TRASH OR TREASURE? RETHINKING POLARIZATION FOR REFLECTION REMOVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Polarization provides physical cues for reflection removal beyond RGB appearance, but how much modern polarization-based methods actually exploit them remains unclear. We revisit this question through a controlled counterfactual analysis that disentangles genuine polarization information from the learned restoration prior. Our analysis reveals that strong polarization-based performance does not necessarily imply strong polarization utilization, and that the benefit of polarization is highly conditional: it is most pronounced when informative physical cues coincide with ambiguity in RGB restoration, but can be redundant or detrimental otherwise. Based on this insight, we derive physically distinct observations from the same polarized capture to expose complementary restoration hypotheses using a shared frozen RGB restorer, and integrate them with polarization evidence through a lightweight fusion module rather than training another dedicated polarization restorer. Across three frozen generative restorers, this consistently improves strong RGB baselines without modifying the underlying restorer. These results recast polarization as *conditional physical evidence* rather than an always-on modality for reflection removal.

1 INTRODUCTION

Reflection removal aims to recover the transmitted scene from an image captured through glass, where the desired transmission is superimposed with unwanted reflections. From a single RGB observation, however, the problem is intrinsically ambiguous: structures from either layer can be visually plausible. Single-image methods address this ambiguity with progressively stronger priors, ranging from hand-crafted statistics (Levin & Weiss, 2007; Li & Brown, 2014) and supervised networks (Zhang et al., 2018b; Hu & Guo, 2023) to more recent pretrained generative restorers (Hu et al., 2026; Li et al., 2026; Zakarin et al., 2026). Despite this progress, they still infer the decomposition from appearance alone. Multi-image approaches instead introduce additional observations through viewpoint or motion (Szeliski et al., 2000; Xue et al., 2015), dual-pixel measurements (Punnappurath & Brown, 2019), or active illumination (Lei & Chen, 2021), but require more restrictive capture settings.

Polarization (Können, 1985) provides a particularly attractive source of additional physical information. Reflected and transmitted light often exhibit different polarization states, such that polarized images captured at different polarizer angles contain cues unavailable in a conventional RGB observation (Wolff, 1989; Nayar et al., 1997). This property has motivated approaches ranging from explicit physical decomposition (Kong et al., 2013; Schechner et al., 1999) to learned models that jointly process multiple polarized images (Lyu et al., 2019; Lei et al., 2020; Yao et al., 2025). The underlying intuition is straightforward: if polarization provides measurements that respond differently to the two superimposed layers, a successful polarization-based model should derive measurable benefit from the variation across polarizer angles.

We find that this intuition does not generally hold. When the four polarized images fed to Polar-Free (Yao et al., 2025), a recent diffusion-based model, are replaced by four identical copies of the unpolarized image, its average PSNR on the PolaRGB (Yao et al., 2025) test split drops by only 0.16 dB (Fig. 1(a)), retaining over 97% of its improvement over the mixed observation; on individual captures, the genuine measurement helps in one case and hurts in another. Yet polarization is far from useless. Feeding the frozen RGB restorer WindowSeat (Zakarin et al., 2026) the minimum-polarizer image, which suppresses the measured polarized component, instead of the ordinary mixed image

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.

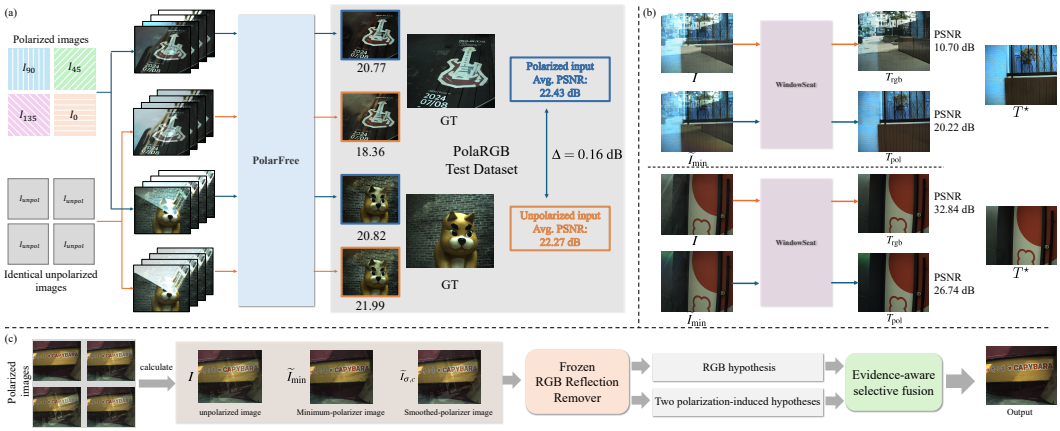


Figure 1: **Rethinking polarization for reflection removal.** (a) PolarFree (Yao et al., 2025) on the 188-image PolaRGB (Yao et al., 2025) test split with genuine polarized inputs (blue) or four copies of the unpolarized image (orange): the average PSNR differs by only 0.16 dB, while individual captures can gain (top) or lose (bottom) from the genuine measurement. (b) A frozen RGB restorer WindowSeat (Zakarin et al., 2026) given I or $\widetilde{I}_{\min}$: suppressing the polarized component rescues a strongly polarized, poorly restored capture (top) but degrades an already accurate one (bottom). (c) Our approach: I, $\widetilde{I}_{\min}$, and the smoothed $\widetilde{I}_{\sigma,c}$ from one capture are restored by the same frozen restorer, and an evidence-aware module selectively fuses the three hypotheses.

raises a strongly polarized capture that the restorer handles poorly from 10.70 to 20.22 dB, but lowers an already well-restored capture from 32.84 to 26.74 dB (Fig. 1(b)). Model performance thus does not establish polarization utilization, and the value of polarization is conditional, raising two questions:

- *When and where does polarization help beyond a strong RGB prior?*
- *How should it be integrated when it helps only conditionally?*

We study these questions by holding the learned prior fixed and varying only the polarization measurement. Released polarization-based models depend weakly on genuine polarization on average, mainly on strongly polarized inputs. Applying three unrelated frozen RGB restorers to physically distinct observations of the same capture then shows when polarization is complementary: the gain concentrates where predominantly reflection-driven polarization coincides with an ambiguous RGB restoration, whereas suppressing the measured polarized component also attenuates polarized transmission and harms restorations that are already accurate. Polarization strength alone therefore cannot decide where polarization should be used.

We accordingly treat polarization as *conditional physical evidence* rather than an always-on input modality (Fig. 1(c)). From one polarized capture we derive the mixed observation, a minimum-polarizer observation that removes the full measured polarized component, and a smoothed observation that removes only a scaled low-frequency part of it; these trade reflection suppression against transmission preservation differently. All three are restored by the *same* frozen RGB restorer, and a lightweight evidence-aware module combines the three hypotheses per pixel from observable cues, without retraining or modifying the restorer. Our contributions are:

- **Conditional value of polarization.** We show that a small average contribution can conceal substantial benefits on strongly polarized inputs, and identify where polarization improves inaccurate RGB restorations or harms already accurate predictions.
- **Complementary polarization-induced hypotheses.** Motivated by this conditional benefit, we construct observations with no, full, and spatially smoothed polarization suppression, providing complementary hypotheses through a shared frozen RGB restorer.
- **Evidence-aware selective fusion.** We design a lightweight fusion module that spatially combines complementary restoration hypotheses using observable polarization cues. Its effectiveness is validated on three different pretrained RGB reflection removal models, improving restoration quality without modifying the underlying restorers.

2 Related Work

RGB-based reflection removal. Single-image reflection removal must separate transmitted and reflected layers from a single mixed RGB observation, and therefore depends strongly on image priors. Early methods exploit hand-crafted assumptions such as gradient sparsity (Levin & Weiss, 2007; Li & Brown, 2014), relative smoothness (Wan et al., 2017), and ghosting (Shih et al., 2015), while learning-based methods replace these priors with supervised networks (Zhang et al., 2018b; Yang et al., 2018) and more structured restoration models (Hu & Guo, 2023; Hu et al., 2024). Recently, language guidance (Zhong et al., 2024; Hong et al., 2024) and pretrained generative models (Zakarin et al., 2026; Hu et al., 2026; Li et al., 2026) provide stronger semantic priors. Multi-image methods instead reduce ambiguity through additional observations from motion or viewpoint changes (Xue et al., 2015), dual-pixel sensing (Punnappurath & Brown, 2019), or active illumination (Lei & Chen, 2021), but require more restrictive capture conditions. Our work starts from the regime where the RGB restoration prior is already strong, and asks what an additional physical measurement can still contribute.

Polarization-based reflection removal. Polarization provides such a physical cue because reflected and transmitted light often exhibit different polarization states. Classical approaches exploit this difference through polarization-space analysis (Wolff, 1989; Nayar et al., 1997), layer decorrelation (Schechner et al., 1999), or explicit physical inversion (Kong et al., 2013). Learned methods instead incorporate polarized observations through paired inputs (Lyu et al., 2019), dedicated multi-stage or recurrent architectures (Lei et al., 2020; Bian et al., 2026), and more recently diffusion-based conditioning (Yao et al., 2025). While effective, these methods typically couple the polarization measurement with the restoration architecture and training procedure, making it difficult to isolate how much of the final performance is attributable to genuine polarization information.

3 Rethinking the Role of Polarization

We disentangle two questions that are often conflated: *utilization*, whether a learned polarization-based model actually depends on the polarization information provided at inference, and *utility*, when that information adds value beyond a strong RGB restoration prior.

3.1 Polarized Image Formation for Reflection Removal

Reflection removal considers a mixed RGB observation $I = T + R$, where T and R denote the transmitted and reflected layers; the goal is to recover T, and we write $T^\star$ for the ground-truth transmission. When a linear polarizer is placed in front of the camera, the observed image varies with its orientation. Let $\mathcal{P} = \{I_\theta\}_{\theta\in\Theta}$ denote four registered polarized images captured at polarizer angles $\Theta = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$. Under the linear Stokes model (Können, 1985), applied independently to each color channel,

$$I_\theta = \tfrac{1}{2}(S_0 + S_1\cos 2\theta + S_2 \sin 2\theta),\ S_0 = I = \tfrac{1}{2}(I_0 + I_{45} + I_{90} + I_{135}), \tag{1}$$

where S_0 is the total intensity before the polarizer, i.e. the mixed observation I, and $S_1 = I_0 - I_{90}$ and $S_2 = I_{45} - I_{135}$ describe linear polarization along the $0^\circ/90^\circ$ and $45^\circ/135^\circ$ bases. Each polarized image therefore consists of an angle-independent component $I/2$ and an angle-dependent variation,

$$I_\theta = \frac{I}{2} + d_\theta, \qquad d_\theta = \frac{1}{2}\left(S_1\cos 2\theta + S_2 \sin 2\theta\right), \qquad \sum_{\theta\in\Theta} d_\theta = 0. \tag{2}$$

Here d_θ isolates the part of the observation that changes with polarizer angle, while the underlying mixed image I remains unchanged. The Stokes representation also provides

$$\rho = \frac{\sqrt{S_1^2 + S_2^2}}{S_0} \quad \text{and} \quad I_{\min} = \frac{1}{2}\left(S_0 - \sqrt{S_1^2 + S_2^2}\right) = \frac{1}{2} I(1-\rho), \tag{3}$$

where ρ is the degree of polarization (DoP), characterizing the strength of linear polarization, and $I_{\min}$ is the minimum intensity attainable over polarizer orientation. When reflection is more strongly polarized than transmission, $I_{\min}$ attenuates part of it, but it is not an estimate of transmission: weakly polarized reflection may remain, and polarized transmission is suppressed too. The decomposition

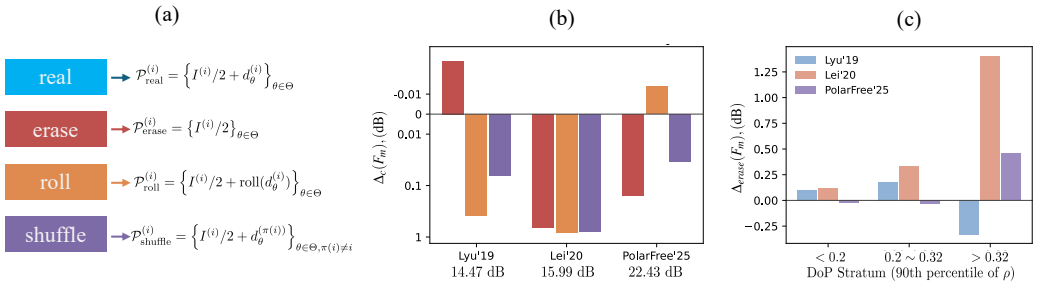


Figure 2: **Counterfactual diagnosis of polarization utilization.** (a) Starting from the same mixed observation I, we erase, spatially shift, or cross-scene replace only the angle-dependent variation d_θ while preserving the shared component $I/2$. (b) Across representative learned polarization-based models, the resulting counterfactual sensitivities differ substantially, and strong restoration performance does not necessarily imply strong dependence on genuine polarization. (c) Polarization utilization is also input-dependent: sensitivity to erasing d_θ increases with polarization strength (the DoP ρ), showing that genuine polarization contributes more strongly to some inputs than to others.

$I_\theta = I/2 + d_\theta$ lets us test whether learned models depend on the angle-dependent information (Sec. 3.2), while ρ and $I_{\min}$ let us study when polarization is complementary to RGB appearance (Sec. 3.3).

3.2 Counterfactual Diagnosis of Polarization Utilization

We first examine how the angle-dependent information d_θ affects the restoration quality of fixed polarization-based models. We consider three representative methods: Lyu et al. (2019), Lei et al. (2020), and PolarFree (Yao et al., 2025), covering paired-input, two-stage, and diffusion-based designs, respectively. For each method m, we use $F_m(\mathcal{P})$ to denote its complete inference pipeline, since different methods may transform the polarized capture into different input representations or derived polarization cues. Under each intervention, all model-specific inputs are recomputed from the modified polarized capture using the original preprocessing procedure. All methods are evaluated on the real-world polarized reflection-removal benchmark PolaRGB (Yao et al., 2025).

A standard ablation that removes polarization inputs and retrains the network cannot directly reveal how much a trained model utilizes polarization at inference time, since retraining may also alter the learned restoration prior. We instead keep each pretrained model fixed and intervene only on the angle-dependent variation d_θ. For a test sample i, we take the genuine polarized capture as reference and first construct its polarization-erased counterpart:

$$\mathcal{P}^{(i)}_{\text{real}} = \left\{ I^{(i)}/2 + d^{(i)}_\theta \right\}_{\theta\in\Theta} \quad \text{and} \quad \mathcal{P}^{(i)}_{\text{erase}} = \left\{ I^{(i)}/2 \right\}_{\theta\in\Theta}. \tag{4}$$

The *erase* construction removes angle-dependent variation while preserving $S_0 = I$. To further test whether a model relies on the correct correspondence of the polarization signal, we introduce two additional interventions. For *roll*, we replace $d^{(i)}_\theta$ with a circularly shifted version $\text{roll}(d^{(i)}_\theta)$, using the same shift across polarizer angles to break its alignment with local image structure. For *shuffle*, we replace $d^{(i)}_\theta$ with $d^{(\pi(i))}_\theta$ from another scene, where $\pi(i) \neq i$, thereby disrupting scene-level correspondence while preserving the mixed observation $I^{(i)}$.[1] In all cases, the underlying mixed observation $I^{(i)}$ and the pretrained model remain unchanged. Fig. 2(a) summarizes the three interventions. For $c \in \{\text{erase}, \text{roll}, \text{shuffle}\}$, we quantify the resulting counterfactual sensitivity of model F_m as

$$\Delta_c(F_m) = \frac{1}{N}\sum_{i=1}^{N}\left[Q\left(F_m(\mathcal{P}^{(i)}_{\text{real}}), T^\star_i\right) - Q\left(F_m(\mathcal{P}^{(i)}_c), T^\star_i\right)\right], \tag{5}$$

where $T^\star_i$ is the i-th ground-truth transmission, Q denotes a restoration-quality metric for which larger values are better, and N stands for the total number of test samples. Δ_{erase} measures sensitivity to

[1] Roll uses a fixed half-frame shift along both image axes; shuffle uses one fixed donor capture per image. Details are given in Appendix A.

the presence of genuine angle-dependent polarization, while Δ_{roll} and $\Delta_{\mathrm{shuffle}}$ further probe whether its correct spatial and scene-level correspondence matters to the final restoration. These quantities characterize polarization utilization through observable changes in restoration quality rather than the model's internal representation.

As shown in Fig. 2(b), the evaluated methods exhibit markedly different sensitivities to the polarization interventions. Erasing genuine angle-dependent variation changes restoration quality by -0.04 to +0.66 dB, while disrupting its spatial and scene-level correspondence produces changes of -0.01 to +0.83 dB and +0.03 to +0.80 dB, respectively. Importantly, these counterfactual sensitivities are not necessarily aligned with the original restoration performance of the models. Strong performance from a polarization-based method therefore does not, by itself, establish strong utilization of the underlying physical measurement. The dependence also varies across inputs (Fig. 2(c)). For Lei et al. (2020) and PolarFree (Yao et al., 2025), the contribution of genuine polarization stays between -0.03 and $+0.33$ dB for images with $\rho_{90} \leq 0.32$ but rises to $+1.41$ and $+0.46$ dB, respectively, in the strongly polarized stratum. For PolarFree (Yao et al., 2025), erasing polarization costs more than 0.5 dB on 42% of strongly polarized images, compared with only 4% of the remaining test images. Lyu et al. (2019) is the exception on this benchmark (-0.33 dB in the high-DoP stratum), but erasing polarization on its own test set reduces PSNR by 4.95 dB. Its weak response on PolaRGB (Yao et al., 2025) should therefore not be generalized to its native domain (Appendix A). Polarization is thus neither uniformly exploited nor simply ignored: its contribution depends on both the model and the input, which raises a more fundamental question: *when and where does polarization actually provide an advantage over an RGB restoration prior?*

3.3 When and Where Does Polarization Help?

To examine utility independently of polarization-specific training, we evaluate three pretrained RGB reflection removers G_k on all 6500 PolaRGB (Yao et al., 2025) images: DAI (Hu et al., 2026), GenSIRR (Li et al., 2026), and WindowSeat (Zakarin et al., 2026), covering one-step latent, flow-matching, and LoRA-adapted diffusion-transformer designs; none is trained on PolaRGB (Yao et al., 2025), and all remain fixed. To probe what polarization can contribute beyond I, we construct a second observation whose deviation from I is determined solely by the measured polarization. A single polarized image I_θ is unsuitable because its appearance depends on the scene-specific polarization orientation; we instead use $I_{\min}$, which maximally suppresses the measured linearly polarized component without additional parameters. Compensating for the factor $1/2$ of an ideal polarizer gives $\widetilde{I}_{\min} = 2I_{\min} = S_0 - \sqrt{S_1^2 + S_2^2} = I(1-\rho)$, which equals I in the absence of linear polarization and deviates from it only according to the measured polarization strength. Each frozen restorer thus produces two hypotheses for sample i,

$$T_{\mathrm{rgb}}^{(i,k)} = G_k(I^{(i)}) \quad \text{and} \quad T_{\mathrm{pol}}^{(i,k)} = G_k(\widetilde{I}_{\min}^{(i)}), \tag{6}$$

the *RGB hypothesis* and the *polarization-induced hypothesis*. Since each pair is produced by the same frozen model, their difference arises solely from the observation presented to the shared prior. We quantify the relative benefit of polarization by the *polarization advantage*

$$A_{\mathrm{pol}}^{(i)}(G_k) = Q(T_{\mathrm{pol}}^{(i,k)}, T_i^\star) - Q(T_{\mathrm{rgb}}^{(i,k)}, T_i^\star), \tag{7}$$

where positive values favor the polarization-induced hypothesis; the same comparison can be averaged over subsets or evaluated over local regions.

Across the three restorers, Fig. 3(a) shows that neither hypothesis is uniformly preferable. Using $\widetilde{I}_{\min}$ for all inputs changes average restoration quality by -1.16 to -0.63 dB, but its effect varies sharply with both polarization strength and RGB restoration quality. When strong polarization coincides with poor RGB restoration (high DoP, $T_{\mathrm{rgb}} < 15$ dB), A_{pol} rises to between +2.62 and +4.57 dB, whereas it falls to between -2.73 and -1.30 dB once the RGB hypothesis exceeds 25 dB. The same conditional structure appears on all three backbones despite their different architectures.

Because regions with stronger reflections can be both harder for RGB restoration and more strongly polarized, this trend could partly reflect reflection severity. We therefore compare regions within the same image under similar reflection contamination (Fig. 3(b)). Among strongly reflected regions of high-DoP images, strongly polarized regions retain a $+2.36$ dB larger polarization advantage than weakly polarized ones on average ($+1.44$ to $+3.16$ dB across backbones), and the difference persists

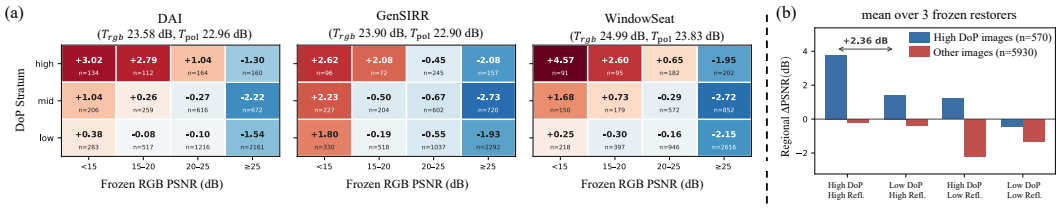


Figure 3: (a) A_{pol} on all 6500 PolaRGB (Yao et al., 2025) images for three frozen restorers, by DoP stratum (rows) and PSNR of T_{rgb} (columns), with image counts; headers give average PSNR of T_{rgb} and T_{pol}. Gains concentrate where strong polarization meets a difficult RGB restoration. (b) Within-image control averaged over the restorers: pixels are split at the within-image 80th percentile of DoP and of reflection strength $|I - T^{\star}|$; bars give regional A_{pol} for high-DoP images (blue, $n = 570$) and the rest (red).

under weak reflection (+1.62 dB); the same holds when reflection strength is estimated without ground truth (Appendix B). The association with DoP persists after this coarse control for reflection strength; the comparison does not hold all aspects of RGB restoration difficulty fixed. High DoP does not by itself guarantee a better T_{pol}, however: when the transmitted layer is also polarized, minimizing the polarizer response suppresses useful scene content, so ρ characterizes the strength of polarization but not whether the cue is beneficial.

Together, these results reveal a complementary relationship: T_{pol} improves regions that remain difficult for RGB restoration, whereas T_{rgb} is preferable where polarization is weak, misleading, or unnecessary, and their relative advantage varies across models, scenes, and regions of one image. This motivates a spatial mechanism that decides where the polarization-induced hypothesis should complement the RGB restoration.

4 Methodology

Sec. 3.3 reveals two complementary restoration hypotheses from the same frozen RGB prior: the RGB hypothesis preserves the ordinary mixed observation, whereas the polarization-induced hypothesis exposes the effect of suppressing the measured polarization component. Since their relative advantage varies spatially, the remaining problem is not to train another polarization-specific restorer, but to determine how these hypotheses should be combined according to the available physical evidence. We therefore keep the RGB restorer frozen and introduce a lightweight evidence-aware fusion module. Besides the restoration hypotheses, the module receives polarization-derived observations and cues that characterize the measured polarization itself. This allows existing RGB restorers to incorporate polarization without retraining or architectural modification. Fig. 4 illustrates the resulting framework.

4.1 Evidence-Aware Selective Fusion

For clarity, we omit the sample and backbone indices used in Sec. 3.3 and retain the hypotheses $T_{\text{rgb}} = G(I)$ and $T_{\text{pol}} = G(\widetilde{I}_{\text{min}})$, where G denotes the frozen RGB restorer.

A third, smoothed observation. The polarization-induced observation can be written as

$$\widetilde{I}_{\text{min}} = I - |S_{\text{pol}}|, \qquad |S_{\text{pol}}| = \sqrt{S_1^2 + S_2^2} = I\rho. \tag{8}$$

Here, $|S_{\text{pol}}|$ is the magnitude of the combined linear-polarization signal from the transmitted and reflected layers. It does not identify which layer produces the measured polarization. Consequently, subtracting this magnitude at every pixel can suppress useful transmitted content as well as reflection.

To expose an alternative restoration hypothesis, we smooth the measured polarization magnitude before subtraction:

$$\widetilde{I}_{\sigma,c} = \mathcal{N}_I \left[I - c\, \mathcal{G}_\sigma \left(|S_{\text{pol}}| \right) \right], \tag{9}$$

where $\mathcal{G}_\sigma$ is an isotropic Gaussian filter with standard deviation σ, c controls the subtraction strength, and $\mathcal{N}_I(\cdot)$ clips to $[0, 1]$ and rescales the result to the mean luminance of I. Smoothing reduces high-frequency variation in the subtraction field, while c and luminance normalization control the resulting photometric change. We use this construction as an empirical regularization of polarization

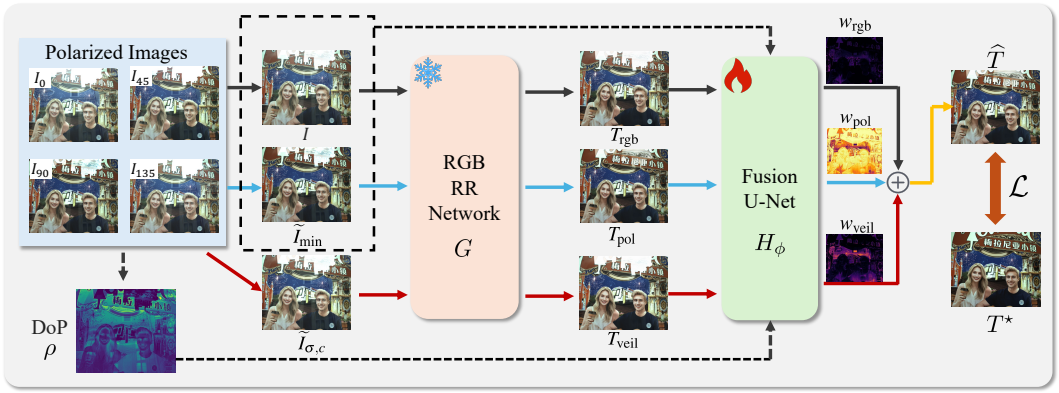


Figure 4: **Evidence-aware selective fusion.** I, $\widetilde{I}_{\min}$, and $\widetilde{I}_{\sigma,c}$ are processed by the same frozen RGB restorer G, yielding T_{rgb}, T_{pol}, and T_{veil}. The fusion U-Net H_ϕ, the only trained component, reads them with observable evidence $\mathcal{E}$ and returns per-pixel convex weights w_{rgb}, w_{pol}, and w_{veil}.

suppression, without assuming that reflection and transmission occupy separable frequency bands. It provides a complementary input whose utility is evaluated through the same frozen restorer and the subsequent selective fusion. We use $c = 0.75$ for all three backbones and set $\sigma = 128$ for DAI (Hu et al., 2026) and GenSIRR (Li et al., 2026), and $\sigma = 384$ for WindowSeat (Zakarin et al., 2026), measured in pixels at the native resolution of 1224×1024. The parameter-selection procedure is provided in Appendix C.

Three hypotheses from one frozen restorer. Applying the same G to the third observation yields

$$T_{\text{rgb}} = G(I), \qquad T_{\text{pol}} = G(\widetilde{I}_{\min}), \qquad T_{\text{veil}} = G(\widetilde{I}_{\sigma,c}). \tag{10}$$

They differ only in the observation presented to a shared prior. T_{pol} reflects pixelwise suppression of the measured polarization magnitude, whereas T_{veil} reflects spatially smoothed suppression with luminance normalization. T_{rgb} provides the restoration of the original mixed observation. These transformations can alter both reflection and transmitted content, so no hypothesis is assumed to be uniformly preferable.

Observable evidence. The decision must be taken without the ground-truth transmission, so for every pixel we form an 18-channel evidence tensor from quantities available at inference,

$$\mathcal{E} = \left[T_{\text{rgb}},\, T_{\text{pol}},\, T_{\text{veil}},\, I,\, \widetilde{I}_{\min},\, \rho,\, D_{\text{pol}},\, D_{\text{veil}}\right], \tag{11}$$

where $D_{\text{pol}} = \|T_{\text{rgb}} - T_{\text{pol}}\|_1$ and $D_{\text{veil}} = \|T_{\text{rgb}} - T_{\text{veil}}\|_1$ are channel-averaged disagreement maps. ρ indicates where appreciable polarization is present and the disagreement maps where changing the observation materially changed the restoration. Neither cue suffices alone—strong polarization may originate from the transmitted layer, and disagreement signals sensitivity rather than correctness—so they serve as evidence for a learned decision rather than as selection rules.

Convex fusion. A lightweight network H_ϕ maps the evidence to one logit per hypothesis and normalizes them across hypotheses:

$$\widehat{T} = \sum_{k \in \mathcal{K}} w_k \odot T_k, \qquad w = \text{softmax}\left(H_\phi(\mathcal{E})\right), \qquad \mathcal{K} = \{\text{rgb}, \text{pol}, \text{veil}\}, \tag{12}$$

so that $\sum_k w_k = 1$ and $w_k \geq 0$ at every pixel. Convexity is a deliberate restriction: every output color lies in the convex hull of predictions the frozen restorer already produced, so the module cannot add a free-form correction and any improvement is attributable to *selecting* among existing hypotheses.

4.2 Training and Implementation

Loss function. The RGB restoration backbone G remains frozen throughout training, and only the parameters ϕ of the fusion module are optimized. Given the estimated transmission $\widehat{T}$ and the corresponding ground-truth transmission $T^\star$, we optimize

$$\mathcal{L} = \left\|\widehat{T} - T^\star\right\|_1 + \lambda_{\text{ssim}} \left[1 - \text{SSIM}\left(\widehat{T}, T^\star\right)\right], \tag{13}$$

Table 1: **Ablation of the hypothesis set on DAI** (Hu et al., 2026) (all 6500 images, out-of-fold). ΔPSNR is over the frozen RGB hypothesis (23.58 dB); "improved" is the share of images with positive ΔPSNR. "Always" rows use one hypothesis everywhere; "fuse" rows share the fusion recipe. Other backbones: Table 4.

			ΔPSNR by DoP stratum		
hypotheses	ΔPSNR	improved	low	mid	high
always T_{pol}	-0.626	33.8%	-0.807	-0.785	$+1.192$
always T_{veil}	$+0.198$	72.1%	$+0.207$	$+0.180$	$+0.194$
fuse $\{T_{\mathrm{rgb}}, T_{\mathrm{veil}}\}$	$+0.263$	**80.4%**	$+0.219$	$+0.281$	$+0.528$
fuse $\{T_{\mathrm{rgb}}, T_{\mathrm{pol}}\}$	$+0.450$	61.9%	$+0.264$	$+0.445$	$+1.824$
fuse $\{T_{\mathrm{rgb}}, T_{\mathrm{pol}}, T_{\mathrm{veil}}\}$ (ours)	$\mathbf{+0.543}$	71.7%	$\mathbf{+0.350}$	$\mathbf{+0.538}$	$\mathbf{+1.975}$

where $\mathrm{SSIM}(\cdot,\cdot)$ denotes the structural similarity index measure and λ_{ssim} balances the pixel-wise reconstruction and structural similarity terms. Gradients are propagated only through H_ϕ, while the three restoration hypotheses are treated as fixed outputs of the frozen backbone.

Implementation details. We instantiate H_ϕ as a lightweight convolutional U-Net (Ronneberger et al., 2015). The backbone predictions are precomputed and detached during fusion training, and the parameters in Eq. 9 are selected on a held-out validation set and fixed thereafter. Training and evaluation use five folds split by ground-truth target, so that captures sharing a target never appear on both sides (Sec. 5.1). At inference time, the frozen RGB restorer is evaluated only on I, $\widetilde{I}_{\min}$ and $\widetilde{I}_{\sigma,c}$, followed by a single forward pass through the fusion module; no test-time optimization or backbone adaptation is required. Architecture, optimization, hyperparameter, and data-split details are provided in Appendix D.

5 Experiments

5.1 Setup

Data and backbones. PolaRGB (Yao et al., 2025) contains 6500 four-angle polarized captures with ground-truth transmissions; we use all of them and additionally report its 188-image official test split. The three frozen restorers of Sec. 3.3—DAI (Hu et al., 2026), GenSIRR (Li et al., 2026), and WindowSeat (Zakarin et al., 2026)—differ in architecture and training, none was trained on PolaRGB (Yao et al., 2025), and none is modified here. **Protocol.** The fusion module is the only trained component. We split all 6500 images into five folds *by ground-truth target*, so that images sharing a target never straddle training and evaluation, and report every image with the module trained on the other four folds. We report native-resolution PSNR (Wang & Bovik, 2009) and SSIM (Wang et al., 2004); perceptual and structural metrics are in Appendix I. Results are further stratified by DoP, placing each image by the 90th percentile of its per-pixel ρ with the cuts of Fig. 2(c): low (< 0.20, 4177 images), mid (0.20–0.32, 1753 images), and high (> 0.32, 570 images).

5.2 Selective Fusion versus Unconditional Polarization

Main results. Using the polarization-induced hypothesis everywhere is harmful on every backbone, and increasingly so as the prior gets stronger (-0.63, -1.00, and -1.16 dB for DAI (Hu et al., 2026), GenSIRR (Li et al., 2026), and WindowSeat (Zakarin et al., 2026)). Selective fusion turns the same measurement into a consistent gain of +0.54, +0.66, and +0.54 dB, improving 71.7%, 76.3%, and 74.3% of images, and it improves every DoP stratum on every backbone (Table 4). Crossing the three DoP strata with the RGB-quality bins of Fig. 3(a), unconditional use is negative in 20 of 36 cells across the three backbones; after fusion only 1 remains (-0.08 dB), while the high-DoP, low-quality cells keep their large gains (Appendix E.1). Cheaper selection does not suffice: on DAI (Hu et al., 2026), a learned image-level router gains $+0.03$ dB between T_{rgb} and T_{pol} and $+0.19$ dB among all three hypotheses, and a spatial threshold on ρ gains $+0.06$ dB (Appendix F).

Which hypotheses are needed. In Table 1, neither standalone transformed hypothesis matches the full fusion system: T_{pol} reduces mean PSNR by 0.63 dB, whereas T_{veil} gains 0.20 dB and improves 72.1% of images. Among the evaluated pairs that retain the RGB hypothesis, $\{T_{\mathrm{rgb}}, T_{\mathrm{veil}}\}$ improves the most images but gives up the strongly polarized peak, while $\{T_{\mathrm{rgb}}, T_{\mathrm{pol}}\}$ keeps that peak but

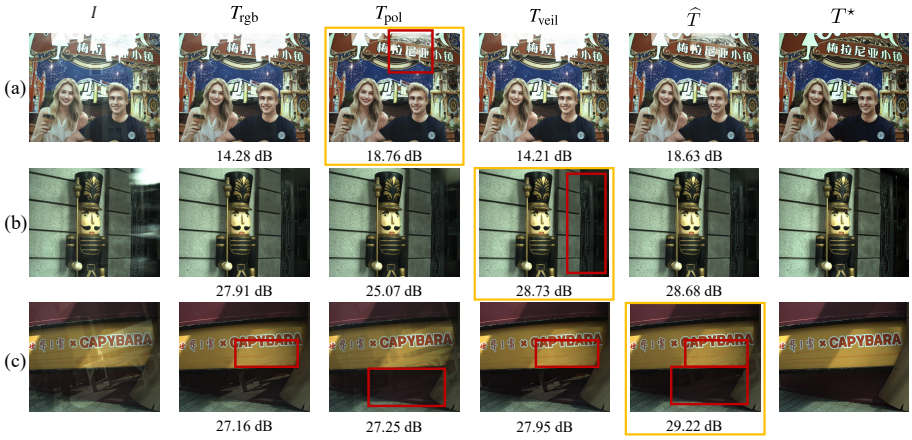


Figure 5: **The hypotheses succeed and fail in different places** (DAI (Hu et al., 2026), out-of-fold; full-frame PSNR below each image, yellow frame = best per row, red boxes = regions where hypotheses differ). (a) Strong polarization: T_{pol} removes the reflection and fusion keeps its gain. (b) Weak polarization: T_{pol} damages the scene, T_{veil} gains, and fusion follows it. (c) Each hypothesis is right in a different region, so fusion exceeds all three.

improves the fewest images. All three together exceed both evaluated pairs in mean gain and in every stratum.

In the learned weights, T_{veil} takes the largest share (0.471; T_{pol} 0.394, T_{rgb} 0.135): the conservative hypothesis absorbs the fallback role, freeing the module to commit to the aggressive one where it pays. Across backbones, the third hypothesis adds +0.06 to +0.16 dB, and its relative contribution is largest in the low-DoP stratum; on WindowSeat (Zakarin et al., 2026) it raises that stratum from 0.24 to 0.35 dB while the high-stratum gain changes slightly from 1.698 to 1.671 dB.

5.3 Transfer and Comparison

Zero-shot transfer across backbones. The full three-hypothesis fusion module transfers across backbones without adaptation: applied to the hypotheses of another frozen restorer, all six ordered directions yield positive mean gains, retaining 32–101% of the target's in-domain gain (Appendix G).

Comparison among methods. On the 188-image official split, scored by one native-resolution harness, the in-domain polarization model PolarFree (Yao et al., 2025) reaches 22.44 dB, whereas the frozen WindowSeat (Zakarin et al., 2026) already reaches 24.75 dB without polarization and selective fusion raises it to 25.83 dB; DAI (Hu et al., 2026) and GenSIRR (Li et al., 2026) improve from 22.16 to 23.77 and from 24.38 to 25.14 dB (Appendix H).

Perceptual and structural quality. Because PSNR (Wang & Bovik, 2009) rewards conservative outputs, we also evaluate LPIPS (Zhang et al., 2018a), DISTS (Ding et al., 2022), CIEDE2000 (Luo et al., 2001; Sharma et al., 2005), and ORB (Rublee et al., 2011) keypoint retention and spurious-keypoint rate; on DAI (Hu et al., 2026) the fused result improves on every individual hypothesis under all five (Appendix I).

6 Conclusion

We revisit the role of polarization in reflection removal under strong learned image priors. Fixed-model interventions show that strong restoration performance can coexist with a small average quality contribution from angle-dependent polarization, while its benefit varies across models and inputs. Motivated by this finding, our evidence-aware selective fusion combines three complementary hypotheses from the same frozen RGB restorer. The results support using polarization as conditional physical evidence, selectively integrating its contribution while preserving the restoration capabilities of pretrained RGB models.

AI use statement. Large language models were used for manuscript polishing. All experiments, numerical results, and scientific claims were produced and verified by us.

Reproducibility statement. Appendix D describes the target-level folds, fusion training setup, and inference cost. The source code will be made publicly available upon acceptance.

REFERENCES

Wenjiao Bian, Yusuke Monno, and Masatoshi Okutomi. Reflection removal using recurrent polarization-to-polarization network. *Machine Vision and Applications*, 2026.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Yuchen Hong, Haofeng Zhong, Shuchen Weng, Jinxiu Liang, and Boxin Shi. L-DiffER: Single image reflection removal with language-based diffusion model. In *Proc. of European Conference on Computer Vision*, 2024.

Jichen Hu, Chen Yang, Zanwei Zhou, Jiemin Fang, Qi Tian, and Wei Shen. Dereflection any image with diffusion priors and diversified data. In *Proc. of the AAAI Conference on Artificial Intelligence*, 2026.

Qiming Hu and Xiaojie Guo. Single image reflection separation via component synergy. In *Proc. of International Conference on Computer Vision*, 2023.

Qiming Hu, Hainuo Wang, and Xiaojie Guo. Single image reflection separation via dual-stream interactive transformers. *Proc. of Advances in Neural Information Processing Systems*, 2024.

Naejin Kong, Yu-Wing Tai, and Joseph S Shin. A physically-based approach to reflection separation: from physical modeling to constrained optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

GP Können. *Polarized light in nature*. CUP Archive, 1985.

Chenyang Lei and Qifeng Chen. Robust reflection removal with reflection-free flash-only cues. In *Proc. of Computer Vision and Pattern Recognition*, 2021.

Chenyang Lei, Xuhua Huang, Mengdi Zhang, Qiong Yan, Wenxiu Sun, and Qifeng Chen. Polarized reflection removal with perfect alignment in the wild. In *Proc. of Computer Vision and Pattern Recognition*, 2020.

Anat Levin and Yair Weiss. User assisted separation of reflections from a single image using a sparsity prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.

Mingjia Li, Jin Hu, Hainuo Wang, Qiming Hu, Jiarui Wang, and Xiaojie Guo. Rectifying latent space for generative single-image reflection removal. In *Proc. of Computer Vision and Pattern Recognition*, 2026.

Yu Li and Michael S. Brown. Single image layer separation using relative smoothness. In *Proc. of Computer Vision and Pattern Recognition*, 2014.

M. R. Luo, G. Cui, and B. Rigg. The development of the CIE 2000 colour-difference formula: CIEDE2000. *Color Research & Application*, 2001.

Youwei Lyu, Zhaopeng Cui, Si Li, Marc Pollefeys, and Boxin Shi. Reflection separation using a pair of unpolarized and polarized images. *Proc. of Advances in Neural Information Processing Systems*, 2019.

Shree K. Nayar, Xi-Sheng Fang, and Terrance E. Boult. Separation of reflection components using color and polarization. *International Journal of Computer Vision*, 1997.

Abhijith Punnappurath and Michael S. Brown. Reflection removal using a dual-pixel sensor. In *Proc. of Computer Vision and Pattern Recognition*, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015.

Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *Proc. of International Conference on Computer Vision*, 2011.

Yoav Y Schechner, Joseph Shamir, and Nahum Kiryati. Polarization-based decorrelation of transparent layers: The inclination angle of an invisible surface. In *Proc. of International Conference on Computer Vision*, 1999.

Gaurav Sharma, Wencheng Wu, and Edul N. Dalal. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application*, 2005.

Yichang Shih, Dilip Krishnan, Frédo Durand, and William T. Freeman. Reflection removal using ghosting cues. In *Proc. of Computer Vision and Pattern Recognition*, 2015.

Richard Szeliski, Shai Avidan, and P. Anandan. Layer extraction from multiple images containing reflections and transparency. In *Proc. of Computer Vision and Pattern Recognition*, 2000.

Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C. Kot. Benchmarking single-image reflection removal algorithms. In *Proc. of International Conference on Computer Vision*, 2017.

Zhou Wang and Alan C. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 2009.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004.

Lawrence B. Wolff. Using polarization to separate reflection components. In *Proc. of Computer Vision and Pattern Recognition*, 1989.

Tianfan Xue, Michael Rubinstein, Ce Liu, and William T Freeman. A computational approach for obstruction-free photography. *ACM Transactions on Graphics*, 2015.

Jie Yang, Dong Gong, Lingqiao Liu, and Qinfeng Shi. Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In *Proc. of European Conference on Computer Vision*, 2018.

Mingde Yao, Menglu Wang, King-Man Tam, Lingen Li, Tianfan Xue, and Jinwei Gu. PolarFree: Polarization-based reflection-free imaging. In *Proc. of Computer Vision and Pattern Recognition*, 2025.

Daniyar Zakarin, Thiemo Wandel, Anton Obukhov, and Dengxin Dai. Reflection removal through efficient adaptation of diffusion transformers. In *Proc. of Computer Vision and Pattern Recognition*, 2026.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. of Computer Vision and Pattern Recognition*, 2018a.

Xuaner Cecilia Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *Proc. of Computer Vision and Pattern Recognition*, 2018b.

Hao Zhao, Mingjia Li, Qiming Hu, and Xiaojie Guo. Reversible decoupling network for single image reflection removal. In *Proc. of Computer Vision and Pattern Recognition*, 2025.

Haofeng Zhong, Yuchen Hong, Shuchen Weng, Jinxiu Liang, and Boxin Shi. Language-guided image reflection separation. In *Proc. of Computer Vision and Pattern Recognition*, 2024.

SUPPLEMENTARY MATERIAL

A COUNTERFACTUAL DIAGNOSIS: DEFINITIONS AND FULL RESULTS

Interventions. Each capture is decomposed as $I_\theta = I/2 + d_\theta$ (Eq. 2), and only the angle-dependent component d_θ is edited; Fig. 6 shows the three edits on one capture. *Erase* sets $d_\theta = 0$: the four polarized images become identical copies of $I/2$ and the degree of polarization ρ re-derived by the pipeline is zero everywhere. *Roll* circularly shifts every d_θ by half the frame along both image axes, keeping its statistics but breaking its alignment with the image content; the misplaced polarization structure is visible in both d_θ and ρ. *Shuffle* replaces d_θ by that of a fixed donor capture from a different scene, resampled to the target size when necessary, so the polarization is genuine but belongs to another scene. All three edits preserve S_0 before clipping. Where the released pipeline expects frames in $[0, 1]$—all interventions for PolarFree (Yao et al., 2025) and the shuffle input of Lei et al. (2020)—the edited frames are clipped to that range; the roll input of Lei et al. (2020) and all inputs of Lyu et al. (2019) are passed unclipped. Erase never leaves the valid range and preserves S_0 exactly; for roll and shuffle, clipped pixels can change the reconstructed S_0, and those results include this effect. For the diffusion-based PolarFree (Yao et al., 2025), sampling noise is reseeded from the sample identifier before each pass, so all conditions of one image share one noise draw; without this, the sampler alone changes PSNR by up to 0.2 dB per image, comparable to the effect being measured.

A polarized example. On the capture in Fig. 6 (DoP stratum high), PolarFree does depend on the genuine measurement: erasing d_θ lowers PSNR from 21.23 to 19.89 dB, and rolling or shuffling it gives 20.62 and 20.46 dB. The restorations nonetheless remain close to one another and far above the mixed observation (8.96 dB), and over the whole test split erasing polarization costs only 0.16 dB while rolling or shuffling it changes PSNR by at most 0.03 dB (Table 2). Most of PolarFree's restoration therefore comes from its image prior, while the genuine polarization contributes on strongly polarized captures such as this one.

Evaluation domain. The 188 common images come from the official test directory of PolaRGB (Yao et al., 2025). We use this subset to avoid evaluating PolarFree (Yao et al., 2025) on its training captures; scene indices restart inside each dataset directory and must be interpreted together with the directory name. Averaging over all 6500 images would include 6312 training captures for PolarFree (Yao et al., 2025), which measures a different quantity. A small signed mean Δ_c can reflect cancellation or an ineffective model, so it must be interpreted together with baseline quality and per-image behavior. Lyu et al. (2019) ends -1.21 dB *below* the mixed observation on PolaRGB (Yao et al., 2025), so we also report it on the test set released with it (40 images), where it is strongly sensitive. Lei et al. (2020) clears the mixed observation by only +0.31 dB and its own dataset is no longer available. PolarFree (Yao et al., 2025) clears the mixed observation by +6.75 dB. Table 2 gives all numbers behind Fig. 2. Polarization changes PolarFree (Yao et al., 2025)'s result by more than 0.5 dB in its favor on 19% of these images and against it on 7%, and even in the high-DoP stratum only +0.21 dB requires the polarization to come from the aligned scene rather than another scene. This is why we describe the cue as conditionally used rather than unused.

B WHEN AND WHERE POLARIZATION HELPS: WITHIN-IMAGE CONTROL

Regional gains are larger where the RGB error is already high, and such regions also tend to be strongly polarized. We therefore compare regions within each image and condition on reflection strength $|I - T^\star|$, splitting both DoP and reflection strength at the within-image 80th percentile; both regions of a pair come from the same capture, the same restorer, and the same split, so scene identity and capture-level exposure are shared, although regional restoration difficulty can still differ. Fig. 7 gives the per-backbone version of Fig. 3(b). Among strongly reflected regions of high-DoP images, strongly polarized regions gain $+3.16$, $+1.44$, and $+2.47$ dB more than weakly polarized ones on DAI (Hu et al., 2026), GenSIRR (Li et al., 2026), and WindowSeat (Zakarin et al., 2026), and $+2.68$, 0.91, and 1.27 dB among weakly reflected regions. Replacing the privileged reflection strength by the observable $|I - T_{\mathrm{rgb}}|$ gives $+2.53$ and $+3.55$ dB on DAI (Hu et al., 2026), so the conclusion does not depend on ground truth defining the strata. The rightmost panel shows the regression-to-the-mean check: within-image quintiles of baseline error (dashed) produce a steeper

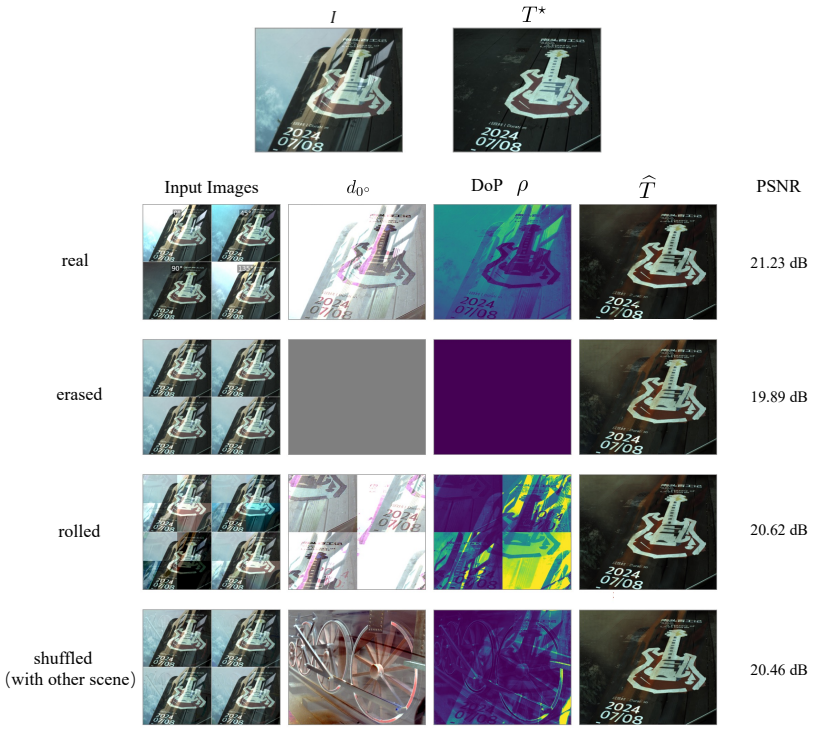


Figure 6: **The three counterfactual interventions on one strongly polarized test capture**, as received by the released PolarFree (Yao et al., 2025) pipeline. Top: the mixed observation I and the ground-truth transmission $T^\star$. Rows: the genuine capture (real) and the three interventions. Columns: the four polarized images passed to the model; the angle-dependent component d_{0°, displayed as $0.5 + 6\, d_{0^\circ}$ so that mid-grey is zero; the DoP ρ that the pipeline re-derives from the edited images; the restoration $\widehat{T}$; and its PSNR against $T^\star$. I is shared by all rows up to clipping, so every difference between rows stems from d_θ.

Table 2: **Counterfactual sensitivity of three released polarization-based models.** Δ_c is the restoration quality *lost* under intervention c, so larger means the model depended more on the genuine polarized signal. All interventions preserve S_0 before clipping (Appendix A). The last three columns stratify Δ_{erase} by DoP as in Fig. 2(c): low ($\rho_{90} < 0.20$, $n = 74$), mid (0.20–0.32, $n = 43$), and high (> 0.32, $n = 71$). Because Lyu et al. does not improve on its input on PolaRGB, the upper row also reports it on its own test set.

			Δ_c (dB)			Δ_{erase} by DoP		
method	evaluated on (n)	matched	erase	roll	shuffle	low	mid	high
in its own domain								
Lyu et al. (2019)	own test set (40)	29.14	+4.95	+6.99	+6.54	—	—	—
on the common benchmark								
Lyu et al. (2019)	PolaRGB held-out (188)	14.47	−0.04	+0.38	+0.06	+0.10	+0.18	−0.33
Lei et al. (2020)	PolaRGB held-out (188)	15.99	+0.66	+0.83	+0.80	+0.12	+0.33	+1.41
PolarFree (2025)	PolaRGB held-out (188)	22.43	+0.16	−0.01	+0.03	−0.02	−0.03	+0.46

gain profile than quintiles of DoP (solid), which is exactly why the error must be held fixed before attributing the gain to polarization.

C The Smoothed Observation

Response over (σ, c). Table 3 gives the response of T_{veil} alone on the DAI (Hu et al., 2026) search subset, with luminance renormalized in every entry. Both evaluated limits of Eq. 9 score below the best interior setting: at $\sigma = 0$ and $c = 1$, the observation before normalization is $\widetilde{I}_{\min}$, and its normalized version scores -0.35 dB; spatially constant subtraction scores $+0.14$ dB. The stronger

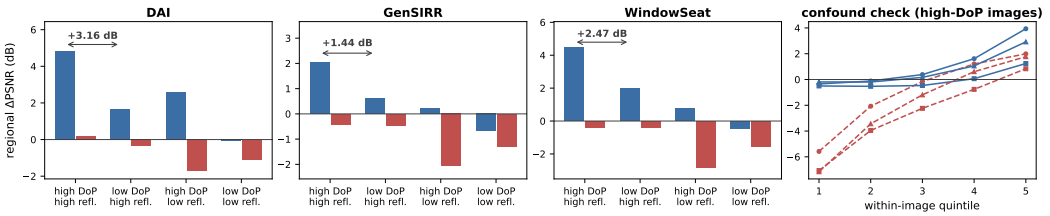


Figure 7: **Within-image analysis per backbone.** Regional A_{pol} for high-DoP images (blue) and all others (red), for the four combinations of within-image DoP and reflection strength; arrows mark the net DoP effect under strong reflection. Right: regional gain by within-image quintile of DoP (solid) and of baseline error (dashed) for high-DoP images; markers denote DAI (Hu et al., 2026) (circle), GenSIRR (Li et al., 2026) (square), and WindowSeat (Zakarin et al., 2026) (triangle).

Table 3: **The response surface over (σ, c), and the input-side parameterizations it was chosen against.** ΔPSNR over the frozen RGB branch on the parameter-search subset, luminance renormalised in every entry. Upper block: the two parameters trade off along a broad ridge — a coarser blur removes less contrast and so tolerates a larger coefficient — and the evaluated settings with $\sigma \geq 96$, $c \leq 1$ lie within 0.062 dB. Over-subtraction is punished far more sharply than under-subtraction. Lower block: the best member of each competing family, all of which condition the input on ρ or on intensity instead. None reaches the two-constant transform, and a learned 0.5M-parameter input gate does not either.

σ	c=0.5	c=0.75	c=1	c=1.25	c=1.5	c=2
4	—	—	-0.282	—	—	—
12	—	—	-0.060	—	—	—
32	—	—	+0.111	—	—	—
64	+0.257	—	+0.231	—	-0.159	-0.735
96	—	+0.313	+0.288	—	—	—
128	+0.258	+0.320	+0.298	+0.184	—	—
192	—	+0.301	+0.315	—	—	—
256	—	+0.270	+0.303	—	—	—
best of each competing input-side parameterization						
threshold on ρ						-0.198
subtract in proportion to ρ						+0.098
gate on intensity *and* ρ						-0.154
learned 0.5M-parameter input gate						+0.296

interior result indicates that this constant-subtraction control does not fully reproduce the gain. The interior is a broad plateau: the evaluated settings with $\sigma \geq 96$ and $c \leq 1$ lie within 0.062 dB, whereas the evaluated $c = 2$ setting reduces PSNR by 0.73 dB. Competing input-side parameterizations that condition the subtraction on ρ or intensity, including a learned 0.5M-parameter input gate, do not reach the two-constant transform. Because σ is a length in pixels, we also rendered 200 images at three sizes: scaling σ with the image height gives +0.187, +0.171, and +0.181 dB at heights 512, 1024, and 2048, and keeping it fixed in pixels costs at most 0.012 dB, so the relative form $\sigma \approx H/8$ transfers across resolution.

Selection subset. The 640 images used to choose (σ, c) are part of the 6500 evaluated images. Excluding them leaves the conclusions unchanged: on the remaining 5860 images, three-hypothesis fusion gains $+0.52$, $+0.64$, and $+0.52$ dB on DAI (Hu et al., 2026), GenSIRR (Li et al., 2026), and WindowSeat (Zakarin et al., 2026) (all images: $+0.54$, $+0.66$, $+0.54$), against $+0.43$, $+0.49$, and $+0.46$ dB for two-hypothesis fusion.

What the scale does. Correlating $\mathcal{G}_\sigma(|S_{\mathrm{pol}}|)$ with the reflection residual and with the transmission on 300 images shows that the measured polarized radiance tracks the reflection in the strongly polarized stratum (correlations 0.71 against 0.05) but tracks the transmission more than the reflection among the most weakly polarized images ($\rho_{90} < 0.12$; 0.43 against 0.61). Per image, a validity margin defined from these correlations predicts which of T_{pol} and T_{veil} is more accurate (correlation +0.50 on 800 images, against +0.24 for DoP itself): the full subtraction wins on 57.0% of images in the top quartile of the margin and only 10.5% in the bottom quartile. Low-pass filtering does not, however, improve this attribution—the margin decreases slightly with σ—so we read the smoothed observation

Table 4: **Adding the veil hypothesis.** All 6500 images from PolaRGB (Yao et al., 2025), out-of-fold, same fusion recipe throughout; Δ is over that backbone's frozen RGB hypothesis, and the DoP strata are those of Sec. 5.1.

					ΔPSNR by DoP stratum		
backbone	fusion	frozen RGB	ΔPSNR	improved	low	mid	high
DAI	$\{T_{\mathrm{rgb}}, T_{\mathrm{pol}}\}$	23.583	+0.450	61.9%	+0.264	+0.445	+1.824
	$\{T_{\mathrm{rgb}}, T_{\mathrm{pol}}, T_{\mathrm{veil}}\}$	23.583	+0.543	71.7%	+0.350	+0.538	+1.975
GenSIRR	$\{T_{\mathrm{rgb}}, T_{\mathrm{pol}}\}$	23.900	+0.499	68.2%	+0.398	+0.595	+0.941
	$\{T_{\mathrm{rgb}}, T_{\mathrm{pol}}, T_{\mathrm{veil}}\}$	23.900	+0.663	76.3%	+0.570	+0.741	+1.108
WindowSeat	$\{T_{\mathrm{rgb}}, T_{\mathrm{pol}}\}$	24.987	+0.473	59.8%	+0.237	+0.635	+1.698
	$\{T_{\mathrm{rgb}}, T_{\mathrm{pol}}, T_{\mathrm{veil}}\}$	24.987	+0.537	74.3%	+0.346	+0.622	+1.671

as a conservative commitment to a partially invalid estimate rather than as a clean band separation of the two layers. This is the argument for keeping both polarization-derived hypotheses and selecting between them.

Additional scales yield no clear gain in the evaluated setting. Adding two further scales ($\sigma = H/32$ and $H/2.7$) as extra hypotheses raises the per-image oracle from +0.749 to +0.797 dB but changes the fused result by only +0.021 $\pm$ 0.040 dB (paired, clustered on the 67 ground-truth targets). This experiment does not establish a measurable advantage for these two additional scales.

D Implementation and Reproducibility

Training. The fusion module uses 384^2 random crops, horizontal flips, batch size 8, AdamW with learning rate 2×10^{-4}, weight decay 10^{-4}, one-cycle scheduling, gradient clipping at 1.0, and 12 epochs, minimizing Eq. 13 with $\lambda_{\mathrm{ssim}} = 0.15$. Every checkpoint is the final epoch; selecting the best epoch on the fold later used for evaluation would inflate the crop-level gain by +0.07 dB on average. Folds are built from the canonical 6500-image manifest at the level of ground-truth targets, greedily assigning the largest target group to the emptiest of five folds, so that every image is predicted by a module that never saw its target. All hypotheses are precomputed once per backbone; inference is fully convolutional at native resolution. An independent repeat of the two-hypothesis DAI (Hu et al., 2026) run reproduces the gain within 0.02 dB and the mean weight on T_{pol} within 0.029; this repeat provides a scale for run-to-run variation in that setting, rather than an uncertainty bound for all configurations.

Cost. On one H800 GPU at 768^2, a DAI (Hu et al., 2026) forward takes 0.90 s and 10.2 GB, and the 1.95M-parameter fusion module 0.07 s and 1.8 GB; the full method (three independent backbone forwards plus fusion) therefore takes about 2.8 s per image, roughly three times the RGB restorer alone. For the multi-step backbones the cost is likewise three restorer calls plus a negligible fusion pass.

E Full Results Across Backbones

Table 4 gives the per-stratum results of the main text for all three backbones. The third hypothesis raises the share of improved images on every evaluated backbone (Table 4). The factor by which it multiplies the two-hypothesis gain is largest in the low-DoP stratum on all three backbones ($\times 1.32$ against $\times 1.08$ in the high stratum on DAI (Hu et al., 2026), $\times 1.43$ against $\times 1.18$ on GenSIRR (Li et al., 2026), and $\times 1.46$ against $\times 0.98$ on WindowSeat (Zakarin et al., 2026)), as expected where there is little polarization to remove and full removal is the wrong move. The mean weights differ across backbones (mean $w_{\mathrm{rgb}}/w_{\mathrm{pol}}/w_{\mathrm{veil}}$ of 0.135/0.394/0.471 on DAI (Hu et al., 2026), 0.296/0.338/0.367 on GenSIRR (Li et al., 2026), and 0.314/0.356/0.330 on WindowSeat (Zakarin et al., 2026)).

E.1 Per-image distributions and the envelope after fusion

Stratum means cannot distinguish a small effect everywhere from a large effect somewhere. Fig. 8 shows that unconditional use of T_{pol} places most of its mass below zero on every backbone, whereas fusion concentrates it at or above zero. Fig. 9 gives the per-cell values on the three DoP strata of

Sec. 5.1 crossed with the RGB-quality bins of Fig. 3(a): negative cells drop from 20 to 1 of 36 (-0.08 dB), and the high-DoP, weak-baseline cells keep their gains.

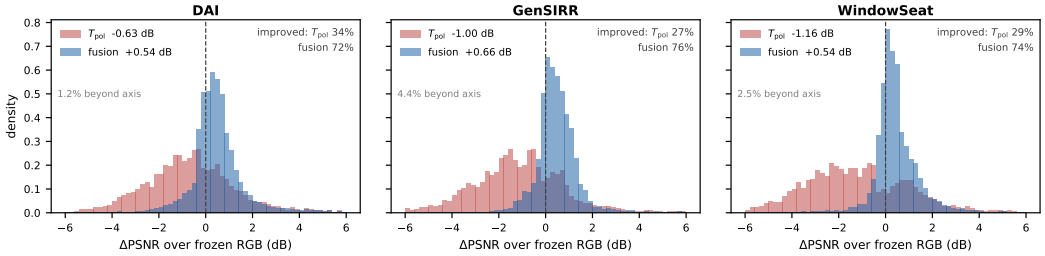


Figure 8: **Per-image ΔPSNR over the frozen RGB hypothesis** on all 6500 images: T_{pol} everywhere (red) and three-hypothesis selective fusion (blue). Values beyond the axis are omitted rather than clipped into the end bins; the means and improvement rates use every image.

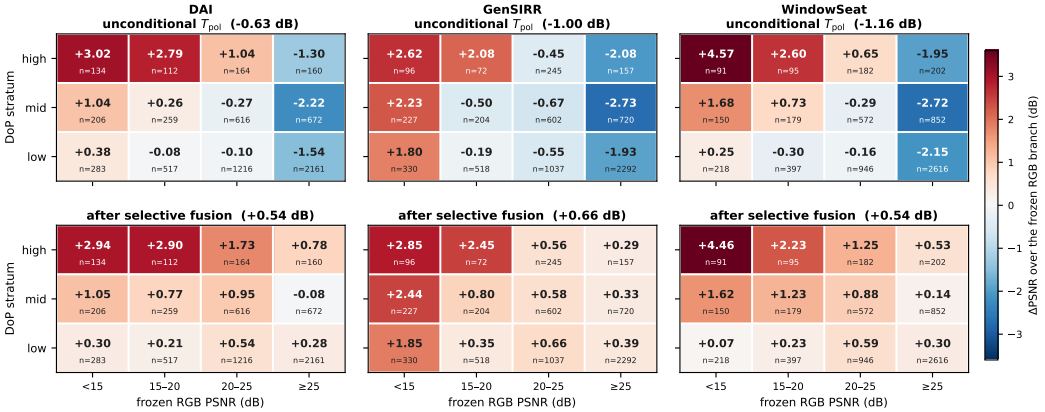


Figure 9: **The conditional structure before (top, unconditional T_{pol}) and after (bottom) selective fusion**, on identical strata (three DoP strata $\times$ RGB quality), so each cell can be read vertically.

F Selection Baselines and Fusion-Module Design

Table 5 places the method on a ladder of increasingly capable ways to use the same observations on DAI (Hu et al., 2026), with ground-truth oracles bounding what a selector of each form could reach. The image-level routers are gradient-boosted models on GT-free per-image features (eight DoP and intensity statistics and six statistics of the disagreement and luminance ratio between hypotheses), trained and evaluated with the same target-level folds as the fusion module. The two-way router predicts whether T_{pol} beats T_{rgb}, with its decision threshold chosen on the training folds (the best threshold in hindsight gives only $+0.03$ dB as well); the three-way router regresses the advantage of T_{pol} and T_{veil} over T_{rgb} and picks the largest. The DoP rule activates T_{pol} where a smoothed ρ exceeds a threshold (best of three threshold pairs).

Fusion-module design. The following comparisons use the two-hypothesis DAI (Hu et al., 2026) system and are paired on images, with standard errors clustered on ground-truth targets. The convolutional U-Net outperforms a matched-budget alternative built from attention and feature-modulation blocks by +0.106 $\pm$ 0.044 dB; removing either of that alternative's specialized mechanisms changes its result by +0.013 $\pm$ 0.047 and -0.029 ± 0.047 dB, i.e. not measurably. Replacing the convex output by a bounded residual head scores +0.299 $\pm$ 0.132 dB higher, but it gains $+0.57$ dB in the low-DoP stratum, where T_{pol} costs -0.81 dB. Such a gain is consistent with a generic post-restorer rather than with selecting polarization-derived hypotheses, so we retain convex fusion, which makes attribution exact at some cost in PSNR.

Table 5: **Selection ladder on DAI (Hu et al., 2026)**, all 6500 images out-of-fold. Δ is relative to the frozen RGB hypothesis. Oracles use the ground truth and are not deployable; the block oracle picks the best hypothesis per 16×16 block.

method	PSNR	ΔPSNR
frozen RGB hypothesis	23.58	—
always T_{pol}	22.96	−0.63
always T_{pol}, color-matched	23.27	−0.31
always T_{veil}	23.78	+0.20
image-level router $\{T_{\mathrm{rgb}}, T_{\mathrm{pol}}\}$	23.61	+0.03
image-level router $\{T_{\mathrm{rgb}}, T_{\mathrm{pol}}, T_{\mathrm{veil}}\}$	23.78	+0.19
spatial DoP rule $\{T_{\mathrm{rgb}}, T_{\mathrm{pol}}\}$	23.65	+0.06
selective fusion $\{T_{\mathrm{rgb}}, T_{\mathrm{pol}}\}$	24.03	+0.45
selective fusion $\{T_{\mathrm{rgb}}, T_{\mathrm{pol}}, T_{\mathrm{veil}}\}$	**24.13**	**+0.54**
image-level oracle, two / three hypotheses	—	+0.51 / +0.75
block oracle, two / three hypotheses	—	+1.04 / +1.41

G Zero-Shot Transfer Between Backbones

The three-hypothesis fusion module trained on one backbone is applied to another without adaptation, using the target's own frozen hypotheses $\{T_{\mathrm{rgb}}, T_{\mathrm{pol}}, T_{\mathrm{veil}}\}$ (with the target's own σ) and the donor's module for the fold that excludes each image's target (Table 6). All six ordered directions yield positive mean gains, retaining 32–101% of the target's in-domain gain. Modules trained on GenSIRR (Li et al., 2026) and WindowSeat (Zakarin et al., 2026) transfer well in every direction (72–101%), whereas the DAI (Hu et al., 2026) module transfers least (44% and 32%). The DAI module relies least on the RGB fallback (mean w_{rgb} of 0.135, against 0.296 and 0.314), which suits its weaker prior but over-commits on stronger ones: on WindowSeat its gain in the low-DoP stratum drops to -0.01 dB while the high-DoP gain remains $+1.36$ dB. The residual gap is therefore one of calibration, not of sign; Fig. 10 compares in-domain and transferred fusion on the same images for GenSIRR→WindowSeat.

Table 6: Zero-shot transfer of the three-hypothesis fusion module on all 6500 images (PSNR, dB). Upper block: each backbone with its own module. Lower block: every ordered donor→target pair without adaptation. Δ is over the target's frozen RGB hypothesis; "improved" is the share of images with positive Δ; "retained" is the fraction of the target's in-domain gain.

module → backbone	frozen RGB	always T_{pol}	fused (Δ)	improved	retained
DAI → DAI	23.58	−0.63	+0.54	71.7%	—
GenSIRR → GenSIRR	23.90	−1.00	+0.66	76.3%	—
WindowSeat → WindowSeat	24.99	−1.16	+0.54	74.3%	—
DAI → GenSIRR	23.90	−1.00	+0.29	65.2%	44%
DAI → WindowSeat	24.99	−1.16	+0.17	55.2%	32%
GenSIRR → DAI	23.58	−0.63	+0.55	77.5%	**101%**
GenSIRR → WindowSeat	24.99	−1.16	+0.47	71.5%	88%
WindowSeat → DAI	23.58	−0.63	+0.48	77.0%	89%
WindowSeat → GenSIRR	23.90	−1.00	+0.48	69.2%	72%

H Released-System Comparison on the Official Split

Table 7 scores released checkpoints with one native-resolution harness on the 188-image official test split. DSRNet, RDNet, DSIT, and the three generative restorers are cross-dataset; PolarFree (Yao et al., 2025) is trained end to end on PolaRGB (Yao et al., 2025), and for our method only the fusion module is trained on PolaRGB (Yao et al., 2025), with the folds of Sec. 5.1. Thus, the fusion results are out-of-fold predictions restricted to the official test subset, rather than the result of a single fusion module trained only on the official training split. Two frozen RGB restorers already exceed the in-domain polarization model before any polarization is used, which is the change in prior strength motivating this study; selective fusion improves all three. The systems differ in capacity and training data, so the table compares released systems, not architectures.

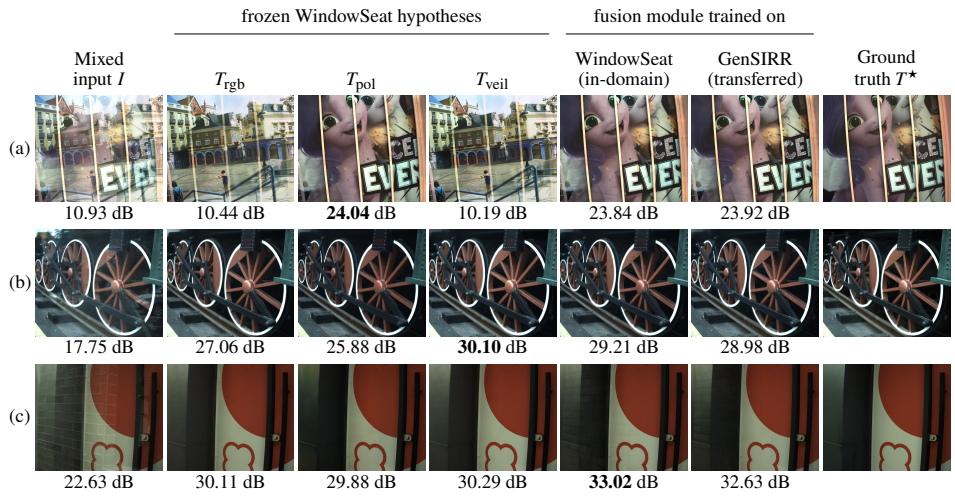


Figure 10: **Zero-shot transfer from GenSIRR (Li et al., 2026) to WindowSeat (Zakarin et al., 2026).** The same frozen WindowSeat hypotheses are fused by in-domain and by the module trained on GenSIRR; both predictions are out-of-fold. (a) Strong polarization: both modules select T_{pol} (mean weight 0.98). (b) Weak polarization, where T_{veil} is best: the transferred module puts most weight on T_{veil} (0.53). (c) No single hypothesis is best: both modules exceed all three. Over all 6500 images the transferred module gains $+0.47$ dB against $+0.54$ dB in-domain.

Table 7: Official PolaRGB (Yao et al., 2025) test split (188 images). "pol." marks methods that use polarization. PolarFree is trained end to end on PolaRGB; for selective fusion only the fusion module is trained on it (Sec. 5.1), and the generative restorers remain frozen.

method	pol.	PSNR	SSIM
mixed observation		15.68	0.7740
DSRNet (Hu & Guo, 2023)		16.76	0.7872
RDNet (Zhao et al., 2025)		17.77	0.8221
DSIT (Hu et al., 2024)		17.78	0.8283
PolarFree (Yao et al., 2025)	✓	22.44	0.8593
frozen DAI (Hu et al., 2026)		22.16	0.8592
DAI (Hu et al., 2026) + selective fusion	✓	23.77	0.8857
frozen GenSIRR (Li et al., 2026)		24.38	0.8642
GenSIRR (Li et al., 2026) + selective fusion	✓	25.14	0.8850
frozen WindowSeat (Zakarin et al., 2026)		24.75	0.8859
WindowSeat (Zakarin et al., 2026) + selective fusion	✓	**25.83**	**0.9028**

I Perceptual and Structural Quality

Table 8: **Perceptual, structural and keypoint quality** on DAI, all 6500 images out-of-fold. PSNR rewards conservative outputs, so these five measures — which fail in different ways — are reported together. LPIPS and DISTS are learned perceptual distances, ΔE_{00} is CIEDE2000 colour error, and the last two are the share of ORB keypoints of the ground-truth transmission that survive in the prediction and the share of predicted keypoints with no ground-truth counterpart, the latter being the measure a hallucinated texture would move.

	LPIPS ↓	DISTS ↓	ΔE_{00} ↓	ORB retained ↑	ORB spurious ↓
frozen RGB	0.1856	0.1329	6.22	78.4%	18.2%
always T_{pol}	0.1807	0.1300	6.32	76.6%	19.4%
always T_{veil}	0.1831	0.1294	6.11	78.8%	18.3%
fuse $\{T_{\text{rgb}}, T_{\text{pol}}\}$	0.1455	0.1110	5.85	80.1%	17.0%
fuse $\{T_{\text{rgb}}, T_{\text{pol}}, T_{\text{veil}}\}$	0.1464	0.1124	5.82	80.2%	17.1%

PSNR (Wang & Bovik, 2009) rewards conservative outputs, so we report five measures that fail in different ways (Table 8). Both fused systems improve on every individual hypothesis under all five, including the spurious-keypoint rate that a hallucinated texture would raise. Compared with

two-hypothesis fusion, adding the third hypothesis improves CIEDE2000 and ORB retention slightly, while LPIPS, DISTS, and the spurious-keypoint rate deteriorate slightly. The additional PSNR gain therefore comes with small metric-dependent trade-offs; both fused systems remain better than the individual hypotheses on all five measures.