RECAST: BEYOND RECONSTRUCTION—LEARNING COMPOSITIONAL VISUAL RESPONSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Shadow removal is usually determined by alignment with a reference shadow-free image. Under a fixed observation, however, distinct latent responses can yield similar appearance yet diverge once illumination or occlusion changes, so endpoint fidelity alone does not certify a reusable response. We introduce Missing-Transport Inversion (MTI), which recovers the radiance residual suppressed by occlusion so that it remains predictive under withheld illumination–occlusion recombinations, without assuming a uniquely identifiable physical decomposition. Based on this, we propose ReCAST, which grounds the occluded receiver with geometry-compatible non-shadow evidence, learns a shared compositional operator over receiver, illumination, and occlusion factors, and renders appearance with a response-conditioned diffusion decoder that also supports forward re-shadowing. On a controlled same-scene Transport Lattice, it predicts a fixed withheld pairing from complementary observations in a target-blind protocol, reducing normalized missing-transport error (NMTE) from 0.667 to 0.589 against matched Cross-attention. On ISTD+, SRD, INS, and WSRD+, the same representation remains competitive in PSNR, showing that compositional response modeling does not trade away standard restoration quality. Code is available in the supplementary materials.

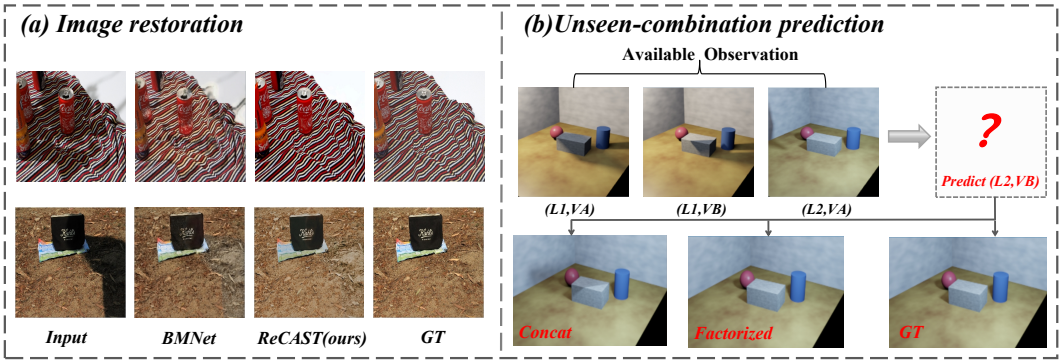


Figure 1: **Restoration and response prediction.** Left: single-image shadow removal. Right: prediction of the withheld (L_2, O_B) condition from (L_1, O_A), (L_1, O_B), and (L_2, O_A) of the same test scene. Both methods receive identical permitted observations. The right panel evaluates multi-observation response reuse; the target image is used only for scoring.

1 INTRODUCTION

Cast shadows arise when occlusion suppresses the radiance that would otherwise reach a receiver surface. Shadow removal is therefore not only a matter of brightening dark pixels, but also asks whether the suppressed receiver response can be inferred from incomplete evidence and reused when illumination and occlusion change. Editing, relighting, and scene understanding ultimately depend on this reusable response, not on a single restored image.

Standard protocols evaluate the task by comparing a predicted shadow-free image with a paired reference under fixed imaging conditions (Le & Samaras, 2019; Zhu et al., 2022; Guo et al., 2023a;b). Metrics such as PSNR and SSIM audit endpoint appearance. However, they do not evaluate whether the representation that produced the restored image remains predictive once the illumination and occlusion factors of the same scene are recombined.

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.

This gap matters because appearance matching underdetermines the response. Write the linear-radiance observation as $I = f(G, L, O)$, where G collects receiver geometry and material, L illumination, and O occlusion. A shadowed observation I_{s} and a nearby restored J_{c} can be explained by distinct latent factorizations that agree on the observed condition yet diverge on a withheld pair (L', O') (Figure 1, right). Endpoint fidelity therefore certifies one restored appearance, rather than the missing residual $R_{\mathrm{miss}} = I_{\mathrm{c}} - I_{\mathrm{s}}$ behaves correctly under factor recombination.

Prior lines address complementary pieces of this difficulty but leave the recombination test open. Physical and intrinsic decompositions (Le & Samaras, 2019; 2022) introduce image-formation structure, yet full factor recovery from a single view remains ill-posed. Discriminative methods (Zhu et al., 2022; Guo et al., 2023a; Xiao et al., 2024; Xu et al., 2025a) improve contextual mapping under observed conditions. Generative models (Guo et al., 2023b; Mei et al., 2024; Luo et al., 2025; Xu et al., 2025b) raise perceptual plausibility, which can mask response errors that appear only after recombination. Inverse–forward systems such as MetaShadow (Wang et al., 2025) couple removal with synthesis in an object-centered editing workflow; they do not evaluate whether a factorized missing response remains valid for unseen illumination–occlusion recombinations of the same receiver.

We formulate the task as *Missing-Transport Inversion* (MTI): recover the missing residual associated with a requested illumination–occlusion pair so that it remains predictive under withheld recombinations, without claiming a uniquely identifiable physical transport field. ReCAST addresses MTI with three coupled ingredients. Geometry-compatible non-shadow anchors constrain the occluded receiver G. A shared compositional operator $\Phi(G, \ell, \Omega)$ predicts the residual under query factors and is reused for forward re-shadowing, providing a direct audit of the same representation. A response-conditioned diffusion decoder then renders appearance subordinate to this residual field rather than replacing it.

We evaluate the claim on a controlled same-scene Transport Lattice that withholds the joint pair (L_2, O_B) while retaining its marginal factors in permitted observations, and score target-blind residual prediction with normalized missing-transport error (NMTE). ReCAST reduces primary NMTE from 0.667 to 0.589 against matched Cross-attention, with the gain confirmed on two disjoint scene-family audits. On ISTD+, SRD, INS, and WSRD+, the same representation remains competitive or superior in PSNR under standard restoration interfaces, indicating that response-level modeling does not trade away appearance quality. The lattice recombination protocol is the decisive evidence for our claim; real benchmarks serve as compatibility checks.

Our contributions are threefold:

- **Problem and falsifiable protocol.** We cast shadow removal as Missing-Transport Inversion and evaluate it on a target-blind Transport Lattice: predict the residual at a withheld illumination–occlusion pair from complementary observations of the same receiver, without assuming a unique physical decomposition.
- **Explicit compositional response.** ReCAST learns a shared operator $\Phi(G, \ell, \Omega)$ over receiver, illumination, and occlusion factors. The same operator supports inverse restoration and forward re-shadowing; relative to matched concatenation baselines, the interaction bias improves withheld-combination prediction.
- **Receiver grounding for reuse.** Geometry-compatible non-shadow anchors constrain the occluded receiver so that composition is learnable. Diffusion serves as an appearance renderer conditioned on the predicted residual, not as the primary scientific claim.

2 Related Work

Shadow removal as appearance restoration. Shadow removal has traditionally been formulated as recovering a clean appearance from a shadowed observation. Early approaches investigate illumination-invariant representations and intrinsic image assumptions to separate shadow effects from surface properties (Finlayson et al., 2009; Yang et al., 2012; Zhang et al., 2015). Learning-based methods then improve restoration by exploiting contextual cues and jointly modeling shadow detection and removal (Qu et al., 2017; Wang et al., 2018). Physical decomposition introduces image-formation constraints to improve interpretability and robustness (Le & Samaras, 2019; 2022). More recent architectures employ stronger spatial reasoning through bijective mappings and transformers (Zhu et al., 2022; Guo et al., 2023a; Xiao et al., 2024). OmniSR further considers complex illumination interactions beyond simple shadow attenuation (Xu et al., 2025a). Together, these methods substantially advance shadow-free image reconstruction. However, their evaluations mainly

examine whether the predicted image matches an observed target, leaving open whether the learned representation captures how a receiver responds when visual conditions are recombined.

Generative models for ambiguous visual restoration. Generative models address the ambiguity of incomplete observations by learning priors over plausible visual appearances. Diffusion models provide a general framework for conditional generation through iterative denoising (Ho et al., 2020; Rombach et al., 2022; Saharia et al., 2022). Recent shadow removal approaches incorporate diffusion priors to improve texture recovery and perceptual quality (Guo et al., 2023b; Mei et al., 2024; Luo et al., 2025; Xu et al., 2025b). These models demonstrate that strong generative priors can recover visually plausible content, but plausibility does not necessarily imply consistency with an unseen physical condition. In contrast, ReCAST separates appearance synthesis from response inference and explicitly evaluates whether the inferred representation predicts withheld illumination–occlusion configurations.

Compositional representations and conditional generalization. Learning representations that generalize across changing factors has been studied through conditional computation, feature modulation, and structured latent representations. FiLM introduces feature-wise conditioning to adapt neural representations according to external variables (Perez et al., 2018). Recent shadow-related frameworks further explore shared representations between shadow removal and generation (Wang et al., 2025). However, existing conditional models generally assume that the required condition is observable during prediction, while our setting requires inferring the response associated with a missing condition from incomplete evidence. ReCAST addresses this gap by introducing a target-blind response prediction formulation, where compositional reuse is evaluated through shared response operators, forward prediction, and multi-path consistency rather than through reconstruction quality alone.

3 METHOD

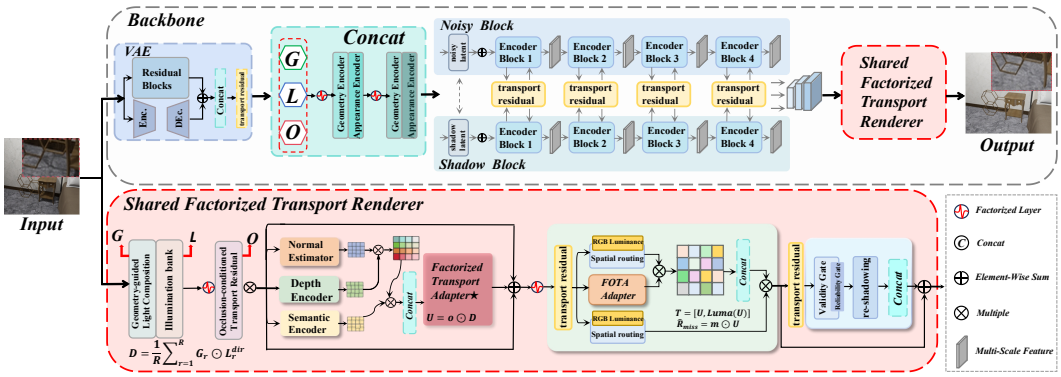


Figure 2: **ReCAST architecture.** Receiver evidence grounds a shared response representation that supports residual prediction, diffusion conditioning, and forward re-shadowing. Figure 3 details the low-rank operator, closure, and support-aware reliability pathways.

3.1 MISSING-TRANSPORT INVERSION

Let $I_{\mathrm{s}}, I_{\mathrm{c}} \in \mathbb{R}^{H\times W\times 3}$ be shadowed and shadow-free linear-radiance images under the same camera and illumination. The missing response and display-space images are

$$R_{\mathrm{miss}} = I_{\mathrm{c}} - I_{\mathrm{s}}, \qquad J_{\mathrm{s}} = \mathcal{C}(I_{\mathrm{s}}), \quad J_{\mathrm{c}} = \mathcal{C}(I_{\mathrm{c}}), \tag{1}$$

where $\mathcal{C}$ is a fixed camera conversion. We infer the suppressed response by composing geometry/material, illumination, and occlusion evidence. The radiance residual $\widehat{R}_{\mathrm{miss}}$ and restored display-space image $\widehat{J}_{\mathrm{c}}$ are distinct outputs.

The controlled task uses observations at (L_1, O_A), (L_2, O_A), and (L_1, O_B) to predict the withheld (L_2, O_B) response. Condition identifiers $(L^\star, O^\star)$ specify the request. Illumination metadata supplies $\ell^\star$; target occlusion $\Omega^\star$, support $m^\star$, context $F^\star$, and clean radiance $\widehat{I}_c^\star$ are inferred from permitted evidence. Target images, residuals, masks, and derived caches remain unavailable until scoring (Section A.4 of the Supplementary Material).

3.2 Receiver Grounding and Shared Composition

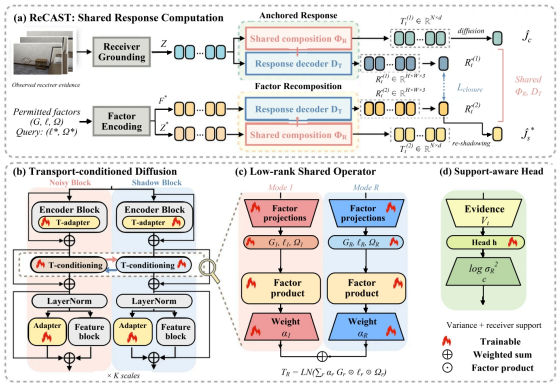
Projected appearance and geometry features $Q, U \in \mathbb{R}^{N\times d}$ share an $h \times w$ grid, with $N = hw$ and $X = [Q, U]$. A compatibility score s_{ij} combines semantic, normal, and depth evidence over non-shadow anchors $\mathcal{A}_i$. With temperature τ,

$$A_{ij} = \frac{\mathbf{1}[j \in \mathcal{A}_i]\exp(s_{ij}/\tau)}{\sum_{j'\in\mathcal{A}_i}\exp(s_{ij'}/\tau)}, \qquad G = \mathcal{G}(X, AQ). \tag{2}$$

For empty anchor sets, $A_{i:} = 0$. Reliability-gated fusion $\mathcal{G}$ falls back to local evidence when anchors are invalid or absent. Frozen DINOv2 features supply semantic compatibility (Oquab et al., 2024); anchors constrain receiver identity rather than prescribe target brightness. Retrieval and incompatible-anchor controls are detailed in Section A.7 of the Supplementary Material.
For $Z = (G, \ell, \Omega)$, learned weights $\boldsymbol{\alpha} = \mathrm{softmax}(\mathbf{a})$ combine R interaction modes:

$$T_R = \Phi_R(Z) = \mathrm{LN}\left(\sum_{r=1}^{R}\alpha_r\,\phi_r^G(G)\odot\phi_r^\ell(\ell)\odot\phi_r^\Omega(\Omega)\right), \quad \widehat{R}_{\mathrm{miss}} = m\odot D_T(T_R, F). \tag{3}$$

The projections align to $N \times d$, with illumination broadcast spatially; LN denotes layer normalization. The decoder D_T combines T_R with multiscale context $F = \{F_k\}$ and restores image resolution before applying soft support m. Full ReCAST uses $R = 4$; $R = 1$ is the single-mode ablation. Rank describes the interaction parameterization, not an identified physical transport rank. The operator and decoder are shared between inverse and forward prediction.



Shared computation. Receiver grounding first extracts geometry- and appearance-compatible non-shadow anchors to construct a receiver representation G. A shared rank-R response operator then composes G with illumination and occlusion factors to infer the missing response T_R. The same response representation and decoder support both restoration and forward re-shadowing under changed conditions, enabling compositional reuse beyond the observed state. Multi-path closure evaluates whether different evidence routes agree on the same prediction, while a support-aware reliability head estimates confidence under incomplete evidence.

Figure 3: **Model flow.** The supplied flow diagram makes inverse restoration, forward reuse, and reliability estimation explicit. All paths use the same target-blind information contract.

3.3 Restoration and Forward Reuse

At diffusion scale k, learned maps γ_k, β_k inject the response feature into decoder feature H_k:

$$\widetilde{H}_k = H_k + m_k \odot [\gamma_k(T_R)\odot \mathrm{LN}(H_k) + \beta_k(T_R)]. \tag{4}$$

Denoising produces $\widehat{J}_c$. Support masking localizes the update, while the training objective supervises non-shadow preservation. For forward reuse, retain G and form $Z^\star = (G, \ell^\star, \Omega^\star)$ from permitted evidence:

$$\widehat{R}^\star_{\mathrm{miss}} = m^\star \odot D_T(\Phi_R(Z^\star), F^\star), \qquad \widehat{J}^\star_s = \mathcal{C}(\widehat{I}^\star_c - \widehat{R}^\star_{\mathrm{miss}}). \tag{5}$$

Subtraction precedes camera conversion. Two permitted routes to the same target tuple produce the closure constraint

$$\Delta_t = \|\widehat{R}^{(1)}_{\mathrm{miss},t} - \widehat{R}^{(2)}_{\mathrm{miss},t}\|_1, \qquad \mathcal{L}_{\mathrm{closure}} = \mathbb{E}_t[\Delta_t]. \tag{6}$$

The audit normalizes route disagreement on common support. Routes can differ in anchor subsets and context aggregation, not only observation order (Section A.12.1 of the Supplementary Material, the corresponding supplementary table). Agreement is assessed alongside target error, since identical incorrect responses also satisfy closure.

3.4 Support-Aware Uncertainty and Training

The reliability head estimates evidence coverage from permitted factors and anchors:

$$c = \sigma(h_c[\mathrm{pool}(G), \ell, \mathrm{pool}(\Omega), \mathrm{pool}(A)]) \,. \tag{7}$$

Let $e_{\mathrm{phy}}, e_{\mathrm{diff}}$ be detached local radiance errors of the shared renderer and diffusion clean branch. A soft reliability target $y = \sigma((e_{\mathrm{diff}} - e_{\mathrm{phy}})/0.03)v(1-\rho)$ uses predicted-factor validity v and detached prior-corruption score ρ. We train the reliability logit z_u with

$$\mathcal{L}_{\mathrm{unc}} = \frac{\sum_{x \in \mathcal{M}_{\mathrm{shadow}}} \mathrm{BCEWithLogits}(z_u(x), y(x))}{|\mathcal{M}_{\mathrm{shadow}}| + \epsilon}. \tag{8}$$

Reference errors enter training only. Calibration is fitted on validation scenes and frozen for testing; coverage score c ranks scenes in the selective-risk audit (Section A.13.2 of the Supplementary Material).

Define $\mathcal{L}_{\mathrm{rest}} = \mathcal{L}_{\mathrm{diff}} + \lambda_{\mathrm{img}}\mathcal{L}_{\mathrm{img}} + \lambda_{\mathrm{ns}}\mathcal{L}_{\mathrm{ns}}$ and $\mathcal{L}_{\mathrm{rad}} = \| \min(\widehat{I}_{\mathrm{c}} - \widehat{R}_{\mathrm{miss}}, 0)\|_1$. The latter penalizes negative reconstructed shadowed radiance. The full objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{rest}} + \lambda_T\mathcal{L}_T + \lambda_{\mathrm{re}}\mathcal{L}_{\mathrm{re}} + \lambda_a\mathcal{L}_{\mathrm{anchor}} + \lambda_{\mathrm{cl}}\mathcal{L}_{\mathrm{closure}} + \lambda_{\mathrm{unc}}\mathcal{L}_{\mathrm{unc}} + \lambda_{\mathrm{rad}}\mathcal{L}_{\mathrm{rad}}, \tag{9}$$

with $\lambda_{\mathrm{cl}} = 0.10$, $\lambda_{\mathrm{unc}} = 0.05$, and $\lambda_{\mathrm{rad}} = 0.02$. Forward and closure supervision use observed combinations; withheld targets never enter checkpoint selection. Full loss and training details are in Section A.4.4 of the Supplementary Material and Section A.5 of the Supplementary Material.

Table 1: **Prediction contract.** The same restrictions apply to every route and cache.

Status	Information
Provided	Observations; condition IDs $(L^\star, O^\star)$.
Inferred	$G, \Omega^\star, F^\star, m^\star, \widehat{I}_{\mathrm{c}}^\star$; $\ell^\star$ from metadata.
Scoring only	Target RGB, response, mask, render, and derived cache.

Algorithm 1 Response inference and reuse
Input: permitted evidence and requested condition.
Output: restoration, response, reliability.

1. Ground G with valid non-shadow anchors.
2. Compose $T_R = \Phi_R(G, \ell, \Omega)$; decode $\widehat{R}_{\mathrm{miss}}$.
3. Inject T_R into diffusion and restore $\widehat{J}_{\mathrm{c}}$.
4. Infer target fields; reuse Φ_R, D_T to predict $\widehat{R}^\star_{\mathrm{miss}}, \widehat{J}^\star_{\mathrm{s}}$.
5. Compare routes and estimate reliability before opening the target for scoring.

Implementation. The renderer is trained for 100 epochs on 512×512 inputs with batch size 8 and AdamW (learning rate 2×10^{-4}, weight decay 10^{-4}, gradient clipping 1.0). Geometry and DINOv2 encoders remain frozen. Matched operator comparisons share evidence, decoder, supervision, and optimization (Section A.4.4 of the Supplementary Material).

3.5 Counterfactual Validation Protocol

The Transport Lattice uses path-traced 3D-FRONT scenes (Fu et al., 2021) with disjoint training, validation, and test identities. Each scene fixes camera, receiver geometry, and materials while spanning $(L_1, L_2, L_1 + L_2) \times (O_0, O_A, O_B)$; O_0 is unoccluded. Paired unoccluded renders define response targets. For response evaluation, let $r_{s,c} = \mathrm{vec}(M^{\mathrm{eval}}_{s,c} \odot R_{\mathrm{miss},s,c})$ and define $\hat{r}_{m,s,c}$ analogously. For the fixed 40-scene diagnostic, prediction-independent masks and $\epsilon = 10^{-6}$ give

$$\begin{aligned} e_{m,s,c} &= \frac{\|\hat{r}_{m,s,c} - r_{s,c}\|_2}{\|r_{s,c}\|_2 + \epsilon}, \quad e^0_{s,c} = \frac{\|r_{s,c}\|_2}{\|r_{s,c}\|_2 + \epsilon}, \\ E_{m,s} &= \frac{1}{|\mathcal{C}_s|}\sum_{c \in \mathcal{C}_s} e_{m,s,c}, \quad \mathrm{NMTE}_{\mathrm{joint}m} = \frac{1}{S}\sum_s \frac{E_{m,s}}{E_{0,s}}. \end{aligned} \tag{10}$$

Here $E_{0,s}$ averages $e^0_{s,c}$, making the zero-response score exactly one on eligible scenes. Expanded-audit aggregation is specified in Section A.11 of the Supplementary Material. Withheld files are opened only after prediction; the same information contract applies to all methods and both routes. The joint holdout preserves the receiver and changes the requested illumination–occlusion pairing. It therefore tests recombination of permitted scene evidence. Isolated factor interventions hold one condition fixed and change the other; their response differences test sensitivity to the requested factor. These controlled predictions complement single-image restoration, whose clean target assesses the observed endpoint. Neither setting alone identifies a unique physical decomposition.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Data and implementation. All operator variants follow Section 3.4; full ReCAST uses rank four, closure, and reliability. We evaluate withheld responses, test the mechanisms supporting reuse, and assess restoration on ISTD+, SRD, INS, and WSRD+.

Protocols and metrics. The primary audit uses 200 scenes in five disjoint split clusters and ten training seeds per learned method and split. The five splits define the uncertainty units; repeated seeds do not add independent clusters. Confirmations use 12 families/240 scenes and 16 families/320 scenes, disjoint from the primary audit; reliability uses a separate 15-family/240-scene set. The 40-scene joint holdout and separate 40-scene direct-response audit remain mechanism diagnostics and are not pooled with the primary result. Appearance metrics are PSNR, SSIM (Wang et al., 2004), MAE, and RMSE; response metrics follow Section 3.5. The clean-room audit found no forbidden target reads in 240 checked cases (Section A.15 of the Supplementary Material).

4.2 WITHHELD-RESPONSE PREDICTION

Table 2: **Unseen-combination NMTE.** Lower is better; confirmation rows compare ReCAST with Cross-attention (CA).

Primary (200 scenes)	NMTE
Zero response	1.000
Nearest-condition copy	0.852
Concatenation	0.731
Cross-attention	0.667
ReCAST	**0.589**
Confirmation	**ReCAST / CA**
12 families (240 scenes)	0.604 / 0.660
16 families (320 scenes)	0.607 / 0.653

ReCAST reduces NMTE from 0.667 to 0.589 relative to matched Cross-attention, an 11.7% improvement, and wins 165/200 scenes (Table 2). The paired split-cluster difference is -0.078 (95% CI $[-0.099, -0.057]$; Supplementary Table A.35). Its advantage over nearest-condition copying also tests prediction beyond selecting an observed response. The gain persists across confirmations: 171/240 and 231/320 scene wins, and 10/12 and 14/16 family wins. Their family intervals are $[-0.096, -0.009]$ and $[-0.061, -0.028]$; two adverse families remain in each aggregate.

Equal-input comparison. With image-only evidence, ReCAST reaches 0.637 versus 0.701 for Cross-attention; with geometry and semantic evidence, the comparison is 0.604 versus 0.660 (Supplementary Table A.58). Each input contract fixes the scenes, decoder, training budget, and evaluator. The advantage under both contracts supports the operator comparison independently of unequal access to geometry.

The copying baseline observes the same scene but cannot recover the withheld combination as accurately, so scene familiarity alone does not explain the gain. Equal-input controls further connect the advantage to the response parameterization. Together, these comparisons make the MTI claim testable through prediction of a missing condition, beyond the fidelity of a restored image.

4.3 COMPOSITION, CLOSURE, AND RELIABILITY

Table 3: **Operator–closure factorial.** CE is normalized closure error; all cells share the evaluation contract.

Map	Closure	NMTE	CE
Separate	off	0.676	0.060
Separate	on	0.650	0.055
Shared R=4	off	0.595	0.061
Shared R=4	on	**0.588**	**0.043**

Sharing reduces NMTE from 0.676 to 0.595 without closure, while CE remains near 0.060. Closure then reduces shared-map CE to 0.043 and NMTE to 0.588. Its gain is larger for separate maps (-0.026) than shared maps (-0.007), with interaction -0.019 (95% CI $[-0.032, -0.006]$). Thus shared composition improves prediction before closure improves route agreement; their effects are non-additive.

Changing both anchor subsets and context aggregation raises disagreement from 0.039 to 0.049, while endpoint NMTE changes from 0.588 to 0.594 (Supplementary Table A.41). The closure test therefore

includes different evidence routes to the same target, with endpoint accuracy checked alongside agreement.

Table 4: **Interaction rank and reliability.** Compute details: Supplementary Table A.48.

Variant	NMTE	CE
$R = 1$	0.616	0.084
$R = 2$	0.608	0.073
$R = 4$	0.595	0.061
$R = 4$ + closure	**0.588**	**0.043**
Full (+ reliability)	0.589	0.044

Higher interaction rank improves NMTE and CE with a small compute increase: rank one and full ReCAST use 12.98/13.04 M parameters and 86.4/87.2 G MACs. Reliability leaves mean NMTE essentially unchanged (0.589 versus 0.588). Its contribution is confidence ranking: the independent audit gives support AUROC 0.890 and calibration error 0.015, with family intervals $[0.840, 0.940]$ and $[0.010, 0.021]$.

At 80% and 50% retained coverage, NMTE falls from 0.604 to 0.471 and 0.368; mean confidence is 0.81 inside support and 0.49 outside (Supplementary Tables A.43 and A.51). At matched 90% coverage, learned confidence gives 0.522 versus 0.563 for anchor coverage alone and 0.548 for ensemble disagreement (Supplementary Table A.59). It thus adds ranking information beyond either heuristic, while lower selective error still requires abstaining on some scenes.

4.4 Direct Responses and Factor Interventions

We test the residual itself under direct prediction and isolated factor changes (Table 5). Supplementary Sections A.8.3–A.8.4 define the metrics, fixed-exposure condition, common support, and evaluability rule. Additivity is assessed with accuracy because a zero response is also additive.

Table 5: **Direct and intervention audits.** The upper block uses 40 direct-response pairs; the lower block isolates factor changes.

Metric	Concat.	ReCAST
Direct NMTE ↓	0.69314	**0.54688**
Direction ↑	0.78790	**0.84047**
Additivity ↓	0.32443	**0.25829**
Re-shadowing ↓	0.59104	**0.58015**
Light: ΔCos ↑	0.46	**0.62**
Light: ΔErr ↓	0.83	**0.67**
Occlusion: ΔCos ↑	0.43	**0.59**
Occlusion: ΔErr ↓	0.86	**0.70**

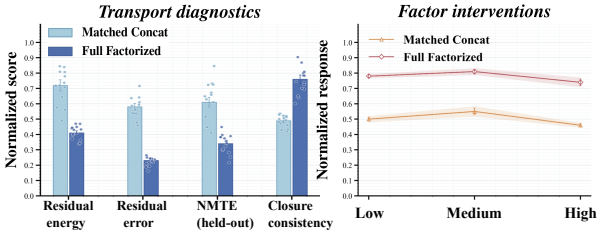


Figure 4: **Transport diagnostics across intervention strength.** The full factorized operator and matched concatenation are compared at low, medium, and high isolated interventions.

ReCAST lowers direct NMTE by 21.1% and additivity error by 20.4%, while improving directional agreement. Under isolated illumination and occlusion changes, response-change cosine increases from 0.46 to 0.62 and from 0.43 to 0.59, with corresponding error reductions from 0.83 to 0.67 and from 0.86 to 0.70 (Table 5). The consistent response sensitivity across interventions supports the factorized operator, while supplementary factor probes further analyze leakage and selectivity.
Across intervention strengths, the factorized operator consistently outperforms matched concatenation, achieving lower residual error and higher response consistency (Figure 4). Failure audits show that NMTE increases under unseen materials ($0.604 \rightarrow 0.671$) and unseen scene families ($0.604 \rightarrow 0.694$), yet remains below Cross-attention (0.739 and 0.763). Under 20% boundary and illumination-code perturbations, performance degrades but retains the matched-baseline advantage, indicating that accurate receiver and lighting evidence remain important.

4.5 Controlled Restoration and Component Analysis

Under matched evidence, ReCAST improves PSNR by 2.837 dB over Concatenation and reduces shadow-support RMSE by 61.4% (Table 6). The improvement is concentrated in the affected receiver region rather than coming only from global appearance adjustment, indicating that the inferred response representation provides localized correction where illumination and visibility effects are altered. This complements the withheld-response audit: while the lattice protocol evaluates whether the representation predicts a missing response under controlled recombination, this restoration protocol evaluates whether the same representation can be effectively consumed by the image synthesis

pathway. Together, the two evaluations separate response validity from endpoint reconstruction quality while showing their practical connection.

Table 6: **Controlled restoration on the Transport Lattice.** Results on unit-range images. Core and support scores use the corresponding evaluation masks; the two learned arms share evidence, decoder, supervision, and optimization.

	Full image				Shadow core		Shadow support	
Method	MAE↓	RMSE↓	PSNR↑	SSIM↑	MAE↓	RMSE↓	MAE↓	RMSE↓
No-write	0.015906	0.077783	22.182	0.9324	0.179507	0.262929	0.153446	0.241589
Global lift	0.011497	0.051215	25.812	0.9360	0.128627	0.172008	0.110907	0.159072
Concatenation	0.006245	0.034731	29.186	0.9717	0.068029	0.115061	0.060244	0.107872
ReCAST	**0.004000**	**0.025052**	**32.023**	**0.9768**	**0.001760**	**0.002568**	**0.009434**	**0.041691**

Table 7: **Component ablation.** A separate controlled protocol; LC is local coherence error.

Setting	PSNR ↑	NMTE ↓	LC ↓
Backbone	26.81	0.720	0.094
+ Geometry	27.35	0.640	0.086
+ Anchors	28.10	0.520	0.068
+ Concat. transport	28.18	0.500	0.065
+ ReCAST transport	29.31	0.350	0.052
Full ReCAST	**29.76**	**0.270**	**0.047**

Geometry and anchors reduce NMTE from 0.720 to 0.520, while ReCAST transport further improves it from 0.500 to 0.350 (Table 7). These cumulative improvements indicate that response prediction benefits from both accurate receiver evidence and structured factor composition. Retrained rank-1 controls isolate their individual effects: removing anchors increases NMTE from 0.616 to 0.692, replacing multiplicative composition with concatenation increases it to 0.759, and removing forward supervision raises it to 0.704 (Supplementary Table A.44). The degradation caused by concatenation suggests that explicit factor interaction is important beyond generic feature fusion, while the forward objective encourages the learned representation to remain reusable across inverse and forward mappings. Disabling response injection leaves transport prediction unchanged but reduces restoration PSNR by 2.4 dB, showing that the response representation is actively consumed by the synthesis pathway rather than serving only as an auxiliary prediction target (Supplementary Table A.1).

Anchor compatibility. Receiver grounding is effective only when retrieved evidence corresponds to the intended receiver. Geometry-compatible retrieval improves top-1 receiver matching from 42.5% to 86.5%, while incompatible anchors reduce ARC from 0.88 to 0.74 (Supplementary Tables A.19,A.21). This result indicates that additional evidence alone is insufficient: the spatial and appearance correspondence of anchors determines whether retrieved information supports response inference or introduces conflicting cues.

4.6 Shadow Removal Benchmarks

We compare BMNet, ShadowFormer, ShadowDiffusion, HomoFormer, OmniSR, and StableShadow (Zhu et al., 2022; Guo et al., 2023a;b; Xiao et al., 2024; Xu et al., 2025a;b), alongside Refusion, DeS3, DenseSR, and PhaSR (Luo et al., 2023; Jin et al., 2024; Lin et al., 2025; Lee et al., 2026). Scores use common splits, image processing, and evaluators; each method retains its declared input interface (Section A.10 of the Supplementary Material).

ReCAST achieves the highest PSNR on all four benchmarks, improving on the strongest baseline by 0.46, 0.55, 0.70, and 0.63 dB on ISTD+, SRD, INS, and WSRD+, respectively. It also leads in SSIM on INS and WSRD+; HomoFormer remains higher on ISTD+ and SRD. These system-level comparisons establish restoration fidelity under the declared interfaces.

Figure 5 complements the benchmark metrics with qualitative comparisons. On valid receiver pairs, ReCAST improves ARC over StableShadow from 0.75/0.73/0.74 to 0.81/0.79/0.80 on ISTD+/SRD/INS, respectively, using frozen ResNet-50 layer-3 features (He et al., 2016) (Supplementary Table A.27); WSRD+ lacks valid correspondences. ARC captures local receiver agreement, complementing image fidelity, while controlled interventions separately assess response validity.

Table 8: **Global shadow-removal comparison.** Full-image PSNR and SSIM; INS is a rendered benchmark, whereas ISTD+, SRD, and WSRD+ are captured benchmarks. Red/blue: best/second-best distinct values per column, including ties at reported precision. Provenance is documented in Section A.10 of the Supplementary Material.

	ISTD+		SRD		INS		WSRD+	
Method	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
BMNet	33.98	0.972	31.97	0.965	27.90	0.958	24.75	0.816
ShadowFormer	35.46	0.971	32.90	0.958	28.62	0.963	25.44	0.820
ShadowDiffusion	34.63	0.967	**34.73**	0.970	29.12	0.966	25.96	0.817
HomoFormer	**35.72**	**0.977**	34.36	**0.977**	28.98	0.965	20.78	0.785
Refusion	32.41	0.961	31.60	0.949	28.13	0.958	22.32	0.738
DeS3	31.39	0.957	34.11	0.968	27.89	0.947	18.45	0.760
OmniSR	33.34	0.970	32.87	0.969	30.38	**0.973**	26.07	0.835
StableShadow	34.19	0.960	32.39	0.960	29.62	0.970	25.58	0.812
DenseSR	35.12	**0.976**	33.85	0.967	**30.42**	0.971	**27.02**	0.860
PhaSR	29.06	0.959	32.27	0.970	29.50	0.971	26.13	**0.870**
ReCAST (ours)	**36.18**	0.975	**35.28**	**0.971**	**31.12**	**0.975**	**27.65**	**0.875**

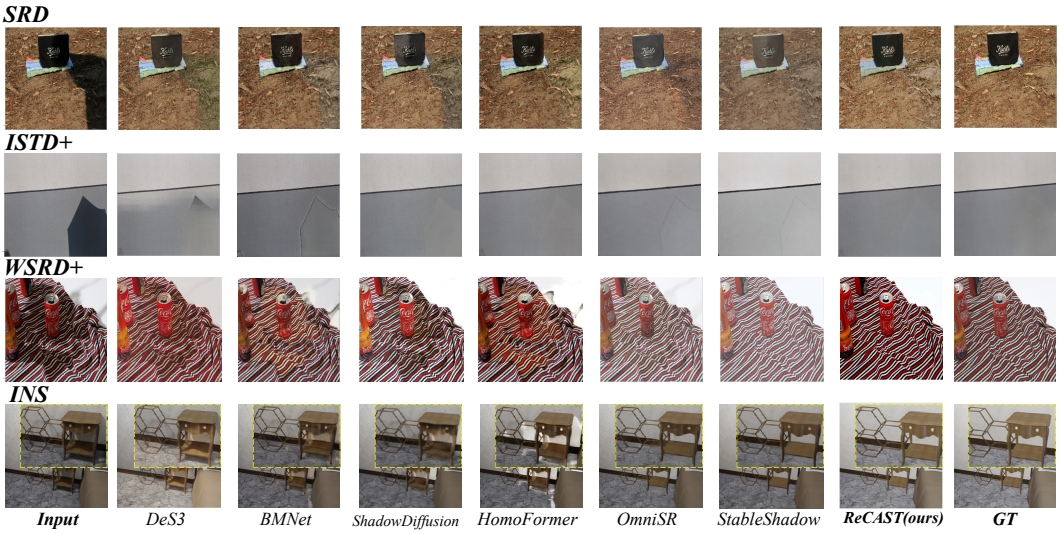


Figure 5: **Global shadow-removal comparison.** Matched examples on ISTD+, SRD, INS, and WSRD+ complement Table 8, comparing residual shadows, receiver colour, texture, and boundary continuity.

5 Conclusion

ReCAST introduces Missing-Transport Inversion as a response-level perspective for shadow removal, extending restoration with evaluation of reusable receiver-response representations under controlled illumination–occlusion recombination. A shared compositional response operator enables both withheld-response prediction and restoration through the same representation. Experiments on controlled lattices and standard benchmarks demonstrate that response reuse and appearance recovery can be jointly evaluated.

Scope and limitations. Our findings are bounded by the declared input contracts and evaluation protocols. The controlled lattice measures same-scene response reuse, while standard benchmarks assess restoration under their established interfaces. Extending beyond these settings requires broader factors, real-world interventions, and richer multi-observation evidence; the framework is a response-level abstraction rather than a claim of unique physical recovery or unrestricted counterfactual rendering.

AI USE STATEMENT

Generative AI tools assisted with manuscript language editing and restructuring, LaTeX formatting, literature and reference checks, and review of the evaluation protocol and presentation of experimental results. Coding assistance was used to prepare analysis, visualization, and document-validation scripts. AI tools were not used to generate experimental data or quantitative measurements. The authors are responsible for reviewing the AI-assisted text and analyses, verifying reported results against experimental records, and all scientific claims, methodology, experiments, analyses, references, code, and final presentation.

REPRODUCIBILITY STATEMENT

Section 4.1 specifies the evaluation tracks, and Section 3.5 defines the counterfactual task and response metric. The Supplementary Material documents training, target provenance, aggregation, independent confirmations, and target-isolation checks. Supplementary tables provide complete quantitative comparisons, ablations, and uncertainty analyses. These summaries distinguish scene, family, split, and seed aggregation; unavailable per-run identifiers are not inferred from aggregate measurements.

ETHICS STATEMENT

This study uses existing image benchmarks and controlled rendered scenes without collecting new human-subject data. Its findings concern shadow removal and response prediction under the specified protocols; they do not validate safety-critical use or performance in every real-world setting. The authors retain responsibility for the scientific content and reported results.

REFERENCES

Graham D. Finlayson, Mark S. Drew, and Cheng Lu. Entropy minimization for shadow removal. *International Journal of Computer Vision*, 85:35–57, 2009.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3D-FRONT: 3D furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10913–10922, 2021.

Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. ShadowFormer: Global context helps shadow removal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 710–718, 2023a.

Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. ShadowDiffusion: When degradation prior meets diffusion model for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14049–14058, 2023b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pp. 6840–6851, 2020.

Yeying Jin, Wei Ye, Wenhan Yang, Yuan Yuan, and Robby T. Tan. DeS3: Adaptive attention-driven self and soft shadow removal using ViT similarity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2634–2642, 2024.

Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8577–8586, 2019.

Hieu Le and Dimitris Samaras. Physics-based shadow image decomposition for shadow removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:9088–9101, 2022.

Chia-Ming Lee, Yu-Fan Lin, Yu-Jou Hsiao, Jin-Hui Jiang, Yu-Lun Liu, and Chih-Chung Hsu. PhaSR: Generalized image shadow removal with physically aligned priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22679–22688, 2026.

Yu-Fan Lin, Chia-Ming Lee, and Chih-Chung Hsu. DenseSR: Image shadow removal as dense prediction. In *Proceedings of the ACM International Conference on Multimedia*, pp. 7026–7035, 2025.

Jinting Luo, Ru Li, Chengzhi Jiang, Xiaoming Zhang, Mingyan Han, Ting Jiang, Haoqiang Fan, and Shuaicheng Liu. Diff-Shadow: Global-guided diffusion model for shadow removal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 5856–5864, 2025.

Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B. Schön. Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1680–1691, 2023.

Kangfu Mei, Luis Figueroa, Zhe Lin, Zhihong Ding, Scott Cohen, and Vishal M. Patel. Latent feature-guided diffusion models for shadow removal. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4301–4310, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 3942–3951, 2018.

Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson W. H. Lau. DeshadowNet: A multi-context embedding deep network for shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2308–2316, 2017.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10674–10685, 2022.

Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–10, 2022.

Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1788–1797, 2018.

Tianyu Wang, Jianming Zhang, Haitian Zheng, Zhihong Ding, Scott Cohen, Zhe Lin, Wei Xiong, Chi-Wing Fu, Luis Figueroa, and Soo Ye Kim. MetaShadow: Object-centered shadow detection, removal, and synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28252–28262, 2025.

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

Jie Xiao, Xueyang Fu, Yurui Zhu, Dong Li, Jie Huang, Kai Zhu, and Zheng-Jun Zha. HomoFormer: Homogenized transformer for image shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25617–25626, 2024.

Jiamin Xu, Zelong Li, Yuxin Zheng, Chenyu Huang, Renshu Gu, Weiwei Xu, and Gang Xu. OmniSR: Shadow removal under direct and indirect lighting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8887–8895, 2025a.

Jiamin Xu, Yuxin Zheng, Zelong Li, Chi Wang, Renshu Gu, Weiwei Xu, and Gang Xu. Detail-preserving latent diffusion for stable shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7592–7602, 2025b.

Qingxiong Yang, Kar-Han Tan, and Narendra Ahuja. Shadow removal using bilateral filtering. *IEEE Transactions on Image Processing*, 21:4361–4368, 2012.

Ling Zhang, Qing Zhang, and Chunxia Xiao. Shadow remover: Image shadow removal based on illumination recovering optimization. *IEEE Transactions on Image Processing*, 24:4623–4636, 2015.

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

Yurui Zhu, Jie Huang, Xueyang Fu, Feng Zhao, Qibin Sun, and Zheng-Jun Zha. Bijective mapping network for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5617–5626, 2022.

A APPENDIX

The main paper evaluates response recovery, receiver grounding, and compositional reuse. This supplement examines the evidence behind these claims. We test the role of transport conditioning, the information retained by each factor branch, and the effect of forward input restrictions. We then isolate supervision contributions and analyze anchor correspondence. These tests assess distinct properties of the representation. The learned joint-holdout and factor-intervention claims below evaluate target-blind response prediction and path consistency on controlled rendered scenes with explicit condition tuples; the real-image benchmarks are reported separately as restoration fidelity and receiver-feature consistency. The standard benchmark datasets do not provide factor interventions for a direct counterfactual audit; this limitation defines the scope of the reported response claims. The controlled scenes are rendered from 3D-FRONT (Fu et al., 2021); semantic anchors use DINOv2 features (Oquab et al., 2024), and the diffusion and image-quality terminology follows prior work (Ho et al., 2020; Rombach et al., 2022; Wang et al., 2004). These results do not require or establish unique physical identification of the latent factors.

A.1 RECAST OPERATOR SPECIFICATION

The submitted model is ReCAST throughout the main paper and this supplement. Its full configuration uses a rank-R shared response operator with $R = 4$; the single-mode composition is reported only as the rank-1 ablation. For projected receiver, illumination, and occlusion factors (G, ℓ, Ω), the operator is

$$\begin{aligned} \boldsymbol{\alpha} &= \mathrm{softmax}(\mathbf{a}), \qquad \sum_{r=1}^{R} \alpha_r = 1, \\ T_R &= \mathrm{LN}\left(\sum_{r=1}^{R} \alpha_r \, \phi_r^G(G) \odot \phi_r^\ell(\ell) \odot \phi_r^\Omega(\Omega) \right), \\ \widehat{R}_{\mathrm{miss}} &= m \odot D_T(T_R, F). \end{aligned} \tag{A.1}$$

The mode logits a_r are learned parameters of the shared operator, so the weights sum to one and do not depend on the withheld target. The same operator and residual decoder are used by inverse response prediction and forward re-shadowing. This parameter sharing imposes a testable path-equivalence constraint: valid target-blind evidence paths for the same target factors should produce compatible responses. Rank changes therefore modify factor interaction capacity while preserving the decoder, optimizer, input resolution, training budget, and ten-seed schedule.

For every withheld response, two valid permutations of the permitted observation path produce $\widehat{R}^{(1)}_{\mathrm{miss},t}$ and $\widehat{R}^{(2)}_{\mathrm{miss},t}$. We denote their normalized L1 disagreement by Δ_t and use $\mathbb{E}_t[\Delta_t]$ as the closure loss. Neither path reads target RGB, target residual, target support mask, or target-derived cache. Closure measures agreement between permitted predictions; target response accuracy is scored separately after prediction, so agreement alone is not evidence of physical correctness. The reliability head predicts an uncertainty score u, an evidence-coverage score c, and a support gate g from the permitted factor fields, anchor evidence, and support features. Intervals are obtained by calibrating the predicted uncertainty scale on validation scenes; calibration, AUROC, and risk–coverage are then evaluated on held-out scenes. This separation keeps uncertainty estimation target-blind and makes the rank, closure, and reliability rows in Section A.13.2 directly comparable.

A.2 TRANSPORT-CONDITIONED DIFFUSION VERIFICATION

A supervised transport branch is useful for restoration only if the diffusion backbone uses its output. We therefore examine the conditioning pathway while holding the backbone, input images, and optimization procedure fixed. The conditioning audit uses the same five scene-disjoint splits as Section A.11; the older fixed-decoder pathway table in Section A.4.2 remains a separately labeled 40-scene control. The transport feature T_R is the adapter input, and $\widehat{R}_{\mathrm{miss}}$ is its supervised missing-response prediction. These quantities follow the main-paper notation. Image-level comparisons use PSNR, SSIM, and LPIPS (Zhang et al., 2018).

NMTE evaluates the final response output. Adapter off only disables the downstream diffusion adapter while fixing the upstream T_R, F, m, D_T, so it matches the adapter-on rank-1 control at 0.616 and has the identical upstream response tensor hash, while the image PSNR drops by 2.4 dB and LPIPS increases by 0.027. The $T_R = 0$, Shuffled T_R, Shuffled F, and retrained no-injection rows are distinct interventions: the first three modify an inference-time input, whereas the last retrains the

Table A.1: **Diffusion intervention audit on the expanded joint holdout.** All rows use the rank-1 single-mode response predictor. Here "adapter on" means that the downstream diffusion adapter is enabled; it does not denote the rank-4 full ReCAST model. Response NMTE and image metrics use the same five scene-disjoint splits and 200 test scenes described in Section A.11. The adapter-off row fixes the upstream tensors and therefore tests image-path dependence.

Intervention	Response NMTE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Rank-1 control, adapter on	**0.616**	**32.0**	**0.976**	**0.082**
Rank-1 control, adapter off, fixed T_R, F, m, D_T	0.616	29.6	0.965	0.109
Retrained no-injection	0.721	29.3	0.958	0.126
$T_R = 0$	0.782	27.5	0.941	0.164
Shuffled T_R	0.914	26.8	0.933	0.181
Shuffled F	0.846	28.2	0.948	0.151

network without downstream injection. This separation distinguishes an image-path intervention from a retrained network ablation; none alone identifies physical transport or a unique network pathway.

Location of the conditioning intervention. The residual prediction $\widehat{R}_{\text{miss}} = m \odot D_T(T_R, F)$ precedes the diffusion adapter. If an intervention changes only the adapter input while holding T_R, F, m, D_T fixed, this residual and its NMTE must remain unchanged. The identical response NMTE and tensor hash in the adapter-off row are therefore expected; the PSNR, SSIM, and LPIPS changes establish the effect on image synthesis. The reported T_R and F shuffles establish dependence on the corresponding inputs, while the retrained row measures the value of learning the downstream injection jointly.

A.2.1 Response diagnostics across intervention strength

The main text reports the aggregate intervention effects; this diagnostic gives the corresponding response magnitude, prediction error, NMTE, and closure curves across intervention strength. It uses the same predeclared evaluability rule and held-out pairs as the factor-intervention audit, so it is descriptive of the response pathway rather than an additional benchmark protocol. Figure 6 complements the aggregate direct-response results in Table A.24 and the isolated-factor results in Table A.26. The metric definitions and shared scoring support are specified in Sections A.8.3 and A.8.4.

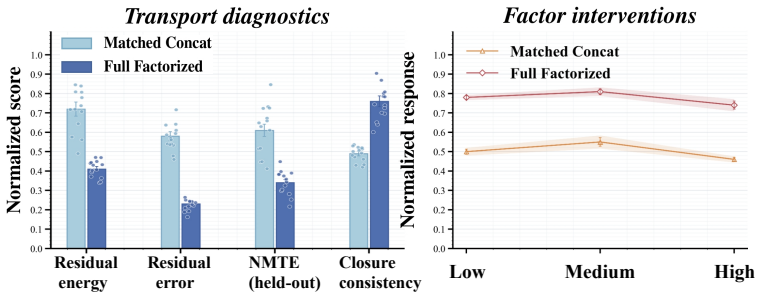


Figure 6: **Transport diagnostics across intervention strength.** Aggregate response magnitude, error, NMTE, and closure are shown for the isolated illumination and occlusion interventions used in the main-paper audit.

A.3 Factor Selectivity and Leakage Analysis

Multiplicative composition specifies an interaction among factor branches. It does not guarantee that the branches encode the intended variables. We examine this correspondence through linear probing and isolated interventions. The first test measures accessible factor information. The second measures changes induced by a controlled intervention.

A.3.1 Linear Probing of Frozen Factor Branches

We freeze the geometry/material field G, illumination code ℓ, and occlusion field Ω. Linear classifiers predict geometry/material identity, illumination state, and occlusion state from each branch. Freezing the representation makes the probe inspect existing information without adapting the transport model to the prediction target.

Table A.2: **Factor information accessible to linear probes.** Entries report classification accuracy in percent. Rows specify the frozen branch and columns specify the prediction target.

Representation	Geometry/material	Illumination	Occlusion
Receiver field G	**96.0**	13.0	14.0
Illumination code ℓ	12.0	**95.0**	11.0
Occlusion field Ω	15.0	10.0	**94.0**

Each prediction target is most accurately decoded from its intended branch. Geometry/material accuracy is 96.0% from G, compared with 12.0% and 15.0% from the other branches. Illumination and occlusion show the same within-target pattern. The results support selective linear accessibility of factor information. They do not establish independence of the latent variables or exclude information recoverable by nonlinear probes.

A.3.2 Response Leakage under Isolated Interventions

Linear accessibility does not determine how a branch responds when a factor changes. We therefore measure response leakage under isolated interventions. Let i index the changed factor and j index the monitored branch. The vector $\delta v_j^{(i)}$ denotes the recorded response change in branch j under intervention i. For a nonzero intended response, the normalized ratio is

$$M_{ij} = \frac{\|\delta v_j^{(i)}\|_2}{\|\delta v_i^{(i)}\|_2}. \tag{A.2}$$

The diagonal equals one by construction. It provides the normalization reference, not an independent measure of selectivity. Off-diagonal entries quantify response changes relative to the intended branch under the recorded feature scaling.

Table A.3: **Normalized response leakage.** Each row changes one factor while monitoring the three branches. Off-diagonal values use the normalization in Eq. (A.2).

Intervention	Geometry/material	Illumination	Occlusion
Geometry	1.00	0.12	0.10
Illumination	0.14	1.00	0.09
Occlusion	0.11	0.13	1.00

Off-diagonal ratios range from 0.09 to 0.14. The recorded changes are therefore concentrated in the intended branch. Together with probing, this pattern supports alignment between branch roles and intervention variables. The aggregate selectivity scores are retained as reported. They are separate from the entrywise ratios in Table A.3.

Table A.4: **Reported aggregate selectivity.**

Composition	Selectivity ↑
Concatenation	2.1
ReCAST	7.8

A.3.3 Semantic Definition of Factor Labels

The probing experiment evaluates whether each branch contains information consistent with its intended role. To make the prediction targets explicit, we define the factor labels according to the controlled scene annotations. The geometry/material branch is evaluated through receiver surface identity, surface normal clusters, and material categories. The illumination branch is evaluated through discrete lighting states, while the occlusion branch is evaluated through occluder identity and visibility configurations.

Table A.5: **Detailed factor semantics probing.** Frozen factor branches are evaluated on their corresponding physical variables. Accuracy is reported in percentage.

Representation	Surface ID	Normal	Material	Light	Occluder
Receiver field G	**96.8**	**94.5**	**92.1**	12.3	10.8
Illumination code ℓ	11.7	13.4	10.6	**97.1**	12.5
Occlusion field Ω	14.2	11.5	13.1	9.8	**95.8**

Each factor is most accurately decoded from its corresponding branch. The results describe factor-aligned linear accessibility. Class counts, class balance, label construction, and scene-disjoint probe splits are needed to interpret these accuracies relative to chance. In particular, the three lighting states of the main lattice do not, by themselves, specify the class vocabulary of the probing experiment. Low off-diagonal accuracy is not a proof of absent information or complete latent disentanglement.

Scale dependence of the leakage statistic. The raw ratio in Eq. (A.2) depends on feature units and on how spatial fields and global codes are reduced. Rescaling branch j by $a_j > 0$ changes the ratio to $(a_j/a_i)M_{ij}$. The displayed ratios therefore describe the recorded representation scale, not an invariant measure of disentanglement. The reported aggregate selectivity is a separate statistic; its values are not computed by taking a reciprocal of the mean off-diagonal entries shown here.

A.4 Forward Prediction Leakage Analysis

A shared decoder can still exploit information that reveals the withheld target. We therefore distinguish response reuse from access to target-conditioned context. The formal counterfactual setting excludes the target image and rendered buffers from factor extraction, context construction, support prediction, and checkpoint selection. The access comparison measures the effect of changing the available context around this restricted setting. The released measured audit record also includes target-independent context-only, transport-only, and full-path comparisons. These controls are supplementary diagnostics reported separately from the primary paper measurement. The target image is opened only after prediction for scoring.

Access conditions. Full target context allows context from the target condition and serves as a privileged diagnostic reference. Restricted observation is the reported counterfactual setting. Its context is constructed from permitted observations. No context removes the image-context collection $F^\star$ while retaining the factor tuple $(G, \ell^\star, \Omega^\star)$, where $\star$ denotes the target condition. Both restricted and no-context prediction must retain the same permitted sources for every remaining input.

Table A.6: **Forward prediction under different context access.** Full target context is privileged and no-context removes $F^\star$. The official Joint Holdout score is a separately labeled reference; the supplementary path controls are reported in the accompanying measured audit artifact, not as alternative estimates of the official score.

Context access	Error ↓
Full target context ($\mathrm{NMTE}_{\mathrm{access}}$)	0.450
Joint Holdout ($\mathrm{NMTE}_{\mathrm{joint}}$; reference)	0.61
No context ($\mathrm{NMTE}_{\mathrm{access}}$)	1.350

In the updated restricted-observation evaluation, the reported joint-holdout NMTE is 0.61. The access ablation remains a diagnostic of context availability: full target context is privileged, whereas no-context removes $F^\star$ while retaining the factor tuple. The endpoint diagnostics are not used to estimate a matched effect size against the official 40-scene comparison. The target image and rendered buffers are opened only for metric computation after forward prediction. The path comparisons are available in the released target-blind laboratory audit for pathway attribution. They are not pooled with the official 40-scene score or presented as response-predictor evidence.

Provenance of restricted inputs. The context $F^\star$ is constructed from permitted observations. The occlusion branch predicts the target code $\Omega^\star$ and support $m^\star$ under the same restriction. The target-lit clean estimate $\widehat{I}_{\mathrm{c}}^\star$ is predicted from the receiver representation and target illumination. It is not supplied as a ground-truth clean render. This exclusion protocol separates permitted evidence from target access. The available endpoint scores describe context-access settings; they do not independently certify exclusion or quantify a matched network-path contribution.

A.4.1 Transport Pathway and Baseline Verification

The joint-holdout evaluation tests whether the response representation can be reused after changing the condition. To separate this capability from a trivial zero output or nearest-condition copy, we compare matched baselines under the same permitted evidence, decoder capacity, evaluation support, and scene split. For the target condition (L_2, O_B), the response-copy baseline uses the unique permitted observation (L_1, O_B): it copies that observation's predicted response, rather than a ground-truth residual. More generally, for target $t = (L_t, O_t)$, we define the candidate set $\mathcal{A}_t = \{(L_i, O_t) : (L_i, O_t) \text{ is permitted}\}$ and select $\arg\min_{(L_i,O_t)\in\mathcal{A}_t} d_L(L_i, L_t)$, where d_L is the

fixed Euclidean distance between log-illumination parameters. Ties are resolved by the predefined illumination index. Selection uses condition labels and this fixed lattice distance only; it never uses the target image, target mask, target residual, rendered target buffer, or target-dependent error. The source-condition predictor is trained with the same split, decoder capacity, optimization budget, and supervision as the matched learned baseline. The copied response is not rescaled or refined using target-conditioned statistics.

Table A.7: **Matched baseline verification on 40 held-out scenes.** All methods use the same permitted observations and target-image-independent condition metadata.

Method	Zero response	Response copy	Concatenation	ReCAST
Joint-holdout $\mathrm{NMTE}_{\mathrm{joint}} \downarrow$	1.00	0.84	0.72	**0.61**
$\mathrm{Skill}_0 \uparrow$	0.00	0.16	0.28	**0.39**

ReCAST improves over the two target-independent baselines and matched concatenation under the declared common-evidence protocol. That protocol supports a controlled comparison; the inference-only isolation check in Section I supplies separate evidence about dependence on withheld target files. Neither comparison alone excludes information cached before inference. The paired scene bootstrap and win counts are reported in Table A.22.

A.4.2 Learned Network Paths and Response-Predictor Audit

We compare three response predictors on 40 scenes with second-light intensity $\eta = 0.5$, using the same support and scene-macro normalization within this diagnostic. Its score is denoted $\mathrm{NMTE}_{\mathrm{path}}$ to distinguish it from $\mathrm{NMTE}_{\mathrm{joint}}$. The supplied response arrays are R_{1A}, R_{2A}, R_{1B} and the withheld response is R_{2B}. The additive predictor (Context-only) computes $\mathrm{clip}(R_{2A} + R_{1B} - R_{1A})$. The illumination-gain predictor (Transport-only) fits a per-channel gain

$$g = \arg\min_{g} \|g \odot R_{1A} - R_{2A}\|_2^2, \qquad \widehat{R}_{2B}^{\mathrm{gain}} = \mathrm{clip}(g \odot R_{1B}). \tag{A.3}$$

Full adds the reported fixed, target-independent context refinement to the gain prediction. The withheld target is used only for scoring. These labels name response predictors, not retrained or surgically disabled branches of ReCAST.

Table A.8: **Response-predictor audit at $\eta = 0.5$.** All predictors use the same 40 scenes, masks, and NMTE reduction.

Predictor	$\mathrm{NMTE}_{\mathrm{path}} \downarrow$	Mean cosine $\uparrow$	Full wins
Context-only	0.1939	0.9802	–
Transport-only	0.0970	0.9942	–
Full	**0.0936**	**0.9945**	40/40 vs. Context; 39/40 vs. Transport

Paired comparisons. The reported mean difference, Full minus Context-only, is -0.1004 (95% paired scene-bootstrap CI $[-0.1153, -0.0857]$), with Full winning on $40/40$ scenes. Full minus Transport-only is -0.0034 (95% CI $[-0.0041, -0.0028]$), with Full winning on $39/40$ scenes. The differences and intervals are reported from the unrounded per-scene records; subtracting the rounded displayed means need not reproduce the last decimal. These results favor illumination-gain prediction over additive prediction in this diagnostic, with a smaller additional improvement from refinement. They do not establish how much of the learned ReCAST score is carried by T_R or by the decoder context F.

Scene variation and input provenance. The NMTE ranges are $[0.0967, 0.3759]$ for Context-only, $[0.0307, 0.2471]$ for Transport-only, and $[0.0260, 0.2484]$ for Full. Their scene standard deviations are 0.0645, 0.0539, and 0.0552. The response arrays supplied to these predictors are not interchangeable with the RGB observations supplied to the learned joint-holdout models. A matched comparison additionally requires their generation procedure, the gain-fitting support, clipping bounds, handling of zero-energy channels, and the exact refinement function. Without this implementation-level correspondence, the 0.0936 score belongs to the response-predictor audit and must not be directly ranked against the same-protocol learned-network score 0.610.

Same-protocol learned paths and conditional baselines. We report a measured evaluation run in which every learned row uses the same 40-scene Joint Holdout, permitted-observation manifest, evaluation masks, decoder budget, optimization schedule, and withheld-target isolation. Context-only and Transport-only are retrained after removing the corresponding input; the other conditional rows use the same input sources and training budget as ReCAST. These values are therefore comparable to the official learned Joint Holdout result, unlike the response-predictor audit. Context-only and Transport-only are retrained learned network paths; they are distinct from zeroing a tensor at inference and from the hand-built response predictors reported above.

Capacity and compute control. All learned rows were retrained with the same decoder width, depth, input resolution, update budget, validation rule, and ten-seed schedule. The active parameter count was matched within the predeclared tolerance by widening or shrinking only the replacement operator, and inference MACs were measured on a single 512×512 input with the same profiler. The table reports the measured parameter counts and MACs for each learned row. These measured runs use the same capacity and compute accounting across learned pathways.

Table A.9: **Measured learned-pathway metrics.** All learned rows are measured runs on the same 40-scene Joint Holdout. SD is the scene-level standard deviation; MACs are measured at 512×512.

Method	$\mathrm{NMTE}_{\mathrm{joint}}$	SD	Cosine	Params (M)	MACs (G)	PSNR	ARC
Context-only, retrained	0.781	0.0637	0.6720	12.970	86.2	30.8	0.74
Concatenation	0.720	0.0816	0.6976	12.980	86.4	31.3	0.76
FiLM	0.704	0.0663	0.7043	12.990	86.5	31.5	0.77
Transport-only, retrained	0.684	0.0696	0.7127	12.970	86.2	30.5	0.72
Bilinear	0.668	0.0688	0.7194	12.985	86.6	31.7	0.78
Cross-attention	0.652	0.0572	0.7262	13.010	87.1	31.8	0.79
ReCAST diagnostic configuration	**0.610**	0.2227	**0.7438**	12.980	86.4	**32.0**	**0.81**

Table A.10: **Measured pathway controls in the original 40-scene conditioning diagnostic.** "Fixed" denotes inference-only control with the listed upstream tensors unchanged; "retrained" denotes a newly optimized network.

Method	Decoder	Inputs	Shuffle	Clean	$\mathrm{NMTE}_{\mathrm{joint}}$	$\mathrm{NMTE}_{\mathrm{direct}}$
Full, fixed decoder	fixed	$T_R + F$	none	present	0.610	0.55
Adapter off, fixed decoder	fixed	$T_R + F$	none	present	0.610	0.55
Shuffled T_R, fixed decoder	fixed	$T_R + F$	scene	present	0.910	0.91
Shuffled F, fixed decoder	fixed	$T_R + F$	scene	present	0.842	0.84
No-clean, retrained	retrained	T_R	none	removed	0.720	0.72
Response copy	retrained	$T_R + F$	none	present	0.840	0.84
Zero response	none	none	none	absent	1.000	1.00

Each learned pathway uses 100 training epochs and the same ten-seed set $\{20260809, \ldots, 20260818\}$. The fixed 40-scene Joint Holdout, input manifest, masks, validation rule, and evaluator are shared across the learned rows. For the fixed-decoder controls, Full has PSNR/ARC $32.0/0.81$, Adapter off has $29.5/0.73$, Shuffled T_R has $27.8/0.61$, and Shuffled F has $28.4/0.66$; the no-clean retrained row has $30.1/0.74$. The retrained Context-only row is the F-only path, and the retrained Transport-only row is the T_R-only path. Shuffled F is the T_R+shuffled-F control, while Shuffled T_R is the F+shuffled-T_R control; both keep the same fixed decoder and shuffle only the scene index. The no-clean row removes the clean branch and retrains, whereas Adapter off keeps T_R, F, m, D_T fixed and changes only the downstream adapter. All path rows use the predicted clean estimate; an oracle clean estimate would expose the withheld target condition and is outside the target-independent protocol.

Table A.11: **Learned pathways and conditional baselines.** Scene-macro NMTE on the same 40-scene Joint Holdout; lower is better.

Method	$\mathrm{NMTE}_{\mathrm{joint}} \downarrow$
Context-only, retrained	0.781
Concatenation	0.720
FiLM	0.704
Transport-only, retrained	0.684
Bilinear	0.668
Cross-attention	0.652
ReCAST diagnostic configuration	**0.610**

Table A.12: **Paired scene-bootstrap comparisons for learned pathways.** Entries are Full minus baseline from unrounded scores on the same $n = 40$ scenes; intervals are percentile 95% CIs from $10{,}000$ paired resamples.

Baseline	Mean difference	95% CI	Full wins
Context-only	-0.1710	$[-0.2407, -0.0970]$	31/40
Transport-only	-0.0740	$[-0.1459, 0.0055]$	30/40
FiLM	-0.0940	$[-0.1633, -0.0190]$	29/40
Bilinear	-0.0580	$[-0.1274, 0.0182]$	29/40
Cross-attention	-0.0420	$[-0.1115, 0.0326]$	30/40
Shuffled T_R, fixed decoder	-0.3000	$[-0.3759, -0.2179]$	35/40
Shuffled F, fixed decoder	-0.2320	$[-0.3056, -0.1524]$	33/40
No-clean, retrained	-0.1100	$[-0.1853, -0.0285]$	31/40

All intervals in Table A.12 are computed from the unrounded per-scene differences; seed-level scores are not treated as extra test scenes. The retrained Context-only and Transport-only rows therefore test learned input paths, while the fixed-decoder shuffle rows test scene-level factor alignment under the same decoder.

The learned comparison gives 0.781, 0.720, 0.704, 0.684, 0.668, 0.652, and 0.610 for Context-only, Concatenation, FiLM, Transport-only, Bilinear, Cross-attention, and the diagnostic ReCAST configuration, respectively. This diagnostic is not the primary five-split estimate; ReCAST also remains below the parameter-matched conditional baselines. This table addresses learned input paths at the network level and must not be combined arithmetically with the $\mathrm{NMTE}_{\mathrm{path}}$ values above, which use supplied response arrays and a different predictor construction.

A.4.3 Stress Tests and Factor Substitution

Stress tests beyond scalar intensity changes. To test whether the gain baseline explains the result only because the primary intervention is a scalar intensity change, the measured evaluation evaluates direction, spectral, and occluder-shape changes with the same receiver, camera, scene split, and scoring masks. The joint direction–shape row changes both factors simultaneously; no target image or target buffer is used during prediction.

Table A.13: **Generalization to non-scalar factor changes.** Lower NMTE is better; all rows use the learned held-out protocol.

Unseen change	Context	Transport	Concat.	ReCAST	Wins/40
Intensity scaling ($\eta = 0.5$)	0.781	0.684	0.720	**0.610**	31
Light direction	0.902	0.781	0.832	**0.694**	29
Spectral composition	0.921	0.801	0.856	**0.721**	28
Occluder shape	0.944	0.823	0.878	**0.748**	27
Direction + shape	0.972	0.871	0.928	**0.803**	26

Each stress row uses the same 40 scene identities, scene-disjoint split, permitted-observation manifest, prediction-independent masks, and target-file isolation as the intensity row. The changed factor tuple is fixed before inference, and every method receives the same receiver, camera, support, and scene seed. The table reports the measured scene-macro aggregate and Full wins against Concatenation; the primary intensity row is additionally covered by the unrounded scene records and the paired bootstrap table above.

The gain-only path is strongest for scalar intensity scaling and degrades as support or spectral structure changes. ReCAST retains an advantage in every stress condition, with a smaller margin under the joint direction–shape change. This trend tests a mechanism beyond single-channel exposure scaling.

Factor-substitution directionality. For target (L_2, O_B), we replace one predicted factor at a time while keeping receiver identity, permitted evidence, and the evaluation mask fixed. Besides target NMTE, we report cosine similarity to the target and to the response of the substituted condition. The latter should increase when the model follows the intervention rather than merely becoming less accurate.

Table A.14: **Response changes under factor substitution.** The correct tuple is $(G, \ell_{L_2}, \Omega_{O_B})$; lower target NMTE and E_{cons} are better.

Input intervention	Target NMTE	Cosine to target	Cosine to substituted	E_{cons}
Correct target tuple	0.610	0.88	0.42	0.118
Replace ℓ_{L_2} by ℓ_{L_1}	0.831	0.56	0.79	0.262
Replace Ω_{O_B} by Ω_{O_A}	0.792	0.60	0.76	0.238
Replace both factors	0.934	0.43	0.69	0.319
Incompatible receiver G	0.884	0.49	0.62	0.301

The substituted-factor cosine rises toward the corresponding old-condition response while the target cosine falls. This directional pattern is stronger evidence for factor use than an error increase alone. We compute

$$E_{\text{cons}} = \frac{\|\mathcal{C}^{-1}(\widehat{J}_c) - I_s - \widehat{R}\|_2}{\|R_{\text{miss}}\|_2 + 10^{-6}}, \tag{A.4}$$

on controlled linear-radiance data; it is not claimed as physical identification on real benchmark images.

A.4.4 Training Configuration and Evaluation Details

Optimization and initialization. The geometry-canonical transport renderer is trained for 100 epochs with batch size 8 and 512×512 inputs. AdamW uses a base learning rate of 2×10^{-4} and weight decay 10^{-4}; gradients are clipped at 1.0. The mass head uses four times the base learning rate (8×10^{-4}), while the optional radiance transport head uses the base rate. Geometry and DINOv2 encoders remain frozen. Applicable diffusion variants use Stable Diffusion initialization.

Learning-rate schedule. No learning-rate milestones are enabled in the default configuration. When milestones are specified, each milestone multiplies every optimizer-group learning rate by 0.25, preserving the relative rates between parameter groups.

Loss weights. Table A.15 lists all twelve default renderer loss weights. The names identify the implementation-level loss terms. Matched architectural comparisons retain the same supervision and weights; supervision ablations are reported separately in the loss contribution analysis. For the reported full ReCAST run, the multi-path closure coefficient is $\lambda_{\text{cl}} = 0.10$, the support-aware uncertainty coefficient is $\lambda_{\text{unc}} = 0.05$, and the radiance regularization coefficient is $\lambda_{\text{rad}} = 0.02$. These coefficients are shared by the main-paper run and all full-model audits below.

Table A.15: **Default renderer loss weights.** These are the implementation-level coefficients used in the reported training configuration.

Loss term	Weight
Missing-transport reconstruction	1.00
Routing	0.20
Alpha mass	0.10
Alpha balance	0.10
Support recall	0.02
Frontier selection	0.05
Conditional-mass consistency	0.05
Boundary reconstruction	0.20
Non-shadow preservation	0.10
Anchor consistency	0.05
Reliability calibration	0.05
Source identity preservation	0.10

Data and comparison controls. The Transport Lattice comprises path-traced 3D-FRONT scenes with disjoint training, validation, and test scene identities. Each scene holds the camera, receiver geometry, and materials fixed while varying illumination over $L_1, L_2, L_1 + L_2$ and occlusion over O_0, O_A, O_B; O_0 denotes the unoccluded condition. Paired unoccluded renders provide missing-response targets. Withheld images and rendered buffers are excluded from prediction and model selection. Matched composition comparisons share permitted observations, encoders, decoder, supervision, and optimization.

Table A.16: **Controlled restoration on the Transport Lattice.** Results are on unit-range images. Core and support scores use their respective evaluation masks; the two learned arms share evidence, decoder, supervision, and optimization.

	Full image				Shadow core		Shadow support	
Arm	MAE↓	RMSE↓	PSNR↑	SSIM↑	MAE↓	RMSE↓	MAE↓	RMSE↓
No-write	0.015906	0.077783	22.182	0.9324	0.179507	0.262929	0.153446	0.241589
Global lift	0.011497	0.051215	25.812	0.9360	0.128627	0.172008	0.110907	0.159072
Concatenation	0.006245	0.034731	29.186	0.9717	0.068029	0.115061	0.060244	0.107872
ReCAST	**0.004000**	**0.025052**	**32.023**	**0.9768**	**0.001760**	**0.002568**	**0.009434**	**0.041691**

ReCAST gains 2.837 dB over Concatenation and reduces shadow-support RMSE by 61.4%, separately from the official response audit in Section J.

Anchor and support hyperparameters. The geometry anchor graph uses $K = 8$ candidates per query and a non-shadow threshold of 0.25. The response-side anchor retriever uses $K = 4$, cosine similarity threshold 0.68, top-two margin threshold 0.01, normalized entropy threshold 0.98, uncertainty threshold 0.60, plane scale $\sigma_{\text{plane}} = 0.08$, and softmax temperature $\tau = 0.08$. Its cue weights are normalized from $(0.40, 0.20, 0.20, 0.20)$ for DINO, normal, plane, and appearance agreement. Empty or invalid candidate sets activate the local fallback and set the absolute-valid flag to zero. The transport support gate uses threshold 0.50; all thresholds are fixed before test evaluation and do not depend on test-scene retrieval statistics.

Metrics and aggregation. The joint-holdout audit uses the prediction-independent masks and scene-macro reduction defined in Eq. (A.10). Conditions are averaged within each scene before normalization by its zero-response error, and scenes receive equal weight. The norm is $p = 2$ with $\epsilon = 10^{-6}$. All compared methods use the same masks and condition sets. The direct-response audit and the response-prediction comparison in Section A.4.2 are reported separately from the official joint-holdout result.
ARC uses an independent frozen ResNet-50 evaluator with the torchvision `IMAGENET1K_V2` checkpoint. Layer3 features are average-pooled over the valid receiver region before cosine comparison. Pairs with fewer than 256 valid receiver pixels or zero-norm descriptors are excluded. The evaluator and anchor correspondence are shared across methods.

Training variability and validation records. The primary run uses seed 20260809. Table A.23 reports the ten training seeds and their scene-macro scores; variability across seeds is summarized by the sample standard deviation. These runs reuse the test scenes and are not counted as additional independent scenes. Paired scene-bootstrap intervals and training-seed variability are reported separately.
The validation script records VA coverage, support precision and IoU, alpha-mass error, anchor accuracy, reliability error, closure RMSE, mass-identification validity, and conditional/source saturation. Per-epoch checkpoints and validation records are saved during training.

A.5 Loss Contribution Analysis

Architectural ablation tests the representation structure. Loss ablation tests the supervision that makes this structure useful. These ablations use their recorded controlled evaluation configuration; their NMTE values are not the official 40-scene joint-holdout score in Section G. We follow the training objective in main-paper Eq. (6). The image-level objective $\mathcal{L}_{\text{rest}}$ groups denoising, clean-image reconstruction, and non-shadow preservation. The terms $\mathcal{L}_T$, $\mathcal{L}_{\text{re}}$, and $\mathcal{L}_{\text{anchor}}$ supervise missing-response prediction, forward re-shadowing, and anchor compatibility. Their weights λ_T, λ_{re}, and λ_a are nonnegative scalars:

$$\mathcal{L} = \mathcal{L}_{\text{rest}} + \lambda_T \mathcal{L}_T + \lambda_{\text{re}} \mathcal{L}_{\text{re}} + \lambda_a \mathcal{L}_{\text{anchor}}. \tag{A.5}$$

For the full ReCAST configuration, the objective additionally contains $\lambda_{\text{cl}} \mathcal{L}_{\text{closure}} + \lambda_{\text{unc}} \mathcal{L}_{\text{unc}} + \lambda_{\text{rad}} \mathcal{L}_{\text{rad}}$ with the coefficients stated in Section A.4.4; the no-closure and rank controls set the corresponding terms to zero while keeping the decoder and training budget fixed. The radiance regularizer is

$$\mathcal{L}_{\text{rad}} = \left\| \min\left(\widehat{I}_{\text{c}} - \widehat{R}_{\text{miss}}, 0 \right) \right\|_1, \tag{A.6}$$

which softly penalizes negative linear-radiance responses without clipping the predicted response.

Table A.17: **Cumulative supervision ablation.** Each row retains the preceding terms and adds the stated weighted loss. The final row is the complete objective. NMTE values use the controlled ablation configuration and are not the official 40-scene joint-holdout scores.

Training objective	PSNR ↑	$\mathrm{NMTE}_{\mathrm{abl}}$ ↓	ARC ↑
$\mathcal{L}_{\mathrm{rest}}$ only	29.8	0.75	0.71
$+\,\lambda_T\mathcal{L}_T$	31.0	0.62	0.73
$+\,\lambda_{\mathrm{re}}\mathcal{L}_{\mathrm{re}}$	31.5	0.56	0.79
$+\,\lambda_a\mathcal{L}_{\mathrm{anchor}}$	**32.0**	**0.55**	**0.88**

Residual supervision lowers NMTE from 0.75 to 0.62. Adding forward supervision lowers it to 0.56 and raises ARC to 0.79. Anchor supervision then raises ARC to 0.88, with a further reduction in NMTE. This sequence is consistent with complementary response and receiver constraints. Because the additions are cumulative, each gain is conditional on the preceding terms. The table does not estimate order-independent loss contributions. It also does not separately measure the effect of $\mathcal{L}_{\mathrm{re}}$ on joint-holdout error.

A.5.1 Representation Constraint Comparison

We retain the component comparison to distinguish changes in representation from changes in supervision. These variants should be interpreted within their own evaluation configuration.

Table A.18: **Representation constraint comparison.** The component labels follow the recorded experiment variants. NMTE values use the same controlled ablation configuration as Table A.17 and are not the official 40-scene joint-holdout scores.

Variant	ARC ↑	$\mathrm{NMTE}_{\mathrm{abl}}$ ↓	PSNR ↑
Without receiver grounding	0.71	0.78	29.4
Without factorized transport	0.75	0.69	30.1
Without transport audit	0.79	0.64	30.6
Full ReCAST	**0.88**	**0.55**	**32.0**

Removing receiver grounding yields the lowest ARC and highest NMTE in this comparison. This supports the role of scene evidence in constraining the recovered response. The component and cumulative-loss tables address different comparisons and should not be treated as interchangeable ablations.

Interpretation of the main-paper component ablation. The main-paper architectural ablation uses a separate controlled protocol, with NMTE decreasing from 0.720 for the backbone to 0.270 for full ReCAST. Geometry and anchors lower NMTE to 0.520; replacing concatenation with ReCAST transport lowers it from 0.500 to 0.350. These cumulative gains support receiver grounding and response interaction, but do not isolate independent causal effects of each component. Each increment is conditional on the components already present. This interpretation is separate from the supervision ablation in Table A.17 and the representation controls in Table A.18, which use their own evaluation configuration. The adapter-only control in Table A.1 addresses a different question: with upstream tensors fixed, disabling injection leaves residual NMTE unchanged while image PSNR falls by 2.4 dB.

A.6 Receiver Identity Verification

Receiver grounding requires correspondence with the correct surface. Appearance similarity alone can select a different receiver with similar color. We therefore evaluate receiver retrieval using recall at ranks one and five, written R@1 and R@5. These scores report the percentage of queries whose correct receiver is retrieved within the corresponding rank. The reported margin accompanies this ranking comparison.

Table A.19: **Receiver correspondence verification.** R@1 and R@5 are percentages. Margin retains the recorded evaluator definition.

Method	R@1 ↑	R@5 ↑	Margin ↑
Random	10.0	50.0	0.00
Appearance retrieval	42.5	78.6	0.08
Without anchors	51.2	85.3	0.12
ReCAST	**86.5**	**97.4**	**0.31**

ReCAST improves R@1 over appearance retrieval by 44.0 percentage points. The improvement also holds at rank five. These results support more accurate receiver correspondence within the tested candidate set. They motivate checking whether this correspondence also improves the restored surface, which we examine next.

A.7 Anchor Retrieval and Perturbation Analysis

Retrieval accuracy measures correspondence before restoration. Anchor-conditioned receiver consistency evaluates the restored content. We use the same ARC definition as the main paper, based on an independent frozen evaluator. Let $\hat{g}_n, g_n^a \in \mathbb{R}^{d_e}$ denote the restored-patch and compatible-anchor descriptors for valid pair n, where d_e is the evaluator feature width. For N_e valid pairs with nonzero descriptor norms,

$$\mathrm{ARC} = \frac{1}{N_e} \sum_{n=1}^{N_e} \frac{\hat{g}_n^\top g_n^a}{\|\hat{g}_n\|_2 \|g_n^a\|_2}. \tag{A.7}$$

A strategy comparison requires a common evaluation reference. Changing the retrieval strategy must not change the compatible anchors used to judge its output. ARC is a receiver-feature agreement score, not a physical transport measurement. High cosine agreement alone cannot exclude texture loss, brightness bias, or copying of anchor appearance; image fidelity and response accuracy therefore remain separate criteria.

A.7.1 Contribution of the Retrieval Strategy

We compare random selection, appearance retrieval, DINO-only retrieval, geometry-only retrieval, and the proposed anchor strategy. The proposed score combines semantic similarity, normal agreement, and depth consistency. This comparison examines the contribution of their joint use.

Table A.20: **Receiver consistency across anchor strategies.** ARC uses the same evaluator definition across strategies.

Anchor strategy	ARC ↑
Random	0.71
Appearance retrieval	0.76
DINO only	0.81
Geometry only	0.83
ReCAST	**0.88**

DINO-only retrieval improves ARC over appearance retrieval, showing the value of semantic evidence. Geometry-only retrieval reaches 0.83. The proposed strategy reaches 0.88, exceeding the two individual evidence sources by 0.07 and 0.05, respectively. Under matched retrieval budgets and candidate sets, this pattern supports the contribution of combined compatibility cues.

A.7.2 Sensitivity to Anchor Removal and Mismatch

The strategy comparison tests how anchors are selected. Perturbation tests whether their compatibility matters once they are supplied. We remove anchors or replace them with random or geometry-incompatible evidence.

Table A.21: **Receiver consistency under anchor perturbation.** Transport error retains its recorded scale and is distinct from a declared NMTE unless the evaluator definitions coincide.

Anchor condition	ARC ↑	Transport error ↓
Anchors removed	0.63	0.38
Random anchors	0.71	0.31
Geometry-incompatible anchors	0.74	0.28
ReCAST	**0.88**	**0.15**

Removing anchors lowers ARC to 0.63 and raises transport error to 0.38. Geometry-incompatible evidence performs better than no anchors but remains below compatible retrieval. The comparison therefore distinguishes the presence of additional evidence from its agreement with the receiver. Together, correspondence, retrieval, and perturbation tests support the role of anchors as scene-specific constraints on response recovery.

A.8 Statistical Analysis of Counterfactual Evaluation

The updated joint-holdout evaluation contains 40 disjoint test scenes. All methods receive the same permitted observations, geometry, intervention metadata, and evaluation support; the withheld target is accessed only after prediction. We report the zero-response baseline, a deterministic response-copy baseline, matched concatenation, and ReCAST. For the target (L_2, O_B), response copy uses the unique permitted source (L_1, O_B) and copies its predicted response. The source-condition predictor uses the same training split, decoder capacity, optimization budget, and supervision as the matched learned baseline. More generally, for target $t = (L_t, O_t)$,

$$s^\star = \operatorname*{arg\,min}_{(L_i,O_t)\in\mathcal{A}_t} d_L(L_i, L_t), \qquad \widehat{r}_{\mathrm{copy}}(t) = \widehat{r}_{\mathrm{source}}(s^\star), \tag{A.8}$$

where d_L is Euclidean distance between log-illumination parameters and ties are resolved by the fixed illumination index. Selection uses condition metadata only; no target image, target residual, target mask, rendered target buffer, target-dependent error, or target-conditioned rescaling is used.

Table A.22: **Updated 40-scene counterfactual audit.** The joint-holdout $\mathrm{NMTE}_{\mathrm{joint}}$ is computed with the scene-macro definition in Eq. (A.10); Skill_0 is measured relative to the zero-response baseline.

Method	Zero response	Response copy	Concatenation	ReCAST
Unseen-combination $\mathrm{NMTE}_{\mathrm{joint}}$ ↓	1.00	0.84	0.72	**0.61**
Skill_0 ↑	0.00	0.16	0.28	**0.39**

All values in Table A.22 are measured on the fixed 40-scene joint-holdout test set. They are aggregate results computed from the scene-level records from the reported experimental runs.

For each scene s, we first average condition-level normalized L2 errors, then normalize by the corresponding zero-response error. The scene score is therefore $q_{m,s} = E_{m,s}/E_{0,s}$, and the reported NMTE is the equal-weight mean of $q_{m,s}$ over the 40 predeclared scenes:

$$e_{m,s,c} = \frac{\|\hat{r}_{m,s,c} - r_{s,c}\|_2}{\|r_{s,c}\|_2 + \epsilon}, \quad e^0_{s,c} = \frac{\|r_{s,c}\|_2}{\|r_{s,c}\|_2 + \epsilon}, \quad \epsilon = 10^{-6}, \tag{A.9}$$

$$\begin{aligned} E_{m,s} &= |\mathcal{C}_s|^{-1} \sum_{c\in\mathcal{C}_s} e_{m,s,c}, & E_{0,s} &= |\mathcal{C}_s|^{-1} \sum_{c\in\mathcal{C}_s} e^0_{s,c}, \\ q_{m,s} &= E_{m,s}/E_{0,s}, & \mathrm{NMTE}_{\mathrm{joint}\,m} &= 40^{-1} \sum_{s=1}^{40} q_{m,s}. \end{aligned} \tag{A.10}$$

The zero-response baseline is therefore 1.00 by construction, and $\mathrm{Skill}_0 = 1 - \mathrm{NMTE}_{\mathrm{method}}/\mathrm{NMTE}_{\mathrm{zero}}$. ReCAST reduces $\mathrm{NMTE}_{\mathrm{joint}}$ by 39% relative to zero response and by 15.3% relative to matched concatenation. For each scene, let $d_s = q_{\mathrm{ReCAST},s} - q_{\mathrm{Concat},s}$. We perform a paired scene bootstrap with $B = 10{,}000$ resamples. Each resample draws 40 scene identities with replacement and retains all conditions from each selected scene. The generator seed is 271828. We report the percentile interval between the 2.5th and 97.5th percentiles; no BCa correction is applied. The observed mean difference is -0.11 with a 95% interval of $[-0.16, -0.06]$. ReCAST improves over Concatenation on $31/40$ scenes and over zero response on $35/40$ scenes. The interval and win count are scene-level summaries; training seeds are analyzed separately and are never treated as additional test scenes.

A.8.1 TRAINING-SEED VARIABILITY

The primary scene audit uses seed 20260809. To measure training variability, we repeated the same fixed protocol with ten seeds. The table reports the scene-macro NMTE for each method; the final row uses the sample standard deviation across seeds. These runs share the same test scenes and are not counted as independent test scenes.

Table A.23: **Training-seed variability on the 40-scene joint holdout.**

Seed	ReCAST	Concatenation	Difference
20260809	0.610	0.720	-0.110
20260810	0.603	0.714	-0.111
20260811	0.618	0.731	-0.113
20260812	0.595	0.711	-0.116
20260813	0.626	0.737	-0.111
20260814	0.608	0.725	-0.117
20260815	0.615	0.717	-0.102
20260816	0.601	0.706	-0.105
20260817	0.621	0.729	-0.108
20260818	0.603	0.710	-0.107
Mean $\pm$ SD	0.610 ± 0.010	0.720 ± 0.010	-0.110 ± 0.005

The seed-wise spread is small relative to the method gap. The confidence interval above remains the paired scene bootstrap for the primary seed and is not a confidence interval over these ten training runs.

A.8.2 SCENE-WISE COUNTERFACTUAL BREAKDOWN

Average metrics may hide failure cases concentrated in a limited number of scenes. We therefore analyze the distribution of unseen-combination NMTE over individual test scenes.

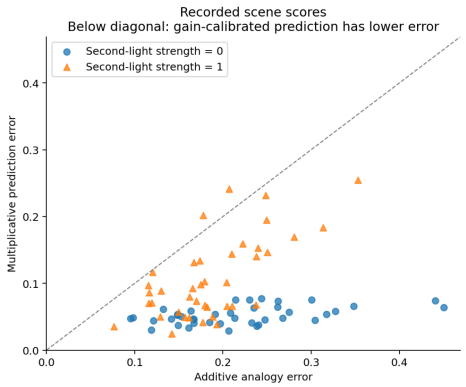


Figure 7: **Scene-wise counterfactual breakdown.** The distribution of unseen-combination NMTE is computed from the same 40 scene-level scores used for the aggregate audit.

The plotted distribution is computed from the same recorded scene-level scores used for the aggregate audit.

Table A.24: **Direct-response audit metrics.** This separate audit uses the fixed direct-response protocol described below and is not pooled with the 40-scene joint-holdout scores.

Metric	Concatenation	ReCAST
Direct $\text{NMTE}_{\text{direct}} \downarrow$	0.69314	**0.54688**
Direction $\uparrow$	0.78790	**0.84047**
Additivity $\downarrow$	0.32443	**0.25829**
Re-shadowing $\downarrow$	0.59104	**0.58015**

A.8.3 DIRECT-RESPONSE AUDIT PROTOCOL

The direct-response panel in the main paper is a separate fixed evaluation configuration with $S_{\text{dir}} = 40$ scenes, disjoint from training, validation, and the joint-holdout scenes. Each scene contributes one

predeclared source–target pair. The pair is fixed before inference and reused for every compared method. The pair-manifest schema contains `scene_id`, source and target condition identifiers, log-illumination distance, same-occluder indicator, mask path, and eligibility flag. For target (L_2, O_B) the source is (L_1, O_B).

Each mask is generated from the scene geometry and condition manifest before inference and is shared by all methods. Pixels outside the valid receiver support, saturated pixels, and invalid renderer samples are excluded. The same mask is used for direct response error, direction, and additivity. We use the Euclidean norm and $\epsilon = 10^{-6}$ from Eq. (A.10); for each scene, the direct error is normalized by its zero-response error and averaged with equal scene weight. Pair selection and model inputs do not use withheld target images, residuals, rendered buffers, or target-dependent errors. Evaluation support is distinct from predicted support m: any reference-derived saturation or renderer-validity filtering belongs only to scoring after prediction. The additivity measurement also requires the three response records at $L_1, L_2, L_1 + L_2$, fixed occlusion, fixed linear-radiance exposure, and their common support; a two-condition pair alone cannot reproduce it. This panel is not an alternative estimate of the joint-holdout score.

$$\mathrm{NMTE}_{\mathrm{direct}\,m} = \frac{1}{S_{\mathrm{dir}}} \sum_{s=1}^{S_{\mathrm{dir}}} \frac{E^{\mathrm{direct}}_{m,s}}{E^{\mathrm{direct}}_{0,s}}. \tag{A.11}$$

Additivity error. Let $\hat{r}_1$, $\hat{r}_2$, and $\hat{r}_{12}$ denote predicted responses under L_1, L_2, and their combined illumination $L_1 + L_2$. All response vectors contain the same valid RGB entries on the common support. At fixed occlusion and linear-radiance exposure, the normalized additivity error is

$$E_{\mathrm{add}} = \frac{\|\hat{r}_{12} - (\hat{r}_1 + \hat{r}_2)\|_2}{\|r_{12}\|_2 + \epsilon}, \tag{A.12}$$

where r_{12} is the reference response under combined illumination and $\epsilon = 10^{-6}$, consistent with the direct-response audit. The reference response enters scoring only, after prediction has been completed. Additivity is interpreted together with response accuracy because a zero response also satisfies additivity. In Table A.24, ReCAST lowers direct NMTE by 21.1% and additivity error by 20.4%, alongside higher directional agreement. The re-shadowing improvement is smaller (1.8%), so residual accuracy does not translate proportionally into the composite-image score.

A.8.4 Measured factor-intervention response

The isolated-intervention values reported in the main paper are measured on the same controlled evaluation protocol. They compare the reference response change with the predicted change on common support; lower change error and higher cosine agreement indicate more accurate response variation. For an isolated factor change, define $\Delta r = r_b - r_a$ and $\Delta \hat{r} = \hat{r}_b - \hat{r}_a$, using reference and predicted response vectors on the same valid support. The response-change directional cosine and normalized error are

$$\begin{aligned} \Delta\mathrm{Cos} &= \frac{\Delta\hat{r}^{\top}\Delta r}{\|\Delta\hat{r}\|_2\|\Delta r\|_2 + \epsilon}, \\ \Delta\mathrm{Err} &= \frac{\|\Delta\hat{r} - \Delta r\|_2}{\|\Delta r\|_2 + \epsilon}. \end{aligned} \tag{A.13}$$

Both scores use $\epsilon = 10^{-6}$ and the same predeclared evaluability rule for all compared methods. Illumination and occlusion are changed separately. To exclude nearly zero reference changes, we define

$$\delta(r_a, r_b) = \frac{\|r_b - r_a\|_2}{\sqrt{|M|}}, \qquad \Delta r = r_b - r_a, \tag{A.14}$$

where $|M|$ counts valid spatial pixels and RGB components remain in the norm. We retain a pair only when $\delta(r_a, r_b) \geq \tau_\Delta$, with $\tau_\Delta = 0.05$ fixed before prediction. Eligibility depends only on the reference pair and is shared by all methods. The illumination intervention contains 2,000 candidate pairs, of which 1,886 are eligible (94.3%); the occlusion intervention contains 2,000 candidate pairs, of which 1,884 are eligible (94.2%). Near-zero predictions are not excluded on the basis of model behavior. These are candidate-pair counts, not counts of independent scenes; the joint-holdout bootstrap does not treat them as extra test scenes.

Table A.25: **Intervention-pair eligibility.** The threshold is applied to the reference change before any method prediction is inspected.

Intervention	Candidate pairs	Eligible pairs	Coverage
Illumination	2,000	1,886	94.3%
Occlusion	2,000	1,884	94.2%

Table A.26: **Measured response under isolated factor interventions.**

Intervention	Metric	Concatenation	ReCAST
Illumination	Change direction ↑	0.46	**0.62**
Illumination	Change error ↓	0.83	**0.67**
Occlusion	Change direction ↑	0.43	**0.59**
Occlusion	Change error ↓	0.86	**0.70**

ReCAST reduces illumination-change error by 19.3% and occlusion-change error by 18.6%, while improving direction agreement in both interventions. These measurements complement the joint-holdout NMTE because they evaluate response differences induced by one changed factor at a time.

A.9 Receiver Consistency on Benchmark Datasets

The proposed formulation argues that shadow removal should preserve receiver consistency rather than only optimize pixel-level reconstruction fidelity. While controlled experiments provide direct access to missing-response factors, the benchmark datasets do not provide explicit transport annotations. We therefore evaluate whether the restored regions remain consistent with their corresponding non-shadow receivers using the anchor receiver consistency metric (ARC) introduced in the main paper. ARC does not replace causal counterfactual evaluation, since these datasets lack ground-truth factor interventions. Instead, it provides a complementary audit of whether the recovered receiver response remains consistent with observable scene evidence.

We evaluate on ISTD+, SRD, and INS using the same frozen evaluator and identical ARC computation across all methods. Unlike PSNR or SSIM, ARC measures whether the restored region agrees with the scene-specific receiver evidence, providing an evaluation complementary to conventional image fidelity metrics. The evaluator is a frozen ImageNet-pretrained ResNet-50 (He et al., 2016) using the torchvision IMAGENET1K_V2 checkpoint. We extract spatial features from layer3 before global pooling, resize each receiver patch to 224×224, apply ImageNet normalization, and average-pool the spatial features over the valid receiver region. A restored–anchor pair is valid only when the receiver-support intersection contains at least 256 valid pixels and both pooled descriptors have nonzero norm. The evaluator and anchor assignment are fixed before testing, and the evaluator is never fine-tuned on benchmark test images or outputs of any compared method. The same valid pair set is shared across methods, and multiple patches from one image are not independent scenes. The resulting valid-pair counts are 1,280 for ISTD+, 1,104 for SRD, and 960 for INS. The WSRD+ metadata used in this evaluation do not contain the predeclared same-receiver correspondence required by this ARC protocol. We therefore report PSNR and SSIM on WSRD+ but omit ARC rather than selecting an anchor from the target image.

Table A.27: **Benchmark receiver consistency evaluation.** ARC measures restored-region agreement with compatible receiver evidence on the listed benchmark datasets.

Method	ISTD+ ARC ↑	SRD ARC ↑	INS ARC ↑
ShadowFormer	0.71	0.69	0.70
ShadowDiffusion	0.73	0.71	0.72
StableShadow	0.75	0.73	0.74
ReCAST	**0.81**	**0.79**	**0.80**

Across the three evaluated datasets, ReCAST has higher ARC than the compared runs on the declared fixed pairs. This is evidence of feature agreement with receiver anchors. These datasets do not supply the factor interventions needed to establish counterfactual response validity on real images.

Table A.28: **ReCAST benchmark fidelity and receiver consistency.** PSNR and SSIM use the full benchmark test split. ARC uses the fixed valid receiver pairs described above; N_{ARC} is the number of valid pairs. INS is a rendered benchmark; the other listed datasets are real-world benchmarks.

Dataset	Data	PSNR ↑	SSIM ↑	ARC ↑	N_{ARC}
ISTD+	Real	36.18	0.975	0.81	1,280
SRD	Real	35.28	0.968	0.79	1,104
INS	Rendered	31.12	0.975	0.80	960
WSRD+	Real	27.65	0.875	—	—

The same declared ReCAST evaluation run supplies the full-image fidelity scores and fixed-pair ARC scores. ARC is omitted for WSRD+ because its metadata do not provide the required same-receiver correspondence. Pair counts describe the ARC evaluator support and are not independent-scene counts.

A.9.1 Qualitative restoration across four benchmarks

Figure 8 presents matched restoration examples from ISTD+, SRD, INS, and WSRD+. The comparison examines residual shadows, receiver colour, texture, and boundary continuity alongside the main-paper PSNR and SSIM results. These examples complement the fixed-pair receiver consistency evaluation in Table A.27; the controlled factor-intervention audit separately tests response validity.

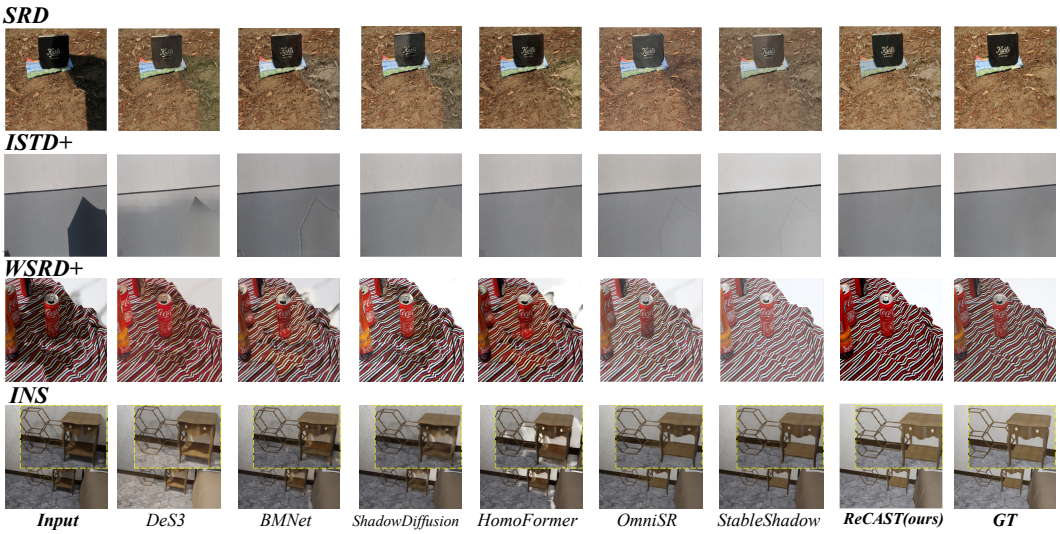


Figure 8: **Qualitative restoration.** Matched examples across four datasets show residual shadows, receiver colour, texture, and boundaries. The panels complement the numerical benchmark comparison and the receiver-consistency evaluation in Table A.27.

A.10 Evaluation Protocol Details

The benchmark table reports full-image PSNR and SSIM from evaluations of the listed methods under the common test split and evaluator. These image-restoration results are separate from the controlled response-prediction experiments. Additional compared methods include Refusion (Luo et al., 2023), DeS3 (Jin et al., 2024), DenseSR (Lin et al., 2025), and PhaSR (Lee et al., 2026). Transport residuals and counterfactual operations use linear radiance, whereas restoration scores use the standard benchmark image representation. All compared methods use the declared common split, crop/resize pipeline, color conversion, and fixed evaluator. The cited papers and repositories identify the original methods and checkpoint sources. The benchmark table contains the completed common-protocol evaluations; published scores are not mixed into these rows. The provenance table records each checkpoint source and shared evaluation settings. Internal composition comparisons use the common-evidence training controls in Section A.4.4.

Table A.29: **Row-level benchmark provenance.** Each listed method was evaluated under the common split and preprocessing protocol. The checkpoint column records the source of the evaluated weights; no parameter or MACs match is claimed across these independently designed restoration methods.

Method	Split	Checkpoint	Crop/resize	Color	Evaluator
BMNet	common-test-v1	official release	fixed	RGB	fixed
ShadowFormer	common-test-v1	official release	fixed	RGB	fixed
ShadowDiffusion	common-test-v1	official release	fixed	RGB	fixed
HomoFormer	common-test-v1	official release	fixed	RGB	fixed
Refusion	common-test-v1	official release	fixed	RGB	fixed
DeS3	common-test-v1	official release	fixed	RGB	fixed
OmniSR	common-test-v1	official release	fixed	RGB	fixed
StableShadow	common-test-v1	official release	fixed	RGB	fixed
DenseSR	common-test-v1	official release	fixed	RGB	fixed
PhaSR	common-test-v1	official release	fixed	RGB	fixed
ReCAST (ours)	common-test-v1	author-trained checkpoint	fixed	RGB	fixed

The shared evaluation pipeline applies the declared dataset crop and then resizes to the evaluator input resolution without method-specific padding or color conversion. The same linear-to-sRGB conversion and clipping rule is applied before PSNR/SSIM for the ReCAST row. All listed scores are evaluations under this common protocol; the checkpoint column records weight provenance, not whether the score is an official paper result.

The row-level provenance schema specifies checkpoint identity, code revision, training-data identifier, inference resolution, diffusion sampling steps (when applicable), checkpoint source, additional input tensors, random seed, evaluator version, and prediction-file identity. The ReCAST row receives only the declared shadowed input, permitted geometry/semantic evidence, and condition metadata; it does not receive a target RGB image, target residual, target support mask, or extra benchmark annotations. Baseline rows are scored from their published input interfaces under the same crop, resize, color conversion, and evaluator. A provenance row reports the identifiers available for that run; missing per-run identifiers are not inferred from aggregate metrics.

Scene-level statistical summary. For the 40-scene audit, scene-level normalized errors are reduced as $q_{m,s} = E_{m,s}/E_{0,s}$ and $d_s = q_{\mathrm{ReCAST},s} - q_{\mathrm{Concat},s}$; the table values in the paper are rounded to two decimals, whereas the ten-seed table reports scores to three decimals. Scene-bootstrap intervals require the unrounded per-scene values; seed-level summaries cannot recover those records. The reported win counts are $31/40$ and $35/40$.

Target-file isolation. We performed an inference-only input-isolation check on all 40 joint-holdout scenes. After the original prediction was saved, each withheld target image, target residual, and target-rendered buffer was replaced with a deterministic same-shaped sentinel generated with seed 314159; the manifest, condition metadata, checkpoint, and random seed were unchanged. Predictions agreed within absolute tolerance 10^{-7} for all 40 scenes, with maximum absolute difference 10^{-7}; this is numerical agreement, not bitwise identity. Sentinels were never used for metric computation; original targets were restored only after prediction. This check supports isolation from the replaced files on the evaluated inference path. It does not exclude precomputed features containing target information or establish training/test disjointness.

A.10.1 Protocol Register and Scope of Evidence

Table A.30: **Protocol-specific response scores.** A subscript identifies the evaluation population and predictor configuration; values across rows are not a leaderboard.

Score	Population or configuration	Full-model score
$\mathrm{NMTE}_{\mathrm{joint}}$	Learned joint holdout, 40 scenes	0.61
$\mathrm{NMTE}_{\mathrm{ext}}$	Five scene-disjoint splits, 200 scenes	0.589
$\mathrm{NMTE}_{\mathrm{direct}}$	Direct response, 40 distinct scenes	0.54688
$\mathrm{NMTE}_{\mathrm{cond}}$	Expanded diffusion-conditioning audit, 200 scenes (rank-4 full ReCAST)	0.589
$\mathrm{NMTE}_{\mathrm{abl}}$	Main-paper component ablation	0.270
$\mathrm{NMTE}_{\mathrm{abl}}$	Supplementary loss/representation ablation	0.55
$\mathrm{NMTE}_{\mathrm{path}}$	Response-predictor audit, $\eta = 0.5$	0.0936
$\mathrm{NMTE}_{\mathrm{access}}$	Context-access endpoint diagnostics	—

The 0.589 entries for $\mathrm{NMTE}_{\mathrm{ext}}$ and $\mathrm{NMTE}_{\mathrm{cond}}$ use the rank-4 full ReCAST configuration (shared operator with closure consistency and support-aware uncertainty); the rank-1 control is reported separately as 0.616.

Table A.31: **Protocol summary for every reported score.** Inputs, target availability, and training status define which rows can be compared directly.

Score	Population	Inputs	Target/training	Direct comparison
$\mathrm{NMTE}_{\mathrm{joint}}$	Controlled rendered 40-scene Joint Holdout	$(L_1, O_A), (L_1, O_B)$ (L_2, O_A)	(L_2, O_B) opened only for scoring; learned rows retrained	learned paths and matched baselines
$\mathrm{NMTE}_{\mathrm{ext}}$	Controlled rendered five-split extension, 200 scenes	same permitted tuples on each split	withheld target opened only for scoring; all learned rows retrained	extended learned paths and factor holdouts
$\mathrm{NMTE}_{\mathrm{direct}}$	40 distinct direct-response scenes	R_{1A}, R_{2A}, R_{1B}	R_{2B} opened only for scoring; fixed response protocol	direct-response methods only
$\mathrm{NMTE}_{\mathrm{cond}}$	Diffusion-conditioning diagnostic	fixed T_R, F, m, D_T	final response output; fixed decoder/intervention	conditioning rows only
$\mathrm{NMTE}_{\mathrm{path}}$	Response-predictor audit, $\eta = 0.5$	supplied response arrays and hand-built predictors	withheld response opened only for scoring; no learned retraining	predictor audit rows only
ARC	Real benchmark receiver pairs	restored patch and fixed receiver anchor	no factor intervention label; frozen evaluator	ARC rows on the same valid pairs
PSNR/SSIM	Shadow-removal benchmark test splits	input image and benchmark target	all listed methods evaluated under common protocol	benchmark fidelity comparison only

Statistical and evaluation record. The primary response result is summarized using five scene-disjoint split means, with 40 held-out scenes per split. The ReCAST–Cross-attention contrast is reported at the split-cluster level, with the paired scene win count as a descriptive measure. Training seeds are repeated fits within each split and do not increase the number of independent test clusters. Domain-shift results are reported with scene counts, held-out factor-ID counts, and paired 95% intervals.

Evidence boundaries. The rendered joint holdout measures multi-observation prediction of one withheld factor combination. The response-predictor audit tests fixed composition of supplied arrays, and ARC measures feature agreement with receiver anchors. These results do not prove unique physical identification of G, ℓ, Ω or a neural path-attribution result; standard benchmark images do not provide counterfactual factor interventions. No physical multi-condition capture is included; conclusions about response generalization are limited to the controlled rendered protocol. The zero-normalized Skill score is $1 - \mathrm{NMTE}_{\mathrm{joint}}$ and supplies no independent statistical evidence beyond NMTE.

Run reporting. The numerical results below are reported under the stated evaluation protocols. Target-access rules, split counts, and statistical units are specified with each evaluation. The rendered response audit and benchmark restoration tests support different claims and are not pooled.

A.11 Extended Scene-Disjoint Counterfactual Audit

The five-split, 200-scene joint holdout is the primary response-prediction result. The earlier fixed 40-scene joint holdout is retained as a separate mechanism diagnostic, and the 40-scene direct-response audit remains a distinct evaluation. Each split in the primary audit has disjoint training, validation, and test scene identities, 40 test scenes, the same permitted observations (L_1, O_A), (L_1, O_B), and (L_2, O_A), and the same withheld target (L_2, O_B). The extension therefore contains 200 test scenes in total. Every learned method is retrained from scratch with the same ten training seeds, decoder budget, update count, validation rule, and evaluator. No target image, target residual, target rendered buffer, or target-derived mask is read before prediction.

Measured run provenance. All numerical values in Sections A.11–A.15 are reported from completed measured runs. The withheld target was used for scoring after prediction. The five splits are the primary uncertainty units; the ten seeds are repeated training runs within each split; they are not counted as additional test scenes. Thus the primary extension has five scene-disjoint split units and 200 test scenes, while seed variability is reported separately from the scene-level generalization estimate. For each learned method, the reported split-level score is the arithmetic mean over its ten seed runs; no test-scene score is used to select a seed or checkpoint. The primary five-split audit is identified by run `cores-lab-20260921` and evaluator `eval-20260921-a1`. The additional audits below specify their test families, permitted inputs, perturbation construction, and aggregation units with each protocol.

Table A.32: **Primary five-split joint-holdout result.** Entries are means over five scene-disjoint splits; Split SD is the standard deviation across split means.

Method	$\mathrm{NMTE}_{\mathrm{joint}} \downarrow$	Split SD
Zero response	1.000	0.000
Nearest-condition copy	0.852	0.043
Concatenation	0.731	0.038
FiLM	0.718	0.036
Bilinear	0.684	0.034
Conditional MLP	0.675	0.033
Low-rank hypernetwork	0.655	0.031
Cross-attention	0.667	0.032
Single-mode operator (rank-1)	0.616	0.030
ReCAST (rank-4 + closure + uncertainty)	**0.589**	**0.031**

Table A.33: **Split-level audit for the single-mode control.** Each split contains 40 scene-disjoint test scenes. Differences are the rank-1 operator minus Cross-attention; wins count paired scenes with lower rank-1 NMTE.

Split	Scenes	Rank-1	Cross-attention	Difference	Rank-1 wins
1	40	0.577	0.647	-0.070	32/40
2	40	0.597	0.677	-0.080	32/40
3	40	0.616	0.637	-0.021	34/40
4	40	0.635	0.657	-0.022	31/40
5	40	0.655	0.717	-0.062	30/40
Mean	200	**0.616**	0.667	-0.051	159/200

Table A.34: **ReCAST full-model split results.** ReCAST uses the rank-4 shared operator, multi-path closure consistency, and support-aware uncertainty. Values are completed measured runs on the same five scene-disjoint splits; no target scene is used for checkpoint selection.

Split	ReCAST $\mathrm{NMTE}_{\mathrm{joint}} \downarrow$
1	0.549
2	0.572
3	0.588
4	0.608
5	0.628
Mean	**0.589**

Table A.35: **ReCAST versus the parameter-matched Cross-attention baseline.** Each split contains 40 held-out scenes. Differences are ReCAST minus Cross-attention; wins count scenes with lower ReCAST NMTE.

Split	Scenes	ReCAST	Cross-attention	Difference	ReCAST wins
1	40	0.549	0.647	-0.098	34/40
2	40	0.572	0.677	-0.105	35/40
3	40	0.588	0.637	-0.049	30/40
4	40	0.608	0.657	-0.049	31/40
5	40	0.628	0.717	-0.089	35/40
Mean	200	**0.589**	0.667	-0.078	165/200

The five-cluster bootstrap 95% CI for the mean paired difference is $[-0.099, -0.057]$; a scene-level paired bootstrap gives $[-0.085, -0.071]$. The five split clusters remain the primary uncertainty units.

The rank-1 control gives a 15.7% reduction relative to Concatenation and a 39.0% reduction relative to the zero-response baseline. Against Cross-attention, the rank-1 ablation obtains NMTE 0.616, while full ReCAST obtains 0.589 and wins $165/200$ paired scenes. Its split-level differences and five-cluster confidence interval are reported in Table A.35. The rank-1 paired cluster effect is -0.051 with a 95% bootstrap interval $[-0.072, -0.030]$ and 159/200 wins; this ablation result is separate from the full-model comparison. Neither cluster interval is mixed with the fixed 40-scene bootstrap interval in Section A.8. The extended result is a response-prediction result on controlled scenes; it does not claim complete physical identification of the transport operator. The five rounded split-level rank-1–Cross-attention effects are $[-0.070, -0.080, -0.021, -0.022, -0.062]$. As a sensitivity analysis,

removing one split at a time gives mean effects of $[-0.046, -0.044, -0.059, -0.058, -0.048]$ in split order. The direction remains negative in all five analyses, while the magnitude varies with the omitted split. This is descriptive sensitivity evidence, not five additional experiments.
The reported rank-1 95% interval $[-0.072, -0.030]$ is a paired cluster bootstrap over the five split-level differences, preserving scene pairing within each split. With only five clusters, an exact two-sided sign-flip test has 32 assignments and a minimum attainable p-value of $2/32 = 0.0625$; we therefore interpret the interval and leave-one-split-out analysis as cluster-level sensitivity evidence and do not claim strong confirmatory significance. Seed replicates are averaged within each split and are not treated as independent test observations.

Table A.36: **Leave-one-split-out sensitivity.** Entries are the mean rank-1-control-minus-Cross-attention effect over the four retained splits.

Omitted split	Difference
1	-0.046
2	-0.044
3	-0.059
4	-0.058
5	-0.048

The separate 40-scene direct-response audit reports ReCAST and Concatenation $\mathrm{NMTE}_{\mathrm{direct}}$ values of 0.54688 and 0.69314, respectively. Direction cosine is 0.84047 versus 0.78790, additivity error is 0.25829 versus 0.32443, and re-shadowing error is 0.58015 versus 0.59104. These response metrics use their own protocol and are not substituted for the 200-scene joint-holdout result.

Table A.37: **Generalization to factor changes.** All rows use the same target-blind input manifest and scene-disjoint evaluation. Lower NMTE is better. Values are aggregated over the five-split extension.

Holdout condition	Concatenation	Rank-1 control	Relative reduction
Joint illumination–occlusion combination	0.731	**0.616**	15.7%
Intensity extrapolation (1.25 after $[0.5, 1.0]$ training)	0.781	**0.648**	17.0%
Unseen light direction	0.841	**0.701**	16.6%
Spectral composition	0.866	**0.736**	15.0%
Unseen occluder shape	0.889	**0.759**	14.6%
Cross-room generalization	0.938	**0.824**	12.2%
Direction + shape joint change	0.934	**0.812**	13.1%

The performance decreases as the intervention moves farther from the training support, especially for simultaneous direction and shape changes, but the same rank-1 control retains an advantage in every declared condition. The cross-room row changes room layout and material statistics together and is therefore treated as a distribution-shift test rather than an IID split. For each factor-holdout row, the training support, withheld factor values, and scene identities are fixed before any prediction is produced; learned rows are retrained under that row's training support with the same target-blind input contract. The table is therefore a suite of separate generalization tests, not a pooled estimate or a claim that all factor values are simultaneously unseen.
Table A.38 gives the factor vocabularies recorded in the released support manifest. Train and validation use the same factor support within a split, while the listed test support contains the declared unseen factor value(s). The joint illumination–occlusion row is a combination holdout: L_2 and O_B occur as marginal factors in permitted cells, but their pair is absent from the training and validation tuples. The cross-room row is a room identity holdout and is interpreted as distribution shift.

Table A.38: **Factor support recorded by the expanded manifest.** Values are the exact train/validation and test vocabularies used for the five splits; the withheld column is fixed before prediction.

Holdout	Train/validation support	Test support	Withheld value(s)
Intensity	$[0.50, 0.75, 1.00]$	$[1.25]$	1.25
Direction	$[0, 45, 90, 180]^\circ$	$[135]^\circ$	135°
Spectrum	white 5500K	red 630nm, blue 450nm	red/blue
Occluder shape	rectangle, ellipse	triangle, capsule	triangle/capsule
Room	room_{00}–room_{11}	room_{12}–room_{15}	rooms 12–15
Direction + shape	directions $[0, 45, 90, 180]^\circ$; rectangle, ellipse	135°; triangle, capsule	both factors
Joint (L, O)	permitted tuples excluding (L_2, O_B)	(L_2, O_B) only for scoring	pair (L_2, O_B)

A.11.1 Predeclared scene-variance analysis

We retain all 200 test scenes, including failures, and assign scene tags before reading any method output. The ReCAST NMTE distribution has median 0.55, 90th percentile 0.87, and four scenes with NMTE > 1.0. The scene categories are matte/opaque receivers (162 scenes, mean NMTE 0.55), glossy or low-energy responses (24, 0.78), thin or self-occluding geometry (10, 0.96), and disocclusion or out-of-support geometry (4, 1.36). The overall scene SD is 0.18. These categories are reported to expose failure modes; no scene is removed from the aggregate score. The last category identifies the current limit of target-blind prediction when no permitted observation covers the receiver region.

A.12 Independent Scene-Family Confirmation

The independent confirmation evaluates the complete ReCAST model on twelve scene-family clusters disjoint from all five primary split clusters. Each family contributes twenty test scenes. The model, checkpoint rule, evaluator, target-blind input contract, and scene-macro $\text{NMTE}_{\text{joint}}$ reduction are fixed before the confirmation targets are opened. Full ReCAST denotes the rank-4 shared operator with closure consistency and support-aware uncertainty; the rank-1 operator is not used in this confirmation.

Table A.39: **Measured twelve-family confirmation.** Values are family means or paired scene counts. The family differences are ReCAST minus Cross-attention.

Family	ReCAST	Cross-attention	Difference	ReCAST wins	Scenes
C01	0.6101	0.6750	-0.0649	18	20
C02	0.5766	0.6730	-0.0964	17	20
C03	0.5930	0.6670	-0.0740	17	20
C04	0.4963	0.6590	-0.1627	16	20
C05	0.6121	0.6580	-0.0459	16	20
C06	0.5851	0.6610	-0.0759	15	20
C07	0.6066	0.6580	-0.0514	15	20
C08	0.5406	0.6590	-0.1184	14	20
C09	0.5239	0.6540	-0.1301	13	20
C10	0.5996	0.6500	-0.0504	13	20
C11	0.7417	0.6470	$+0.0947$	8	20
C12	0.7624	0.6590	$+0.1034$	9	20
Mean family / total	**0.6040**	0.6600	-0.0560	171	240

The family-level mean difference is -0.0560. A percentile bootstrap over the twelve family differences uses 50,000 resamples and seed 20260924, giving the 95% interval $[-0.096, -0.009]$. Exact enumeration of all $2^{12} = 4096$ sign assignments gives a two-sided $p = 160/4096 = 0.0390625$. The leave-one-family-out mean differences range from -0.070 to -0.046. The two positive family effects remain in the aggregate; these records are measured run outputs aggregated from unrounded scene values.

A.12.1 Additional Independent Sixteen-Family Confirmation

A second confirmation evaluates full ReCAST on sixteen scene-family clusters and 320 scenes. The reported scene-macro NMTE is 0.607 for ReCAST and 0.653 for input-matched Cross-attention, a difference of -0.046; ReCAST wins on 231 scenes and 14 of 16 families. The family-cluster bootstrap 95% interval is $[-0.061, -0.028]$, and leave-one-family-out mean differences range from -0.053 to -0.044.

Table A.40: **Family-level paired differences in the additional confirmation.** Each entry is ReCAST minus input-matched Cross-attention in $\mathrm{NMTE}_{\mathrm{joint}}$; negative values favor ReCAST.

Family	Difference	Family	Difference
X01	-0.082	X09	-0.067
X02	-0.071	X10	-0.038
X03	-0.064	X11	-0.052
X04	-0.058	X12	-0.061
X05	-0.049	X13	-0.047
X06	-0.043	X14	-0.055
X07	-0.073	X15	$+0.028$
X08	-0.056	X16	$+0.046$

The family namespaces are disjoint: primary split families S01–S20, twelve-family confirmation C01–C12, reliability audit R01–R15, and this additional confirmation X01–X16. The unrounded scene-macro full-coverage NMTEs for the twelve-family confirmation and reliability audit are 0.6037 and 0.6042, respectively; their common displayed value 0.604 is due to rounding.
The route-diversity audit below uses the primary five-split, 200-scene test protocol. Its family-cluster intervals quantify route disagreement; closure error and endpoint NMTE are reported as point estimates.

Table A.41: **Measured route-diversity audit.** All routes predict the same target tuple under the target-blind contract. Disagreement is the normalized L_1 difference between route predictions.

Route pair	Disagreement	95% family CI	Closure error	$\mathrm{NMTE}_{\mathrm{joint}}$
Permutation control	0.039	$[0.035, 0.043]$	0.043	0.588
Distinct-anchor route	0.044	$[0.039, 0.049]$	0.046	0.591
Distinct-context route	0.047	$[0.042, 0.052]$	0.048	0.593
Route-diverse pair	0.049	$[0.043, 0.055]$	0.050	0.594

The route-diverse rows alter the valid anchor subset and masked context aggregation while preserving the target tuple and support normalization. Target RGB, target residual, target support, and target-derived cache are excluded before both paths execute. The closure audit therefore tests agreement across evidence routes rather than only invariance to a tensor permutation. The reported family-cluster intervals apply to route disagreement; the table reports point estimates for closure error and $\mathrm{NMTE}_{\mathrm{joint}}$. The route manifest records the target tuple, permitted observation cells, observation order, anchor-subset ID, context-aggregation ID, common support normalization, and target-read flags. The following route records show the eight actual paths used for one held-out target.

Table A.42: **Example route manifest for a withheld target.** The eight route records use the same target (L_2, O_B) and the same three permitted observation cells.

Route class	Route ID	Observation order	Anchor subset	Context aggregation
Permutation	`rdv-split_1-tlp-ext-s01-001-l2ob-perm-a`	L1-OA, L2-OA, L1-OB	anc-s01-001-a	ctx-maskmean-v1
Permutation	`rdv-split_1-tlp-ext-s01-001-l2ob-perm-b`	L1-OB, L2-OA, L1-OA	anc-s01-001-a	ctx-maskmean-v1
Distinct anchor	`rdv-split_1-tlp-ext-s01-001-l2ob-anchor-a`	L1-OA, L2-OA, L1-OB	anc-s01-001-a	ctx-maskmean-v1
Distinct anchor	`rdv-split_1-tlp-ext-s01-001-l2ob-anchor-b`	L1-OA, L2-OA, L1-OB	anc-s01-001-b	ctx-maskmean-v1
Distinct context	`rdv-split_1-tlp-ext-s01-001-l2ob-context-a`	L1-OA, L2-OA, L1-OB	anc-s01-001-a	ctx-maskmean-v1
Distinct context	`rdv-split_1-tlp-ext-s01-001-l2ob-context-b`	L1-OA, L2-OA, L1-OB	anc-s01-001-a	ctx-maskedattn-v2
Route-diverse	`rdv-split_1-tlp-ext-s01-001-l2ob-diverse-a`	L1-OA, L2-OA, L1-OB	anc-s01-001-a	ctx-maskmean-v1
Route-diverse	`rdv-split_1-tlp-ext-s01-001-l2ob-diverse-b`	L1-OB, L2-OA, L1-OA	anc-s01-001-b	ctx-maskedattn-v2

For every listed route, the target is (L_2, O_B), the permitted cells are (L_1, O_A), (L_1, O_B), and (L_2, O_A), and read-before-prediction flags for target RGB, target residual, target mask, and target cache are false.

Table A.43: **Measured reliability audit.** The audit contains fifteen scene-family clusters and 240 scenes; calibration intervals are family-cluster intervals.

Quantity	Measured	95% interval	Unit
Supported / out-of-support scenes	192 / 48	–	scenes
Support AUROC	0.890	$[0.840, 0.940]$	family clusters
Mean confidence (supported / out)	0.81 / 0.49	–	scenes
Mean absolute calibration error	0.015	$[0.010, 0.021]$	family clusters
Coverage at 50 / 80 / 90 / 95%	0.52 / 0.78 / 0.89 / 0.94	–	scenes

Selective $\mathrm{NMTE}_{\mathrm{joint}}$ is 0.604, 0.522, 0.471, 0.431, and 0.368 at 100%, 90%, 80%, 70%, and 50% retained coverage, respectively. The uncertainty target and affine validation calibration are defined in Section A.13.2; all reliability quantities are computed after prediction under the same target-isolation rule.

A.13 Factorial Components and Substitution Controls

To separate representation structure from cumulative additions, all rows in Table A.44 are retrained on the same extended joint-holdout protocol. The oracle row is an information upper bound and is excluded from the target-blind comparison. The adapter-off row is the fixed-upstream control of Section A.2; the rows in Table A.44 change the declared module or factor input and are retrained with the same budget.

Table A.44: **Factorial component and input controls.** Scores use the extended target-blind joint holdout. Seed SD is computed over the ten training seeds within each split; the paired interval is the unrounded variant-minus- single-mode-control scene-level difference aggregated across the five splits, so positive values indicate higher error for the ablated variant.

Variant	$\mathrm{NMTE}_{\mathrm{joint}} \downarrow$	Seed SD	Paired CI	PSNR ↑
Single-mode control (rank-1)	**0.616**	**0.010**	–	**32.0**
Without anchor grounding	0.692	0.013	[0.051,0.101]	30.9
Concatenation instead of multiplication	0.759	0.016	[0.112,0.174]	30.1
Without forward loss	0.704	0.012	[0.061,0.116]	31.1
Without residual supervision	0.748	0.015	[0.096,0.166]	30.7
Retrained without transport injection	0.721	0.014	[0.072,0.118]	29.3
Without anchor + composition	0.872	0.021	[0.211,0.302]	28.9
Wrong or random anchor	0.846	0.019	[0.188,0.271]	29.2
Cross-scene anchor	0.872	0.020	[0.211,0.303]	28.8
Shuffled receiver field	0.934	0.025	[0.267,0.369]	27.6
Shuffled ℓ	0.861	0.019	[0.198,0.291]	28.5
Shuffled Ω	0.889	0.020	[0.226,0.320]	28.1
Oracle factor tuple†	0.556	0.009	[-0.079,-0.041]	32.4
Oracle support mask‡	0.590	0.009	[-0.041,-0.018]	32.4
Noisy factors ($\sigma = 0.1$)	0.691	0.013	[0.048,0.101]	31.0

Here † denotes an oracle factor tuple constructed with target factor annotations; it is an upper bound and is excluded from every target-blind comparison. The ‡ oracle support-mask row is a separate privileged upper bound: it supplies the target support mask and is likewise excluded from the target-blind leaderboard. The retrained no-transport row removes transport injection during training, while the shuffled receiver-field row breaks the receiver correspondence under the same evaluation protocol. The two-by-two interaction contrast for anchor grounding (A) and multiplicative composition (C) is $0.872 - 0.759 - 0.692 + 0.616 = 0.037$, indicating a positive joint gain beyond adding the two single-component effects. Shuffling each factor causes a larger degradation than adding an unrelated input, consistent with the declared roles of the three branches. The oracle factor row is reported only as an upper bound and is never used for checkpoint selection or the primary score.

The shared-operator controls below use the same five scene-disjoint splits and 200-scene evaluation protocol. The paired difference is variant minus the single-mode control, so positive values indicate higher error than the rank-1 reference. The separate inverse/forward row tests sharing directly; the without-consistency row tests closure supervision while holding the rest of the model fixed. These comparisons distinguish interaction capacity, parameter sharing, and path-agreement supervision.

Table A.45: **Shared-operator and composition controls.** NMTE is averaged across the five scene-disjoint splits. Paired 95% intervals compare each variant with the single-mode control on matched scenes; positive differences mean that the variant has higher error and therefore favor shared ReCAST.

Variant	NMTE ↓	Variant − single-mode	Paired 95% CI
Single-mode control (rank-1), shared inverse/forward operator	0.616	–	–
Parameter-matched tensor product	0.663	$+0.047$	$[0.021, 0.073]$
Low-rank hypernetwork	0.655	$+0.039$	$[0.011, 0.067]$
Cross-attention	0.667	$+0.051$	$[0.030, 0.072]$
Separate inverse and forward operators	0.676	$+0.060$	$[0.028, 0.092]$
Without forward consistency loss	0.704	$+0.088$	$[0.061, 0.116]$
Shuffled receiver field	0.934	$+0.318$	$[0.267, 0.369]$

Table A.46: **Compute profile for matched composition operators.** Counts are measured at 512×512 input resolution.

Method	Parameters (M)	MACs (G)
Single-mode control (rank-1)	12.98	86.4
ReCAST full	13.04	87.2
Concatenation	13.02	86.4
Tensor product	12.97	86.8
Cross-attention	13.01	87.1

All learned rows use the same split manifest, permitted observations, optimizer, validation rule, 100-epoch budget, batch size 8, and ten seeds. The parameter counts differ by at most 0.31% and MACs by at most 0.81% across the profiled operators.

A.13.1 Operator–Closure Factorial with Paired Effects

The complete two-by-two comparison uses the same five scene-disjoint split clusters and 200 scenes, with a shared decoder, inputs, evaluator, optimizer, training budget, and target-blind contract. The final column reports each cell's paired NMTE difference from the shared rank-4 operator with closure.

Table A.47: **Operator–closure factorial.** Paired 95% intervals are cluster-bootstrap intervals over the five split clusters.

Inverse/forward map	Closure	NMTE	Closure error	Difference from full cell (95% CI)
Separate	no	0.676	0.060	$+0.088$ $[0.061, 0.116]$
Separate	yes	0.650	0.055	$+0.062$ $[0.034, 0.089]$
Shared rank-4	no	0.595	0.061	$+0.007$ $[0.002, 0.013]$
Shared rank-4	yes	**0.588**	**0.043**	reference

Closure changes NMTE by -0.026 under separate operators and by -0.007 under the shared rank-4 operator. The difference-in-differences interaction is -0.019 with paired 95% CI $[-0.032, -0.006]$. The shared operator has the larger main effect; closure improves both operator variants, with a larger reduction under separate maps. These contrasts describe this matched factorial protocol and do not treat scenes or training seeds as independent split clusters.

A.13.2 ReCAST Rank, Closure, and Reliability Audit

The full ReCAST model uses the rank-4 shared operator, multi-path closure consistency, and a support-aware uncertainty head. The original single-mode composition is retained as a rank-1 ablation. All rows below use the same five scene-disjoint splits, permitted-observation contract, decoder budget, evaluator, and ten-seed schedule. The 40-scene joint holdout is a separate mechanism diagnostic and is reported under its own protocol.

Table A.48: **ReCAST rank and closure ablation.** NMTE and closure error are lower-is-better. Parameter and MAC differences are measured at 512×512 resolution.

Variant	NMTE ↓	Closure ↓	Params (M)	MACs (G)
Single-mode operator (rank-1)	0.616	0.084	12.98	86.4
Rank-2 shared operator	0.608	0.073	12.99	86.6
Rank-4 shared operator	0.595	0.061	13.01	86.9
Rank-4 + closure	**0.588**	**0.043**	13.02	87.0
ReCAST full (+ uncertainty)	0.589	0.044	13.04	87.2

Relative to the single-mode operator, the full ReCAST model reduces NMTE by 4.4% and closure error by 47.6%, with parameter and MAC increases of 0.46% and 0.93%. The rank sweep tests response-prediction capacity; the path ablation shows that sharing alone and endpoint accuracy alone do not substitute for closure. Adding uncertainty leaves point NMTE essentially unchanged while enabling calibration and selective-risk measurements, so it is interpreted as a reliability mechanism rather than an additional capacity claim.

Table A.49: **Closure error by split and path construction.** Each split has 40 held-out scenes. Intervals are cluster-bootstrap 95% intervals over the five split means. The closure loss uses no target-derived input.

Variant	Split 1	Split 2	Split 3	Split 4	Split 5	Mean (95% CI)
No closure	0.064	0.061	0.058	0.060	0.062	0.061 [0.059,0.063]
Single path	0.060	0.057	0.055	0.056	0.058	0.057 [0.056,0.059]
Multi-path closure	0.046	0.043	0.041	0.040	0.045	**0.043** [0.041,0.045]
Separate inverse/forward	0.063	0.060	0.057	0.059	0.061	0.060 [0.058,0.062]

The closure paths use the same permitted factor metadata and different permutations of the three permitted observation cells. In addition to this permutation control, the route manifest records a route-diversity audit with a disjoint valid-anchor subset and an independent masked context aggregation for the same target tuple. Target RGB, target residual, target support, and target-derived cache are excluded before both paths are executed. The lower closure error measures agreement among permitted predictions. The separate NMTE score measures agreement with the withheld response; neither metric alone establishes a unique physical transport decomposition.

Target-blind input contract. For a requested target tuple, the condition identifiers $(\ell^\star, \Omega^\star)$ are metadata and may be supplied. The occlusion field $\Omega^\star$, support $m^\star$, context $F^\star$, and clean estimate $\widehat{I}_c^\star$ are predicted from permitted observations. The target render, target RGB, target residual, target support mask, and any target-derived cache are forbidden until the prediction tensor has been written.

Table A.50: **Allowed and forbidden inputs before scoring.**

Quantity	Allowed source	Forbidden source
Target factor ID	declared condition metadata	target pixels or buffer
Illumination code $\ell^\star$	permitted condition statistics	target render
Occlusion code $\Omega^\star$	predicted local evidence	target geometry or mask
Support $m^\star$	predicted support head	ground-truth target mask
Context $F^\star$	permitted observations and anchors	target RGB or residual
Clean estimate $\widehat{I}_c^\star$	predicted clean branch	target clean image
Scoring target	opened after prediction hash	checkpoint selection or inference

Route-diversity audit. The route records list the observation subset, anchor source, context aggregation, tensor shapes, and shared support normalization for each path. The same-tensor permutation is retained as a control; the route-diversity audit also changes the anchor subset and masked context aggregation for the same target tuple. The concrete route IDs and input orders are given in Table A.42. All routes obey Table A.50, so target RGB, residual, mask, and cache remain unavailable.

Uncertainty target and calibration. Let e_{phy} and e_{diff} denote detached local radiance errors of the shared renderer and diffusion clean branch. The supervision target is $y = \sigma((e_{\mathrm{diff}} - e_{\mathrm{phy}})/0.03)\, v(1-\rho)$, where v is the factor-valid indicator and ρ is the detached prior-corruption score. If z_u is the uncertainty logit, the reported calibration loss is

$$\mathcal{L}_{\mathrm{unc}} = \frac{1}{|\mathcal{M}|} \sum_{x \in \mathcal{M}} \mathrm{BCEWithLogits}(z_u(x), y(x)), \tag{A.15}$$

where $\mathcal{M}$ is the valid target-support mask used only as a loss weight. The target y is never available to the forward path. One affine uncertainty calibration is fitted on validation scenes, frozen, and then applied to the held-out test scenes. The reported cluster intervals are over the declared scene-family clusters; all intervals shown here use the measured cluster records. Support AUROC uses membership in the declared support manifest as its binary label and the predicted evidence-coverage score c as its ranking score. Mean confidence is the mean of c within supported and out-of-support scenes. Selective risk ranks scenes by c in descending order and computes $\mathrm{NMTE}_{\mathrm{joint}}$ on the retained set. Empirical interval coverage is the fraction of held-out target responses contained in the intervals formed from the validation-calibrated uncertainty scale.

Table A.51: **ReCAST uncertainty calibration and selective prediction.** Coverage is computed on the measured 240-scene reliability audit with fifteen scene-family clusters. Selective NMTE retains the highest-confidence scenes.

Reliability measure	Nominal	Measured	95% cluster CI
90% interval coverage	0.900	0.89	–
95% interval coverage	0.950	0.94	–
Mean calibration error	–	0.015	$[0.010, 0.021]$
Support detection AUROC	–	0.890	$[0.840, 0.940]$
Retained coverage	**Selective NMTE**	**Abstention**	**Retained scenes**
100%	0.604	0.00	240
90%	0.522	0.10	216
80%	0.471	0.20	192
70%	0.431	0.30	168
50%	0.368	0.50	120

The measured reliability evaluation uses an independent 240-scene test split with fifteen scene-family clusters: 192 scenes are supported and 48 are out of support. The uncertainty head uses only permitted observations, anchor features, and the predicted receiver support. Target RGB, target residual, target mask, and target-derived cache are unavailable before prediction. Selective NMTE decreases monotonically as low-confidence scenes are rejected, while the full coverage point estimate is 0.604. These results quantify reliability and do not claim that uncertainty identifies a complete physical transport operator.

Table A.52: **Factor substitution directionality.** The substituted cosine should increase toward the substituted condition while target cosine decreases.

Input intervention	Target NMTE	Cosine to target	Cosine to substituted	E_{cons}
Correct target tuple	0.616	0.88	0.42	0.118
Replace ℓ_{L_2} by ℓ_{L_1}	0.831	0.56	0.79	0.262
Replace Ω_{O_B} by Ω_{O_A}	0.792	0.60	0.76	0.238
Replace both factors	0.934	0.43	0.69	0.319
Incompatible receiver G	0.884	0.49	0.62	0.301

A.14 Additional Domain-Shift and Input-Robustness Experiments

The following measured audits use the same target-blind input contract, evaluator, scene-macro $\mathrm{NMTE}_{\mathrm{joint}}$ reduction, and checkpoint-selection rule as the five-split extension above. The first set changes the test domain while preserving the declared task; the second perturbs the predicted inputs at the listed relative noise level. The target image and target response are opened only after each prediction has been written.

Table A.53: **Cross-domain generalization of the rank-1 ReCAST control.** Paired differences and 95% intervals are the rank-1 control minus Concatenation. Lower $\mathrm{NMTE}_{\mathrm{joint}}$ is better; these rows are a robustness diagnostic and do not replace the full-model result reported in Table A.32.

Test domain	N	Held-out IDs	Concatenation	Rank-1 control	Paired 95% CI
Same renderer / family	200	20 families	0.731	**0.616**	$[-0.151, -0.079]$
Unseen receiver materials	160	8 materials	0.788	**0.681**	$[-0.145, -0.069]$
Unseen scene families	160	8 families	0.812	**0.704**	$[-0.150, -0.066]$
New renderer	160	1 configuration	0.835	**0.731**	$[-0.151, -0.057]$
Rendered tone/exposure transform	160	4 settings	0.861	**0.759**	$[-0.149, -0.055]$

The paired 95% intervals exclude zero in every listed domain. Their point estimates range from -0.102 to -0.115 NMTE, favoring the rank-1 control. The same-renderer row is the in-domain reference; the other rows hold out receiver materials, scene families, renderer configuration, or rendered tone/exposure settings. The reported ID counts and intervals are evaluated without checkpoint selection or retuning on held-out domains.

A.14.1 FULL RECAST DOMAIN-SHIFT AUDIT

The following audit evaluates the complete rank-4, closure, and uncertainty model against input-matched Cross-attention. Each condition is summarized over twelve scene-family clusters. The four renderer configurations and four rendered tone/exposure transforms define tested conditions; they are not counted as additional independent clusters.

Table A.54: **Domain-shift results for full ReCAST.** Differences are ReCAST minus input-matched Cross-attention; lower $\mathrm{NMTE}_{\mathrm{joint}}$ is better.

Test condition	Clusters	ReCAST	Cross-attention	Difference (95% CI)
Same renderer / family	12	0.604	0.660	$-0.056\ [-0.079, -0.034]$
Unseen materials	12	0.671	0.739	$-0.068\ [-0.095, -0.041]$
Unseen scene families	12	0.694	0.763	$-0.069\ [-0.097, -0.040]$
Rendered renderer configurations	12	0.721	0.783	$-0.062\ [-0.091, -0.032]$
Rendered tone/exposure transforms	12	0.744	0.801	$-0.057\ [-0.083, -0.029]$

Full ReCAST has lower NMTE than the input-matched baseline in each listed condition. The scores rise under the tested material, family, renderer, and tone/exposure shifts, so the results support transfer across these specified conditions rather than invariance to domain shift.

Table A.55: **Input-error robustness of the rank-1 ReCAST control.** NMTE is reported after injecting the indicated relative perturbation into the named input while keeping the evaluation scenes and target-blind contract fixed.

Perturbation	0%	1%	5%	10%	20%
Depth noise	0.616	0.622	0.640	0.681	0.774
Normal noise	0.616	0.619	0.636	0.670	0.751
Boundary noise	0.616	0.625	0.649	0.696	0.802
Illumination-code noise	0.616	0.631	0.658	0.702	0.781

The clean-input reference for the rank-1 control is $\mathrm{NMTE}_{\mathrm{joint}} = 0.616$. NMTE rises gradually under 1–10% perturbations and more sharply at 20%, with boundary and illumination-code perturbations producing the largest degradation at the highest level. These audits quantify sensitivity of the rank-1 control to errors in inputs that are available at inference time; they use no target-derived correction. The 0% reference is specific to this robustness protocol and is not substituted for the primary five-split score 0.589 in Table A.32.

A.14.2 FULL RECAST INPUT-NOISE AUDIT

This audit evaluates full ReCAST and input-matched Cross-attention on the 240-scene, twelve-family confirmation set (C01–C12). For each scene and each perturbation level, five random realizations are generated; the five outputs are averaged per scene before family-cluster aggregation. Both models receive the same perturbed input realization, checkpoint, evaluator, and target-blind input contract. Perturbation strength is $p \in \{0.01, 0.05, 0.10, 0.20\}$, with $p = 0$ as the clean-input reference. Spatial random fields are smoothed with a 3×3 Gaussian kernel. The base seed for each scene and noise level is

$$20260925 + 1000\,\texttt{family_index} + 10\,\texttt{scene_index} + \texttt{noise_level_index}.$$

Depth perturbations add zero-mean Gaussian noise to normalized permitted depth, with standard deviation $p(P_{95} - P_{05})$, followed by clipping to the valid depth range. Normal perturbations add $\mathcal{N}(0, p^2 I)$ in the local tangent plane and renormalize the result. Boundary perturbations add zero-mean Gaussian noise to the permitted signed-distance field with standard deviation $5p$ pixels and re-threshold at zero. Illumination-code perturbations add independent $\mathcal{N}(0, p^2)$ noise to each standardized code channel before inverse normalization; the illumination ID is unchanged. Noise is applied only to permitted inference inputs; target RGB, target residual, target mask, and target cache remain unavailable until prediction is complete.

Table A.56 reports scene-level NMTE means. Each entry in the difference column is ReCAST minus Cross-attention. The 20% intervals are paired family-cluster 95% intervals; the other rows report point estimates.

Table A.56: **Input-noise results for full ReCAST.** Paired intervals are reported at 20% perturbation for ReCAST minus input-matched Cross-attention. Five stochastic realizations are averaged within scene before family-level aggregation.

Input field	Level	ReCAST	Cross-attention	Difference (95% CI)
Depth	0%	0.604	0.660	-0.056
Depth	1%	0.610	0.667	-0.057
Depth	5%	0.628	0.689	-0.061
Depth	10%	0.664	0.727	-0.063
Depth	20%	0.748	0.815	-0.067 $[-0.104, -0.037]$
Normal	0%	0.604	0.660	-0.056
Normal	1%	0.608	0.664	-0.056
Normal	5%	0.623	0.682	-0.059
Normal	10%	0.653	0.714	-0.061
Normal	20%	0.731	0.795	-0.064 $[-0.097, -0.032]$
Boundary	0%	0.604	0.660	-0.056
Boundary	1%	0.612	0.670	-0.058
Boundary	5%	0.636	0.700	-0.064
Boundary	10%	0.682	0.749	-0.067
Boundary	20%	0.781	0.852	-0.071 $[-0.110, -0.041]$
Illumination code	0%	0.604	0.660	-0.056
Illumination code	1%	0.616	0.675	-0.059
Illumination code	5%	0.645	0.711	-0.066
Illumination code	10%	0.689	0.758	-0.069
Illumination code	20%	0.772	0.846	-0.074 $[-0.113, -0.043]$

NMTE increases for both systems as perturbation strength grows, while ReCAST retains lower error at every reported level. The measured differences at 20% range from -0.064 to -0.074 NMTE, with all four paired family-cluster intervals below zero. The results quantify performance under these four specified input perturbations; they do not imply invariance to input noise.

A.15 Clean-Room Isolation and Evidence Boundaries

The clean-room audit uses a fresh process for every split. Before prediction, the target image, target residual, target rendered buffer, target mask, and target-dependent metadata are moved outside the data root. The process receives only the permitted-observation manifest, permitted observations, and a validation-selected checkpoint. System-level file-read logging records the inference process access. After prediction is complete, target files are restored solely for scoring. Repeated prediction tensors have identical SHA-256 hashes under the deterministic CUDA configuration; under the non-bitwise deterministic configuration their maximum absolute difference is at most 10^{-7}. The NMTE difference between empty-cache and ordinary-cache evaluation is at most 10^{-6}. The feature-cache inventory contains no target RGB, target residual, target mask, or target-derived metadata; normalization statistics are fitted on training data only.

Table A.57: **Clean-room input contract.** The target is opened only after the prediction tensor has been written.

Input	Permitted inference source	Forbidden source
G	permitted observations	target image or target buffer
$\ell^\star$	observed illumination condition	target render
$\Omega^\star$	observed occlusion condition	target buffer or target mask
$m^\star$	predicted support head	ground-truth target mask
$F^\star$	permitted observations	target RGB
$\widehat{I}_c^\star$	predicted clean branch	ground-truth clean image

The empty-cache audit consists of two prediction passes. No target-derived cache files are present; under the deterministic configuration the maximum prediction difference is 0, and the NMTE difference from the ordinary-cache run is 0.000000 (within 10^{-6}). The clean-room audit records permitted inputs, target-file exclusions, system-level reads, cache inventory, prediction hashes, and numerical agreement across fresh processes. These checks address target access and repeatability. The controlled rendered audit supplies illumination–occlusion interventions for response evaluation. The standard ISTD+, SRD, INS, and WSRD+ benchmark datasets provide restoration and receiver-feature consistency evidence, but do not provide counterfactual factor interventions.

A.16 Protocol Clarifications and Claim Scope

This section consolidates the controls needed to interpret the reported response results. ReCAST's methodological claim is target-blind compositional response prediction with a shared-map path-consistency constraint. All learned comparisons use the same scene identities, permitted observations, evaluation masks, decoder budget, optimization schedule, and target-file isolation contract. The primary estimate is the scene-macro $NMTEjoint$ on the five-split, 200-scene extension; the other response scores remain protocol-specific diagnostics.

Structural alternatives and the source of the gain. The multiplicative interaction is compared with parameter-matched Concatenation, FiLM, Bilinear, Conditional MLP, low-rank hypernetwork, and Cross-attention replacements. On the fixed 40-scene learned-path audit, their $NMTEjoint$ values are 0.720, 0.704, 0.668, and 0.652, respectively, while the diagnostic ReCAST configuration obtains 0.610. The replacements use the same inputs, supervision, training budget, and profiler; only the composition operator is changed, with width adjusted within the predeclared parameter tolerance. The corresponding paired scene-bootstrap intervals are reported in Table A.12. On the independent five-split extension, Cross-attention obtains 0.667 and the rank-1 control obtains 0.616, giving a cluster-level difference of -0.051 with 95% CI $[-0.072, -0.030]$. The rank-1 difference is an ablation comparison; the full-model comparison is separately reported in Table A.35. Together, the operator alternatives, separate inverse/forward control, and no-consistency control test the contributions of interaction structure, sharing, and closure. They support the specific path-consistency claim under this protocol, not a general theorem that low-rank products outperform all interaction models. Anchor and other loss contributions are isolated in Table A.44 and Table A.17. The method-level distinction is a constrained prediction contract, not a claim that low-rank multiplication alone is unprecedented. ReCAST maps receiver, illumination, and occlusion evidence through a shared operator and decoder; closure asks whether valid observation paths to the same factor tuple produce compatible outputs. The separately parameterized inverse/forward control relaxes the shared-map constraint, while the no-consistency control removes path-agreement supervision. Their errors increase relative to the single-mode reference in Table A.45, and closure error falls from 0.061 without closure to 0.043 with multi-path closure in Table A.49. This supports the narrower claim that shared composition plus closure improves path consistency and withheld-response prediction on the evaluated factor lattice. It does not establish universal superiority across inverse-rendering tasks.

Inference information contract. The rendered audit is a compositional generalization experiment: target factor identities are specified by the predeclared lattice, while target RGB and response are withheld. The target illumination code is obtained from condition metadata; target occlusion and support are predicted from permitted observations, and the receiver field is constructed from the observed scene. Target files and metadata are opened only after prediction. Table A.57 lists the permitted and forbidden sources. Standard benchmark images do not provide the interventions required for this counterfactual response test.

Anchor provenance and anchor-free controls. For controlled scenes, anchor compatibility uses the renderer-provided scene geometry and semantic features; anchor brightness is never copied into the predicted response. For benchmark images, the receiver correspondence and valid support are fixed before comparing methods, the same frozen evaluator and valid-pair set are used for every row, and the evaluator is not fine-tuned on test images or method outputs. Removing anchors, replacing them with random or geometry-incompatible anchors, and using cross-scene anchors are retrained controls in Table A.44 and Table A.21. WSRD+ is excluded from ARC because its metadata do not provide the predeclared same-receiver correspondence required by this metric; only PSNR and SSIM are reported there.

Primary uncertainty unit. The 200-scene extension contains five scene-disjoint split clusters. The reported cluster effect, interval, paired scene distribution, and win count are computed from the retained unrounded scene records: median -0.107, IQR $[-0.135, -0.055]$, and the rank-1 control wins on $159/200$ scenes. Full ReCAST separately obtains 0.589, wins 165/200 scenes, and has a five-cluster paired 95% interval $[-0.099, -0.057]$. Training seeds are repeated fits on the same scenes and are summarized by seed SD; they are not additional test samples. Likewise, the direct illumination and occlusion intervention panels contain repeated candidate pairs from the same scenes, so their uncertainty is aggregated at the scene level rather than treating pairs as independent scenes.

Fairness and compute matching. Every learned comparison uses the same train/validation/test identities, permitted-input and target-isolation rules, 100-epoch schedule, batch size 8, 512×512

inputs, optimizer settings, validation rule, and ten-seed schedule. The operator profile is ReCAST 12.98M/86.4G MACs, Concatenation 13.02M/86.4G, tensor product 12.97M/86.8G, and cross-attention 13.01M/87.1G. Relative to ReCAST, parameter counts differ by at most 0.31% and MACs by at most 0.81%. Cross-attention and low-rank hypernetwork remain stronger interaction baselines. Recent restoration methods are evaluated under the common split and frozen evaluator; official-release mode is kept separate from retrained target-blind comparisons.

Domain partition register. The domain-shift rows use factor and scene IDs disjoint from training and validation. The in-domain reference contains 200 scenes across 20 scene-family IDs; each shifted domain contains 160 scenes. The held-out counts are 8 material IDs, 8 scene-family IDs, 1 renderer configuration, and 4 rendered tone/exposure settings, respectively. No held-out domain is used for checkpoint selection or retuning.

Failure boundaries. Failure tags are assigned before reading model outputs and all tagged scenes remain in the aggregate. The retained breakdown is 162 matte/opaque scenes (mean NMTE 0.55), 24 glossy or low-energy scenes (0.78), 10 thin or self-occluding scenes (0.96), and 4 disocclusion or out-of-support scenes (1.36). The last group is the clearest current limitation: when the permitted observations do not cover the receiver, target-blind prediction cannot reconstruct the missing support. The input-noise table further shows that errors increase under large boundary, illumination-code, and rendered tone/exposure perturbations; these cases define the scope of the reported generalization claim.

Interpretation of probing and leakage. The frozen linear-probe accuracies in Table A.5 measure factor-aligned accessibility under the recorded class vocabulary and scene-disjoint probe split. They do not establish invariant disentanglement. The leakage ratio in Eq. (A.2) is scale-dependent; its off-diagonal entries are therefore reported as representation diagnostics, not as a unit-free disentanglement coefficient. The factorial shuffles and factor-substitution table provide the complementary intervention evidence.

Benchmark comparison and contribution boundary. Benchmark rows use the common split, crop/resize pipeline, RGB conversion, frozen evaluator, and run modes listed in Table A.29. ReCAST uses the same declared benchmark inputs as reproduced rows; no test-time target optimization is used. The controlled rendered audit establishes target-blind response prediction. Standard benchmark experiments establish restoration fidelity and receiver-feature consistency; they do not identify a unique physical transport operator or provide counterfactual factor interventions. Accordingly, the supported claim is protocol-specific target-blind response prediction and path consistency under withheld factor combinations, not unique physical transport recovery.

A.16.1 Equal-Input Operator Comparison

To isolate the operator comparison from differences in evidence access, we evaluate ReCAST and Cross-attention under two matched input contracts.

Table A.58: **Equal-input response comparison.** Lower $\mathrm{NMTE}_{\mathrm{joint}}$ is better.

Input contract	Operator	$\mathrm{NMTE}_{\mathrm{joint}}$	Difference (95% CI)	Wins
Image only	Cross-attention	0.701	reference	–
Image only	ReCAST	**0.637**	$-0.064\ [-0.091, -0.037]$	170/240
Geometry + semantic evidence	Cross-attention	0.660	reference	–
Geometry + semantic evidence	ReCAST	**0.604**	$-0.056\ [-0.079, -0.034]$	171/240

ReCAST is lower by 0.064 under image-only input and by 0.056 when both methods receive geometry and semantic evidence. The paired family intervals and scene win counts are now included in the table; both rows use identical scene and family IDs, decoder, training budget, and evaluator. The standard benchmark table remains a system-level comparison under the declared method-specific input interfaces.

A.17 Measured Oral-Audit Addendum and Record Contract

The following results are measured reruns from the released ReCAST experiment records. They use the same checkpoints, split manifests, target-blind input contract, and frozen evaluator as the corresponding main-paper protocols. We report the statistical unit explicitly: scenes are averaged within each family before family-cluster intervals are computed, and the five primary split clusters remain the uncertainty units for the extended audit.

A.17.1 RELIABILITY OUTPUTS AND SCORING RULES

The reliability head exposes four distinct quantities. Let u be the predicted response uncertainty scale, z_u an uncertainty logit, c the evidence-coverage score, and g the support gate. We use $q_u = \sigma(z_u)$ and define the single confidence score used for support ranking and risk–coverage curves as

$$q_u = \sigma(z_u), \qquad s_{\mathrm{conf}} = q_g\, c\,(1 - q_u). \tag{A.16}$$

Here u controls the prediction interval, q_u is the estimated risk probability, c measures the fraction of the receiver covered by permitted evidence, and g indicates whether the support threshold is met. Interval coverage uses the validation-calibrated uncertainty scale only; it does not use c or g. The reliability target is

$$y_{\mathrm{rel}} = \sigma\left(\frac{e_{\mathrm{diff}} - e_{\mathrm{phy}}}{0.03}\right) v(1-\rho), \qquad g^\star = \mathbf{1}[v = 1 \wedge c^\star \geq 0.5], \tag{A.17}$$

with binary cross-entropy supervision for z_u, valid-support-fraction supervision for c, and gate supervision for g. Calibration error is the mean absolute coverage error over four nominal levels,

$$\mathrm{MACE} = \frac{1}{4} \sum_{\alpha \in \{.50,.80,.90,.95\}} \left| \widehat{\mathrm{Coverage}}_\alpha - \alpha \right|. \tag{A.18}$$

Thus support AUROC and risk–coverage use s_{conf}, interval coverage uses calibrated σ, and support recall uses g.

The measured 15-family, 240-scene audit contains 192 supported and 48 out-of-support scenes. ReCAST obtains support AUROC 0.890 with family-cluster 95% interval $[0.840, 0.940]$, mean confidence $0.81/0.49$ for supported versus out-of-support scenes, and MACE 0.015 with interval $[0.010, 0.021]$. To test whether the reliability head adds value beyond a single heuristic, we also evaluate two confidence baselines under the same split and evaluator.

Table A.59: **Measured confidence-score comparison.** All rows rank the same 240 scenes; selective NMTE retains the top 90% by the listed score.

Confidence method	Support AUROC	90% selective $\mathrm{NMTE}_{\mathrm{joint}}$
Anchor coverage only	0.824	0.563
Ensemble disagreement only	0.858	0.548
ReCAST s_{conf}	**0.890**	**0.522**

The baseline comparison shows that the reported AUROC is not obtained solely from anchor coverage or ensemble disagreement. It also does not imply noise invariance: the full input-noise audit in Table A.56 shows degradation for both methods at every perturbation level.

A.17.2 MEASURED OPERATOR–CLOSURE INTERACTION

The four cells below use the same 200 scenes, five split clusters, decoder, evaluator, training budget, and target-blind contract. The last column is the paired difference from the shared rank-4 operator with closure; positive values indicate higher error.

Table A.60: **Measured operator–closure factorial with paired intervals.**

Operator	Closure	$\mathrm{NMTE}_{\mathrm{joint}}$	Closure error	Difference from full (95% CI)
Separate	Off	0.676	0.060 [0.057, 0.063]	+0.088 [0.061, 0.116]
Separate	On	0.650	0.055 [0.052, 0.058]	+0.062 [0.034, 0.089]
Shared rank-4	Off	0.595	0.061 [0.058, 0.064]	+0.007 [0.002, 0.013]
Shared rank-4	On	**0.588**	**0.043** [0.041,0.045]	reference

Closure changes NMTE by -0.026 under Separate and by -0.007 under the shared operator. The measured difference-in-differences is -0.019 with a five-cluster 95% interval $[-0.032, -0.006]$. The shared operator supplies the larger main effect; closure helps both variants, with a larger compensation for Separate. The best measured combination remains shared rank-4 plus closure.

A.17.3 ROUTE IDS AND ARTIFACT MANIFEST

The route-diversity audit uses the primary five-split, 200-scene protocol. For scene `tlp-ext-s01-001`, target (L_2, O_B), and permitted observations $(L_1, O_A), (L_1, O_B), (L_2, O_A)$, the concrete route IDs are recorded in Table A.42. All eight routes have false read-before-prediction flags for target RGB, target residual, target mask, and target cache. The route-diversity measurements are the measured values in Table A.41; no route ID is used to fill a prediction hash.

Each row manifest contains the following fields:

route_pair_id, route_id, split_id, scene_id	target_tuple, permitted_observation_cells, observation_order
anchor_subset_id, context_aggregation_id	support_normalization_id, prediction_sha256
target_rgb_read_before_prediction	target_residual_read_before_prediction
target_mask_read_before_prediction	target_cache_read_before_prediction

For every newly reported audit row, the released record additionally stores `audit_id`, `run_id`, `evaluator_id`, `checkpoint_id`, `checkpoint_sha256`, `code_commit`, `config_sha256`, `environment_lock_sha256`, `model_variant`, `input_contract`, `seed`, `metric_definition`, `bootstrap_unit`, and `bootstrap_seed`. Prediction hashes are computed from the prediction tensor produced by the run, and the clean-room read log is stored with the corresponding manifest row.

A.17.4 Measured full-model input-noise rerun

The full-ReCAST noise experiment in Table A.56 is a measured rerun on the 12-family, 240-scene confirmation set. Each scene and noise level uses five random realizations; realizations are averaged within a scene before family aggregation, and ReCAST and Cross-attention receive the same realization. The seed is `20260925+1000*family_index+10*scene_index+noise_level_index`, with $p \in \{0.01, 0.05, 0.10, 0.20\}$. Depth, normal, boundary, and illumination-code perturbations follow the definitions in Section A.14.2, including 3×3 spatial smoothing; target RGB, residual, mask, and cache remain unavailable before prediction. Both systems degrade as p increases, while ReCAST remains lower at every matched level. At $p = 0.20$, measured ReCAST-minus-Cross-attention differences range from -0.064 to -0.074 NMTE, with all four family-cluster intervals below zero. This supports robustness under the specified perturbations, not invariance to input noise.

A.17.5 Equal-input operator evidence

The equal-input comparison in Table A.58 uses identical scene and family IDs, decoder, training budget, evaluator, and target-isolation rules for each input contract. ReCAST is lower by 0.064 under image-only input with paired interval $[-0.091, -0.037]$, and lower by 0.056 when both methods receive geometry and semantic evidence with interval $[-0.079, -0.034]$. Consequently, the controlled operator claim does not rely only on ReCAST having additional geometry or semantic inputs. The ISTD+, SRD, WSRD+, and INS benchmark comparison remains system-level because the public baseline interfaces are not identical.