ALBEDO DRIFT: WHAT SINGLE-IMAGE INTRINSIC DECOMPOSITION EVALUATION DOES NOT MEASURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Albedo is illumination-invariant by definition, yet the standard evaluation of single-image intrinsic decomposition never tests invariance: WHDR on IIW scores relative reflectance *within* one image, so a model that bakes illumination into its albedo is never penalized. We name this failure *albedo drift* and introduce a protocol that measures it directly: the ratio of cross-illumination albedo variation to input variation on scenes photographed under many lighting conditions. Across three datasets and six methods we find that (i) a state-of-the-art model absorbs half of the illumination change into its albedo even within its training domain (*drift ratio* 0.55), and remains at 0.50 on photometrically diffuse pixels, ruling out non-Lambertian ambiguity as the cause; (ii) outside the training domain, decomposition *amplifies* illumination variation (*drift ratio* 1.5, worse than returning the input); (iii) under simulated illuminant color changes, learned models amplify chromatic variation up to $10\times$; and (iv) what determines drift is not the architecture family but the match between training and test distributions — a domain-matched diffusion model is as invariant as regression models, while a mismatched one is $6\times$ worse. Invariance alone is trivial (a constant albedo is perfectly invariant), so we evaluate it jointly with fidelity and observe that no existing method occupies the region that is both faithful and invariant. This corner is reachable: a 1M-parameter consistency adapter trained on held-out scenes halves drift (-49%, sign test $p<10^{-9}$) while slightly improving WHDR, generalizes across capture domains, and cannot be replicated by smoothing. Albedo drift is not a fundamental trade-off; it is an axis the field has never optimized because it has never been measured.

1 INTRODUCTION

For a decade, progress in single-image intrinsic decomposition (Narihira et al., 2015; Fan et al., 2018; Li & Snavely, 2018b; Liu et al., 2020; Das et al., 2022; Luo et al., 2020) has been reported as a falling WHDR score on IIW (Bell et al., 2014). WHDR asks whether the predicted albedo agrees with sparse human judgments of relative reflectance between pairs of points *in the same image*. It does not ask — and no widely used metric asks — whether the albedo of a scene stays the same when only the lighting changes. Yet this is the defining property of albedo: an intrinsic image method that absorbs illumination into its reflectance output has, in a precise sense, failed at its own task, regardless of its benchmark score.

We show that this failure is real, large, structural, and invisible to current evaluation. Given a scene photographed under L illuminations $\{I_\ell\}$, a model's albedo predictions $\{A_\ell\}$ should be constant. We measure the *drift ratio*: the mean per-pixel variation of $\{A_\ell\}$ divided by that of $\{I_\ell\}$ (§3), so that 0 is perfect invariance and 1 means no illumination is removed at all. On the MIT Multi-Illumination test set (Murmann et al., 2019), the state-of-the-art regression model of Careaga & Aksoy (2023) has a *drift ratio* of 0.55: half of the lighting change survives in the "intrinsic" output. A diffusion-based model evaluated off its training distribution reaches 3.2 — its albedo varies three times *more* than the input. Qualitatively the failure is unmistakable: under a dark lighting direction, the model returns the dark image as albedo, illumination baked in wholesale (Fig. 2).

One might hope this is measurement noise, non-Lambertian ambiguity, or an artifact of our normalization. It is none of these. Drift is unchanged under diffusion ensembling and across sampling

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

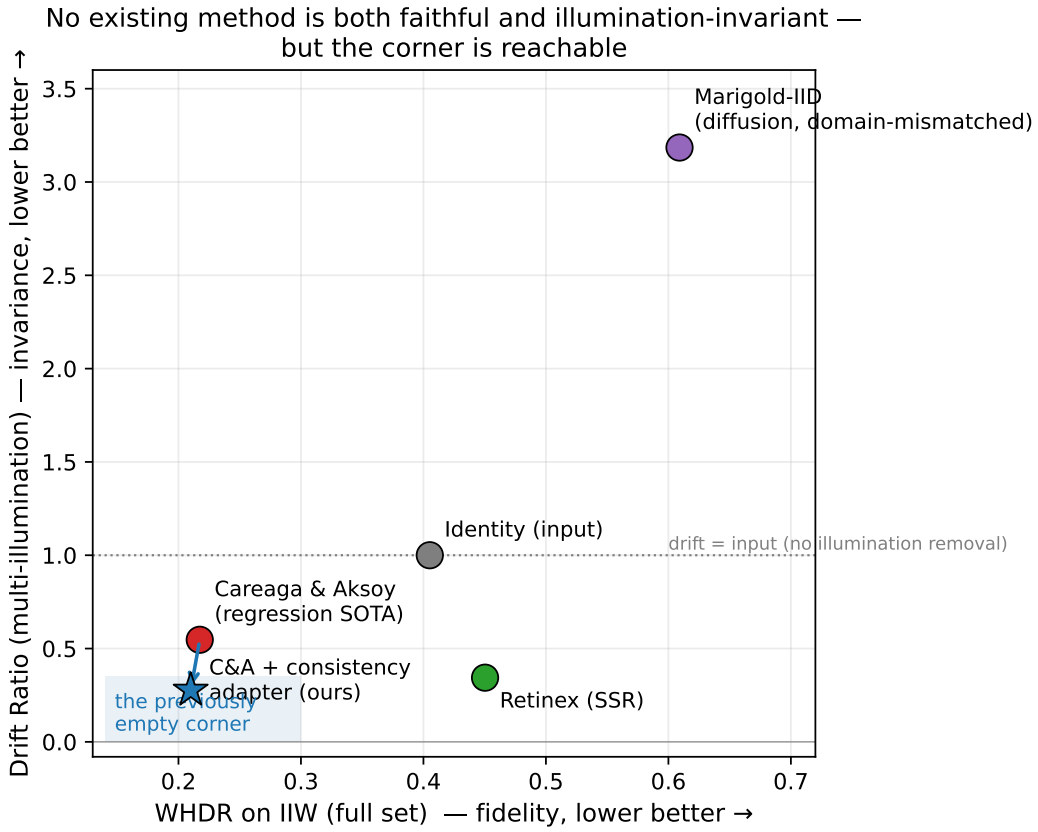

Figure 1: **The invariance–fidelity plane.** X: WHDR on the full IIW set (fidelity). Y: *drift ratio* on 30 multi-illumination test scenes (invariance; identity $= 1$ by construction). No existing method is both faithful and invariant; a 1M-parameter consistency adapter (§5) moves the state of the art into the previously empty corner.

seeds (noise explains 4% of it), survives per-channel and linear-RGB renormalization, persists at 0.50 on photometrically diffuse pixels (saturation and moving highlights masked out), and is spatially correlated with shadow and specularity patterns — a structural leak, not jitter. Nor is it benign: composited relighting flickers in direct proportion to drift (§6).

Invariance alone would be a broken target — a constant image is perfectly invariant. The correct battlefield is two-dimensional (Fig. 1): invariance jointly with fidelity. On this plane the literature collapses onto a frontier with an empty corner. A classical center-surround Retinex (single-scale (Jobson et al., 1997), with lineage to (Land & McCann, 1971)) is more invariant than the 2023 state of the art but far less faithful; learned models are faithful but leak; no method is both. That emptiness is the central fact this paper contributes, and our final result is that it is not a law: a 1M-parameter consistency adapter, trained once on held-out multi-illumination scenes, moves the state of the art into the corner — drift $0.55 \rightarrow 0.28$ with WHDR unchanged — and generalizes to capture setups it never saw. The axis was reachable all along; it was simply never optimized, because it was never measured.

From a learning perspective, albedo drift is an instance of a broader failure mode: when a task-defining invariance appears in neither the training loss nor the benchmark, models are free to satisfy the benchmark with representations that violate the task's semantics — a relative of shortcut learning (Geirhos et al., 2020) and of underspecification, where many solutions fit the training signal but differ off-distribution (D'Amour et al., 2020). Intrinsic decomposition makes this failure unusually crisp because the violated property is not a heuristic desideratum but the definition of the output itself, and because it becomes measurable with existing data and checkpoints alone. We use it as

a case study with three questions of general interest: how large is the violation, what governs it (architecture or distribution), and is it an optimization gap or a trade-off?

Contributions. (1) We name albedo drift and define a normalized, dataset-agnostic protocol for measuring illumination invariance, covering both lighting direction and illuminant color. (2) We measure it across three datasets and six methods, with controls that eliminate sampling noise, normalization choices, gamma conventions, and non-Lambertian ambiguity as explanations. (3) We trace the dominant factor to training-distribution match rather than architecture family. (4) We show the invariance–fidelity corner is reachable with a small consistency adapter, establish its non-triviality against smoothing baselines, and map the controllable trade-off it exposes.

2 Related Work

Evaluation of intrinsic decomposition. The problem dates to Barrow & Tenenbaum (1978), with classical treatments from learned reflectance–shading edge discrimination (Tappen et al., 2005) to joint shape–illumination inference (Barron & Malik, 2015); see Garces et al. (2022) for a survey, and Bonneel et al. (2017) for editing-driven evaluation. IIW/WHDR (Bell et al., 2014) scores sparse relative reflectance within a single image; MIT Intrinsics (Grosse et al., 2009) provides object-level ground truth in a lab setting; MAW (Wu et al., 2023) adds intensity, chromaticity and texture axes for absolute albedo quality. All are within-image axes: none compares predictions of the same scene across illuminations. Our axis is complementary — and empirically near-orthogonal: drift correlates only weakly (Spearman 0.18) with ground-truth albedo error where ground truth exists.

Consistency across illumination as a learning signal. Using illumination variation of a fixed scene to *derive* intrinsics is classical: Weiss (2001) recovered intrinsic images from sequences by a median over filtered frames, and Sunkavalli et al. (2007) factored time-lapse video. The learning generation turned this into supervision: Bi et al. (2018) enforce reflectance equality across lighting during training; Lettry et al. (2018) train on illumination-varying sequences; BigTime (Li & Snavely, 2018a) learns from time-lapse sequences with the same motivation. These works treat cross-illumination consistency as supervision for their own method, not as an evaluation axis for the field; the generation of models since then has not been measured against it. We turn the folklore into a metric, apply it to the current generation, and show the problem is unsolved and in some regimes worse.

Diffusion-based decomposition. Recent generative approaches (Kocsis et al., 2024; Ke et al., 2024; 2025; Zeng et al., 2024; Chen et al., 2024) revived intrinsic decomposition with strong image priors, typically trained on synthetic indoor corpora (Roberts et al., 2021; Zhu et al., 2022). IDArb (Li et al., 2025) observes qualitatively that such models leave lighting in albedo; we quantify that observation, show it is dominated by distribution mismatch rather than the generative formulation itself, and show a domain-matched diffusion model (Kocsis et al., 2024) is as invariant as regression baselines.

Evaluation blind spots in learned models. Benchmarks shape what models learn: ImageNet training induces texture bias (Geirhos et al., 2019), and properties outside the benchmark are underspecified and drift freely (D'Amour et al., 2020; Geirhos et al., 2020). Our results give this abstract concern a physical, dataset-independent handle: the invariance is dictated by the image formation model, so the audit transfers to any method claiming to output albedo.

Color constancy. Illumination-invariant descriptions of surface color are a classical goal of color constancy, from gray-world (Buchsbaum, 1980) and Retinex (Land & McCann, 1971) to von Kries chromatic adaptation and learned estimators (Gijsenij et al., 2011; Barron, 2015); that literature treats invariance to the *illuminant* as the quantity of interest, while decomposition benchmarks never measure it. Our spectral probe (§4.1) connects the two literatures: modern decomposition models fail exactly the invariance that constancy research spent decades pursuing, and amplify illuminant color changes up to $10\times$.

3 ALBEDO DRIFT: DEFINITION AND PROTOCOL

3.1 IMAGE FORMATION AND THE INVARIANCE CONSTRAINT

Under the Lambertian intrinsic model, an image of scene s under illumination condition ℓ factorizes as

$$I_{s,\ell} \;=\; A_s \odot S_{s,\ell}, \tag{1}$$

where A_s is the albedo, $S_{s,\ell}$ the shading, and $\odot$ the per-pixel product. The object of this paper is the constraint that Eq. (1) imposes by construction: A_s carries no subscript ℓ. Writing an estimator's prediction as

$$\hat{A}_{s,\ell} \;=\; \hat{A}(I_{s,\ell}) \;=\; A_s + \varepsilon_{s,\ell}, \tag{2}$$

the cross-illumination variance isolates exactly the illumination-dependent part of the error,

$$\sigma_\ell^2\Big[\hat{A}_{s,\ell}\Big] \;=\; \sigma_\ell^2[\varepsilon_{s,\ell}]\,, \tag{3}$$

because the true albedo term is constant in ℓ. This is the quantity that no single-image metric can see: a within-image score constrains $A_s + \varepsilon_{s,\ell}$ for one ℓ at a time, and is blind to how ε co-varies with the illumination.

3.2 SCALE AMBIGUITY AND NORMALIZATION

Eq. (1) is invariant to the gauge transformation

$$(A_s,\, S_{s,\ell}) \;\longmapsto\; (\alpha A_s,\, S_{s,\ell}/\alpha), \qquad \alpha > 0, \tag{4}$$

so predictions are comparable only after fixing the scale. We fix it per prediction by the median luminance,

$$\tilde{A} \;=\; A \,/\, \mathrm{med}(Y(A))\,, \qquad Y(A) = \tfrac{1}{3}\textstyle\sum_c A^{(c)}, \tag{5}$$

and use a per-channel variant, $\tilde{A}^{(c)} = A^{(c)}/\mathrm{med}(A^{(c)})$, when the gauge group is extended to diagonal (von Kries) color scalings; §4 shows our conclusions are unchanged under either choice. If $\mathrm{med}(Y(A)) \to 0$ (a near-black prediction), Eq. (5) diverges; such degenerate outputs are detected by thresholding the median, counted, reported separately, and excluded from ratio aggregation, with medians used as the central statistic on in-the-wild data.

3.3 THE DRIFT RATIO

Averaging the per-pixel standard deviation over space (denoted by the bar) and normalizing by the same functional applied to the inputs gives our metric:

$$\mathrm{Drift}(s) \;=\; \frac{\overline{\sigma_\ell\Big[\tilde{A}_{s,\ell}\Big]}}{\overline{\sigma_\ell\Big[\tilde{I}_{s,\ell}\Big]}}. \tag{6}$$

Three properties calibrate its reading. (i) For the identity estimator $\hat{A} = I$, numerator and denominator coincide and $\mathrm{Drift} \equiv 1$: a score of 1 means no illumination is removed. (ii) For an oracle $\hat{A} = A_s$, the numerator vanishes and $\mathrm{Drift} = 0$. (iii) For any constant prediction $\hat{A} \equiv c$, $\mathrm{Drift} = 0$ as well — invariance alone is trivially gameable, which is why every result in this paper reports invariance *jointly* with fidelity. The denominator makes scores comparable across scenes with different lighting ranges, at the cost of a dataset-dependent absolute scale, which we state openly.

What drift can and cannot claim. In scenes with interreflections, single-image decomposition is not unique, and different illuminations may resolve the ambiguity differently; zero drift may be unattainable for any single-image method. We therefore treat measured drift as an upper bound on invariance failure and pair every claim with controls: (i) *diffuse-only* evaluation, restricting the spatial mean in Eq. (6) to the mask

$$M \;=\; \neg\,\mathrm{dilate}_7\Big[\max_\ell Y(I_\ell) \geq 0.97 \;\vee\; \max_\ell Y(I_\ell) - \mathrm{med}_\ell\, Y(I_\ell) \geq 0.5\Big]\,, \tag{7}$$

which removes saturated pixels and moving highlights (the Lambertian model holds on the remainder); (ii) seed and ensemble controls for stochastic models; (iii) per-channel and linear-RGB renormalization; (iv) an empirical reachability result (§5) showing drift 0.28 is attainable, and 0.05 if fidelity is sacrificed.



Figure 2: **Drift is structural.** Top: albedo predictions of the same scene under two lighting directions — brightness and color of the "intrinsic" output change with the light. Bottom: the spatial variation map of the predicted albedo (right) correlates with the input shadow/specularity variation map (left). Scene: multilum `everett_kitchen2`.

3.4 Fidelity axis

We use WHDR on the full IIW set (5,230 images) as the fidelity axis of Fig. 1. Given human judgments $J_i \in \{1, 2, \mathrm{E}\}$ with confidences w_i over point pairs (p_i, q_i), the predicted judgment thresholds the luminance ratio at $\delta = 0.1$,

$$\hat{J}_i = \begin{cases} 1 & Y_{q_i}/Y_{p_i} > 1+\delta \\ 2 & Y_{p_i}/Y_{q_i} > 1+\delta \\ \mathrm{E} & \text{otherwise,} \end{cases} \qquad \mathrm{WHDR} = \frac{\sum_i w_i \, \mathbf{1}\left[\hat{J}_i \neq J_i\right]}{\sum_i w_i}, \tag{8}$$

with Y evaluated on the predicted albedo treated as linear; conventions differ across the literature and shift absolute values (e.g., $0.217 \to 0.316$ for C&A if outputs are treated as sRGB), so we state ours explicitly. On MIT Intrinsics we additionally validate against ground-truth albedo error.

4 Measurement Study

Setup. Datasets: MIT Multi-Illumination (30 test scenes $\times$ 25 directional flashes) (Murmann et al., 2019); MIT Intrinsics (20 objects $\times$ 10 lights, with ground truth) (Grosse et al., 2009); BigTime in-the-wild time-lapses (30 sequences, provided validity masks) (Li & Snavely, 2018a). Methods: C&A ordinal-shading pipeline (Careaga & Aksoy, 2023) (2023 weights and the v2 weights of Careaga & Aksoy (2024)); Intrinsic Image Diffusion (Kocsis et al., 2024) (indoor-trained diffusion); Marigold-IID (Ke et al., 2025); single-scale center-surround Retinex (Jobson et al., 1997) (σ=32, fixed, not tuned); identity. Single fixed seed; all configurations and per-scene numbers are released.

Finding 1: in-domain failure. On multilum, C&A scores 0.55 — and its v2 weights were trained with supervision derived from this very dataset's train split, so this is a best-case, in-distribution number — and in 29 of 30 scenes the model absorbs at least a third of the illumination change

Table 1: Main results. Drift: mean over scenes (95% bootstrap CI); lower is better; identity = 1 by construction. Spectral drift: simulated illuminant color change (§4.1). WHDR: full IIW. Relight: flicker of composites with fixed target shading. Unmeasured cells shown as —.

Method	Train dist.	Drift (dir.)	Drift (diffuse)	Drift (spectral)	WHDR	Relight
C&A v2 (Careaga & Aksoy, 2023)	synth.+MID†	0.547 [0.45,0.66]	0.500	6.77	0.217	0.498
C&A 2023 (Careaga & Aksoy, 2023)	synth. indoor	0.661 [0.54,0.80]	—	—	—	—
IID (Kocsis) (Kocsis et al., 2024)	indoor (IV)	0.442 [0.40,0.49]	—	8.41	0.293‡	—
Marigold-IID (Ke et al., 2025)	mismatched	3.185 [2.91,3.48]	2.955	9.83	0.609	4.178
Retinex (SSR) (Jobson et al., 1997)	—	0.344 [0.32,0.37]	0.392	0.62	0.450	0.330
Identity	—	1.000	1.000	1.000	0.405	1.000
C&A+adapter (ours)	+multilum train	**0.277**	**0.303**	3.98	**0.210**	**0.421**

Table 2: Cross-domain *drift ratio*: multilum (mean), MIT Intrinsics (mean), BigTime (median; degenerate-output protocol). Ranking is domain-dependent; no method is invariant everywhere.

Method	multilum	MIT	BigTime
C&A v2	0.547	0.500	1.514
Marigold-IID	3.185	1.902	1.748
Retinex (SSR)	0.344	2.473	0.806
Identity	1.000	1.000	1.000
C&A+adapter (ours)	**0.277**	**0.316**	**1.081**

(per-scene minimum 0.13). Drift concentrates spatially at specular, glass, and shadowed regions (correlation of variation maps 0.43–0.55), and on diffuse-only pixels remains 0.50: the failure is not an artifact of non-Lambertian ambiguity.

Finding 2: controls. For the diffusion model, sampling-seed variation accounts for $1/25$ of cross-illumination variation; ensembling (n=10, recommended settings) changes drift by under 1%. Per-channel normalization (removing global color casts) and linear-RGB inputs leave conclusions unchanged (0.55→0.60 and →1.30 respectively — the latter shows the model is bound to the sRGB input convention, not helped by "more physical" inputs). Test-time photometric augmentation does not reduce drift ($+3\%$): the leak is spatially structured, not a global sensitivity.

Finding 3: out-of-domain amplification. On BigTime, C&A's median drift is 1.51 — worse than the identity. Outside its training distribution, decomposition does not remove illumination; it amplifies it. Ranking is domain-dependent too: Retinex is the most invariant method indoors yet the worst on lab objects (2.47), evidence that a single-dataset invariance benchmark would mislead.

Finding 4: the cause is distribution match, not family. Two comparisons disentangle architecture from domain. Same model, changed domain: C&A goes $0.55 \to 1.5$. Same domain, changed model: the indoor-trained diffusion model scores 0.44 — as invariant as regression — while the mismatched diffusion model scores 3.15. The generative formulation is not the culprit; distribution mismatch is, and it expresses itself as illumination baking. (The two diffusion models differ in more than training data; we state this as consistent evidence, not proof; see §7)

4.1 Spectral drift

The 25 multilum flashes are spectrally identical, so the results above measure *directional* invariance only. We probe the spectral axis by holding the geometry fixed (one frame per scene) and re-rendering it under a family of simulated illuminant colors: for color temperature K,

$$I_K \;=\; f^{-1}(D_K\, f(I_0))\,, \qquad D_K = \mathrm{diag}\!\left(\frac{c(K)}{\bar{c}(K)}\right), \tag{9}$$

where f is the sRGB-to-linear map, $c(K)$ the Planckian RGB gains, and the unit-mean normalization $\bar{c}(K)$ changes chromaticity while approximately preserving mean luminance. Since the true albedo is invariant to the diagonal family $\{D_K\}$, the drift ratio of Eq. (6) computed across $K \in \{2500, \ldots, 9000\}$ K — with per-channel normalization, so that any *global* von Kries cast is forgiven — measures spatially non-uniform illuminant-color baking. Gains follow the Tanner–Helland approximation — an engineering approximation, not a colorimetric rendering; a von Kries

transform on the Planckian locus would be more principled, and we flag this in §7. Learned models amplify the chromatic change 7–10× in the mean and 6–8× in the median (Table 1) — including the domain-matched diffusion model (8.4), whose *directional* drift is the lowest of all learned methods. Domain match, which governs directional drift (Finding 4), does not rescue the spectral axis: spectral fragility appears universal across current learned models; since the metric already forgives global color casts (per-channel normalization), this reflects *spatially non-uniform* color baking. Directional and spectral invariance are distinct axes — a method can hold one and fail the other — and current evaluation measures neither. Interestingly, the adapter of §5, trained only on directional variation under a fixed illuminant spectrum, transfers partially to the spectral axis ($6.8 \to 4.0$); we hypothesize that penalizing shadow-chromaticity baking shares a mechanism with illuminant-color baking, and leave the test of this hypothesis to future work. The simulation ignores camera spectral sensitivity; we state this in §7.

5 Reaching the Empty Corner

If drift were a fundamental trade-off, no operating point could improve invariance without paying fidelity. We test this with the smallest intervention we could design: a UNet adapter g_θ (≈1M parameters; base width 32, three scales) on top of the frozen base model, acting as a residual refinement

$$\hat{A}^{\text{ref}} \;=\; \max\Big(0,\; \hat{A} + g_\theta(I, \hat{A})\Big)\,, \tag{10}$$

with g_θ zero-initialized so that training starts exactly at the base model. For a scene s and a sampled pair of illuminations (a, b), the objective couples a consistency term with a fidelity anchor,

$$\mathcal{L} \;=\; \mathbb{E}_{s,(a,b)}\Big[\big\|\tilde{A}^{\text{ref}}_{s,a} - \tilde{A}^{\text{ref}}_{s,b}\big\|_1 \;+\; \tfrac{\lambda}{2}\big(\|\hat{A}^{\text{ref}}_{s,a} - \hat{A}_{s,a}\|_1 + \|\hat{A}^{\text{ref}}_{s,b} - \hat{A}_{s,b}\|_1\big)\Big], \tag{11}$$

where the consistency term is computed after the normalization of Eq. (5) (so the gauge of Eq. (4) is not penalized). The two limits make the design transparent: $\lambda \to \infty$ recovers the base model, while $\lambda \to 0$ admits the degenerate constant solution of §3; the anchor is what forces invariance to be achieved without abandoning fidelity (Adam, lr 2×10^{-4}, 20k steps, batches of four 256^2 crops). Training uses only multilum *train* scenes (never the test scenes, IIW, or any other evaluation data) with two losses: cross-illumination consistency between scale-normalized refined albedos of the same scene, and an L_1 anchor to the frozen prediction, weighted by λ.

Against a pre-registered criterion (drift -30% with WHDR degradation $\leq 10\%$), the adapter achieves drift $0.547 \to 0.277$ (-49%, CI $[-58, -40]$; all 30 scenes improve, sign test $p{=}9 \times 10^{-10}$) while WHDR on the full IIW set *improves* from 0.217 to 0.210 (-3.4%). The refined model is more invariant than Retinex while achieving the best fidelity of all evaluated methods — the corner of Fig. 1 is occupied.

It is not smoothing. Bilateral filtering does not reduce drift (0.55); Gaussian blur trades 13% drift for 23% WHDR. Neither approaches the adapter's operating point.

It is not rig memorization. Zero-shot on MIT Intrinsics the adapter cuts drift 37% and slightly *improves* ground-truth albedo error (si-MSE $0.0318 \to 0.0302$); on BigTime it cuts median drift 29%. Fidelity was measured on IIW — a domain the adapter never saw — and did not move.

The trade-off is controllable. Sweeping λ traces a Pareto curve: $\lambda{=}0.3$ reaches drift 0.05 (near-perfect invariance) at $+18\%$ WHDR; $\lambda{=}1$ is the knee (-49% drift, WHDR -3.4% on the full set); $\lambda{=}3$ is conservative (-25%, -2.6%). Removing the image input matches the drift reduction (-50%) but costs fidelity ($+3.4\%$): the image conditioning exists to preserve fidelity, not to create invariance.

6 Downstream: drift is visible in applications

Relighting composites a (supposedly invariant) albedo with new shading. We fix a target shading from a reference frame r and composite it with each illumination's predicted albedo,

$$R_{s,\ell} \;=\; \tilde{A}_{s,\ell} \odot S^{\text{ref}}_s, \qquad S^{\text{ref}}_s \;=\; I_{s,r} \oslash \hat{A}_{s,r}, \tag{12}$$

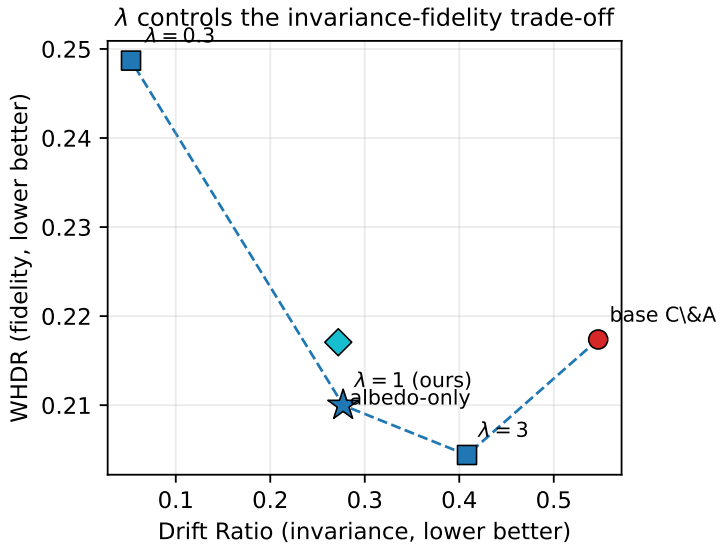

Figure 3: **The trade-off is controllable.** Sweeping the anchor weight λ traces an invariance–fidelity curve; λ=1 is the knee. Base C&A and λ=1 use full-IIW WHDR; $\lambda \in \{0.3, 3\}$ and the albedo-only variant use a deterministic 150-image subset (10k training steps), stated here to avoid mixing protocols silently.

and report the instability of $\{R_{s,\ell}\}_\ell$ normalized as in Eq. (6); a perfectly invariant albedo makes $R_{s,\ell}$ constant in ℓ. The resulting flicker tracks drift across methods (C&A 0.50, ours 0.42, mismatched diffusion 4.18) and is robust to the choice of reference frame. Qualitative sequences are shown in Fig. 4. We also evaluated an edit-stability metric based on shading reuse and rejected it: true shading must vary with illumination, so "lower is better" does not hold (the identity solution degenerates to zero). We report this to spare the axis from metrics that would game it.

7 Limitations

(1) Our measurements use tone-mapped JPEG inputs; HDR-linear validation remains open. (2) The spectral probe uses an approximate Planckian rendering (Tanner–Helland gains in linear RGB) and ignores camera spectral sensitivity; a colorimetric simulation could shift absolute spectral-drift values. (3) Interreflections impose an unquantified lower bound on achievable drift; our reachability result bounds it empirically at ≤ 0.05 under fidelity sacrifice. (4) The adapter is validated on one base model; we claim a drop-in refinement exists, not universality. (5) The two diffusion models differ beyond training data; Finding 4 is consistent evidence, not a controlled proof.

8 Conclusion

Albedo drift renames a folklore complaint into a measured, controlled, reducible quantity. The field's evaluation measured fidelity and called it progress; the invariance axis was left unmeasured, and models regressed on it — absorbing, and off-domain amplifying, the very signal they exist to remove. The axis family (directional and spectral) is now cheap to measure with existing datasets and checkpoints. We expect consistency to move from folklore loss term to reported metric, and invariance-aware training to close the spectral gap our probe exposes.

AI Use Statement

AI coding and writing assistants were used throughout this work, under the authors' direction, for implementing the measurement and training code, orchestrating experiment runs, and drafting and editing the manuscript text. All experimental designs and success criteria were specified by the

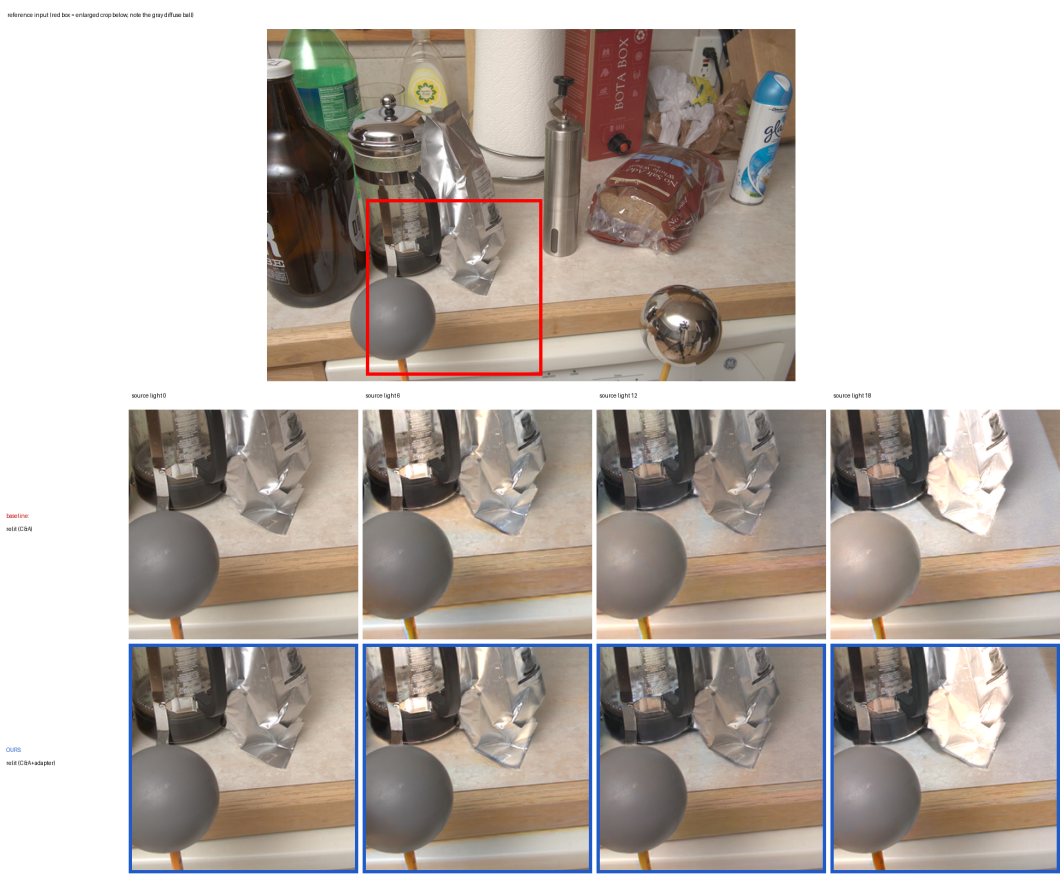

Figure 4: **Relighting with drifting albedo flickers.** Top: reference frame; the red box marks the crop enlarged below, containing a gray *diffuse probe ball* whose true albedo is constant by construction. Middle row (baseline C&A): as the source illumination changes (columns), the relit composite changes — the ball brightens and takes a warm color cast under light 18, and the countertop hue shifts — although the target lighting is fixed; the drifting albedo carries the source light into the composite. Bottom row (**ours**, blue frame): the same crops stay nearly constant across columns; the ball keeps a single gray, and the foil bag — a genuinely non-Lambertian object — remains the main residual failure for both methods. Scene: multilum `everett_kitchen6`.

authors (and pre-registered in code docstrings before results were computed); all reported numbers were produced by the released code on the stated datasets and were verified by the authors against the archived result files. The authors take full responsibility for the entire content of the paper.

Reproducibility Statement

All metrics are defined in closed form in §3 with normalizations and masks stated explicitly (Eqs. 2–8); datasets are public (multi-illumination test set, IIW); model checkpoints are the official public releases, with versions and seeds fixed (seed 0 throughout). Every number in the paper is written by the measurement scripts to a JSON results file, and the appendix reports the full per-scene tables. Code and result files sufficient to reproduce every table and figure will be released.

References

Jonathan T. Barron. Convolutional color constancy. In *ICCV*, 2015.

Jonathan T. Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE TPAMI*, 37 (8), 2015.

Harry G. Barrow and Jay M. Tenenbaum. Recovering intrinsic scene characteristics from images. In *Computer Vision Systems*, 1978.

Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (SIGGRAPH)*, 33(4), 2014.

Sai Bi, Nima Khademi Kalantari, and Ravi Ramamoorthi. Deep hybrid real and synthetic training for intrinsic decomposition. In *EGSR*, 2018.

Nicolas Bonneel, Balazs Kovacs, Sylvain Paris, and Kavita Bala. Intrinsic decompositions for image editing. *Computer Graphics Forum*, 36(2), 2017.

Gershon Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin Institute*, 310(1), 1980.

Chris Careaga and Yağız Aksoy. Intrinsic image decomposition via ordinal shading. *ACM Transactions on Graphics*, 43(1), 2023.

Chris Careaga and Yağız Aksoy. Colorful diffuse intrinsic image decomposition in the wild. *ACM Transactions on Graphics*, 43(6), 2024.

Xi Chen, Sida Peng, Dongchen Yang, Yuan Liu, Bowen Pan, Chengfei Lv, and Xiaowei Zhou. IntrinsicAnything: Learning diffusion priors for inverse rendering under unknown illumination. In *ECCV*, 2024.

Alexander D'Amour, Katherine Heller, Dan Moldovan, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

Partha Das, Sezer Karaoglu, and Theo Gevers. PIE-Net: Photometric invariant edge guided network for intrinsic image decomposition. In *CVPR*, 2022.

Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. Revisiting deep intrinsic image decompositions. In *CVPR*, 2018.

Elena Garces, Carlos Rodriguez-Pardo, Dan Casas, and Jorge Lopez-Moreno. A survey on intrinsic images: Delving deep into lambert and beyond. *International Journal of Computer Vision*, 130, 2022.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2, 2020.

Arjan Gijsenij, Theo Gevers, and Joost van de Weijer. Computational color constancy: Survey and experiments. *IEEE Transactions on Image Processing*, 20(9), 2011.

Roger Grosse, Micah K. Johnson, Edward H. Adelson, and William T. Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *ICCV*, 2009.

Daniel J. Jobson, Zia-ur Rahman, and Glenn A. Woodell. Properties and performance of a center/surround retinex. *IEEE Transactions on Image Processing*, 6(3), 1997.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024.

Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. Marigold: Affordable adaptation of diffusion-based image generators for image analysis. *arXiv preprint arXiv:2505.09358*, 2025.

Peter Kocsis, Vincent Sitzmann, and Matthias Nießner. Intrinsic image diffusion for indoor single-view material estimation. In *CVPR*, 2024.

Edwin H. Land and John J. McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 61(1), 1971.

Louis Lettry, Kenneth Vanhoey, and Luc Van Gool. Unsupervised deep single-image intrinsic decomposition using illumination-varying image sequences. *Computer Graphics Forum*, 37(7), 2018.

Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In *CVPR*, 2018a.

Zhengqi Li and Noah Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *ECCV*, 2018b.

Zhibing Li, Tong Wu, Jing Tan, Mengchen Zhang, Jiaqi Wang, and Dahua Lin. Idarb: Intrinsic decomposition for arbitrary number of input views and illuminations. In *ICLR*, 2025.

Yunfei Liu, Yu Li, Shaodi You, and Feng Lu. Unsupervised learning for intrinsic image decomposition from a single image. In *CVPR*, 2020.

Jundan Luo, Zhaoyang Huang, Yijin Li, Xiaowei Zhou, Guofeng Zhang, and Hujun Bao. NIID-Net: Adapting surface normal knowledge for intrinsic image decomposition in indoor scenes. *IEEE TVCG*, 26(12), 2020.

Lukas Murmann, Michaël Gharbi, Miika Aittala, and Frédo Durand. A dataset of multi-illumination images in the wild. In *ICCV*, 2019.

Takuya Narihira, Michael Maire, and Stella X. Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *ICCV*, 2015.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*, 2021.

Kalyan Sunkavalli, Wojciech Matusik, Hanspeter Pfister, and Szymon Rusinkiewicz. Factored time-lapse video. *ACM Transactions on Graphics (SIGGRAPH)*, 26(3), 2007.

Marshall F. Tappen, William T. Freeman, and Edward H. Adelson. Recovering intrinsic images from a single image. *IEEE TPAMI*, 27(9), 2005.

Yair Weiss. Deriving intrinsic images from image sequences. In *ICCV*, 2001.

Jiaye Wu, Sanjoy Chowdhury, Hariharmano Shanmugaraja, David Jacobs, and Soumyadip Sengupta. Measured albedo in the wild: Filling the gap in intrinsics evaluation. In *IEEE International Conference on Computational Photography (ICCP)*, 2023.

Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. Rgb↔x: Image decomposition and synthesis using material- and lighting-aware diffusion models. *ACM SIGGRAPH Conference Papers*, 2024.

Jingsen Zhu, Fujun Luan, Yuchi Huo, Zihao Lin, Zhihua Zhong, Dianbing Xi, Rui Wang, Hujun Bao, Jiaxiang Zheng, and Rui Tang. Learning-based inverse rendering of complex indoor scenes with differentiable monte carlo raytracing. In *SIGGRAPH Asia*, 2022.

A Additional Quantitative Results

Table 3: Adapter ablations and trivial mitigation baselines (multilum test; WHDR on the deterministic IIW subsets stated in §5). Smoothing either fails to reduce drift or pays fidelity; only the consistency adapter reaches the empty corner, and λ traces the trade-off.

Variant	Drift	WHDR
base C&A	0.547	0.217
+ bilateral filter	0.550	0.200
+ Gaussian blur	0.474	0.256
adapter, λ=3 (10k)	0.408	0.204
adapter, λ=1 (ours, 20k)	**0.277**	**0.210**
adapter, λ=0.3 (10k)	0.053	0.249
adapter, albedo-only input (10k)	0.272	0.217

B Additional Qualitative Results

Table 4: Control experiments: none of the following explains the measured drift. All numbers on multilum test (30 scenes unless stated).

Control	Result
Sampling seed (diffusion, 3 scenes $\times$ 5 seeds)	seed variation $= 1/25$ of illumination variation
Ensembling (n=10, recommended settings)	drift $3.185 \rightarrow 3.153$
Per-channel normalization	$0.547 \rightarrow 0.596$ (C&A), $3.185 \rightarrow 3.150$ (Marigold)
Linear-RGB input	$0.547 \rightarrow 1.302$ (worse; sRGB convention dependence)
Test-time photometric augmentation	drift $+3\%$ (no reduction)
Diffuse-only masking (Eq. 7)	$0.547 \rightarrow 0.500$ (persists)
Reference-frame choice (relighting, $r \in \{0, 12, 24\}$)	ranking unchanged



Figure 5: Full panel for multilum `everett_kitchen2` (the scene with the largest base drift, 1.37). Rows: lighting directions. Columns: input, C&A albedo, adapter albedo, relit (C&A), relit (+adapter). Albedo columns show the drift directly: brightness and color of the C&A albedo follow the illumination.



Figure 6: Full panel for multilum `everett_dining1`, same layout.