energy split). Across all seven scenes, disagreement concentrates overwhelmingly on the energy-argmax boundaries and spares the predicate buckets. On the delta-heavy scenes the dominant flicker is `direct`↔`diffuse-indirect` (Fig. 3), and weighted κ — which forgives exactly that swap — rises above Cohen's κ ($0.828 \rightarrow 0.885$ on `snooker`; $0.786 \rightarrow 0.859$ on `ring_caustic`). On the glossy-rich `kitchen_counter` the dominant flicker is instead `diffuse-indirect`↔`glossy` ($2{,}575$ of $7{,}412$ disagreeing pixels at SPP=64), which our weighting matrix scores as structural — hence the one weighted κ that drops ($0.789 \rightarrow 0.718$) and the 81% struct share. This is the glossy *energy* boundary, the same boundary the estimator-perturbation analysis isolates below (§4.4), not a failure of a predicate bucket: `kitchen_counter`'s `delta-mediated` share drifts by only 0.01% across the $64\times$ budget change, while its `diffuse-indirect`/`glossy` shares exchange up to 2.9% of pixels. Across all scenes, structural-bucket pixel shares drift by at most 2.9% — the bound is set by the glossy energy boundary on the glossy-rich scene, and falls to $\leq 1.2\%$ when that scene is excluded.

Against the same metric, the variance signal does not come close. Discretising $\hat{\sigma}^2$ onto quantile bins and scoring it with the identical agreement metric (Table 3) — the apples-to-apples comparison the raw "89.5% vs ρ=0.26" juxtaposition does not provide — it agrees with itself across the $64\times$ change at only 0.42–0.65 (3-bin) or 0.21–0.34 (7-bin), against the mechanism label's 0.87–0.996: on every scene the mechanism label is 1.4–$2.3\times$ more stable than coarse-binned variance and 2.5–$4.5\times$ more than fine-binned. Bin edges are per-scene quantiles computed *independently at each SPP level*, which removes the global $1/N$ rescaling of $\hat{\sigma}^2$ and scores only the stability of the field's shape — the reading most favourable to variance (§1, Problem 4); class imbalance is handled by the chance-corrected κ above, not by raw agreement.

`snooker`, with low per-pixel agreement (89.55%), has among the smallest bucket-fraction drifts (0.58%): boundary pixels flicker along the direct/diffuse-indirect edge of the felt, but the bucket sizes are unmoved. Residual predicate-bucket swaps occur — the largest single transition is 0.8% (`glass_caustic`, `specular-direct`→`direct`) — but stay small everywhere. The structural content of each scene is fixed at SPP$= 64$; the energy-argmax boundaries are what need more budget, and the analysis identifies which ones.

The continuous sidefields drift slightly with SPP, as expected for continuous quantities. The *purity* median drifts by 0.94% to 2.4% across SPP levels; the *glossy energy fraction* is more stable, drifting by less than 0.5% in the `snooker` scene where it is non-zero. The continuous information layered on top of the main partition moves only slightly with budget, while the discrete labels are fully stable in size and overwhelmingly stable in location.

4.4 Robustness to estimator completeness

Cross-SPP stability establishes that the labels do not move when the *sample budget* changes. A second, independent perturbation is available: changing the *sampling strategy* itself. Our pilot integrator resolves the `glossy` and `delta-mediated` buckets through next-event estimation, with the BSDF-sampling-into-emitter contribution suppressed by an emitter-acceptance rule (§3.1) that avoids double-counting in the absence of multiple importance sampling. This means the energy accounting on those two buckets is, by construction, only half of a complete MIS estimator. If the `dominant_class` argmax were sensitive to this, the label would inherit an estimator dependence of exactly the kind we fault $\hat{\sigma}^2$ for. We test it directly.

We implement a restricted power-heuristic MIS on the emitter-hit branch of the `glossy` and `delta-mediated` buckets only — the only two buckets whose end-vertex is non-delta and therefore admits a genuine NEE-versus-BSDF competition; the `emitter-direct` and `specular-direct` buckets have no NEE counterpart and are left at unit weight — and recompute `dominant_class` at SPP$=4096$ against the original NEE-only labelling. Two checks confirm the implementation is clean. 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`↔`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`→`delta-mediated` transition per scene is a known artefact of an upper-bound test rather than a classification ambiguity.

Table 5: Robustness of `dominant_class` to estimator completeness. "MIS agree" is the per-pixel argmax agreement between the NEE-only pilot and the MIS-completed estimator at SPP$=4096$. "cross-SPP (MIS)" is the 64 vs 4096 agreement recomputed under MIS, to be compared against the NEE-only range of Table 2.

scene	MIS agree (4096)	cross-SPP (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%

We emphasise the scope of this experiment before its result: it is robustness to *one controlled perturbation* of the estimator — restoring the suppressed MIS half — not a demonstration of general estimator independence. The per-pixel energy argmax can in principle still depend on the path-sampling strategy, MIS heuristic, lobe-sampling probabilities, and omitted path families; what the experiment isolates is the perturbation our own pilot construction makes testable, which is also the largest one it introduces.

Within that scope, the argmax is nearly insensitive to restoring the missing estimator half (Table 5): five of seven scenes are unchanged to the pixel — including `ring_caustic`, whose $6{,}949$-pixel `delta-mediated` population is untouched — `snooker` moves by 0.14%, and the glossy-rich `kitchen_counter` by 1.48%. The cross-SPP stability recomputed under MIS lands within a whisker of the NEE-only values of Table 2 ($89.55 \to 89.58\%$ on `snooker`; $88.69 \to 89.87\%$ on `kitchen_counter`): the structural content of each scene is determined regardless of whether the estimator is NEE-only or MIS-complete. The predicate buckets carry this robustness most strongly — `delta-mediated` agrees at 99.2% or better on every scene, because its membership is decided by the presence of a δ-S event, a discrete predicate that no amount of energy re-weighting can move. What flickers is, once again, the glossy energy-argmax boundary (`glossy` ↔ `diffuse-indirect`), which responds to energy perturbation whatever its source. This is the same split we observe under SPP perturbation, reproduced on a second, independent perturbation axis.

The cost of the boundary's sensitivity. The one place the perturbation does move the label is informative. On `snooker`, of the 83 pixels the NEE-only pilot calls `glossy`-dominant, 44 re-label as `diffuse-indirect` under MIS, leaving 37. These 44 are the narrowest-lobe pixels in the bucket: on a near-specular lobe the BSDF PDF dominates the NEE PDF, so the power heuristic correctly drives the NEE weight toward zero, and the NEE-only estimate — which retained that half at full weight — had overstated the glossy energy by a factor that pushed the argmax across the `glossy`/`diffuse-indirect` boundary. The swap is therefore not noise but a correction: the NEE-only pilot systematically over-assigned the narrowest-lobe pixels to `glossy`. We return to the consequence of this for the time-to-quality figures in §6.

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 `glossy` bucket spans $10{,}137$ pixels across roughnesses $\alpha \in [0.03, 0.16]$, and under the same MIS completion it is 95.8% stable, versus 37% for `snooker`'s 83-pixel population concentrated at $\alpha \approx 0.03$. The `snooker` fragility is thus the narrow-lobe boundary behaviour the mechanism analysis says it is, not a property of the glossy class per se: at population scale, with lobe widths spread across the bucket's range, membership is overwhelmingly stable under the perturbation.

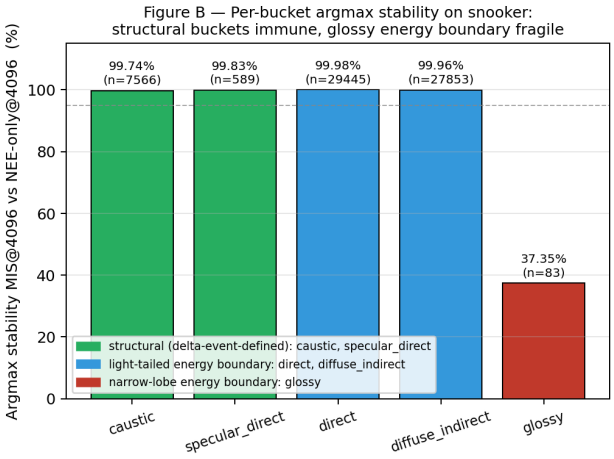

Figure 4: Per-bucket argmax stability on `snooker` under estimator completion (MIS@4096 vs NEE-only@4096). The structural buckets, whose membership is fixed by a δ-event predicate (`delta-mediated`, `specular-direct`), and the light-tailed energy boundaries (`direct`, `diffuse-indirect`) are all stable above 99.7%; only the narrow-lobe `glossy` energy boundary moves, for the reason analysed below. The dashed line marks 95%.

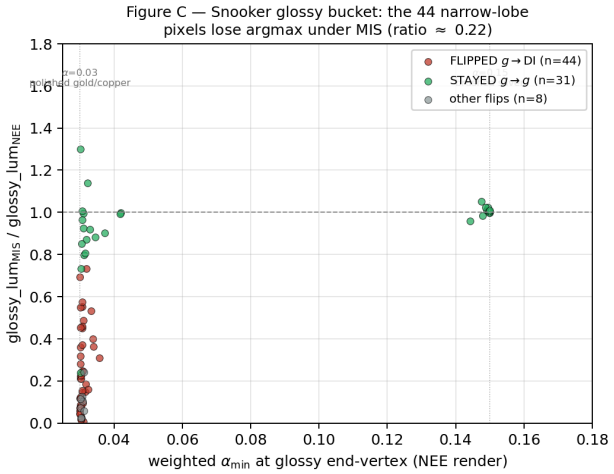

Figure 5: Mechanism of the `glossy` re-labelling on `snooker`. Each point is a pixel from the NEE-only `glossy` bucket, plotted by its end-vertex lobe width (weighted $\alpha_{\min}$) against the ratio of its glossy energy under MIS to under NEE-only. The 44 FLIPPED pixels (red) cluster at the narrowest lobe ($\alpha \approx 0.03$, polished gold/copper), where MIS down-weights the NEE half and the glossy energy contracts to ≈ 0.22 of its NEE-only value — enough to lose the argmax. The 31 STAYED pixels (green) span wider lobes and keep a ratio near unity. The swap is a correction concentrated on exactly the pixels the kurtosis argument flags as worst-behaved.

4.5 The variance baseline

The same data, judged by the variance-derived reference that current practice would use, is self-inconsistent in a way no amount of post-processing repairs — but, crucially, only in the regime the thesis is about. We measure two split-half self-consistency quantities. First, a per-pixel convergence exponent $\hat{\alpha}(p)$ fit from a four-level SPP sweep ($\hat{\sigma}^2$ at SPP $\in \{64, 256, 1024, 4096\}$) with ten independent seeds per level ($\sim$40,000 samples per pixel): we partition the ten seeds into two halves, refit $\hat{\alpha}(p)$ on each half, and Spearman-correlate the two fields, averaging over 20 random $5{+}5$ partitions (a single fixed split under-reports partition variance; the partition-to-partition standard deviation we measure is ≤ 0.003, so the earlier single-split values were in fact representative). Second, and more directly relevant to our thesis, the same procedure applied to $\hat{\sigma}^2$ itself at each fixed SPP level. For a noiseless reference either correlation would be 1.0.

Table 6: Split-half self-consistency of $\hat{\sigma}^2$: Spearman ρ between the $\hat{\sigma}^2$ fields of two disjoint five-seed halves, per fixed SPP level, reported as mean over 20 random $5{+}5$ partitions of the ten seeds (partition-to-partition standard deviation ≤ 0.003 in every cell, so we omit it from the table). A noiseless reference quantity would score 1.0. Self-consistency is *tail-dependent*: it collapses on the caustic-dominated scene (where the hardest pixels live) and is stable on light-tailed scenes.

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 two measurements agree, and localise the instability rather than asserting it globally. The $\hat{\alpha}$-derived ranking on `cubes_dof` and `glass_caustic` has split-half $\rho = 0.23$–0.29 — the Problem 3 reliability figure. The $\hat{\sigma}^2$ measurement (Table 6) is sharper: its self-consistency is *a function of the scene's tail*. On caustic-dominated `glass_caustic` it sits at 0.18–0.26 — at or below the $\hat{\alpha}$ floor, and non-monotone in SPP (more samples do not buy stability, itself a heavy-tail signature). On light-tailed `tabletop_separated` it sits at 0.69–0.71, and on Lambertian `cubes_dof` a flat ~ 0.31 across the whole range. The contrast between `cubes_dof` (flat) and `glass_caustic` (low and jittering) is exactly the §1 prediction: $\hat{\sigma}^2$-field shape is preserved on the light-tailed subset and least informative where the hardest pixels are.

This is the precise, narrower-and-stronger form of "variance is unstable." 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, the one a difficulty reference cannot afford to be wrong about — $\hat{\sigma}^2$ fails even to agree with itself across a seed split, and evaluations against a $\hat{\sigma}^2$- or $\hat{\alpha}$-derived target there are attenuated by this unreliability (§1, Problem 3).

The contrast with the mechanism label is decisive on the same scenes. With a single SPP$= 64$ pilot, our discrete labels agree with the SPP$= 4096$ reference at 87.1–99.6% (Table 2); on `glass_caustic` the label agrees with itself across the $64\times$ change at 95.6% while $\hat{\sigma}^2$ agrees across a seed split at $\rho \leq 0.26$. Discrete labels are stable exactly where $\hat{\sigma}^2$ is not.

Table 7: Split-half self-consistency of $\hat{\sigma}^2$ at SPP$= 4096$, resolved *per mechanism bucket* (mean over 20 seed partitions; "—" marks buckets with fewer than 50 pixels). The `delta-mediated` (d-m) bucket floors at 0.13 on `glass_caustic` — and at 0.03 there at SPP$= 64$, where the two seed halves are essentially uncorrelated — sits at 0.38–0.41 where delta-mediated transport is prominent but dim or mixed (`snooker`, `kitchen_counter`), and at 0.50–0.54 where it is a minority mechanism (tabletops). `ring_caustic` sharpens the reading: its bright, focused cardioid reaches 0.72 at SPP$= 4096$ (from 0.32 at 64) — within-bucket $\hat{\sigma}^2$ reliability tracks the *brightness* of the delta-mediated energy, not merely its prevalence, which is exactly the distinction the continuous `caustic_frac` sidefield resolves.

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

The tail-dependence of the previous paragraph is a scene-level statement; resolving it per bucket localises it to the pixel level and closes the loop with the mechanism partition (Table 7). The single sharpest measurement in this paper is the `delta-mediated` bucket of `glass_caustic`: there, $\hat{\sigma}^2$ has a split-half Spearman of 0.13 at SPP$= 4096$ and 0.03 at SPP$= 64$ — on the pixels the mechanism label marks as `delta-mediated` in a caustic-dominated scene, the variance estimator does not agree with itself across a seed split at all. This is $\hat{\sigma}^2$ failing as a reference signal in its most complete form, and it happens precisely on the pixels the partition isolates into a bucket of their own, computed without ever looking at $\hat{\sigma}^2$.

The effect is not that `delta-mediated` is uniformly the least-self-consistent bucket — in the tabletop scenes `specular-direct` is comparable or lower, and on `ring_caustic` the bright focused cardioid makes the bucket the scene's *most* self-consistent at high SPP (Table 7). The pattern is that the bucket's self-consistency tracks how heavy-tailed the delta-mediated energy actually is at its pixels: it collapses where that energy is dim and firefly-borne

(`glass_caustic`, 0.03 at SPP$= 64$), recovers where it is bright and focused (`ring_caustic` at high SPP), and sits in between where it is prominent but mixed — the scene-level tail-dependence of Table 6 localised to bucket granularity, with the residual within-bucket variation carried by the continuous `caustic_frac` and luminance sidefields. A variance-based reference has no representation to flag, in advance, which pixels carry this risk; the mechanism label does, by construction: the `delta-mediated` bucket is a pre-computed map of where the variance baseline is *at risk*, with the sidefields resolving where within it the risk is realised.

One entry appears, at first reading, to contradict our cost analysis (§6): the `glossy` bucket of `snooker` has the *highest* split-half self-consistency (0.82) yet §6 reports glossy as *slowest-converging*. There is no contradiction at the level of failure modes: self-consistency measures whether the $\hat{\sigma}^2$ *estimator* agrees with itself, time-to-quality whether the *pixel* converges. A glossy pixel can have a stably-estimated, consistently-high variance — $\hat{\sigma}^2$ reliably reports "hard" (high self-consistency), consistent with slow convergence — whereas the `delta-mediated` failure is that $\hat{\sigma}^2$ cannot even report a consistent value. The partition separates these two failure modes into two buckets where a single continuous number would conflate them. We caution, though, that the bucket-level 0.82 is itself an average over two opposite sub-populations: as §5.4 shows, the narrowest-lobe glossy pixels have *negative* ρ, and the high mean is carried by a wider-lobe, multi-path sub-population. The `glossy` bucket is not internally homogeneous in $\hat{\sigma}^2$ stability — itself a reason to prefer the discrete label over any bucket-averaged continuous statistic.

5 Cross-Scene Correlation Structure

The partition equips each pixel with not only a discrete label but a small continuous vector — the caustic, glossy, and bounce-depth energy fractions and the inverse purity — defined deterministically on the same path-tracer state. We examine how these components *correlate across pixels* in each scene and how that structure *varies across scenes*. Two questions drive the analysis. First, local to §3.4: does the empirical caustic–glossy relationship justify keeping them as independent buckets, or should they merge? Second, structural: does the correlation structure expose scene-level geometric variables a scalar variance signal cannot see?

We compute per-pixel Spearman correlations between the four continuous components — `caustic_frac` (the delta-mediated share), `glossy_frac`, `depth` (energy-weighted average bounce count to the end-vertex), and `inv_purity` $(1 - \pi)$ — on the seven scenes of §4.1. Table 8 reports the six pair-wise correlations. One pair is independent across every non-degenerate scene; one is scene-dependent with stable sign; and the remainder vary across scenes, two with their *sign* reversing in a way an identifiable scene-level geometric variable controls. Those two are the substantive content of this section, promoted to named findings in §5.1 and §5.2; the rest are reported as observations (§5.3).

Table 8: Per-pixel Spearman ρ among the four continuous components of the descriptor vector, across seven scenes. Components are abbreviated as C $=$`caustic_frac`, G $=$`glossy_frac`, D $=$`depth`, P $=$`inv_purity`. Scene columns are `cubes_dof` (cb), `glass_caustic` (gc), `snooker` (sn), `tabletop_mixed` (tm), `tabletop_separated` (ts), `ring_caustic` (rc), `kitchen_counter` (kc). Entries marked "*deg.*" have one component identically zero in that scene. Symbol key: ★ independent ($|\rho| < 0.4$) across all non-degenerate scenes; † scene-dependent, sign-stable; ‡ sign-reversal across scenes (Findings 1 and 2); ∘ open observation (§5.3). All entries regenerate directly from the released analysis pipeline.

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‡	F1
$D \leftrightarrow C$	*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	$+0.60$‡	F2

Stable and scene-dependent structure. One pair reads as independent across every non-degenerate scene: `caustic_frac` versus `inv_purity` ($|\rho| \leq 0.35$). One is scene-dependent but sign-stable ($D \leftrightarrow G$, always positive). The remaining pairs vary; the two whose sign reversal is controlled by an identifiable geometric variable are examined below. Each sign-reversal exposes a scene-level structural variable that controls the relationship between two mechanism components — a variable a per-pixel scalar variance has no representation for and therefore cannot detect.

5.1 Finding 1: caustic/glossy sign reversal with object-space separation

The decision to keep glossy independent of caustic (§3.4) was motivated by a falsifiable structural prediction. If the two buckets reflected a fixed relational structure, the per-pixel correlation between `caustic_frac` and `glossy_frac`

should have a consistent sign across scenes. The mechanism representation predicts otherwise. A contribution event lands in the `delta-mediated` bucket if its light-to-end-vertex sequence contains a δ-S event; the same path may also pass through a glossy interaction, in which case γ, the glossy energy fraction (§3.3), is non-zero on the same pixel. The two fractions therefore *co-occur* when the scene's object-space topology makes such combined light paths common, and *compete* when geometry forces a path to pick a specular or a glossy interaction but not both. The sign of the correlation should depend on the object-space separation between specular and glossy materials, and should cross zero somewhere along the separation axis.

"Object-space separation" must be a measured quantity, not a verbal label, so we declare two: (i) the minimum surface-to-surface distance between any delta-S object and any glossy object, and (ii) the *mutual visibility fraction* — the fraction of unoccluded straight segments between surface points sampled on the two object families (64 points per surface, ray-tested against the full scene). The second is the mechanism-relevant one: co-occurrence requires a path segment connecting the two families, so coupling should track sight lines, not metric distance. We report both.

The controlled test is the scene pair. `tabletop_mixed` has glass and metal balls in close proximity (entanglement regime); `tabletop_separated` shares its geometry and materials exactly but adds an opaque partition that blocks delta-S $\leftrightarrow$ glossy sight lines — mutual visibility drops from 0.094 to 0.000 — as a designed test of the prediction: if the sign does not reverse there, the mechanism account of the caustic–glossy relationship is wrong. `snooker`, an uncontrolled scene of irregular packing, sits between them in visibility (0.078).

Table 9: Per-pixel Spearman correlation between the delta-mediated and glossy energy fractions, with the two declared separation measures. The correlation 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), which sharpens the account: the controlling variable is whether one path can encounter both surface families. `kitchen_counter` is an out-of-family point: high visibility but large metric separation, and $\rho \approx 0$ — sight lines are necessary but coupling decays with distance. The full matrix appears in Table 8.

scene	min dist. (m)	mut. vis.	$\rho_{C\text{–}G}$
`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

Across the tested configurations the correlation ordering follows mutual visibility and crosses zero between `tabletop_mixed` and `snooker` (Table 9, Fig. 6); on the controlled pair the reversal is causal, since the partition is the only change. We state the claim at the strength the data supports: an observed ordering across the tested configurations with a controlled two-point contrast — not a continuous monotonic law, which would require a parametric separation sweep (§8). `kitchen_counter` adds an honest boundary to the account: with sight lines present but the families metrically distant, the correlation sits at $\rho \approx 0$ — visibility is necessary for co-occurrence, and the coupling strength decays with distance. What the finding establishes for the design question of §3.4 is unchanged by these qualifications: the sign of the delta-mediated–glossy relationship is a function of scene geometry that passes through zero, so any rule fixing glossy's relationship to the delta-mediated bucket in advance — merging them, or imposing a dominance order — would mislabel at least one observed regime. Glossy stays its own bucket.

5.2 Finding 2: depth/purity sign reversal with material topology

A natural objection to the descriptor vector is that bounce-`depth` alone might already carry most of what the other components encode — "deeper paths are messier paths." The correlation between `depth` and `inv_purity` provides the direct test: if depth predicted mechanism mixing monotonically, $\rho(\texttt{depth}, \texttt{inv\_purity})$ would be positive on every scene. The seven-scene measurement (Table 8, last row) refutes this: ρ is $+0.66$, $+0.52$, $+0.24$, $+0.50$, $+0.60$ on `cubes_dof`, `glass_caustic`, `tabletop_mixed`, `tabletop_separated`, and `kitchen_counter`, but -0.37 on `snooker` and -0.40 on `ring_caustic`.

The data supports a single consistent mechanism. How a path accumulates *depth* determines what depth does to *purity*. In scenes whose 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 metal rings — every added vertex is the same delta event, so paths gain length without gaining lobe diversity: the deepest pixels are the mechanically *purest*, and $\rho(\texttt{depth}, \texttt{inv\_purity})$ is *negative*. In scenes whose deep paths accumulate through lobe-mixing interreflection — diffuse bounce chains (`cubes_dof`, $+0.66$), glass transmission into diffuse interiors (`glass_caustic`, $+0.52$), glossy–diffuse mixing (`kitchen_counter`, $+0.60$) — every added vertex can change the lobe class, so depth accumulates *mixture* and the correlation is *positive*. The sign of $\rho(\texttt{depth}, \texttt{inv\_purity})$ thus separates specular-chain-dominated

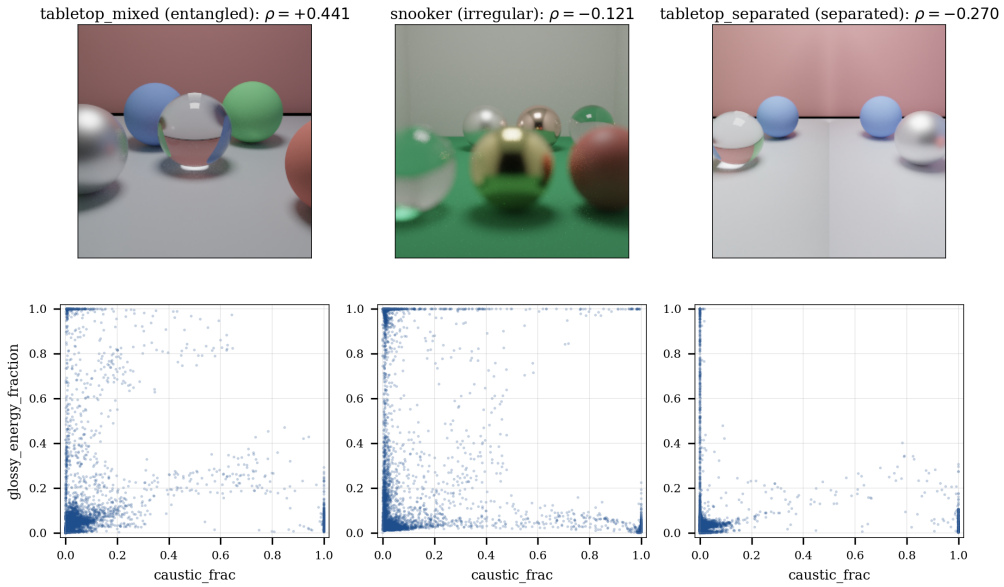

Figure 6: Sign reversal of the delta-mediated/glossy correlation with object-space separation. Top: RGB renders of the three scenes ordered by mutual visibility between the delta-S and glossy object families. Bottom: per-pixel scatter of `caustic_frac` vs `glossy_energy_fraction`. ρ falls from $+0.441$ (entangled) through -0.121 (snooker, irregular) to -0.270 (separated). `tabletop_separated` is the controlled test: identical geometry and materials except for the sight-line-blocking partition — had the sign not reversed there, the mechanism account of the relationship would be wrong.

scenes from lobe-mixing-dominated ones — with `ring_caustic`, a scene whose only non-diffuse transport is delta chains, landing on the negative side as the account requires.

The implication for the descriptor is direct: any reduction that absorbs `depth` into `inv_purity`, or treats them as redundant, will mislabel entire scenes by sign. `depth` and `inv_purity` are independent dimensions *by data*, not merely by construction.

5.3 Observation and open item

Two additional pairs in Table 8 vary in sign and are recorded as observations. `depth` ↔ `caustic_frac` 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 confidently: the partition geometry constrains both `depth` (paths must detour around the wall) and `caustic_frac` (which surfaces the partition occludes) simultaneously, and a single-scene anomaly does not distinguish a genuine mechanism from a partition-induced artefact. `inv_purity` ↔ `glossy_frac` 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 Finding 1's scene pair to be promoted to findings; we leave them as recorded structure.

5.4 Finding 3: lobe width controls the glossy heavy tail

The `glossy` bucket is the one on which continuous cost saturates (§6), and the robustness analysis (§4.4) showed that its narrowest-lobe members are also the ones whose membership is least stable under estimator completion. Both point at the end-vertex roughness α as the controlling variable. We test this directly with a controlled single-sphere experiment that removes the confounds present in a multi-object scene like `snooker`, where lobe width and path multiplicity vary together.

Controlled roughness sweep. We render a single isotropic `roughconductor` sphere in an otherwise diffuse box — no glass or smooth specular surfaces, so no δ-S contaminates the bucket — and sweep the lobe width α over eight values while holding geometry, illumination, and camera fixed. The pre-registered prediction was that two quantities

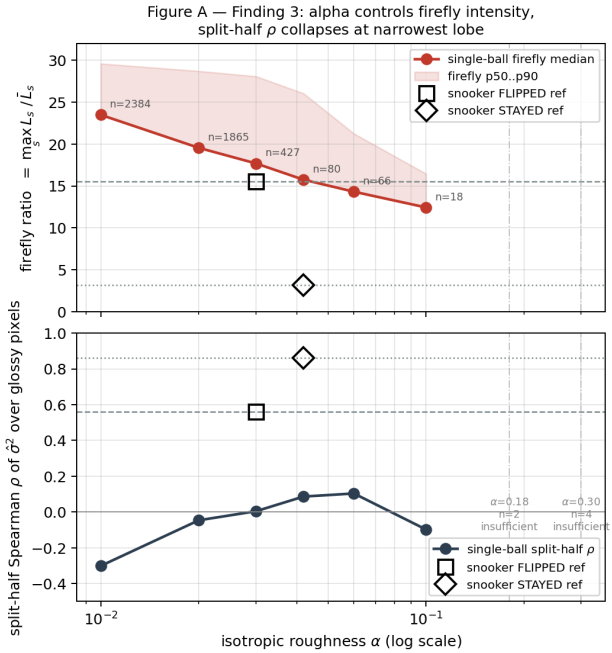

Figure 7: Finding 3 on a controlled single-sphere roughness sweep. *Top:* the median firefly ratio falls monotonically as the lobe widens, from 23.5 at α=0.01 to 12.4 at α=0.10 (shaded p50–p90 band); the `snooker` FLIPPED reference (narrow-lobe, □) lands on the curve while the STAYED reference (wide-lobe, multi-path, ◊) sits far below it, a regime the single sphere does not reproduce. *Bottom:* the split-half ρ of $\hat{\sigma}^2$ is *not* monotone in α — it collapses to -0.30 at the narrowest lobe (seed halves anti-correlated) and otherwise sits in a noise band — while the STAYED reference's $\rho \approx 0.86$ floats well above the entire single-sphere curve. Lobe width controls the firefly tail; it does not, on its own, account for the stable-glossy regime. The two widest α levels yield too few `glossy`-dominant pixels to measure.

would co-vary with α: the per-pixel firefly ratio (the seed-wise ratio of the maximum to the mean glossy contribution) and the split-half self-consistency ρ of $\hat{\sigma}^2$. The firefly prediction is confirmed cleanly: the median firefly ratio falls monotonically from 23.5 at α=0.01 to 12.4 at α=0.10 across six populated α levels (the two widest lobes produce too few `glossy`-dominant pixels to measure, the bucket's energy having dispersed; we report the firefly trend over $\alpha \in [0.01, 0.10]$). Narrow lobes carry a heavier per-pixel contribution tail — exactly the signature the kurtosis bound of §1 predicts, with the narrowest lobe being the highest-kurtosis integrand.

$\hat{\sigma}^2$ self-consistency fails on the same pixels, but not monotonically. The split-half ρ of $\hat{\sigma}^2$ does not rise smoothly with α. It is most informative at the narrow end: at α=0.01 it is -0.30 — the two seed halves rank the pixels in *anti*-correlation, the unambiguous fingerprint of an estimator dominated by rare large samples — and elsewhere it sits in a noise band (-0.10 to $+0.10$) without a clean trend. We do not claim a monotone $\rho(\alpha)$ relationship; the honest reading is narrower and still on-thesis: on a controlled glossy bucket, $\hat{\sigma}^2$ fails to agree with itself across a seed split at every α we can measure, and on the narrowest lobe it agrees *in reverse*. A signal that cannot reproduce its own per-pixel ranking on a clean single-material sphere is not a usable reference on that material, irrespective of budget. This is the glossy-bucket counterpart of the delta-mediated-bucket failure reported in §4.5.

What lobe width does not explain. A second controlled experiment isolates the limits of the lobe-width account. The `snooker glossy` bucket contains a sub-population (the 37 pixels that survive MIS, §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 sweep never reproduces at any α. We tested whether per-pixel path multiplicity was the missing variable by holding α=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. We record it as an open item rather than over-claim a mechanism we have not isolated.

What survives controlled test is the narrower, on-thesis statement: lobe width monotonically controls the firefly intensity of the glossy bucket, and $\hat{\sigma}^2$ is self-inconsistent across that bucket, most severely on the narrowest lobes.

Implications for the descriptor. The same continuous handles (γ, `depth`, `inv_purity`) that decorate the mechanism partition at the pixel level expose, when correlated across pixels and scenes, the geometric variables that govern interactions between the mechanism buckets. A scalar variance signal, which has no representation of light-path topology, could not have produced these predictions, and would not detect that the descriptor vector's effective dimensionality is scene-dependent. Two *specific* variables — mutual visibility between specular and glossy object families (Finding 1) and whether depth accumulates through delta chains or lobe-mixing bounces (Finding 2) — are what the mechanism representation makes visible. Both reverse the sign of a pair-wise correlation; neither can be detected by inspecting per-pixel $\hat{\sigma}^2$.

6 Closure: Continuous Cost Saturates Where Mechanism Does Not

We close with a second route to the same thesis. The variance instability of §1 concerned a sample *statistic*; the failure was that an estimator of σ^2 cannot be stable on heavy-tailed integrands. The same heavy-tailed regime should also make any continuous *cost* measurement saturate — pixels that fail to converge within a budget cap show up as right-censored, not as informative cost values — while the mechanism representation should remain functional on the same pixels. We test this directly, in two parts: a per-bucket median cost (§6.1) that identifies the bucket in which continuous measurement saturates, and a per-dimension correlation between cost and the difficulty vector (§6.2) that quantifies the consequences.

For each pixel we measure a per-pixel time-to-quality cost $c(p)$, defined as the smallest SPP at which $|\mathrm{lum}(I_N(p)) - \mathrm{lum}(I_{\mathrm{ref}}(p))| \,/\, \max(\mathrm{lum}(I_{\mathrm{ref}}(p)), \epsilon) < \tau$, with I_{ref} the SPP=4096 reference image and τ a fixed relative-error threshold. Pixels that fail to reach the threshold by $\mathrm{SPP}_{\max} = 2048$ are right-censored — their true cost is unknown but lies above the cap.

6.1 Per-bucket cost and the glossy long tail

Table 10 reports the median $c(p)$ per `dominant_class` bucket on each of the seven scenes.

Table 10: Median time-to-quality cost (SPP) per `dominant_class` bucket. "—" marks buckets with no pixels. Medians whose bucket is heavily right-censored at $\mathrm{SPP}_{\max} = 2048$ are marked † with the censoring rate below; a censored median is a lower bound, not a measured value. The `glossy` population is NEE-defined and estimator-dependent (§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

† right-censored at $\mathrm{SPP}_{\max}$=2048: `glossy` on sn 51.8%; `specular-direct` on rc 53.0%, on kc 48.8%.

Three entries are right-censored, and they identify the two regimes in which continuous cost measurement fails. On `snooker` the NEE-defined `glossy` bucket has a censored median of 2048 SPP with 51.8% of its pixels unconverged at the cap (Fig. 8); on `ring_caustic` and `kitchen_counter` the `specular-direct` bucket — deep delta chains along the ring interiors, mirror paths among the polished metals — is censored at 53.0% and 48.8%. To be precise about what censoring means: these pixels have not become intrinsically constant in cost; the experiment has ceased to resolve costs above the cap, so the reported medians are lower bounds. The mechanism classifier, on the exact same pixels, remains operational: discrete labels are still discriminating in a regime where time-to-quality has ceased to. The magnitude of the `snooker` censoring rate is, however, sensitive to how the `glossy` bucket is defined, which we examine next.

The 51.8% figure is bucket-definition-dependent. The cost above is measured over the population of pixels our pilot integrator labels `glossy`-dominant. That label is an energy argmax (§3.1), and an energy argmax is only as trustworthy as the energy accounting beneath it. Our pilot integrator resolves the `glossy` and `delta-mediated` buckets through next-event estimation only; the BSDF-sampling-into-emitter half of the estimator, which carries substantial energy on glossy end-vertices, is suppressed by the emitter-acceptance rule that avoids double-counting in the absence of MIS. The robustness analysis of §4.4 quantifies the consequence: when the missing half is restored through a power-heuristic

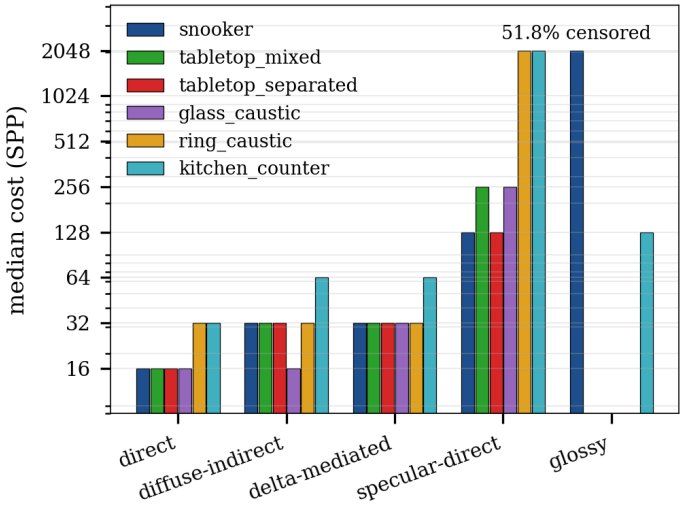

Figure 8: Time-to-quality median cost per `dominant_class` bucket, across the six scenes with non-trivial transport. Three censored medians sit at the SPP$=2048$ cap: the NEE-defined `glossy` bucket on `snooker` (51.8% right-censored) and the deep-chain `specular-direct` buckets on `ring_caustic` (53.0%) and `kitchen_counter` (48.8%) — while `kitchen_counter`'s population-scale `glossy` bucket converges at a median of 128 SPP. The continuous cost signal saturates on these buckets while the discrete mechanism label remains operational on the same pixels; §4.4 shows the `snooker` population is itself estimator-dependent.

MIS computed on the `glossy` and `delta-mediated` buckets, the `snooker` `glossy` population contracts from 83 to 37 pixels — the 44 pixels that leave are the narrowest-lobe pixels, on which MIS correctly down-weights the high-variance NEE half, and they re-label as `diffuse-indirect`. The 51.8% censoring rate is therefore measured over an NEE-only-defined population that is itself a superset of the MIS-defined `glossy` bucket. We report it as the convergence cost of the *NEE-only-defined* `glossy` bucket and flag the definition dependence explicitly. `kitchen_counter` now provides the population-scale counterpoint directly: its $10{,}137$-pixel `glossy` bucket, spread across lobe widths rather than concentrated at the narrowest, has a median cost of 128 SPP with only 5.9% censoring — confirming that `snooker`'s $2048^{\dagger}$ characterises the narrow-lobe NEE-defined sub-population, not the glossy class as such. The argument of this section does not rest on any particular value: under every definition tested, there exist buckets on which $c(p)$ saturates while the discrete label keeps discriminating.

A second consequence of the table is that the common assumption that delta-mediated transport is the worst case is, at least on these scenes, not borne out: `delta-mediated` converges at 32–64 SPP median on every scene where it exists — including `ring_caustic`, whose bright cardioid it contains — while the right-censored long tails live in the narrow-lobe `glossy` and deep-chain `specular-direct` buckets. The mechanism partition exposes this because the regimes live in different buckets by construction; a continuous difficulty signal computed from $\hat{\sigma}^2$ or from $c(p)$ blends them into one tail. We are deliberately more cautious than to call any bucket's pixels the single "hardest" in absolute terms: as the robustness analysis shows, part of the apparent hardness of the NEE-only `glossy` population is the high variance of the NEE estimator on narrow lobes rather than an intrinsic property of the transport — itself an instance of the estimator-dependence this paper argues against in the variance signal.

6.2 Cost versus the difficulty vector, dimension by dimension

The per-bucket view answers "which mechanism is most expensive." A complementary question is whether the continuous components of the difficulty vector correlate with cost in a stable way — whether `depth`, `caustic_frac`, `inv_purity`, and `glossy_frac` would each serve, on their own, as a cheap proxy for time-to-quality. The answer, shown in Table 11, is that they would not, and the reasons are informative about both the cost measurement and the structure of the vector itself.

Three observations follow. We label them to mirror the structure of §5.

Observation A: `glossy_frac` is the only sign-stable dimension, and its magnitude is structurally capped. $\rho(\texttt{glossy\_frac}, c)$ 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$4–8% of the variance in c, and on `snooker` this number is not the

Table 11: Per-pixel Spearman correlation between time-to-quality $c(p)$ and the four continuous components of the descriptor vector, across the seven scenes. "n/a" marks components with no energy in that scene.

dim	sn	gc	tm	ts	cb	rc	kc
`depth`	$+0.27$	$+0.22$	$+0.44$	$+\mathbf{0.04}$	$+0.35$	$+0.22$	$+0.34$
`caustic_frac`	$+0.07$	$+0.17$	$+0.21$	$+0.01$	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	$+0.29$

intrinsic strength of the relationship: with 51.8% of the NEE-defined glossy bucket right-censored, Spearman ρ on c 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*. The discrete-vs-continuous contrast of §4.5 recurs at the level of a second continuous signal: where mechanism labels let a consumer isolate the glossy long-tail by bucket, cost has nothing left to say about it.

Observation B: `inv_purity` versus cost is sign-unstable. $\rho(\texttt{inv\_purity}, c)$ runs -0.16 to $+0.26$ across the seven scenes — mostly small, signs disagreeing. The instability is not noise (it 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 (Table 10). The component does the work of §3.3 — flagging mixture — but is not, alone, a stable cost proxy.

Observation C: `depth` versus cost is unstable, and the failure is geometry-driven. $\rho(\texttt{depth}, c)$ sits at $+0.22$ to $+0.44$ on six scenes but **collapses to** $+0.04$ on `tabletop_separated` — the same scene where $\rho(\texttt{depth}, \texttt{caustic\_frac})$ inverts sign (§5.3), 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. (The collapse is not an instance of the Finding-2 sign structure: Finding 2's negative-ρ scenes are `snooker` and `ring_caustic`, where depth–cost remains positive; the companion anomaly is the depth–`caustic_frac` inversion, which shares the partition geometry as its cause.) `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 of Table 11 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 (A), sign-unstable on the mixture component (B), and geometry-varying on the depth component (C). 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.

Closing the loop. Two consequences. First, the variance-instability argument of §1 is not specific to sample variance: any continuous difficulty measurement reaches a saturation regime on the hardest pixels — here, the narrow-lobe glossy and deep specular-chain tails — exactly where reliable guidance matters most. Second, the mechanism classifier does not hit this regime change: bucket boundaries are computed from path structure, not estimated cost, so they do not saturate when cost does. A discrete representation built on the integrand's deterministic structure is robust where the continuous estimators are not, at both ends of the argument this paper makes. The natural next question — whether this robustness has *operational* value — is the subject of the next section.

7 Downstream Demonstration: Mechanism-Conditioned Pilot Correction for Sample Allocation

The preceding sections establish that the descriptor is stable where variance estimates are not. The natural question is whether that stability buys anything operational. This section answers with a controlled equal-budget sample-allocation experiment, pre-registered in the released protocol with hypotheses, metrics, and decision rules fixed before the runs; the full protocol, two disclosed protocol corrections it required, and every measured cell are in the released artifact.

Protocol. Two-stage adaptive rendering at 256^2 on all seven scenes. A pilot of $N_0 \in \{4, 16, 64\}$ SPP is rendered with the §4.1 integrator, yielding both a pilot variance $\hat{\sigma}^2(p)$ (luminance, unbiased, per pixel) and pilot bucket energies from which the pilot label $\hat{b}(p)$ is read — label estimation shares the pilot budget, costing nothing extra. 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 for every strategy), 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. Error is measured against an *independent* 4096-SPP reference (no shared samples with pool or pilots). Every number is reported as the paired outcome over $K{=}10$ disjoint pilot draws — single-pilot adaptive-sampling comparisons can reverse sign under re-piloting, so we treat multi-pilot pairing as mandatory methodology. The robust-variance incumbent and hypotheses of Result 3 were added by a protocol amendment registered during revision, with estimator definition, block-count rule, and decision criteria fixed before any robust-variance run was executed; the amendment and its timestamps are part of the released artifact, kept separate from the original pre-registration. The label's overhead is negligible at equal time: bucket-energy capture shares the pilot's path tracing, the floor is a per-bucket median and a maximum (< 1 ms per image), and the median-of-means estimate costs $\sim$60 ms per 256^2 pilot — $\sim 60\times$ the naive sample variance in relative terms, negligible against rendering.

Strategies. Baselines: *uniform*; *observed-var* ($D = \hat{\sigma}^2$, the standard pilot-variance allocator); and an *oracle* using the pool's own converged variance (upper bound). The proposed method 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 $\max(\hat{\sigma}^2(p), \mathrm{med}_b)$, where med_b is the median pilot variance within the pixel's own bucket; light-tailed pixels are untouched. The mechanism rationale is §4.5'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, and the bucket median supplies a defensible repair. Controls: a *random-partition placebo* (the same floor applied to a label map with identical class sizes but shuffled pixel assignment, fresh per draw) tests whether coarse partitioning alone explains any gain; a *label-oracle* variant (labels from the reference) prices the pilot-label estimation error of §3.1. Two further treatments — cross-scene transfer of per-bucket scale factors, and reliability-weighted within-bucket shrinkage — are reported in the artifact; our own controls disqualify them (the first fails the no-harm criterion under leave-one-scene-out transfer, the second's gains are largely reproduced by a bucket-free global-shrinkage placebo), and we present them as ablations, not methods. Finally, the registered amendment adds a *robust-variance* incumbent: $D = \hat{\sigma}^2_{\mathrm{MoM}}$, 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; block-count sensitivity is swept in the artifact and does not change any conclusion below), together with *robust-floor* — the same per-bucket median floor with the statistic recomputed from $\hat{\sigma}^2_{\mathrm{MoM}}$, never reused from the naive pilot — and its own matched random-partition placebo. Any global multiplicative bias of the MoM variance cancels under Neyman normalisation; only cross-pixel ordering acts. MoM is not meaningfully defined at $N_0{=}4$ (two-sample blocks), a limitation with content of its own: robust estimation has a minimum sample size, so at the smallest pilots — exactly where the heavy-tailed pilot most needs repair — mechanism conditioning is the only correction of the two that still applies.

Table 12: Equal-budget allocation at pilot $N_0{=}4$ SPP: PSNR (dB, mean over 10 disjoint pilot draws; seed-to-seed std 0.1–3.0) and the per-draw win rate of bucket-floor against observed-var on paired MSE and paired p99 tail error. Bold marks the five scenes on which the standard pilot-variance allocator is at or below *uniform* sampling at this pilot budget: all three scenes with dominant heavy-tailed transport (`glass_c.`, `kitchen`, `ring_c.`), the Lambertian `cubes`, and `t._sep.` within seed noise (-0.04 dB) — see §7, Result 1, for the two distinct failure components. Bucket-floor improves on observed-var with high per-draw consistency wherever heavy-tailed buckets exist, and on `cubes_dof` (no heavy-tailed buckets) it reduces to observed-var exactly — no-harm by construction. The robust-variance incumbent of Result 3 is n/a at this pilot budget (median-of-means needs a minimum sample size; see text) and is reported at $N_0 \in \{16, 64\}$ in Table 13.

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	—	—

Result 1: the pilot fails, for two distinguishable reasons. At $N_0{=}4$, pilot-variance allocation sits at or below *uniform* sampling on five of seven scenes (Table 12), and the failures decompose into two components that must not be conflated. The first is *generic weight-estimation noise*: Neyman weights estimated 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. The Lambertian `cubes_dof` is the pure case — observed-var loses 2.22 dB to uniform with no heavy-tailed transport in the scene at all, so the heavy-tail argument plays no role there; and `tabletop_separated`'s -0.04 dB deficit is within seed noise of uniform. This component is a small-sample property of four-sample Neyman

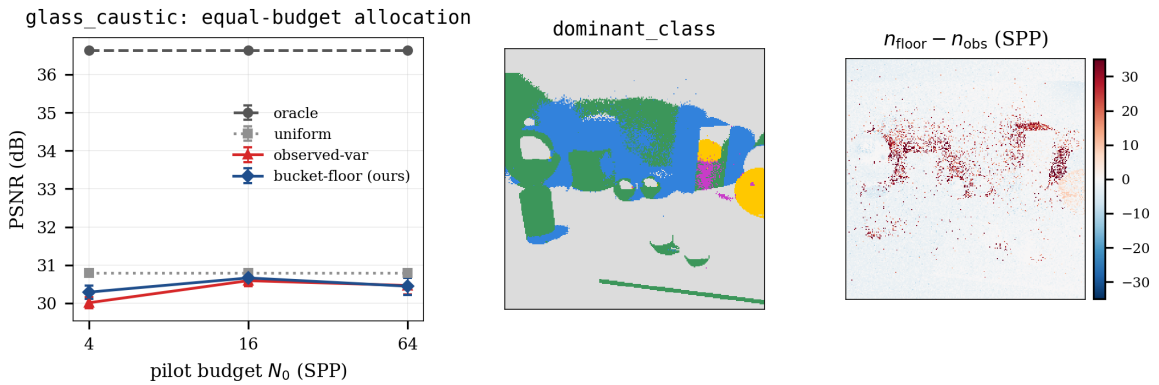

Figure 9: The downstream experiment at a glance. *Left:* equal-budget PSNR versus pilot budget on `glass_caustic` (mean ± std over 10 disjoint pilots). The standard pilot-variance allocator (red) sits *below uniform sampling* (dotted) at every pilot budget — the §1 instability surfacing as a rendering outcome — 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 per-pixel allocation difference $n_{\text{floor}} - n_{\text{obs}}$ (right; red = more samples under bucket-floor) concentrates precisely on the pixels the mechanism label marks `glossy` and `delta-mediated` (middle; colours as in Fig. 2) — the pixels whose pilot variance is systematically underestimated, and the correction a random partition cannot replicate.

weights (here amplified by defocus noise), not a contradiction of §4.5's finding that $\hat{\sigma}^2$ is a serviceable *reference* on light-tailed scenes at measurement budgets; it shrinks as the pilot grows (`cubes_dof` is within 0.15 dB of uniform by N_0=64) and is not what this paper's descriptor addresses. The second component is the *heavy-tail pathology* of §1: on the three scenes with dominant delta-mediated or glossy transport the pilot systematically underestimates variance exactly on the tail pixels, and the loss to uniform does *not* vanish with pilot budget — on `glass_caustic` it persists at N_0=64 (30.47 vs 30.79 dB), with the oracle showing +5.9 dB of unrealised headroom. It is this second, systematic component that the mechanism label identifies a priori and that bucket-floor repairs (Result 2); on scenes carrying only the first component, bucket-floor deliberately does nothing. Three single-variable sweeps in the artifact separate the two components experimentally. Every gap in Table 12 is unchanged when the test is repeated at 512^2, ruling out a resolution artifact. 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 every positive gap compresses): a noisy pilot does more damage the more samples it is trusted with, so the deficit is not a fixed cost of small pilots but a liability that scales with commitment. And a per-pixel sample floor of 4 — the simplest guard against starving mis-ranked pixels — repairs the small-sample component almost entirely (`ring_caustic`@4: $-1.12 \rightarrow +0.02$ 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.

This decomposition also answers a fair question about the baseline: why measure no-harm against observed-var rather than against uniform? Because the experiment isolates the label's *marginal* contribution to the incumbent signal. Whether a system should fall back toward uniform when its pilot is weak is an orthogonal control addressing the generic-noise component — and not a trivially solved one: our bucket-free global-shrinkage placebo, which is exactly such a fallback-by-regularisation, underperforms observed-var on all seven scenes at N_0=4. Since bucket-floor improves on observed-var wherever heavy-tailed buckets exist at small pilots and stays within ±0.1 dB of it everywhere else (H3, released protocol), adding the label is compatible with whatever fallback policy is layered on top.

Result 2: mechanism conditioning repairs part of it, with specificity. On paired MSE, bucket-floor improves on observed-var in 7–10 of 10 pilot draws on every scene containing heavy-tailed buckets (+0.28 to +0.41 dB at N_0=4; delta-mediated-bucket MSE halved on `snooker`), while on `cubes_dof` — no heavy-tailed pixels — it reduces to observed-var exactly. We report win counts as paired descriptive statistics, not significance tests; under the stricter pre-registered criterion ($\geq 8/10$ wins *and* mean gain above one pooled seed standard deviation) the N_0=4 MSE wins on `glass_caustic`, `tabletop_separated`, and `kitchen_counter` qualify; `snooker`'s 10/10 win passes the consistency bar but not the magnitude bar (mean gain 0.0034 against a pooled seed std of 0.0064 — wins that are consistent in sign but individually small; its qualifying win at N_0=4 is on the tail metric), and `ring_caustic`'s 7/10 stays within seed noise at N_0=4 and reaches 10/10 at N_0=16. The tail metric behaves the same way on the delta- and glossy-heavy scenes (p99 wins of 8–10/10 on `snooker`, `kitchen_counter`, `glass_caustic`) with one

instructive exception: on `tabletop_mixed` at N_0=4, bucket-floor *lowers* the delta-mediated-bucket MSE by 33% (0.0136 → 0.0092) yet slightly raises the image-wide p99 in $8/10$ draws (by ∼2%) — the floor spreads budget across the whole bucket, while observed-var concentrates its budget on the few extreme pixels that set the global quantile whenever its pilot happens to catch them. The trade is visible only because we report both metrics; a single scalar would hide it. The spatial structure of the correction is exactly the predicted one (Fig. 9, right): extra samples flow to the pixels labelled `glossy` and `delta-mediated`. The random-partition placebo reproduces none of these gains (its paired-MSE deltas are an order of magnitude smaller or negative), 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 (e.g. 38.15 vs 38.17 dB on `snooker`@4): estimating the label from the same 4-SPP pilot costs essentially nothing, quantifying the finite-sample-label concern of §3.1 operationally. Gains shrink monotonically as the pilot grows (by N_0=64 all deltas are within ±0.03 dB) — the repair targets pilot noise, and vanishes when the pilot no longer needs repairing.

Result 3: orthogonality to robust estimation. The concession of §2.1 — that robust estimators qualify the critique of the naive sample variance — is here made experimental, and it cuts both ways. As an incumbent, $\hat{\sigma}^2_{\mathrm{MoM}}$ helps only where its assumptions hold: it beats observed-var under the pre-registered criterion in exactly one cell (`glass_caustic`@64, $9/10$ draws; mean PSNR can favour MoM without meeting the paired criterion, as for `glass_c.`@16 and `kitchen`@16 in Table 13, both at $7/10$ paired wins with gains below one pooled seed std), and on the most delta-mediated scene it is actively harmful (−2.35 dB vs observed-var on `ring_caustic`@16) — median-of-means buys its 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 containing heavy-tailed buckets (Table 13), qualifying under the strict criterion on `tabletop_mixed`@16 and `tabletop_separated`@16/64, and its matched random-partition placebo reproduces none of it (placebo deltas ≈0 or negative in every qualifying cell). The artifact's block-count sweep localises the interaction (Fig. 10): within-bucket rank fidelity against the converged variance is flat in k for the delta-mediated bucket but collapses for `specular-direct` (Spearman 0.52 → 0.20 on `ring_caustic`@16 as $k = 2 \to 8$), the error explosion localises in that same bucket (0.99 → 13.1 bucket MSE), and at the pre-registered k the bucket line reads: naive 0.87 → MoM 2.6 → MoM+floor 0.70. Two properties make the floor an *independent* information source rather than damage control: it already improves the naive pilot (0.87 → 0.67 on the same bucket), and the repaired level is essentially estimator-invariant — 0.66–0.70 across the naive pilot and $k \leq 4$, with residual degradation to 0.895 only under the most aggressive blocking (k=8: two-sample blocks whose variances carry a single degree of freedom, the same minimum-sample-size limit that rules out N_0=4 entirely). Nor can the failure be dismissed as a poor choice of k: choosing k safely requires knowing in advance which pixels are single-firefly dominated — precisely what $\hat{\sigma}^2$ is least able to report about itself (§4.5) 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 → 0.74 on `snooker`@64) while rare-tail buckets degrade. A per-bucket adaptive block count is the natural composition (§8).

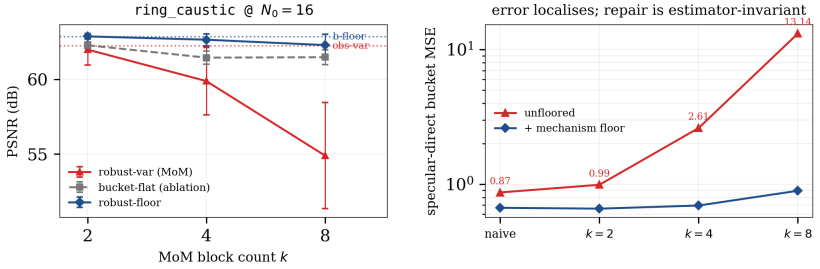

Figure 10: Result 3 mechanism on `ring_caustic` at N_0=16. *Left:* PSNR versus the median-of-means block count k. The robust incumbent (red) degrades monotonically as blocking becomes more aggressive; the per-bucket-uniform ablation (grey) recovers most of the loss — the partition component of the repair, growing with k — and robust-floor (blue) holds an approximately constant level above it: the per-pixel component. Dotted lines: the naive incumbent and its floored variant. *Right:* the damage localises in the `specular-direct` bucket, whose within-bucket ranking is the one that collapses (MSE 0.99 → 13.1 as $k = 2 \to 8$), while the mechanism-floored level holds at 0.66–0.70 across the naive pilot and $k \leq 4$, with residual degradation (0.895) only at k=8's one-degree-of-freedom blocks — the floor is an independent information source, not damage control.

What this demonstrates — and what it does not. To be fully explicit about the strength of the result: no strategy in Table 12 — ours included — beats uniform sampling on all seven scenes at N_0=4, and on the three scenes where the pilot fails outright, bucket-floor narrows but does not close the gap to uniform. The demonstrated claim is *dominance*

Table 13: Result 3 at pilot N_0=16: PSNR (dB, mean over 10 disjoint draws) for the naive incumbent (obs-var), the robust incumbent (MoM, k=4), and their mechanism-floored variants; the last two columns give robust-floor's per-draw paired-MSE win rate against robust-var at N_0=16 and 64. Bold marks the cells where the robust incumbent falls below the naive one, for two distinct reasons: on scenes containing heavy-tailed buckets, blocking discards tail evidence; on the light-tailed `cubes_dof`, small-sample blocking is itself costly (four-sample blocks at N_0=16) — the same minimum-sample-size limitation that makes MoM n/a at N_0=4. 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	—	—

over the pilot-variance allocator: bucket-floor improves on observed-var wherever heavy-tailed buckets exist and degenerates to it exactly where they do not, so a system already running pilot-variance allocation loses nothing and gains where it hurts most. The oracle's $+5.9$ to $+7.5$ dB headroom locates the remaining loss in pilot *quality*, not in the variance-proportional principle itself; Result 3's simplest robust hybrid closes part of it at the larger pilots, with the sharper lesson that estimator robustness and mechanism conditioning repair *different* failures — the former cannot substitute for the latter on rare-tail pixels. The experiment thus shows the descriptor functioning in exactly the role argued for throughout: a stable conditioning variable that identifies, a priori, the pixels on which a continuous estimate needs repair — not a replacement difficulty scalar, and not yet a complete allocator. It is one demonstration on one task with deliberately simple estimators; it does not claim state-of-the-art adaptive sampling, and 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).

8 Limitations and Scope

Taxonomy scope. The descriptor covers surface path tracing under a stated D/G/S decomposition (§3.5). 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. Coverage of the named mechanisms is established on the tested scene class only.

Convention dependence. The per-event label depends on renderer conventions (NEE lobe rule, flag semantics); the per-pixel label is a finite-sample estimator (§3.1). We measured robustness to one controlled estimator perturbation and to budget changes; broader estimator families (bidirectional, guided, regularised) are untested.

Findings. Finding 1 rests on a controlled pair plus ordered observations, not a parametric separation sweep; Finding 2's two-family account is consistent with seven scenes but derived from uncontrolled scene comparisons except for `ring_caustic`; the two recorded observations (§5.3) remain open. The stable-glossy sub-population of `snooker` resists our controlled account (§5.4).

Downstream. The allocation experiment uses a simple NEE-only pilot and a two-stage budget split; production adaptive samplers are more sophisticated on both axes. The demonstrated claim is that mechanism conditioning repairs a noisy pilot where it is least reliable — not that it outperforms modern adaptive pipelines end-to-end.

Robust estimation. Result 3 tests one robust incumbent — median-of-means, chosen for having no free parameter beyond the block count. Catoni-type M-estimators and other sub-Gaussian constructions (§2.1) remain untested, as does the natural composition the block-count sweep motivates: a per-bucket adaptive block count, aggressive in mid-tail buckets where blocking improves rank fidelity and conservative in delta-mediated ones where it destroys it.

Denoising. We deliberately do not evaluate the label as a denoiser auxiliary channel. Such an evaluation would be dominated by the choice of denoiser family (kernel-predicting, guided filtering, temporal), and a conclusion supportable across families requires a comparison well beyond this paper's scope; reporting a single-family result would invite exactly the overclaiming this revision removes. The split-half methodology of §4 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 and no agreed-upon reference signal that remains reliable across transport regimes. The default — the finite-sample per-pixel variance of the path-traced estimator — is governed by the integrand's fourth moment, empirically unstable on the hardest pixels of typical scenes, and of low split-half reliability ($\rho \approx 0.25$) as an evaluation target. The community's response has been to suppress the symptom at the mean estimator and refine the variance estimator itself; we instead take the instability as evidence that a stable *structural* signal should be conditioned on alongside the statistical one.

We describe each pixel by the discrete transport mechanism of its contribution events, defined by a two-axis partition — end-vertex BSDF lobe by presence of δ-S in the light-to-end-vertex sequence — yielding seven mutually exclusive labels with no continuous threshold on the main axis, together with continuous sidefields retaining the mechanism mixture, and with the deterministic per-event label carefully separated from its finite-sample per-pixel estimator. On seven scenes the named mechanisms receive all observed energy; the dominant label agrees at 87.1–99.6% across a $64\times$ budget change where 7-bin quantile-discretised $\hat{\sigma}^2$ agrees at 21–34%; and restoring the MIS half of the estimator moves the argmax on at most 1.5% of pixels, with the flicker isolated to identified energy-argmax boundaries and the predicate buckets essentially immune. The descriptor's correlation structure exposes scene-level geometric variables — mutual visibility between specular and glossy object families, and whether depth accumulates through delta chains or lobe-mixing bounces — that reverse pair-wise correlation signs and that no scalar per-pixel variance can represent. Where time-to-quality itself right-censors — narrow-lobe glossy and deep specular-chain buckets with $\sim 50\%$ of pixels unconverged at 2048 SPP — the discrete labels remain discriminating. And the descriptor earns its keep operationally: conditioning a noisy pilot variance on the label consistently improves on pilot-variance allocation at equal budget, exactly on the scenes where the pilot is least reliable — narrowing, though not yet closing, the gap to uniform sampling where the raw pilot loses even to uniform — while acting only where heavy-tailed buckets exist and surviving a placebo control. Under heavy-tailed transport, finite-sample variance is an unstable measurement; the transport mechanism of each contribution is a stable, deterministic structure to condition on — and doing so already pays.

Future work. Several extensions follow naturally. First, one regime resists our controlled account: the stable-glossy sub-population of `snooker` (low firefly, high $\hat{\sigma}^2$ self-consistency) is reproduced by neither the single-sphere roughness sweep nor the light-count sweep of §5.4, which together rule out lobe width and light-source count as its cause; isolating the responsible scene-topology variable — multi-object inter-reflection is our leading candidate — requires a controlled two-object experiment we leave to future work. Second, the cost-saturation argument of §6 was made against L_2 image error; a perceptual-MSE variant (FLIP, FovVideoVDP) might saturate differently, and an analogous analysis under perceptual ground truth is a clean follow-up. Third, combining the stability of mechanism labels with the local responsiveness of sample-derived signals in a Bayesian shrinkage formulation is a natural direction, with mechanism providing the prior and the sample-derived quantity serving as the likelihood. Finally, the bounce-depth structure consumed implicitly by our partition admits an explicit path-length axis — for instance, a per-pixel Sobol-axes effective dimension — that would complement the lobe-based labels along a quantitative dimension orthogonal to mechanism type.

A Reproducibility

All results were produced with Mitsuba 3.8.0 (`cuda_ad_rgb` variant) on a single NVIDIA RTX 5090; the complete pipeline re-runs in under one hour. We release the custom integrator, all procedural scene builders, analysis scripts, seed layouts, and the pre-registered downstream protocol; every table and figure regenerates from released entry points.

Classifier. Per-event classification follows Table 1 with the sampled-lobe cascade Delta $>$ Diffuse $>$ Glossy for BSDF-sampled bounces and the flag rule of §3.1 for NEE events; the complete BSDF-model-to-lobe mapping used by our scenes: `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-negative emission times non-negative throughput), so no signed-energy convention is needed; non-finite values do not arise in the bucket accumulators, and rare non-finite pixels in raw float32 RGB frames (fireflies) are sanitised to zero by the analysis loaders and counted as background.

Estimator. NEE at every Smooth-flagged 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 (§4.4): power heuristic (β=2, solid-angle measure) on the NEE and emitter-hit branches of the `glossy` and `delta-mediated` buckets only.

Statistics. $\sigma^2_{\text{intrinsic}} = K \cdot \text{Var}$ across 10 seeds (ddof 1) of per-seed luminance at budget K; 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, computed independently at each SPP. Weighted κ: agreement weight $1-$penalty, penalty 0.2 within the light-tailed family, 1.0 otherwise (Table 4). Time-to-quality: $\tau = 0.10$, $\epsilon = 10^{-4}$, SPP grid $\{16, \ldots, 2048\}$ (doubling), independent renders per level, first-crossing rule, reference = the seed-disjoint SPP-4096 stability render. Convergence exponent $\hat{\alpha}(p)$: per-pixel unweighted ordinary least squares of $\log_{10}$ luminance variance on $\log_{10}$ SPP over the four levels, $\hat{\alpha} = -$slope; a pixel is excluded if its variance at *any* level is non-positive or non-finite; 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). 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. Downstream protocol: seed layout pool $= 1000 + i$, pilot draw k frame $j \;= 100000 + 1000k{+}j$, reference chunks $\;= 900000 + c$; disjoint by construction. Result-3 amendment: median-of-means over $k = \lceil\sqrt{N_0}\rceil$ contiguous render-order blocks (per-block variance ddof 1; trailing remainder discarded and counted); 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 verify every pool and pilot statistic bit-identical to the original run, so robust-var and observed-var are computed from exactly the same samples. The amendment text (registered before execution), both manifests, and the block-count mechanism sweep are included in the artifact.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[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] George Casella and Roger L. Berger. *Statistical Inference*. Duxbury / Thomson Learning, Pacific Grove, CA, 2nd edition, 2002.

[10] A. W. van der Vaart. *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[11] 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.

[12] Eric Veach. *Robust Monte Carlo Methods for Light Transport Simulation*. Ph.D. dissertation, Stanford University, December 1997. Technical Report STAN-CS-TR-98-1610.

[13] 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.

[14] 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.

[15] Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[16] Charles Spearman. Correlation calculated from faulty data. *British Journal of Psychology*, 3(3):271–295, 1910.

[17] 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.

[18] 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.

[19] Luc Devroye, Matthieu Lerasle, Gábor Lugosi, and Roberto I. Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

[20] Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

[21] 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.

[22] 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.

[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 Röthlin, 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] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[42] Jacob Cohen. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968.

[43] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.