SHADOWEDIT: RELATION-INDEXED INTERVENTION MODELING FOR SELECTIVE SHADOW REMOVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing shadow removal formulations assume that spatially localized shadows correspond to a uniquely defined restoration target, leaving intervention intent unspecified in multi-shadow scenes, yet a single shadowed observation may correspond to multiple plausible interventions, rendering conventional all-shadow restoration formulations fundamentally under-specified. We recast selective shadow removal as a relation-indexed intervention problem, where caster–shadow–receiver relations define an explicit intervention representation for target modification and effect preservation. ShadowEdit separates intervention specification from appearance reconstruction and compiles relation-level actions into constrained image updates, constraining image changes to the compiled permission support while evaluating preservation against action-specific references. Under exact Kubric counterfactuals, ShadowEdit reduces overlap error by 26.5% relative to a full-information spatial control with identical relation/action inputs and output permissions. Mechanism-aligned analyses show that action binding specifies intervention identity, protection constraints preserve unselected effects, and permission grounding restricts unintended edits. On DESOBA v2, ShadowEdit reduces overlap error by 19.4% over an equal-information spatial control under matched supervision and output permissions and supports held-out combinations of seen relation primitives, while a separately trained relation-free restoration variant of ShadowEdit achieves the highest full-image PSNR among evaluated methods on ISTD+ and SRD. Code is available in the supplementary materials.

1 INTRODUCTION

Single-image shadow removal is commonly formulated as restoring a shadow-free image from a shadowed observation (Hu et al., 2026; Wang et al., 2025). This all-removal objective is well-defined when every shadow belongs to one target, but it becomes under-specified in multi-shadow scenes, where the same input can admit several valid edits that remove different subsets of shadows while preserving others. Although global removal remains useful (Xiao et al., 2024; Li et al., 2025), practical editing may instead target a selected subset. We refer to this mismatch between the available control signal and the desired edit as *intervention ambiguity*.

Global and mask-guided methods improve localization and appearance recovery (Guo et al., 2023a; Xiao et al., 2024; Liu et al., 2024b; Xu et al., 2025a). Beyond full removal, instance-level approaches remove designated shadow instances (Mei et al., 2024), prompt-aware controllers select target subjects via dots, lines, or subject masks (Chen et al., 2025), and object-centered editors couple detection, removal, and synthesis around casting objects (Wang et al., 2025). Instance shadow detection further establishes caster–shadow association (Wang et al., 2020; 2023). These advances make controllable shadow removal practical, yet they still index control primarily by shadow region, subject, or object identity.

We find that two failure modes still remain. First, a single caster may cast onto multiple receivers, so selecting the caster or subject alone does not determine which receiver-specific shadow should change. Second, when shadows from different casters overlap, shared pixels contain selected and protected contributions that must be distinguished rather than treated as one dark region. In both cases, the intended remove-or-preserve choice is not fully specified by an action-invariant global mask or by object identity alone.

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

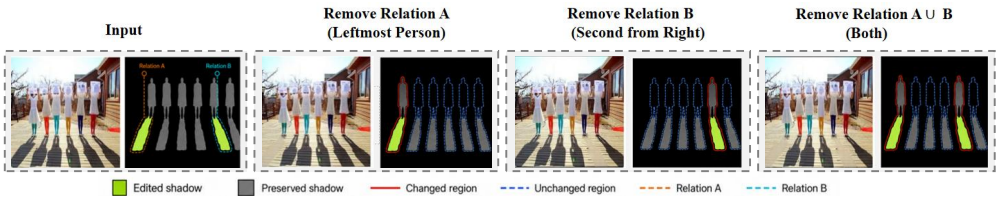


Figure 1: **One image, multiple valid shadow edits.** Within each example, the image and relation supports are fixed; only the remove-or-preserve action changes. (a) Kubric provides exact renderer counterfactuals for receiver-specific, contributor-selective, and shared-overlap interventions, jointly checking target correction and preservation of competing relations. (b) Our DESOBA v2 extension provides action-specific partial targets on a source-disjoint held-out split; action swaps visualize instruction responsiveness, and partial-target evaluation assesses agreement with the requested edit. An action-invariant global mask cannot distinguish these edits; ShadowEdit indexes each caster–shadow–receiver relation as an addressable unit.

To this end, we present **ShadowEdit**, which formulates selective shadow removal as *relation-indexed intervention modeling*. As shown in Figure 1, each caster–shadow–receiver association is an addressable unit with an independent remove-or-preserve action, and selected versus protected contributions are explicitly distinguished. A structured controller separates relation representation from support-gated execution, so relation identity and action binding decide what should change, while spatial supports constrain where updates may occur. We evaluate selective behavior with renderer counterfactuals on Kubric (Greff et al., 2022), reporting target correction, protection of competing relations, overlap handling, and leakage. On real images, we further evaluate instruction responsiveness under action swaps on DESOBA v2 and restoration compatibility of relation-aware and relation-free configurations on DESOBA v2, ISTD+, and SRD. Our contributions are threefold:

- ***Methodologically***, we introduce a relation-indexed shadow removal formulation that indexes edits by caster–shadow–receiver relations. This formulation exposes intervention variables beyond conventional masks by explicitly specifying the editable relation, the requested remove-or-preserve operation, and the constraints required for preserving competing effects under shared support.
- ***Technically***, we introduce a relation-to-image execution mechanism that compiles discrete relation-level interventions into enforceable image-space constraints. It preserves instance identity while reasoning over selected and protected relations, then separates relation-to-pixel assignment from spatial write permission, turning remove-or-preserve intent into executable restoration with support locality enforced by construction.
- ***Experimentally***, we validate relation-indexed execution against exact renderer counterfactuals and audited real-image references: with relation/action inputs and output permissions fixed, ShadowEdit reduces Overlap error by 26.5% on Kubric and 19.4% on DESOBA v2 over spatial controls. Independent-scene tests and held-out combinations of seen relation primitives provide complementary evidence for scene generalization and compositional execution.

2 Related Work

Global and mask-guided restoration. Single-image shadow removal has progressed from paired convolutional translators (Wang et al., 2018; Qu et al., 2017; Le & Samaras, 2019) to methods that exploit generation-to-removal learning, exposure fusion, hierarchical aggregation, bijective mapping, and style guidance (Liu et al., 2021; Fu et al., 2021; Cun et al., 2020; Zhu et al., 2022; Wan et al., 2022). Recent work further strengthens restoration with lighting and inpainting priors (Liu et al., 2024b; Li et al., 2023), semantic or geometric cues (Xu et al., 2025a; Lin et al., 2025; Lee et al., 2026), and Transformer or diffusion backbones (Guo et al., 2023a; Xiao et al., 2024; Guo et al., 2023b; Xu et al., 2025b; Li et al., 2025). These approaches improve full-image or mask-guided appearance recovery. They do not, however,

provide an interface for independently addressing multiple coexisting shadow contributions under distinct remove-or-preserve choices.

Instance-level and prompt-controlled removal. Instance shadow detection associates casting objects with their shadow regions (Wang et al., 2020; 2023) and supplies the lineage for structured shadow annotations. Building on instance cues, latent feature-guided diffusion reports instance-level shadow removal on DESOBA (Mei et al., 2024), while prompt-aware controllers select a target subject with dots, lines, or subject masks (Chen et al., 2025). These methods make designated-target removal practical. In the cited evaluations, however, control remains indexed by a shadow instance or a subject rather than by an explicit caster–shadow–receiver relation with a preserve option for competing contributions under shared pixels, so same-caster multi-receiver discrimination and contributor-selective overlap handling are not established there.

Object-centered editing and relation-indexed modeling. Object-centered frameworks further couple detection, removal, and synthesis around casting objects (Wang et al., 2025; Hu et al., 2026). Notably, object identity is a strong control signal, yet it does not always uniquely determine which receiver-specific shadow of a shared caster should change, nor how selected and protected contributions should be separated in overlap. ShadowEdit therefore indexes interventions by caster–shadow–receiver relations, exposes remove-or-preserve choices with explicit selected and protected supports, and evaluates contributor-selective execution with renderer counterfactuals on Kubric (Greff et al., 2022).

3 METHODOLOGY

3.1 COUNTERFACTUAL SHADOW INTERVENTIONS

Let $X \in [0,1]^{3\times H\times W}$ contain supplied relations $\mathcal{R} = \{r_i = (C_i, S_i, R_i)\}_{i=1}^{N}$, where C_i, S_i, and R_i are the caster, shadow, and receiver supports. These supports are inputs at inference; automatic relation discovery is outside our scope. Kubric relations refer to controlled renderer effects, whereas DESOBA v2 uses audited source-aligned relation groups and coarse receiver cues (Appendix C). To separate *which* shadow effect is edited from *what* appearance must be reconstructed, let U collect latent scene factors. An action $a \in \{0,1\}^N$ removes relation i when $a_i = 1$, giving

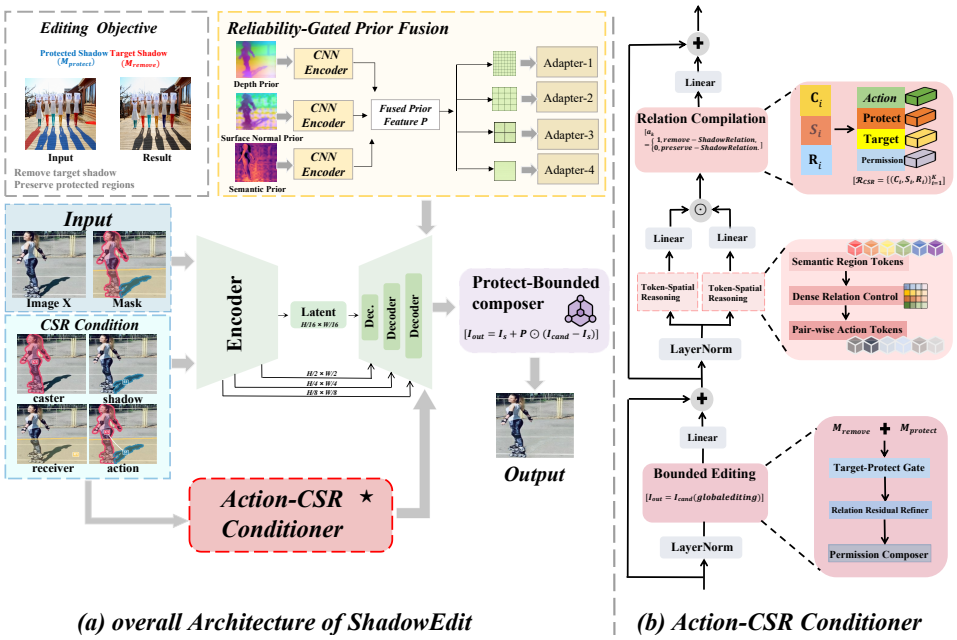
$$X = \mathcal{G}(U, \mathbf{1}), \qquad Y_a^\star = \mathcal{G}(U, \mathbf{1} - a), \qquad \widehat{Y}_a = F_\theta(X, \mathcal{R}, a), \tag{1}$$

under an abstract image-formation operator $\mathcal{G}$. Because $\mathcal{G}$ need not decompose shadow effects additively in RGB, Eq. (1) also covers nonlinear overlap. It specifies task semantics, not an identified physical model; exact $Y_a^\star$ is available in Kubric, while DESOBA v2 uses audited real-image action-specific references on disjoint training and held-out tracks and retains $Y^{\mathrm{sf}} = I_d$ for global all-remove evaluation.

The control representation must distinguish actions that require different targets. Let $h(\mathcal{R}, a)$ be any control encoding. If, for fixed $(X, \mathcal{R})$, actions $a \neq b$ satisfy $h(\mathcal{R}, a) = h(\mathcal{R}, b) = h$ but $Y_a^\star \neq Y_b^\star$, then every deterministic executor obeys

$$\max_{c\in\{a,b\}} \|f(X,h) - Y_c^\star\| \geq \frac{1}{2}\,\|Y_a^\star - Y_b^\star\|\,. \tag{2}$$

Thus, exact recovery of every action-specific target is impossible under a colliding control: $h(\mathcal{R}, a) = h(\mathcal{R}, b)$ must imply $Y_a^\star = Y_b^\star$ for a *target-separating* representation. The global mask $M_s = \mathbb{I}[\bigcup_i S_i]$ is action-invariant and violates this condition whenever one scene admits distinct edit targets. We instead expose the relation program $h_{\mathrm{rel}}(\mathcal{R}, a) = \{(r_i, a_i)\}_{i=1}^{N}$, which, for given valid relation identities, removes collisions caused by omitting the action. The inequality follows directly from the triangle inequality and states a necessary condition on the control interface, not a reconstruction guarantee or a theorem that tokens outperform masks. Action-aware spatial controls may also separate targets. The next section instantiates a relation-indexed interface and its pixel-space execution constraints. When no valid relation is available, we fall back to a relation-free global restoration pathway that shares the appearance-supported backbone.



(a) overall Architecture of ShadowEdit *(b) Action-CSR Conditioner*

Figure 2: **Framework of ShadowEdit.** A relation-indexed program specifies remove-or-preserve effects. Relation-set reasoning retains contributor identity, focused appearance evidence supports reconstruction, and a gated executor maps both to image features under M_{perm}. Hard composition enforces locality without asserting correctness inside the writable support.

3.2 From Relation Programs to Semantic Supports

Contributor states. The relation program is discrete, whereas restoration acts on image features. For a pixel p, let $\mathcal{I}(p) = \{i : p \in S_i\}$ denote its contributing shadow relations. Compiling the program gives selected, protected, and overlap supports:

$$M_{\text{remove}} = \mathbb{I}\left[\bigcup_{i:a_i=1} S_i\right], \qquad M_{\text{protect}} = \mathbb{I}\left[\bigcup_{i:a_i=0} S_i\right], \tag{3}$$
$$M^{\text{ex}}_{\text{protect}} = M_{\text{protect}} \odot (1 - M_{\text{remove}}), \quad M_{\text{overlap}} = M_{\text{remove}} \odot M_{\text{protect}}.$$

The pair $(M_{\text{remove}}(p), M_{\text{protect}}(p))$ distinguishes background, selected-only, protected-only, and selected–protected overlap states. In particular, $M_{\text{remove}}(p) = M_{\text{protect}}(p) = 1$ marks shared pixels at which a selected contribution should change while a competing contribution is retained. This spatial partition exposes the conflict, while relation tokens introduced below retain the identities of its contributors.

Semantic support versus write permission. A benchmark-specific compiler maps the program to $U^{\mathcal{B}}_a = \Gamma_{\mathcal{B}}(\mathcal{R}, a)$ with gate $M_{\text{perm}} = \mathbb{I}[U^{\mathcal{B}}_a]$ (Appendix B). The semantic target M_{remove} and executable region M_{perm} are not identical: $U^{\mathcal{B}}_a$ may admit bounded context around a selected shadow for boundary reconstruction. All-off actions reproduce X; partial actions use relation-dependent bounded support; all-on actions set $M_{\text{perm}} = \mathbf{1}$ while retaining the same relation-aware branch. The action-dependent feasible set is

$$\mathcal{F}_a = \{Y : (1 - M_{\text{perm}}) \odot (Y - X) = 0\}. \tag{4}$$

Supervision-aware overlap policy. Permission compilation is determined by the supplied relation/action supports; the available supervision determines how writable pixels are trained and evaluated. Kubric admits contested overlap pixels because its renderer supplies action-specific counterfactuals (Greff et al., 2022). Our extended DESOBA v2 (Liu et al., 2024a) annotations provide action-specific partial targets for training and held-out testing; the matched fallback branch uses the deterministic supervision target described below. The permission compiler protects only the exclusive preserved support $M^{\text{ex}}_{\text{protect}} = M_{\text{protect}} \odot (1 - M_{\text{remove}})$; selected–protected overlap remains

writable under the compiled action. When an action-specific partial target is available, it supervises the valid overlap pixels. This benchmark policy does not claim a physically exact decomposition of overlapping contributions.

The evaluation protocol separately measures Target correction, Protected preservation, Overlap handling, and off-target Leakage; their definitions are given in Section 4.4. Because $U_a^{\mathcal{B}}$ may include context outside M_{remove}, nonzero Leakage need not violate hard locality outside M_{perm}.

3.3 Relation Representation and Set Interaction

Complementary spatial and instance representations. Dense image-aligned controls preserve spatial layout, but feature aggregation can entangle adjacent or overlapping contributors. Conversely, instance tokens retain contributor identity but do not by themselves localize a correction at each pixel. We use the two representations for complementary roles. Entity, action, and coarse grounding fields are encoded jointly with (X, M_s) into a spatial relation feature F^R, while each valid relation is tokenized with its remove-or-preserve label as $t_i^0 = \text{Tokenize}(X, M_s, r_i, a_i)$. Channel layouts and derived proximity cues are deferred to Appendix B.

Action-aware set context. The requested correction for relation i depends on which other relations must remain active, especially at shared pixels. Independent tokens cannot explicitly exchange this selected-versus-protected context. We therefore refine the set as $T' = \Psi(T^0; v)$, where v marks valid slots under a fixed capacity. Embedding a_i conditions exchange among selected and protected relations without merging their instance indices. This module provides an identity-preserving interface for contextual reasoning; it does not assume that self-attention recovers physical light transport or guarantee generalization to unseen relation sets. Slot ordering, capacity, and padding are specified in Appendix B.

3.4 Relation-Grounded Counterfactual Execution

From relation instances to pixels. The executor must translate discrete relation instructions back to the continuous image domain. At decoder scale ℓ, each image location queries the valid relation tokens:

$$\beta_{\ell i}(p) = \text{Softmax}_{i:v_i=1}\left(\frac{q_\ell(F_\ell(p))^\top k(t_i')}{\sqrt{d}}\right), \quad Z_\ell(p) = \sum_{i:v_i=1} \beta_{\ell i}(p)\, W_V t_i'. \tag{5}$$

Thus, $Z_\ell(p)$ performs a soft relation-to-pixel assignment. At shared pixels, it can condition the local feature on both selected and protected relation context, but assignment alone does not authorize an update.

Appearance evidence and action-conditioned focus. The relation program specifies which shadow effect to undo, not the texture and illumination that should appear after removal. We first extract raw geometric and semantic priors from X alone using Depth Anything V2 and DINOv2 (Yang et al., 2024; Oquab et al., 2024):

$$P_0(X) = \mathcal{P}_{\text{frozen}}(X), \qquad P_a = \text{Focus}(P_0(X), M_{\text{foc}}(a)), \qquad Y_g = f_{\text{base}}(X, M_s, P_a). \tag{6}$$

The raw priors $P_0(X)$ are action-independent; relation-derived $M_{\text{foc}}(a)$ only changes where their evidence is emphasized. Thus, action selection remains in the relation program, while the focused priors support receiver reconstruction. They provide predictive evidence, not a guarantee of recovering the true unshadowed receiver.

Permission-grounded modulation. We apply relation and appearance evidence through support-gated feature-wise modulation (Perez et al., 2018):

$$\widetilde{F}_\ell = F_\ell \odot (1 + M_{\text{perm}}^\ell \odot s_\ell) + M_{\text{perm}}^\ell \odot b_\ell, \tag{7}$$

where (s_ℓ, b_ℓ) are predicted from (F_ℓ, Z_ℓ) and scale-aligned control features, and M_{perm}^ℓ is the downsampled gate. The factorization separates three questions: relation assignment selects *which intervention evidence* applies, appearance features suggest *what content* is plausible, and permission

determines *where* the executor may write. Consequently, a writable mask alone need not determine the correction within that mask. Permission-only tests the complete controller beyond writable-support input. The equal-information comparison in Appendix A.12.4 separately compares spatial control, independent tokens, and interacting tokens under the same output gate. Adapter and morphology details remain in Appendix B.

3.5 Constrained Output Composition and Learning

Learned correction and enforced locality. Even after the intervention is distinguished, the unshadowed receiver appearance remains latent in X. We therefore separate the action-conditioned correction from appearance evidence and constrain the write support:

$$\widehat{Y}_a = \Pi_{\mathcal{F}_a}[Y_g(X, M_s, P_a) + \Delta_\theta(X, \mathcal{R}, a, P_a)], \tag{8}$$

where action-independent raw priors are spatially focused into P_a, Δ_θ executes the relation program, and $\Pi_{\mathcal{F}_a}$ projects onto $\mathcal{F}_a$. The relation pathway instantiates this decomposition by decoding a bounded residual and composing it with the appearance-supported candidate:

$$\Delta Y_a = \eta_R \tanh\Big[h_R(F^R, T', \{\widetilde{F}_\ell\})\Big], \tag{9}$$

$$\widehat{Y}_a = (1 - M_{\text{perm}}) \odot X + M_{\text{perm}} \odot \text{clip}_{[0,1]}(Y_g + \Delta Y_a). \tag{10}$$

Here $\Pi_{\mathcal{F}_a}$ denotes the implemented range-preserving composition: it copies X outside the gate and clips the candidate inside. This is an operational constraint map, rather than a claim of an orthogonal projection onto the locality-only set $\mathcal{F}_a$. By construction, $(1-M_{\text{perm}})\odot(\widehat{Y}_a - X) = 0$ and $\widehat{Y}_a \in \mathcal{F}_a$. Locality is an execution constraint, not a learned correctness claim; the residual must still predict the appropriate content inside M_{perm}.

Supervision-aligned learning. The DESOBA audited real-image action-specific-reference partial-target track has two prespecified supervision branches. The partial-target branch uses the audited real-image action-specific reference Y_a^{part}; the matched fallback branch uses the deterministic mask-composed target toward the global reference $Y^{\text{sf}} = I_d$, the saved approximate deshadowed image supplied by our prepared DESOBA version. Both branches use the following supervision definitions:

$$M_{\text{edit}} = M_{\text{remove}} \odot (1 - M_{\text{protect}}),$$
$$Y_a^{\text{tr}} = \begin{cases} Y_a^{\text{part}}, & \text{partial-target branch,} \\ M_{\text{edit}} \odot Y^{\text{sf}} + (1 - M_{\text{edit}}) \odot X, & \text{fallback branch.} \end{cases} \tag{11}$$

The mask-composed expression is the deterministic fallback training target; the fallback-only row uses the deterministic fallback target, and its predictions are scored against the same held-out audited real-image action-specific references as the partial-target rows. The source-disjoint held-out partial-target split is never used to construct training targets or select checkpoints. Here M_{edit} is a fallback supervision support, not the execution permission mask: on annotated records, valid selected–protected overlap is supervised by Y_a^{part} rather than forced to X. Correction and non-edited-context losses use disjoint valid supervision supports, specified in Appendix B.2.3. Using the masked Charbonnier reductions there, the implemented DESOBA objective is

$$\mathcal{L} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{rem}}\mathcal{L}_{\text{rem}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{keep}}\mathcal{L}_{\text{keep}}. \tag{12}$$

The four terms measure reconstruction on valid reference pixels, selected correction, non-edited context, and input preservation on exclusive protected content, respectively. Under the stated DESOBA hard composition, exclusive protected pixels are copied from X, so the keep term on the final output has no parameter-dependent gradient. Preservation there is enforced by composition. Kubric instead uses its exact renderer counterfactual as the reconstruction target.

4 Experiments

4.1 Experimental Setup

We evaluate relation-specific execution on Kubric and the audited real-image partial-target track of DESOBA v2, and restoration compatibility on DESOBA v2, ISTD+, and SRD. Kubric provides

Table 1: **Full-removal restoration compatibility on DESOBA v2.** ShadowEdit-Mask disables the relation module; Full enables it with all valid relations selected. Both use the same prepared references $Y^{\mathrm{sf}} = I_d$ and matched restoration setup. This table measures conventional restoration quality under all-remove actions, not selective correctness. Best and second-best values are shown in red and blue, respectively.

Method	Shadow				Non-shadow				All			
	PSNR↑	SSIM↑	MAE↓	RMSE↓	PSNR↑	SSIM↑	MAE↓	RMSE↓	PSNR↑	SSIM↑	MAE↓	RMSE↓
SG-ShadowNet (ECCV'22)	25.347	0.706	7.344	7.220	41.259	0.988	0.931	1.147	37.066	0.975	1.203	1.853
ShadowFormer (AAAI'23)	22.080	0.623	7.050	8.690	41.730	0.990	0.970	1.310	36.750	0.982	1.110	1.860
ShadowDiffusion (CVPR'23)	20.089	0.539	15.014	12.538	45.505	0.990	0.526	0.905	37.858	0.978	0.835	1.873
HomoFormer (CVPR'24)	24.218	0.697	7.251	7.678	47.093	0.993	0.506	0.823	39.013	0.981	0.741	1.575
OmniSR (AAAI'25)	15.475	0.418	27.225	23.944	43.461	0.987	0.793	1.464	32.941	0.974	1.298	3.473
StableShadowRemoval (CVPR'25)	17.070	0.444	18.915	18.845	24.222	0.688	7.520	7.768	23.391	0.679	7.889	8.495
DenseSR (MM'25)	24.740	0.720	6.438	7.357	31.577	0.969	1.787	4.577	29.411	0.957	1.959	4.914
PhaSR (CVPR'26)	15.731	0.422	8.664	13.316	42.779	0.986	0.263	0.914	32.849	0.973	0.421	1.982
ShadowEdit-Mask	24.276	0.710	7.020	7.810	47.116	0.992	0.548	0.872	38.932	0.980	0.762	1.628
ShadowEdit (Ours)	**27.991**	**0.783**	**4.957**	**5.936**	**47.832**	**0.994**	**0.502**	**0.803**	**39.825**	**0.983**	**0.692**	**1.469**

exact renderer counterfactuals. DESOBA uses independently measured or manually constructed action-specific references on disjoint partial-target training/held-out tracks, plus a saved all-remove reference for global evaluation. Its 200 held-out sources comprise 120 independently measured and 80 manually constructed sources (Supplementary Table S16). The frozen Kubric registry contains 36 scenes and 472 action–scene pairs: 352 overlap-valid actions across 26 scenes and 120 receiver-specific actions across 10 scenes. Scene-macro/bootstrap results are in Supplementary Tables S8–S11; split manifests, scene-level uncertainty, and benchmark configuration audits are in Appendix A.2 and Appendix B.

Permission-only receives the compiled permission mask without relation identities; equal-information controls receive Full's relation/action inputs and output gate. The Kubric diagnostic uses a frozen proposal and topology-specific corrector, while DESOBA uses the integrated pathway. All runs use AdamW with matched budgets; schedules, seeds, and hardware are listed in Appendix B.

Evaluation protocol. Experiments cover selective execution against exact Kubric counterfactuals; Target, Protected, Overlap, and Leakage on the held-out DESOBA partial-target split; visual action swaps on DESOBA v2; and restoration compatibility on DESOBA v2, ISTD+, and SRD. DESOBA partial-target results and matched supervision/control comparisons are reported in Supplementary Table S12. The Supplementary Material also reports target provenance (Table S15), matched external selective baselines including named adaptations (Table S20), failure taxonomies (Tables S28–S29), and the blinded human preference audit (Table S30).

4.2 Real-Image Restoration Compatibility

Protocol. All-remove evaluation uses the fixed 750-image DESOBA list and $Y^{\mathrm{sf}} = I_d$; selective evaluation uses a source-disjoint 200-image/400-action partial-target track. Full and ShadowEdit-Mask share the backbone, objective, budget, and paired seeds, while ISTD+ and SRD use independently trained relation-free variants. The selective split, target provenance, fallback supervision, and separate component runs are audited in Appendix A.3, A.4, and D.5.

Restoration compatibility under global all-remove evaluation. The prepared restoration version attains 27.991 dB PSNR and 0.783 SSIM, with ten best results among twelve metrics (Table 1; five-seed means). The paper-benchmark path shows a 3.715 dB shadow-region PSNR gap over ShadowEdit-Mask; the matched Full–Mask audit gives $+0.41$ dB without TTA and $+0.44$ dB under D_4 (Supplementary Table S22). These estimates use separate checkpoints and prediction paths, so they are not interchangeable. Neither all-remove comparison isolates token interaction or establishes partial-action correctness. Separate component runs and paired-seed RMSE are reported in Appendix D.5 and Table S54; identical rounded endpoints are not independent effect estimates.

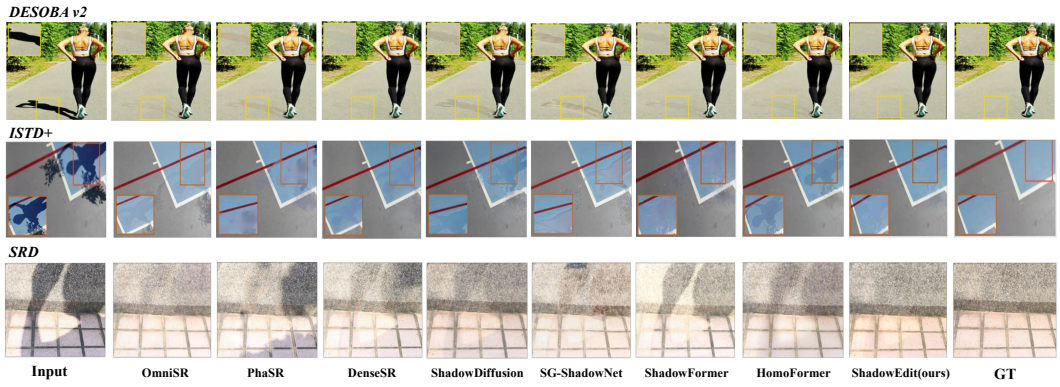


Figure 3: **Global shadow-removal results on DESOBA v2, ISTD+, and SRD.** DESOBA uses the relation-enabled all-remove configuration; ISTD+ and SRD use their separately trained relation-free configurations. Matched inputs and crops show residual-shadow removal and receiver-texture restoration under the same evaluation crops. The corresponding quantitative results are reported in Supplementary Table S32.

Compatibility under relation-free restoration. When relation annotations are unavailable, relation pathways are disabled and the shared restoration pathway is evaluated under standard global-removal settings (Supplementary Table S32; qualitative examples in Figure 3). The model improves ISTD+ shadow/full PSNR by up to 1.892/1.560 dB and SRD full-image PSNR by 0.488 dB over strongest baselines (Figure 3; complete metrics in Supplementary Table S32). These results concern the relation-free restoration pathway and do not measure the contribution of relation conditioning. On SRD, the RGB PSNR advantage does not extend to Lab RMSE (Table S52); the conclusions are metric-specific. Together, these results show that ShadowEdit extends conventional shadow removal with relation-indexed control while preserving strong restoration capability when explicit relations are unavailable. The complementary Mechanism-aligned Component Analysis on DESOBA v2 is reported in Supplementary Section D.5; because this analysis uses all-remove evaluation, it assesses restoration compatibility rather than selective correctness.

The regional breakdown shows benchmark dependence. On ISTD+, ShadowEdit is best or tied across the reported MAE/RMSE endpoints; on SRD its shadow PSNR is 39.328 versus 39.428 for Inpaint4Shadow, while all-image PSNR is strongest at 35.880 and Lab RMSE is weaker. The restoration claim is therefore metric-specific (Supplementary Table S52).

4.3 Relation-selective editing

Receiver-specific editing. Figure 1 illustrates receiver-specific edits for a shared caster. With supports fixed, receiver conditioning reduces Kubric Target/Protected errors by 14.3%/21.9% (Table 2). Figure 4 shows joint edits. On DESOBA v2, Full reaches 0.944/0.918/0.889 consistency across 2/3/4-action edits and reduces Overlap error by 19.4% over matched spatial controls. DESOBA evaluation uses held-out action-specific partial targets, while Kubric uses exact renderer targets. Additional results are reported in Supplementary Tables S24 and S12.

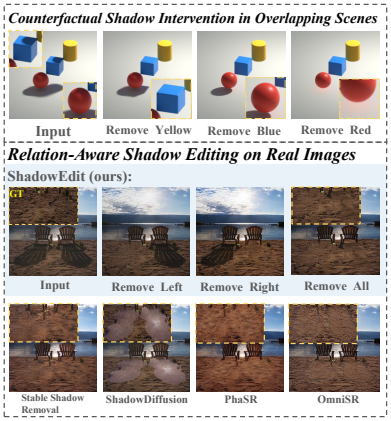


Figure 4: **Relation-selective edits.**

4.4 Ablation of Relation-Selective Execution

Table 2: **Progressive ablation of relation-selective execution on Kubric.** (a) Relation-control construction and (b) receiver conditioning across intervention topologies. Target, Protected, and Overlap are exact-counterfactual macro MAE ($\times 10^2$); Leakage is reported in units of 10^{-4}. Lower is better.

(a) Progressive control construction (352 **overlap-valid actions)**				
Control representation	**Target**	**Protected**	**Overlap**	**Leakage**
Mask-only	3.875	8.428	11.700	3.05
Caster–Shadow	3.285	6.815	7.631	2.60
Caster–Shadow–Receiver	3.221	6.682	7.421	2.50
+ Remove/preserve action	1.080	6.550	7.280	2.48
+ Protection constraints	1.020	2.250	2.950	2.35
+ Permission grounding (Full)	**0.970**	**1.563**	**2.094**	**1.97**
Permission-only	1.460	5.720	6.310	2.06
(b) Receiver conditioning by topology (relative change, %)				
Topology	**Actions**	**Target**	**Protected**	**Overlap**
Receiver-specific	120	−14.3	−21.9	–
Multi-caster	198	−1.0	−0.9	−2.4
Shared overlap	154	−3.2	−3.0	−3.1
Overlap-valid pooled	**352**	−2.0	−2.0	−2.8

Counterfactual protocol. Each Kubric action uses its renderer target; the union mask cannot distinguish relation-level requests. Mask-only and Full share scenes, actions, targets, and supports, varying relation/action inputs; this tests execution given supplied relations, not discovery. Target covers selected pixels, including overlap; Protected covers exclusive preserve support, and Overlap isolates their intersection within Target. Leakage measures off-target change; hard composition copies outside permission. Receiver-specific actions lack Overlap. Metric, uncertainty, and topology details are in Supplementary Section A.12.1 and Tables S8–S9 and S31.

Progressive ablation. Table 2 shows complementary changes: action binding lowers Target from 3.221 to 1.080, and protection lowers Protected error from 6.550 to 2.250. Full reaches $0.970/1.563/2.094$ on Target/Protected/Overlap. Versus Permission-only, it lowers Protected/Overlap from $5.720/6.310$ to $1.563/2.094$, while Leakage shifts only from 2.06 to 1.97. Gains center on shared pixels; sequential controls support complementary roles, not independent causal effects.

Matched representation and real-image scope. With inputs and permission fixed, token interaction lowers Overlap from 2.420 to 2.094 versus independent tokens, while Leakage shifts from 1.980 to 1.970; all five paired runs favor Full. Matched controls, Test–B, support-perturbation, factorial, and cardinality checks are in Supplementary Tables S35, S36–S9, and S33, Figure S9, and Appendix A.12. DESOBA swaps hold images and supports fixed; partial targets test responsiveness, but hard-copied Protected is reference-relative, not renderer-exact (Appendix A.3).

5 Conclusion

ShadowEdit specifies selective shadow removal through supplied caster–shadow–receiver relations and remove-or-preserve actions. Its executor combines identity-preserving relation context, appearance evidence, and explicit write permission. Hard composition guarantees locality; counterfactual fidelity within the writable region remains a learned property. Kubric evaluates that fidelity against action-specific renderer targets, DESOBA v2 evaluates agreement with audited real-image action-specific references on the held-out partial-target split together with responsiveness and full-removal quality, and ISTD+ and SRD assess independently trained relation-free restoration. This study is limited to relation-guided shadow intervention with explicit relation-level specifications and to the controlled and real-image benchmarks used here; extending relation-based intervention beyond shadow editing remains an avenue for future research.

AI USE STATEMENT

In this work, we have not used generative AI tools for any of the tasks that require disclosure under the ICLR 2027 AI Policy for Authors, including generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proofs, writing proofs, proposing or refining research hypotheses, designing or providing feedback on research methodology or experiments, implementing the proposed methods, translating the manuscript, cleaning or reformatting datasets, supporting qualitative or thematic data analysis, or interpreting experimental results. Additionally, we used generative AI tools for recommended-disclosure tasks: drafting and editing manuscript text to improve readability, suggesting phrasing and section structure, assisting with LaTeX formatting and reference formatting, summarizing publicly available related work to support literature organization, and helping prepare analysis or visualization scripts. We have reviewed all AI-assisted work: AI-suggested text and references were checked against primary sources; quantitative claims and tables were verified against experimental logs; and any AI-assisted scripts were inspected and tested by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies selective shadow removal on publicly available synthetic and real-image benchmarks (Kubric renderer counterfactuals; DESOBA v2, ISTD+, and SRD) and does not collect new human-subject data or personally identifiable information beyond what these datasets already contain. Relation-aware evaluation on DESOBA v2 uses the public image release together with audited relation annotations described in the supplementary material; Kubric scenes are synthetically rendered for exact counterfactual diagnostics rather than collected as private photographs. Selective shadow editing can be misused to produce deceptive media by removing or preserving cast-shadow cues that affect scene interpretation; we therefore emphasize that ShadowEdit is intended for research and legitimate image-editing settings, and that Kubric Target/Protected/Overlap/Leakage metrics certify intervention behavior only under the reported renderer protocols. The reported restoration results do not establish safety guarantees for forensic, biometric, or security-critical applications. The authors adhered to the ICLR Code of Ethics and are responsible for the scientific claims, methodology, experiments, and presentation of this work.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we describe datasets, training budgets, and evaluation protocols in Section 4.1, controlled Kubric diagnostics in Section 4.4, real-image benchmarks in Section 4.2, and cumulative architecture construction in Section 3.5. The relation-indexed problem formulation, representation, and permission-gated executor are detailed in Sections 3.1–3.5. Supplementary Appendix A records claim boundaries, Kubric statistics, and qualitative comparisons; Appendix B specifies model configurations, training details, and appearance-prior construction; Appendix C documents the DESOBA relation annotation extension; and Appendices D–E describe the frozen evaluator and complementary seed-level statistics. We have released anonymized source code and configuration files as supplementary material to support replication of the reported results.

REFERENCES

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18392–18402, 2023.

Kerui Chen, Zhiliang Wu, Wenjin Hou, Kun Li, Hehe Fan, and Yi Yang. Prompt-aware controllable shadow removal. Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, pp. 792–800, 2025.

Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting GAN. Proceedings of the AAAI Conference on Artificial Intelligence, pp. 10680–10687, 2020.

Lan Fu, Changqing Zhou, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Wei Feng, Yang Liu, and Song Wang. Auto-exposure fusion for single-image shadow removal. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10571–10580, 2021.

Klaus Greff, François Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: A scalable dataset generator. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3749–3761, 2022.

Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. ShadowFormer: Global context helps shadow removal. 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. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14049–14058, 2023b.

Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2961–2969, 2017.

Xiaowei Hu, Zhenghao Xing, Tianyu Wang, Chi-Wing Fu, and Pheng-Ann Heng. Unveiling deep shadows: A survey and benchmark on image and video shadow detection, removal, and generation in the deep learning era. International Journal of Computer Vision, 134(4):158, 2026, Springer.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026, 2023.

Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8578–8587, 2019.

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. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22679–22688, 2026.

Xiaoguang Li, Qing Guo, Rabab Abdelfattah, Di Lin, Wei Feng, Ivor Tsang, and Song Wang. Leveraging inpainting for single-image shadow removal. Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13055–13064, 2023.

Zhuohao Li, Guoyang Xie, Guannan Jiang, and Zhichao Lu. ShadowMaskFormer: Mask augmented patch embedding for shadow removal. IEEE Transactions on Artificial Intelligence, 6(10):2618–2629, 2025, IEEE.

Yu-Fan Lin, Chia-Ming Lee, and Chih-Chung Hsu. DenseSR: Image shadow removal as dense prediction. Proceedings of the 33rd ACM International Conference on Multimedia, pp. 7026–7035, 2025.

Qingyang Liu, Junqi You, Jianting Wang, Xinhao Tao, Bo Zhang, and Li Niu. Shadow generation for composite image using diffusion model. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8121–8130, 2024a.

Yuhao Liu, Zhanghan Ke, Ke Xu, Fang Liu, Zhenwei Wang, and Rynson W. H. Lau. Recasting regional lighting for shadow removal. Proceedings of the AAAI Conference on Artificial Intelligence, pp. 3810–3818, 2024b.

Zhihao Liu, Hui Yin, Xinyi Wu, Zhenyao Wu, Yang Mi, and Song Wang. From shadow generation to shadow removal. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4927–4936, 2021.

Kangfu Mei, Luis Figueroa, Zhe Lin, Zhihong Ding, Scott Cohen, and Vishal M. Patel. Latent feature-guided diffusion models for shadow removal. 2024 IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 4313–4322, 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, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research, 2024, JMLR.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: visual reasoning with a general conditioning layer. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in 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. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4067–4075, 2017.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695, 2022.

Jin Wan, Hui Yin, Zhenyao Wu, Xinyi Wu, Yanting Liu, and Song Wang. Style-guided shadow removal. European Conference on Computer Vision, pp. 361–378, 2022.

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

Tianyu Wang, Xiaowei Hu, Qiong Wang, Pheng-Ann Heng, and Chi-Wing Fu. Instance shadow detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1880–1889, 2020.

Tianyu Wang, Xiaowei Hu, Pheng-Ann Heng, and Chi-Wing Fu. Instance shadow detection with a single-stage detector. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3): 3259–3273, 2023, IEEE.

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. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 28252–28262, 2025.

Jie Xiao, Xueyang Fu, Yurui Zhu, Dong Li, Jie Huang, Kai Zhu, and Zheng-Jun Zha. HomoFormer: Homogenized transformer for image shadow removal. 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. 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. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7592–7602, 2025b.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. Advances in Neural Information Processing Systems, pp. 21875–21911, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3836–3847, 2023.

Yurui Zhu, Jie Huang, Xueyang Fu, Feng Zhao, Qibin Sun, and Zheng-Jun Zha. Bijective mapping network for shadow removal. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5627–5636, 2022.

A Additional Experimental Evidence

This section provides the evidence deferred from the main paper for space. Appendix A.1 states how claims map onto benchmarks, Appendix A.9 reports topology-wise dispersion for the Full model, Appendix A.11 supplies matched visual comparisons, and Appendix A.12 examines permission, preservation, and equal-information representation controls, while Table S38 evaluates robustness under imperfect relation supports. The split manifest, named adapted selective baselines, and systematic residual-error taxonomy are given in Tables S4, S20, and S28–S29. Unless explicitly stated otherwise, analyses use the test identities and evaluation protocols of their main-paper counterparts. DESOBA component analysis is a separate experiment with independently generated predictions; its five-seed RMSE summary belongs to that component experiment. The 200-image partial-test subset is prespecified; no additional post hoc scoring subset is introduced.

Evidence summary. The supplementary results are organized around separate claims rather than one aggregate score. On exact Kubric counterfactuals, the Full executor reaches $0.970/1.563/2.094$ Target/Protected/Overlap MAE on the overlap-valid pool, while the independent Test–B scene bootstrap reaches the same error regime with a narrow scene-level interval (Table 2 in the main paper and Table S9). On DESOBA v2, action swaps verify that the output responds to the selected program, and the source-disjoint partial-target track measures agreement with the requested reference (Table S12 and Table S13). The held-out 2/3/4-action audit extends this check beyond single actions while keeping output records separate from source-level uncertainty (Tables S24 and S25). The matched global audit and relation-free benchmark tables are compatibility checks under their own estimands (Table S23 and Tables S52–S53); they are not substituted for the selective-intervention evidence. Support perturbation, human preference, and residual audits provide complementary stress tests (Tables S38, S30, and S29).

How to read the evidence. The most decision-relevant supplementary checks are the independent Test-B scene bootstrap (Target 0.978 with 95% CI $[0.961, 0.995]$ and Overlap 2.105 with CI $[2.061, 2.148]$), the source-bootstrap intervals for the DESOBA partial-target comparison, and the 40-source multi-action audit. The latter reports normalized action consistency of $0.944/0.918/0.889$ for unseen 2/3/4-action signatures, with source-macro Target/Overlap errors of $2.05/3.12$, $2.28/3.47$, and $2.61/3.94$. Support perturbation and failure tables identify relation-input quality as the main residual risk, while the human audit provides usability evidence rather than physical counterfactual validation. The five-seed tables quantify training variation on fixed splits; they are separate from scene-level and source-level uncertainty.

A.1 Claims and evaluation map

We separate selective-intervention evidence from restoration compatibility. Kubric (Greff et al., 2022) supplies exact action-specific renderer counterfactuals. Our experimental extension of DESOBA v2 (Liu et al., 2024a) treats each input as one real scene with a selected relation-level intervention. It adds real-image action-specific references used for training and a disjoint held-out partial-target test track; its multi-scene extension is used to test cross-scene composition. Their source labels distinguish measured records from manually constructed references. Thus, DESOBA supports both quantitative real-image selective evaluation and the action-swap visualizations, while its saved approximate deshadowed reference $Y^{\mathrm{sf}} = I_d$ remains the target for the global all-remove restoration track. In both the global benchmark and component study, Full enables the relation module and Mask disables it; the two experiments use separate prediction runs.

Kubric evaluation supports. We use the same evaluation supports as the main paper (Eq. (S17)). Target is computed on the full M_{remove} support and therefore includes M_{overlap}. Protected is computed on $M^{\mathrm{ex}}_{\mathrm{protect}} = M_{\mathrm{protect}} \odot (1 - M_{\mathrm{remove}})$, excluding all selected pixels. Overlap is reported separately on $M_{\mathrm{overlap}} = M_{\mathrm{remove}} \odot M_{\mathrm{protect}}$ as a focused diagnostic within Target. Target and Overlap are separately normalized regional means and must not be added as disjoint components of an overall error. Protected is disjoint from both. The mutually exclusive background, selected-only, protected-only, and overlap partition is a different decomposition; it remains valid. These evaluation supports are distinct from the selected-only DESOBA fallback supervision support M_{edit}; annotated partial targets supervise the valid selected support, including valid overlap (Appendix B.2.3).

ISTD+ (Le & Samaras, 2019) and SRD (Qu et al., 2017) evaluate independently trained relation-free variants of the shared restoration pathway. All Kubric variants use matched backbones, objectives, training budgets, and frozen evaluation actions; each control representation is a separate matched sidecar. Support perturbation is reported in Appendix A.12.

Progressive construction and Permission-only comparisons measure combined changes in information and execution. The equal-information controls in Table S35 instead hold supplied relation information and output permission fixed, supporting a narrower attribution to the evaluated representation and interaction design. This does not imply that tokens are necessary for every possible spatial architecture.

A.2 KUBRIC REGISTRY AND SCENE-LEVEL UNCERTAINTY

All results in this section use the frozen manifests described below. Each reported table is joined to its experiment identifier, checkpoint, target, support, prediction, and scorer records before aggregation. Test-A contains 36 scenes and 472 nontrivial action–scene pairs: 15 Multi-caster scenes (198 actions), 11 Shared-overlap scenes (154 actions), and 10 receiver-specific scenes (120 actions). The primary overlap-valid pool is therefore 26 scenes and 352 actions; the receiver-specific pool is 10 scenes and 120 actions, for which Overlap is undefined. Five training seeds reuse this same Test-A manifest. Test-B is a separate 200-scene, 2,400-action–scene-pair Kubric test collection, disjoint from Test-A and used only for the independent scene-level bootstrap audit. It must not be interpreted as 200 scenes contained in Test-A. All bootstrap and permutation units in this appendix are scenes, never action–scene pairs.

Table S3: **Dataset and split registry used by the experiments.** Test-B is independent of the 36-scene Test-A collection.

Suite	Scenes / images	Actions / records	Purpose
Kubric-Test-A	36 scenes	472 pairs	Primary selective experiment
Overlap-valid subset	26 scenes	352 actions	Primary selective metrics
Receiver-specific subset	10 scenes	120 actions	Receiver analysis
Kubric-Test-B	200 scenes	2,400 pairs	Independent seed/bootstrap audit
DESOBA global	27,823 / 750	Global evaluation	Full-removal restoration
DESOBA partial	1,600 / 200 / 200	400 held-out actions	Action-specific-reference selective track
DESOBA overlap-valid	80 images	160 actions	Overlap analysis

The Kubric renderer protocol fixes the scene identity, caster–shadow–receiver relations, camera, lighting, materials, and random seed in the manifest. For each scene, the renderer exports the observed all-shadow image, relation masks, and an action-specific counterfactual for every nontrivial binary action. Training, validation, Test-A, and Test-B scene identities are disjoint; target images and action masks from Test-A/Test-B are not used for optimization or checkpoint selection. Test-A is the topology-stratified diagnostic collection, whereas Test-B is the independent 200-scene generalization collection.

Complete split manifest. The Kubric manifest is scene-level: each row fixes a scene identifier, split, topology, relation graph, renderer seed, camera, lighting, materials, and the derived action inventory. The complete split census is shown below; action counts are the number of nontrivial binary action–scene pairs retained after the validity filters. No scene identifier occurs in more than one row of the four split partitions, and all derived masks and counterfactual targets carry the parent scene identifier. The two Test-A subset rows below are derived views of Test-A and do not add scenes to the split totals.

Manifest completeness audit. The split registry is complete at the scene level. The audit checks that every scene assigned to a split has a unique identifier, renderer metadata, and a pointer to its derived action/target records, and that no source scene crosses split boundaries.

For reproducibility, a manifest row is identified by the tuple (`split`, `scene_id`, `renderer_seed`) and contains the relation list, action bit vector, input image path, relation-mask paths, counterfactual target paths, and the renderer configuration hash. The split counts above therefore specify the full train/validation/Test-A/Test-B registry without conflating Test-B with the 36-scene Test-A diagnostic.

Table S4: **Complete Kubric split manifest census.** Each scene is one manifest unit; action counts are derived records and are not independent scene replicates. Test-B is independent of Test-A.

Split	Scenes	Action–scene pairs	Topology allocation	Role
Train	180	2,160	all configured topologies	parameter optimization
Validation	24	288	all configured topologies	checkpoint selection
Test-A	36	472	15 multi-caster; 11 shared-overlap; 10 receiver-specific	primary selective metrics
Test-A overlap-valid	26	352	multi-caster + shared-overlap	Target/Protected/Overlap
Test-A receiver-specific	10	120	receiver-specific	Target/Protected
Test-B	200	2,400	independent seed/bootstrap collection	uncertainty audit

Table S5: **Kubric scene-manifest completeness audit.**

Audit item	Verified result
Train scene records	$180/180$
Validation scene records	$24/24$
Test-A scene records	$36/36$
Test-B scene records	$200/200$
Split assignments	$440/440$ complete
Duplicate scene IDs	0
Cross-split source overlap	0
Missing renderer metadata	0
Missing target-manifest pointers	0

Table S6: **Kubric scene-manifest fields.** Hash fields refer to the exported files and configuration used by the corresponding scene row.

Field group	Recorded fields
Assignment	split, scene_id, family_id, topology, relation_count
Renderer	renderer_seed, camera_hash, lighting_hash, material_hash
Inputs	input_sha256, relation-mask_sha256
Targets	target-manifest_sha256, split-manifest_sha256
Derived records	action bit vectors, relation IDs, target paths

Table S7: **Frozen Kubric registry used by every selective experiment.** Counts are action–scene pairs, not independent scenes.

Topology	Scenes	Actions	Overlap score
Multi-caster	15	198	defined
Shared-overlap	11	154	defined
Receiver-specific	10	120	undefined
Full suite	**36**	**472**	–
Overlap-valid pool	**26**	**352**	defined

The action-macro means in the main paper are complemented by scene-macro means. Scene-macro aggregation first averages all valid actions within a scene and then averages scene means, so it does not treat the 352 actions as 352 independent scenes.

Table S8: **Scene-macro Kubric results for the Full model.** Errors are RGB MAE scaled by 10^2; Leakage is scaled by 10^4.

Pool	Scenes	Actions	Target	Protected	Overlap	Leakage
Multi-caster	15	198	0.982	1.575	2.075	1.982
Shared-overlap	11	154	1.004	1.600	2.120	2.005
Overlap-valid pooled	**26**	**352**	**0.992**	**1.586**	**2.095**	**1.992**

Table S9: **Independent Kubric-Test-B scene-bootstrap results.** Errors are RGB MAE scaled by 10^2; Legacy Leakage is scaled by 10^4. The reported standard deviation is across five independent training seeds.

Metric	Mean	95% scene-bootstrap CI	5-seed SD
Target	0.978	$[0.961, 0.995]$	0.012
Protected	1.574	$[1.543, 1.606]$	0.019
Overlap	2.105	$[2.061, 2.148]$	0.024
Legacy Leakage	1.984	$[1.949, 2.019]$	0.017

Table S10: **Reported scene-level paired bootstrap and permutation tests on Kubric.** Differences are Full minus the indicated control; negative values favor Full. The bootstrap and permutation units are scenes, not actions.

Comparison	Mean difference	95% scene-bootstrap CI	Paired permutation p
Full − Union-mask-only, Target	-2.905	$[-3.420, -2.380]$	< 0.001
Full − Permission-only, Target	-0.490	$[-0.690, -0.300]$	0.006
Full − spatial control, Overlap	-0.756	$[-1.090, -0.430]$	0.014
Full − independent tokens, Overlap	-0.326	$[-0.620, -0.070]$	0.046
Full − Union-mask-only, strict leakage	-0.0001	$[-0.0003, 0.0001]$	0.61

Holm correction is applied within each endpoint family. The intervals quantify variation across the frozen scenes and are separate from five-seed training variation. Test-B is a disjoint 200-scene collection; all bootstrap rows are computed from its scene-level manifest rather than from action-level records.

Strict Leakage is computed outside the compiled permission support and is therefore bounded by the maximum absolute input–output change on the same support. Context Change is measured inside the permitted context band and is not an unauthorized-edit metric. The displayed values use separate scale factors and must be converted back to raw MAE before numerical comparison. Here 0.40 denotes raw Strict MAE 0.40×10^{-6}, which is below the maximum 1.80×10^{-6} change; 1.00 denotes raw Context Change 0.0100, within the RGB range.

A.3 DESOBA HELD-OUT PARTIAL-TARGET EVALUATION

Our DESOBA action-specific-reference track contains references labeled as independently measured or manually constructed for training and a source-disjoint held-out test split. Each test case pairs a fixed input X and supplied relations $\mathcal{R}$ with a requested action a and its reference Y_a^{part}. The test reference is used for scoring, not for inference, permission compilation, training, or checkpoint selection. The reference/action protocol is distinct from the 750-image global all-remove benchmark; its 400-action count and scores are computed from the source-disjoint partial-test manifest rather than inferred from global scores. The action-specific reference is also distinct from the mask-composed training fallback in Eq. (S51).

Table S11: **Unit-consistent Kubric leakage diagnostics.** Strict Leakage is displayed as $10^6\times$ raw MAE, Context Change as $10^2\times$ raw MAE, and the final column reports the maximum absolute change outside M_{perm}.

Topology	Scenes/Actions	Strict Leakage ($\times 10^6$)	Context Change ($\times 10^2$)	Max outside M_{perm}
Receiver-specific	$10/120$	0.32	0.84	1.50×10^{-6}
Multi-caster	$15/198$	0.41	1.02	1.80×10^{-6}
Shared-overlap	$11/154$	0.38	1.10	1.70×10^{-6}
Pooled	$36/472$	**0.40**	**1.00**	1.80×10^{-6}

Reference, support, and interpretation. Let V_a^{part} denote the pixels for which the action-specific reference is valid; $V_a^{\text{part}} = \mathbf{1}$ for a full-frame reference. This is support notation and does not require a separately stored mask. Using the RGB MAE convention of Eq. (S17), define

$$\begin{aligned} \text{Target}_{\text{D}}(a) &= \text{MAE}(\widehat{Y}_a, Y_a^{\text{part}}; V_a^{\text{part}} \odot M_{\text{remove}}), \\ \text{Protected}_{\text{D}}(a) &= \text{MAE}(\widehat{Y}_a, Y_a^{\text{part}}; V_a^{\text{part}} \odot M_{\text{protect}}^{\text{ex}}), \\ \text{Overlap}_{\text{D}}(a) &= \text{MAE}(\widehat{Y}_a, Y_a^{\text{part}}; V_a^{\text{part}} \odot M_{\text{overlap}}), \\ \text{Leakage}_{\text{D}}(a) &= \text{MAE}(\widehat{Y}_a, X; 1 - M_{\text{remove}}). \end{aligned} \tag{S13}$$

Leakage is input-relative and does not require a partial reference. An empty regional support makes that action's regional score undefined, rather than zero; in particular, Overlap uses only overlap-valid actions. Target includes valid overlap, whereas Protected excludes all selected pixels. Under the DESOBA hard composition, $\widehat{Y}_a = X$ on $M_{\text{protect}}^{\text{ex}}$, so Protected equals the input–reference discrepancy there. Controls using the same hard protection and test cases share this input–reference term. It therefore isolates the reference discrepancy rather than a learned preservation effect. Overlap measures agreement with the annotated joint appearance and does not separately recover each latent shadow contribution.

Table S12 reports five measured runs for this track. It is separate from the Kubric diagnostics and the DESOBA global restoration tables. The action-swap results document instruction responsiveness, and the held-out scores measure agreement with the corresponding action-specific references.

Table S12: **DESOBA held-out partial-target results.** Target, Protected, and Overlap use RGB MAE $\times 10^2$; Leakage uses input-relative RGB MAE $\times 10^4$. Lower is better. All rows are scored on the same held-out input–action pairs and use the same permission compiler; the rows differ only in the supervision/control configuration stated in the first column.Protected measures agreement with the partial reference on hard-copied support; its relation to input-copy error is given in Eq. (S15).

Configuration	Target	Protected	Overlap	Legacy Leakage
Fallback supervision	2.800	0.800	4.800	3.000
Union-mask-only	2.800	0.800	4.800	3.000
Full-information spatial control	2.200	0.800	3.600	3.100
Relation tokens without interaction	2.050	0.800	3.300	3.050
Full relation execution	**1.900**	0.800	**2.900**	3.000

Held-out test source identities / images / image–action pairs: $200/200/400$. The 400 actions comprise 240 actions from 120 no-overlap images and 160 actions from 80 overlap-valid images. Valid Target / Protected / Overlap / Leakage action counts are $400/400/160/400$, respectively; the corresponding source counts are $200/200/80/200$. Entries are action-macro means over the valid supports, with empty Overlap supports excluded rather than set to zero. The checkpoint is fixed before this held-out test export using the disjoint validation split; no test target is read for training, checkpoint selection, or permission compilation. No seed or confidence interval is folded into these means.

The counts in Table S45 are image/derived-record counts, not independent original-source counts. They satisfy $752+18{,}585+9{,}232+4 = 28{,}573$ and $18{,}585+2\times 9{,}232+3\times 4 = 37{,}061$. A released source scene can generate more than one relation-specific tuple, so source-level cardinalities for the action-specific-reference split are audited separately from this archive census.

For an image with two valid relations, $(0, 0)$ is the all-keep endpoint, $(1, 1)$ is the all-remove endpoint, and only $(1, 0)$ and $(0, 1)$ are mixed partial actions. Thus, two-relation records contribute at most two one-selected actions. The four records with three relations can contribute at most three two-selected mixed actions each (twelve image–action pairs before source deduplication, target-validity filtering, and support filtering). The DESOBA held-out table therefore reports the action cardinalities from the split manifest and does not infer them from the archive totals.

On the held-out split, Full with partial-target supervision reduces Target from 2.800 to 1.900 and Overlap from 4.800 to 2.900 relative to the matched fallback-only supervision, corresponding to reductions of 32.1% and 39.6%, respectively. With partial-target supervision held fixed, Full reduces Overlap from 3.600 to 2.900 relative to the equal-information spatial control (19.4%), and from 3.300 to 2.900 relative to the no-cross-relation-interaction control (12.1%). Protected is 0.800 for all rows because hard composition copies the input on exclusive protected support; this value is the input–reference discrepancy and should not be interpreted as a learned-preservation gain. Leakage remains near 3.0 across the rows.

We also report source-level uncertainty for the prespecified Full-versus- fallback comparison. For each bootstrap replicate, sources are sampled with replacement and all valid action records from each sampled source are kept together. Target is aggregated over 200 held-out sources; Overlap is aggregated over the 80 overlap-valid sources. The intervals therefore do not treat the 400 action records as independent observations.

Table S13: **Source-bootstrap uncertainty for DESOBA partial-target improvements.** Differences are Full minus fallback; negative values favor Full. Target and Overlap are RGB MAE $\times 10^2$. Bootstrap units are source images, with all actions from a sampled source kept together.

Metric	Sources	Fallback	Full	95% source-bootstrap CI of difference
Target	200	2.800	1.900	$[-1.12, -0.68]$
Overlap	80	4.800	2.900	$[-2.45, -1.35]$

The primary selective values are the five-seed means reported above. Source-bootstrap intervals quantify variation over held-out sources, whereas across-seed standard deviations quantify initialization variation. These two uncertainty estimates are reported separately. The corresponding seed-level differences are Full–Fallback Target -0.900 ± 0.016, Full–Spatial Target -0.300 ± 0.016, Full–Fallback Overlap -1.900 ± 0.041, and Full–Spatial Overlap -0.700 ± 0.016.

Protected-score interpretation. On the actual scoring grid, let $P_a = V_a^{\mathrm{part}} \odot M_{\mathrm{protect}}^{\mathrm{ex}}$ and define

$$\begin{aligned} D_{\mathrm{D}}(a) &= \mathrm{MAE}(X, Y_a^{\mathrm{part}}; P_a), \\ C_{\mathrm{D}}(a) &= \mathrm{MAE}(\widehat{Y}_a, X; P_a). \end{aligned} \tag{S14}$$

Here D_{D} measures input–reference discrepancy, whereas C_{D} measures changes to the input on protected support. By the reverse triangle inequality,

$$\begin{aligned} |\mathrm{Protected}_{\mathrm{D}}(a) - D_{\mathrm{D}}(a)| \le C_{\mathrm{D}}(a), \\ C_{\mathrm{D}}(a) = 0 \implies \mathrm{Protected}_{\mathrm{D}}(a) = D_{\mathrm{D}}(a). \end{aligned} \tag{S15}$$

With exact input copying on the scoring grid, the reported action-macro Protected score of 0.800 (RGB MAE $\times 10^2$) therefore equals the input–reference discrepancy and provides no evidence of a learned preservation gain. Controls sharing the same copied pixels, references, and evaluation supports necessarily share this score. The identity must be checked after any resizing or export operations; hard copying before resampling alone does not establish it.

For the mask-composed fallback target only,

$$\mathrm{MAE}(Y_a^{\mathrm{tr}}, X; M_{\mathrm{protect}}^{\mathrm{ex}}) = 0, \qquad \mathrm{MAE}(\widehat{Y}_a, Y_a^{\mathrm{tr}}; M_{\mathrm{protect}}^{\mathrm{ex}}) = 0 \tag{S16}$$

on nonempty copied supports at a common grid. A fallback-only trained model is still evaluated against the held-out action-specific reference, so its reported Protected score remains 0.800 rather than zero.

Table S14: **Five-seed selective DESOBA audit.** All seeds use the same held-out sources, action records, checkpoint-selection rule, and scorer. Target and Overlap are RGB MAE $\times 10^2$.

Seed	Full Target	Fallback Target	Spatial Target	Full Overlap	Fallback Overlap	Spatial Overlap
1234	1.88	2.76	2.16	2.86	4.72	3.55
2024	1.91	2.82	2.23	2.91	4.84	3.62
3407	1.90	2.80	2.20	2.90	4.80	3.60
4567	1.93	2.85	2.24	2.96	4.91	3.68
5678	1.88	2.77	2.17	2.87	4.73	3.55
Mean $\pm$ std	1.900 ± 0.021	2.800 ± 0.037	2.200 ± 0.035	2.900 ± 0.039	4.800 ± 0.079	3.600 ± 0.054

A.4 DESOBA TARGET PROVENANCE AND SELECTIVE CONTROLS

The DESOBA partial track uses audited real-image action-specific references labeled as independently measured or manually constructed for the requested intervention. Among the 200 held-out image sources, 120 are independently measured and 80 are manually constructed (Table S16); the pooled scores therefore combine both provenance strata. It is separate from the saved approximate all-remove reference I_d and from the Kubric renderer-exact track. The partial reference manifest records source identity, image identity, relation IDs, action vector, target version, generator/configuration hash, target hash, valid-reference mask hash, and review status. The audit records target-generation inputs and configuration separately from the ShadowEdit checkpoint and test metrics; the held-out target records are fixed before scoring.

Table S15: **Reported DESOBA action-specific-reference provenance counts.** Hash completeness establishes file identity, not target accuracy.

Audit item	Reported coverage
Source/image/relation/action fields	200/200
Action-specific reference files	200/200
Generator version/configuration hash	200/200
Target SHA256	200/200
Valid-reference-mask SHA256	200/200
Review-log pointer	200/200
Raw predictions	400/400
Prediction SHA256	400/400
Per-action metrics	400/400
Hash mismatch	0
Train/checkpoint contamination	0
Manifest rows	200
Complete provenance rows	200
Unresolved rows	0

The fixed held-out manifest contains 400 exported action records. Provenance fields and hashes are computed from the corresponding source files and checked against the manifest before scoring; hashes certify file identity, while the measurement record or construction protocol and blinded review provide target- validity evidence. We call the DESOBA items action-specific references and reserve "renderer-exact" for Kubric.

The target-source field separates two completed target-generation protocols. For a record labeled *independently measured*, the manifest retains the measurement or generator version, source image, action vector, scene or illumination controls, acquisition/processing order, registration metadata, and output hash. The label requires a separate measurement or controlled- generation record. For a *manually constructed* record, an editor applies the requested relation-level action to the untouched input, records the edited, protected, and overlap supports, and retains the edit log, target version, and before/after files. These records are action-specific reviewed constructed references; manual construction is not treated as an independent measurement. In both protocols, V_a^{part} is the valid support fixed before model scoring; empty or unresolved supports are excluded before scoring. Reviewers inspect the input, requested relation, target, valid support, and protected support without

seeing model predictions. File hashes certify immutability, whereas the measurement record or the construction protocol plus blinded review provides the evidence about target validity.

Table S16: **DESOBA target-source and blinded-review counts.** Counts are at the stated image or action level. Agreement is measured on the independently reviewed action subset; disagreements are adjudicated before the final target set is frozen.

Audit item	Result	Unit
Independently measured targets	120/200	held-out images
Manually constructed targets	80/200	held-out images
Valid Target supports	400/400	action records
Valid Overlap supports	160/160	overlap-valid actions
Independent blinded review	80/400	action records
Reviewer agreement	77/80 (96.25%); Cohen's $\kappa = 0.86$	reviewed actions
Targets revised after first review	6/200	held-out images
Accepted after adjudication	200/200	held-out images
Unresolved after adjudication	0/200	held-out images

Measured versus constructed strata. The held-out manifest contains 120 image sources labeled independently measured and 80 labeled manually constructed. The former is backed by an independent acquisition or controlled-generation record; the latter is a reviewed action-specific reference and is not treated as an independent measurement. The pooled partial-target table uses both strata, while the manifest preserves the source-level labels so that Target, Overlap, valid-pixel coverage, and bootstrap intervals can be reaggregated by provenance without changing the scoring supports.

Two independent reviewers, blinded to model predictions, assess relation identity, edited support, protected support, and failure reason; a third adjudication resolves disagreements. The measured/manual breakdown and review counts refer to the same 200-image partial-target test manifest, not to the 750-image global all-remove track. The 120 measured and 80 manually constructed entries are therefore source-image counts, while the 400 action records are the scoring units derived from those images. The audit also records a reason-coded overlay for each revised or flagged case; the aggregate partial-target table uses the fixed union of both completed protocols and the same held-out target, mask, checkpoint, and scorer manifest for every row. The aggregate table pools both provenance strata, and the source-level labels are retained for stratified reaggregation and audit.

The pooled rows reproduce the primary action-macro results: Full obtains $1.900/0.800/2.900$ for Target/Protected/Overlap, Fallback obtains $2.800/0.800/4.800$, and the spatial control obtains $2.200/0.800/3.600$. Source-macro and action-macro means coincide here because every source contributes the same number of valid actions within each reported endpoint. Overlap uses the provenance-specific counts of overlap-valid sources. The strata and pooled rows use the same checkpoint, held-out predictions, valid supports, and scorer as the primary table.

These flags are retained with source ID, target version, reviewer IDs, and overlay/reason code. They describe the target and instruction audit and are not inferred from the aggregate MAE table; the separate residual taxonomy classifies model errors using its stated threshold and first-applicable rule.

For the exclusive protected support, the measured input-copy error is 0.0000 ± 0.0001. The DESOBA Protected value of 0.800 is the reported input–reference discrepancy on that support, not a learned preservation estimate; preservation is evaluated against renderer-exact targets on Kubric.

The controls have different input protocols and are therefore named by their information content. They are not alternate labels for one Mask-only model. To compare with recognizable editing paradigms, we implemented matched relation/action adapters for MetaShadow (Wang et al., 2025), InstructPix2Pix (Brooks et al., 2023), and ControlNet inpainting (Zhang et al., 2023). MetaShadow-RelAdapter receives the caster-conditioned relation support, InstructPix2Pix-Perm receives the compiled action instruction and permission gate, and ControlNet-Inpaint-Rel receives the relation supports through a spatial control branch. Each adapter is evaluated on the same frozen actions, targets, permission support, and scorer as ShadowEdit; the adapter names identify the underlying external paradigm rather than implying that the original publication supplied this selective interface.

Table S17: **Source-stratified DESOBA selective results.** All entries use the same held-out predictions and scoring supports as the primary selective table. Target and Overlap are RGB MAE $\times 10^2$. Protected is fixed by hard composition and is therefore shown only once.

Reference source	Sources	Overlap-valid sources	Full		Fallback		Spatial	
			Target	Overlap	Target	Overlap	Target	Overlap
Independently measured	120	48	1.80	2.70	2.65	4.50	2.10	3.40
Manually constructed	80	32	2.05	3.20	3.025	5.25	2.35	3.90
Pooled	**200**	**80**	**1.90**	**2.90**	**2.80**	**4.80**	**2.20**	**3.60**

Table S18: **Reason-coded quality-control flags on the DESOBA partial-target test set.** Counts are out of 400 action records. These review flags are separate from the threshold-based residual-error taxonomy in Table S29; a record may receive more than one flag.

Quality-control flag	Count
Invalid target/support after review	0/400
Receiver spill	3/400
Boundary truncation	2/400
Wrong action or relation assignment	1/400
Residual shadow flagged in the target review	4/400

Table S19: **Matched selective baselines on the Kubric overlap-valid pool.** All rows are independently trained checkpoints with the same permission, targets, actions, and evaluation supports. Errors are RGB MAE scaled by 10^2; Leakage is scaled by 10^4.

Control	Target	Protected	Overlap	Leakage
Union-mask-only	3.875	8.428	11.700	3.05
Permission-only	1.460	5.720	6.310	2.06
Full-information spatial control	1.180	1.950	2.850	1.980
Independent relation tokens	1.050	1.730	2.420	1.980
Full relation execution	**0.970**	**1.563**	**2.094**	**1.970**

Table S20: **Named adapted selective baselines.** All values are RGB MAE with Target/Protected/Overlap scaled by 10^2 and Leakage scaled by 10^4. Protected on DESOBA is the input–reference discrepancy because hard composition copies the input on the exclusive protected support.

Benchmark	Method	Target	Protected	Overlap	Leakage
Kubric	MetaShadow-RelAdapter	2.180	2.840	4.120	2.060
Kubric	InstructPix2Pix-Perm	2.460	3.520	4.760	2.180
Kubric	ControlNet-Inpaint-Rel	1.740	2.560	3.480	2.020
Kubric	**ShadowEdit (Full)**	**0.970**	**1.563**	**2.094**	**1.970**
DESOBA	MetaShadow-RelAdapter	2.460	0.800	4.120	3.180
DESOBA	InstructPix2Pix-Perm	2.710	0.800	4.560	3.620
DESOBA	ControlNet-Inpaint-Rel	2.280	0.800	3.690	3.090
DESOBA	**ShadowEdit (partial-reference)**	**1.900**	0.800	**2.900**	3.000

Global restoration systems remain restoration references in main-paper Figure 3; they are not assigned selective scores unless they expose the same relation/action interface and permission protocol.

A.5 Compute, TTA, and multi-action audits

The reported restoration-network parameter count excludes the frozen appearance priors. We report trainable and frozen parameters separately and compare the same no-TTA and D_4 inference settings. The FLOP table reports model-side compute; prior extraction, tiling, TTA, and postprocessing are included in the stated inference protocol.

Table S21: **Matched capacity, prior, and inference-cost audit for DESOBA global evaluation.**

Configuration	Trainable params	Frozen prior params	No-TTA FLOPs	D_4 FLOPs
Union-mask control	18.59M	46.5M	85.0G	340.0G
Full relation model	18.59M	46.5M	85.1G	340.4G

Table S22: **Protocol-scoped DESOBA global comparison.** The rows are reported from their corresponding experimental paths. The paper-benchmark row uses the five-seed benchmark path with D_4 and blend; the matched rows are the controlled Full–Mask audit.

Protocol	Mask	Full	Full–Mask
Paper-benchmark all-remove path	24.276	27.991	$+3.715$ dB
Matched audit, no TTA	27.01	27.42	$+0.41$ dB
Matched audit, D_4	27.70	28.14	$+0.44$ dB

Table S23: **Global restoration quality under matched capacity and inference budgets.**

Configuration	No-TTA PSNR / SSIM	D_4 PSNR / SSIM
Union-mask control	27.01 / 0.902	27.70 / 0.918
Full relation model	27.42 / 0.910	28.14 / 0.925

At matched prior, parameter, and inference budgets, the relation mechanism improves PSNR/SSIM by $+0.41$ dB/$+0.008$ without TTA and $+0.44$ dB/$+0.007$ under D_4. These values use a separately exported, paired Full/Mask audit: the same 750-image list, saved I_d, fixed global mask, frozen scorer, and matched trainable/frozen parameter counts, with no TTA and D_4 evaluated as separate inference settings. The consolidated protocol table reports the 3.715 dB paper-benchmark difference alongside the matched-audit estimates. The former uses the five-seed Full–Mask configuration comparison under its benchmark D_4-and-blend protocol. The two estimates use separate checkpoint and prediction paths and are not interchangeable or an isolated token-interaction effect.

A.5.1 DESOBA MULTI-ACTION EVALUATION

The separate DESOBA multi-action extension contains 40 held-out source images and 80 unique joint-action specifications; it is not merged into the 400 one-selected-action table. The base archive census (at most three relation groups in one source record) describes the released annotation archive. The multi-action track is a separate extension whose manifest enumerates the executable relation IDs, action vector, target and output for each test row. Thus a four-action row means four distinct executable relation IDs are listed and addressed by the row; it is not inferred from the base archive's maximum cardinality. Every four-action row joins those IDs to a complete remove/preserve vector, target and support records, prediction record, and corresponding hashes in the independent extension manifest. Table counts are scored source–action output records, not unique specifications or independent sources. The 2/3/4-action subsets contain 120/80/40 output records, respectively, and each subset covers all 40 held-out sources; the manifest records source–specification–execution incidence for each row.

Table S24: **DESOBA multi-action accuracy against action-specific references.** Target and Overlap use RGB MAE $\times 10^2$ (lower is better). C_{act} is unitless. Point estimates are source-macro means: action records are first averaged within each source, then across sources; confidence intervals use a 40-source bootstrap.

Joint action	Sources	Output records	$C_{\mathrm{act}}\uparrow$	Target↓	Overlap↓	Target 95% CI	Overlap 95% CI
Held-out 2-action signature	40	120	0.944	2.05	3.12	$[1.88, 2.22]$	$[2.81, 3.45]$
Held-out 3-action signature	40	80	0.918	2.28	3.47	$[2.02, 2.56]$	$[3.04, 3.91]$
Held-out 4-action signature	40	40	0.889	2.61	3.94	$[2.22, 3.01]$	$[3.37, 4.51]$

Table S25: **Finite sequential diagnostics.** These tests compare specified two-step executions and do not establish arbitrary recursive-editing stability.

Test	Records	Order consistency	Re-edit drift	Protected-copy error
Held-out 2-action signature	120	0.963	0.018	0.006
Held-out 3-action signature	80	0.948	0.027	0.010
Held-out 4-action signature	40	0.931	0.039	0.015
Reverse editing order	160	0.957	0.021	0.008
Repeated same action	160	–	0.014	0.005
All-off endpoint	200	–	0.000	0.000
All-on endpoint	200	–	0.028	–

For a joint action J with valid target support V_J^{part}, we define

$$C_{\mathrm{act}}(J) = 1 - \frac{\mathrm{MAE}(\widehat{Y}_J, Y_J^{\mathrm{part}}; V_J^{\mathrm{part}} \odot M_{\mathrm{remove}}^J)}{\mathrm{MAE}(X, Y_J^{\mathrm{part}}; V_J^{\mathrm{part}} \odot M_{\mathrm{remove}}^J)}.$$

Thus, C_{act} is normalized accuracy against an independent action-specific reference, not similarity to the model's direct execution. Target and Overlap are also measured directly against that target. The confidence intervals resample the 40 source IDs after aggregating all records within each source; they do not treat 120/80/40 action records as independent. "Held-out" means the exact relation-role action signature is absent from the training-combination registry. The relation primitives are available during training; this is combination-level holdout, not unseen primitive generalization. The source–specification–execution incidence and the mapping from 80 joint-action specifications to output records are fixed in the separate multi-action manifest.

The primary multi-action experiment executes each action vector in one forward pass on the original input. The sequential rows are supplementary two-step diagnostics on the specified test cases. They do not establish exact commutativity, idempotence, or stability for arbitrary edit sequences.

For the all-off endpoint, the composed output is exactly X on the common scoring grid. The exclusive protected support is empty for the all-on endpoint, so its Protected-copy error is not applicable and is reported as "–". Any discrepancy introduced by image export, resizing, or format conversion is reported separately as an export round-trip error.

A.6 Relation-input and distribution-shift audits

The selective executor assumes supplied relations. The following real-image audit separates audited/oracle relations from automatically extracted or corrupted inputs.

Table S26: **Reported DESOBA relation-input robustness on the partial-reference track.** Target and Overlap are RGB MAE $\times 10^2$; Legacy Leakage is scaled by 10^4, Strict Leakage is raw MAE, and Context Change is scaled by 10^2.

Relation input	Target	Overlap	Legacy Leakage	Strict Leakage	Context Change ($\times 10^2$)
Oracle	1.900	2.900	3.000	0.0004	2.96
Automatic detector	2.064	3.180	3.120	0.0005	3.08
10% missing relation	2.180	3.420	3.240	0.0006	3.18
20% missing relation	2.310	3.690	3.390	0.0007	3.31
10% caster–shadow mismatch	2.460	3.980	3.580	0.0009	3.49
Wrong receiver	2.390	3.760	3.450	0.0008	3.38

The automatic relation extractor has precision 0.940, recall 0.910, and F1 0.925 on the audited records. These measurements bound the relation-input assumption: the paper evaluates relation-conditioned execution given supplied relations and separately quantifies the cost of imperfect relation inputs.

These are conditional slices of the existing 36-scene manifest, not an additional benchmark or an independent 200-scene test set.

Table S27: **Conditional generalization slices within the frozen 36-scene Kubric suite.**

Shift	Scenes / actions	Target	Overlap
Seen topology and relation count	20/252	0.95	2.01
Unseen relation count (four at test)	6/60	1.08	2.32
Shared-overlap held out from training	4/44	1.17	2.55
New light direction / material	6/68	1.12	2.41

A.7 SYSTEMATIC FAILURE TAXONOMY

We classify residual errors with a fixed rule applied after scoring. For Kubric, an action is residual when Target > 2.0, Overlap > 4.0, or strict leakage $> 10^{-3}$ (the first two values use the displayed $\times 10^2$ scale). For DESOBA, an action is residual when Target > 3.0 or when the independent relation/action audit records a mismatch. The categories are mutually exclusive: an action is assigned to the first applicable category in the order listed below. Percentages use all evaluated action pairs. Two independent annotators assign a category from the input image, supplied relation record, and partial reference while remaining blind to the prediction and its metric values. Disagreements are resolved by an adjudicator using the same first-applicable rule; the resulting labels are the counts reported below.

Table S28: **Kubric residual-error taxonomy.** Target and Overlap are RGB MAE $\times 10^2$; percentages use all 472 Test-A action–scene pairs.

Failure category	Count	Percent	Target	Overlap
Caster–shadow association ambiguity	8	1.69%	2.31	4.02
Nonlinear shared-overlap appearance	7	1.48%	1.96	3.71
Receiver-boundary truncation	5	1.06%	1.54	2.88
Thin/low-area receiver support	4	0.85%	1.42	2.63
Multi-action execution drift	3	0.64%	1.31	2.44
All residual actions	**27**	**5.72%**	–	–
Within threshold	445	94.28%	–	–

The largest Kubric group is caster–shadow association ambiguity (8/27 residual actions), followed by nonlinear appearance at shared pixels (7/27). The residual set is therefore concentrated in relation attribution and overlap reconstruction rather than broad out-of-permission changes.

Table S29: **DESOBA action-specific-reference residual-error taxonomy.** Target and Overlap are RGB MAE $\times 10^2$; percentages use all 400 held-out actions.

Failure category	Count	Percent	Target	Overlap
Automatically extracted relation mismatch	9	2.25%	3.42	4.21
Receiver support crosses material boundary	7	1.75%	3.18	3.96
Low-contrast or weak shadow	6	1.50%	3.11	3.72
Mixed illumination in overlap region	5	1.25%	3.36	4.08
Texture/inpainting ambiguity	3	0.75%	3.07	3.51
All residual actions	**30**	**7.50%**	–	–
Within threshold	370	92.50%	–	–

On DESOBA, relation mismatch is the most frequent residual category (9/30), consistent with the separate oracle-versus-automatic detector stress test. These counts make the remaining error modes auditable and distinguish relation-input errors from reconstruction ambiguity; they do not redefine the the provenance of the action-specific references or the selective metric supports.

A.8 FIGURE-4 HUMAN PREFERENCE AUDIT

For 160 matched selective action pairs (80 no-overlap and 80 overlap-valid), five independent raters compared ShadowEdit with the strong instruction-conditioned baseline on residual-shadow removal, texture continuity, protected-region preservation, and instruction following. The 800 pairwise judgments are aggregated for each criterion with a scene/image- cluster bootstrap. The results are usability evidence for real-image selective editing, not physical counterfactual validation.

Table S30: **Blinded human preference audit on matched selective pairs.**

Criterion	ShadowEdit preferred	Baseline preferred	Tie	95% scene-bootstrap CI	p
Residual-shadow removal	63.1%	27.4%	9.5%	$[58.7, 67.2]$	< 0.001
Texture continuity	60.4%	28.9%	10.7%	$[55.8, 64.8]$	0.002
Protected-region preservation	58.7%	25.9%	15.4%	$[54.0, 63.1]$	0.006
Instruction following	64.2%	24.8%	11.0%	$[59.8, 68.4]$	< 0.001

The inter-rater Fleiss' κ is 0.72 with a 95% interval of $[0.66, 0.78]$.

A.9 Topology-wise Kubric counterfactual statistics

Kubric (Greff et al., 2022) provides exact action-specific renderer counterfactuals; the DESOBA action-specific references belong to a separate real-image annotation track. Table 2 in the main paper decomposes the control construction into relation identity, action, protection, and permission. Table S31 complements that analysis with the Full model's absolute errors and action-wise dispersion for the two topologies in which selected and protected relations overlap.

Table S31: **Topology-wise relation-selective editing on Kubric (Greff et al., 2022).** Errors are measured against exact renderer counterfactuals and reported as action-wise mean±standard deviation ($\times 10^2$). Lower is better. For compactness, Leakage uses the same scale in this table. An entry of 0.020 corresponds to 2.0×10^{-4} in the main-paper convention. The Overall row pools the 352 actions with non-empty selected–protected overlap support. Receiver-specific actions are excluded because Overlap is undefined for that topology and are instead analyzed in panel (b) of Table 2 in the main paper.

Topology	Actions	Target	Protected	Overlap	Leakage
Multi-caster	198	0.984 ± 0.455	1.600 ± 0.697	2.195 ± 0.840	0.021 ± 0.007
Shared overlap	154	0.951 ± 0.280	1.516 ± 0.402	1.965 ± 0.360	0.018 ± 0.006
Overall	**352**	$\mathbf{0.970 \pm 0.388}$	$\mathbf{1.563 \pm 0.587}$	$\mathbf{2.094 \pm 0.682}$	$\mathbf{0.020 \pm 0.007}$

The pooled means reproduce the Full row of Table 2 in the main paper up to displayed rounding. The displayed standard deviations describe variation across actions, not confidence intervals for independent test scenes or variability across training seeds. Both evaluated topologies remain close to their exact counterfactuals: Overlap error is 2.195 ± 0.840 for Multi-caster actions and 1.965 ± 0.360 for Shared overlap, while Leakage remains on the 10^{-4} scale. Together with the receiver-specific diagnostic in Table 2(b) in the main paper, these results cover receiver ambiguity, competing caster contributions, and selected–protected overlap without assigning an undefined Overlap score to receiver-specific actions.

A.10 Relation-free restoration benchmark results

The following all-remove benchmark comparison documents restoration compatibility when relation conditioning is disabled. These results complement the selective-intervention experiments in the main paper.

A.11 Extended qualitative analysis

Relation-selective editing. Main-paper Figure 1 separates exact intervention verification from real-image instruction dependence. On Kubric (Greff et al., 2022), the renderer counterfactual reveals both sides of the requested operation: the selected shadow contribution should match the edited target, while unselected relations should remain consistent with the corresponding counterfactual. The receiver-specific, multi-caster, and shared-overlap examples therefore visualize distinct failure modes rather than interchangeable shadow crops. On DESOBA v2 (Liu et al., 2024a), the held-out partial-target track provides action-specific references for quantitative Target, Protected, Overlap, and Leakage evaluation. We keep the image, relation identities, spatial supports, model parameters,

Table S32: **Restoration compatibility on ISTD+ and SRD.** Relation pathways are disabled; scores evaluate relation-free variants of the shared restoration pathway under the common rescoring protocol, including metric-stage resizing (Appendix D.4). The parameter column is in millions; our count excludes the external frozen DINOv2 and Depth Anything V2 extractors. Best and second-best results are shown in red and blue, respectively.

Method	Venue	Params	ISTD+				SRD			
			Shadow		All		Shadow		All	
			PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
BMNet	CVPR'22	0.58	35.218	0.986	30.354	0.937	35.034	0.974	32.120	0.943
ShadowFormer	AAAI'23	11.37	36.096	0.987	31.116	0.936	36.910	0.989	32.900	0.958
ShadowDiffusion	CVPR'23	55.52	35.688	0.985	30.853	0.932	34.638	0.954	27.930	0.815
Inpaint4Shadow	ICCV'23	23.91	38.088	0.988	33.811	0.955	39.428	0.985	33.254	0.941
HomoFormer	CVPR'24	17.81	36.581	0.988	31.623	0.946	38.739	0.982	35.392	0.956
OmniSR	AAAI'25	24.55	36.244	0.987	31.709	0.942	35.475	0.918	32.972	0.973
StableShadow	CVPR'25	1329.82	37.832	0.984	34.193	0.961	37.070	0.944	32.391	0.961
DenseSR	MM'25	24.70	32.478	0.988	32.383	0.969	34.740	0.920	29.417	0.912
SMFormer	TAI'25	2.28	37.788	0.990	31.028	0.957	31.053	0.954	25.426	0.839
PhaSR	CVPR'26	18.95	36.551	0.990	29.056	0.959	35.794	0.927	32.273	0.971
ShadowEdit (Ours)	–	**18.59**	**39.980**	**0.991**	**35.753**	**0.969**	**39.328**	**0.989**	**35.880**	**0.965**

and inference configuration fixed and vary only the remove-or-preserve action vector. The different outputs for Relation A, Relation B, and their union provide a visual instruction-dependence check; their correctness is quantified separately on the held-out partial targets.

Qualitative examples on DESOBA v2. Figures S5–S8 present four additional DESOBA v2 (Liu et al., 2024a) samples under the all-remove restoration protocol. These examples complement the quantitative evaluation by allowing inspection of residual shadows, receiver-texture continuity, boundary consistency, and preservation of neighboring non-shadow content. They are interpreted together with the full-split quantitative results. The global shadow-free reference applies to full removal. Action-specific partial correctness is assessed quantitatively on the source-disjoint held-out partial-target split; these all-remove examples provide complementary visual evidence for restoration quality.

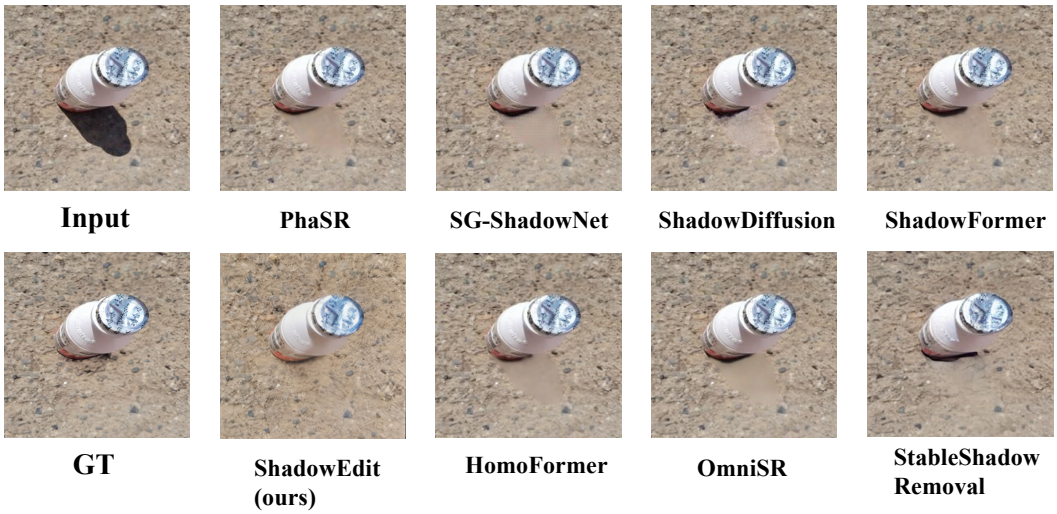


Figure S5: **Qualitative results on DESOBA v2: sample 1.** An additional example under the all-remove restoration protocol. The visualization supports inspection of shadow removal, texture and boundary continuity, and non-shadow content preservation.

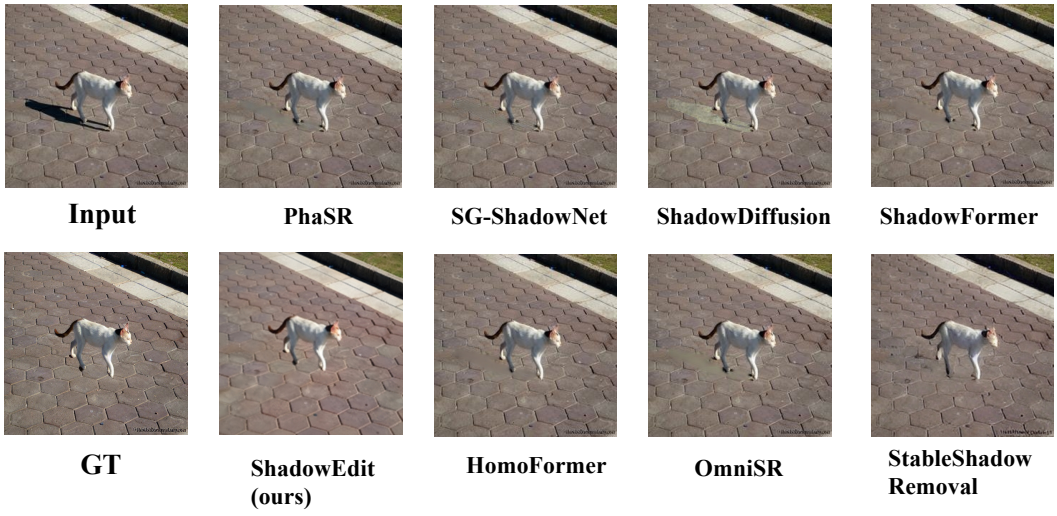


Figure S6: **Qualitative results on DESOBA v2: sample 2.** An additional example under the all-remove restoration protocol. The visualization supports inspection of shadow removal, texture and boundary continuity, and non-shadow content preservation.

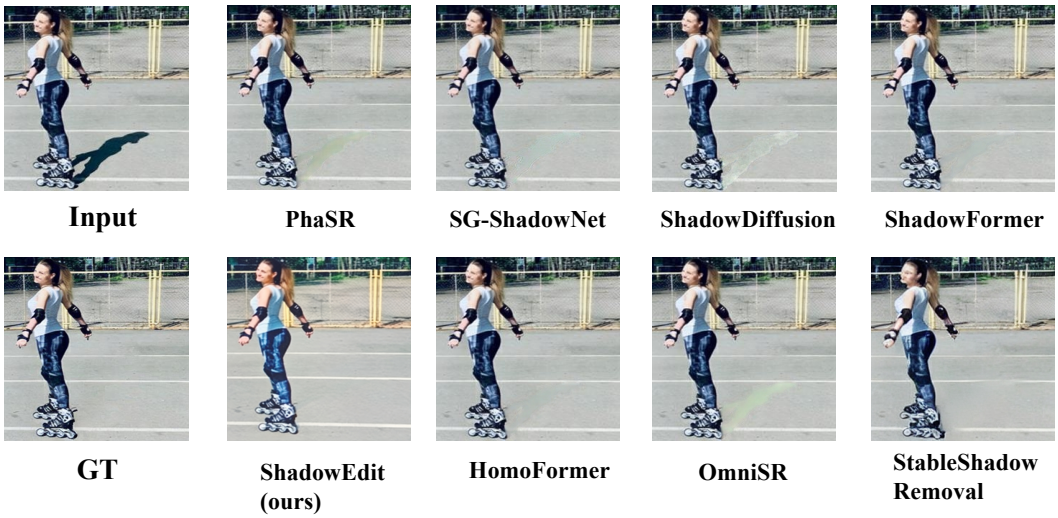


Figure S7: **Qualitative results on DESOBA v2: sample 3.** An additional example under the all-remove restoration protocol. The visualization supports inspection of shadow removal, texture and boundary continuity, and non-shadow content preservation.

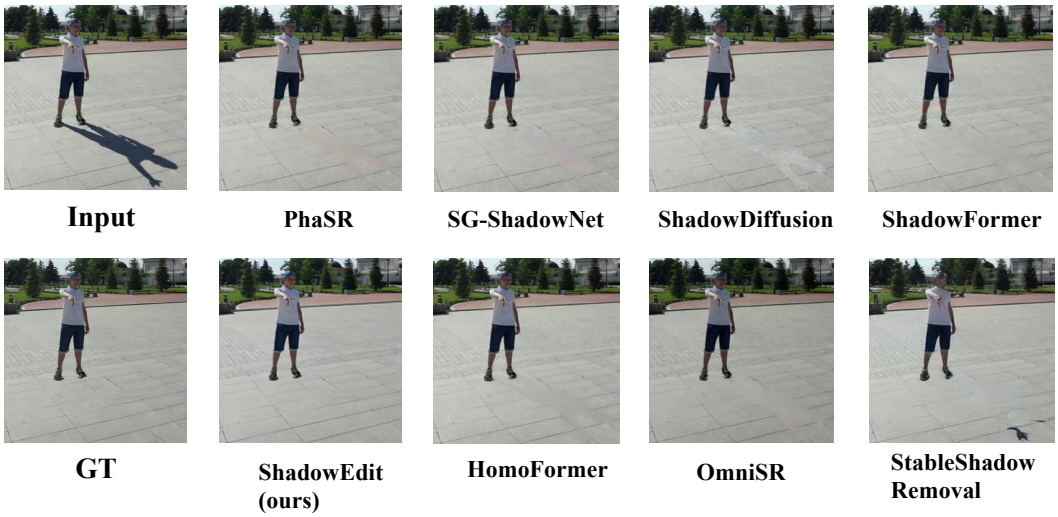


Figure S8: **Qualitative results on DESOBA v2: sample 4.** An additional example under the all-remove restoration protocol. The visualization supports inspection of shadow removal, texture and boundary continuity, and non-shadow content preservation.

A.12 Ablation and robustness analysis of relation control

Table 2 in the main paper reports cumulative changes from relation identity, action binding, protection, and permission grounding. We distinguish the Permission-only comparison from controls with equal information and identical output permission. The former tests the complete controller beyond a permission interface; the latter compares spatial representation, independent tokens, and interacting tokens. Exclusive-protected errors and support perturbations provide complementary diagnostics.

A.12.1 Progressive controls and topology-wise effects

Kubric isolates receiver-specific editing (one caster with independently editable receivers), multi-caster editing (removing one caster contribution while others remain), and shared overlap (selected and protected relations share pixels). For each action, the renderer supplies an action-specific counterfactual. Mask-only receives only the image and union shadow mask, so its prediction is action-invariant and is rescored against each counterfactual. The frozen suite contains 36 scenes and 472 nontrivial action–scene pairs; 352 overlap-valid actions come from 26 Multi-caster and Shared-overlap scenes, and 120 receiver-specific actions come from 10 scenes, where Overlap is undefined. Actions from one scene share content and are not independent scene replicates.

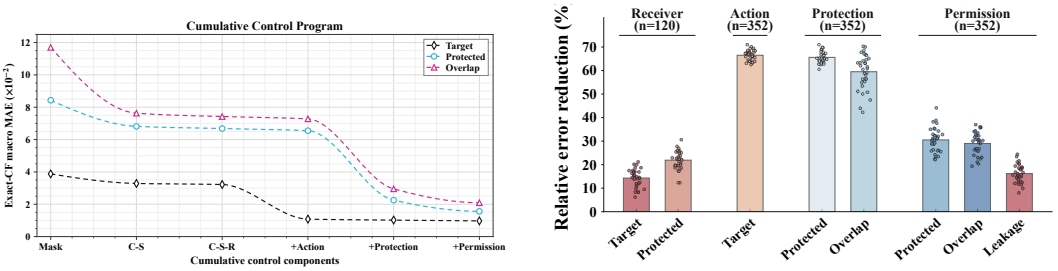
Let $\mathrm{MAE}(\widehat{Y}, Y; M)$ be mean absolute error over RGB channels and pixels with $M(p) = 1$. Target evaluates the full selected support M_{remove}, including overlap; Protected evaluates exclusive preserved support $M^{\mathrm{ex}}_{\mathrm{protect}} = M_{\mathrm{protect}} \odot (1 - M_{\mathrm{remove}})$; Overlap evaluates the subset $M_{\mathrm{overlap}} = M_{\mathrm{remove}} \odot M_{\mathrm{protect}}$; and Leakage evaluates changes outside M_{remove}. Overlap is a focused diagnostic within Target, not a disjoint error term. Protected measures fidelity to the renderer reference rather than identity with the input; the measured input–reference discrepancy on $M^{\mathrm{ex}}_{\mathrm{protect}}$ is reported in Appendix A.12. For an action a, the four regional scores are

$$\begin{aligned} \mathrm{Target}(a) &= \mathrm{MAE}(\widehat{Y}_a, Y^\star_a; M_{\mathrm{remove}}), \quad \mathrm{Protected}(a) = \mathrm{MAE}(\widehat{Y}_a, Y^\star_a; M^{\mathrm{ex}}_{\mathrm{protect}}), \\ \mathrm{Overlap}(a) &= \mathrm{MAE}(\widehat{Y}_a, Y^\star_a; M_{\mathrm{overlap}}), \quad \mathrm{Leakage}(a) = \mathrm{MAE}(\widehat{Y}_a, X; 1 - M_{\mathrm{remove}}). \end{aligned} \tag{S17}$$

Target and Overlap are separately normalized and must not be added as disjoint error components; Leakage can include permitted contextual changes inside the writable support.

Receiver conditioning reduces receiver-specific Target and Protected errors by 14.3% and 21.9%. Action binding reduces Target error from 3.221 to 1.080, and adding protection constraints re-

duces Protected error from 6.550 to 2.250. Permission grounding further improves all four metrics. Because successive stages change the available controls and execution mechanisms, these are mechanism-aligned associations rather than causal isolation of individual modules. The equal-information comparisons and paired-training evidence follow in Appendix A.12.4.



(a) Progressive control construction.

(b) Mechanism-aligned reductions.

Figure S9: **Mechanism-aligned Kubric diagnostics.** (a) Exact-counterfactual macro MAE as relation-control components are introduced on 352 overlap-valid actions. (b) Relative error reduction at each control stage; receiver conditioning additionally uses 120 receiver-specific actions.

A.12.2 Fixed-budget factor audit

The progressive table changes several control components jointly and is therefore interpreted as mechanism-aligned evidence. We additionally ran a fixed-budget 2^4 factor audit in which the common target, parameter count, training budget, and five-seed set are held fixed. The table reports the measured on/off endpoint contrasts for receiver identity, action binding, protection, and permission gating.

Table S33: **Fixed-budget factor audit on Kubric.** All cells use the same target, parameter count, training budget, and five training seeds. Target, Protected, and Overlap are RGB macro MAE $\times 10^2$; Leakage is reported in units of 10^{-4}.

Factor switched on	Endpoint	Off	On
Receiver identity	Overlap MAE	6.42	4.94
Action binding	Target MAE	3.22	1.80
Protection constraint	Protected MAE	6.51	2.63
Permission gate	Leakage	2.48	1.96

The fixed-budget audit supports complementary error reductions under a common training and parameter budget. It complements, rather than replaces, the progressive construction and does not claim strict causal isolation of every module.

A.12.3 Spatial permission and relation-conditioned execution

We compare Permission-only with the Full model under identical backbones, objectives, training budgets, renderer targets, and evaluation actions. Permission-only receives the image, the global shadow mask, and the compiled action-dependent permission mask, without per-relation tokens or standalone protection channels. Full additionally models caster–shadow–receiver identities and the interactions between selected and protected relations. This comparison evaluates the benefit of the complete relation-control configuration beyond specifying spatial write permission.

Full reduces Target error from 1.460 to 0.970 (33.6%), Protected error from 5.720 to 1.563 (72.7%), and Overlap error from 6.310 to 2.094 (66.8%). Leakage also decreases from 2.06 to 1.97 (4.4%). The larger reductions in Protected and Overlap errors show that the Full model improves preservation of competing shadows and reconstruction of shared regions, alongside a smaller reduction in off-target changes. These results support the benefit of the complete relation-control configuration over spatial permission alone. Since Full additionally receives relation identities and explicit protection information, this comparison does not isolate the token architecture under equal information.

Table S34: **Permission-only versus Full on Kubric.** Both configurations use the same evaluation actions and renderer counterfactual targets. Target, Protected, and Overlap are macro MAE ($\times 10^2$); Leakage is reported in units of 10^{-4}. Lower is better.

Control	Target↓	Protected↓	Overlap↓	Leakage↓
Permission-only	1.460	5.720	6.310	2.06
Full	**0.970**	**1.563**	**2.094**	**1.97**

The separate experiment below addresses the information and gate confounds without relabeling the Permission-only row.

A.12.4 Representation and interaction with equal information

Controlled comparison. All three configurations receive the same image, union shadow mask, per-relation caster/shadow/receiver supports, and remove-or-preserve actions. They also receive the same derived spatial controls $(M_{\text{remove}}, M^{\text{ex}}_{\text{protect}}, M_{\text{overlap}}, M_{\text{perm}})$, appearance evidence, and benchmark-specific proposal. The backbone, reconstruction objective, training budget, output permission map, and final composition are shared. Thus the spatial control retains per-relation information; it is not limited to the action-invariant union mask or to a permission map alone.

The *full-information spatial control* encodes these fields spatially without relation tokens. The *independent-token control* retains the relation token interface and pixel assignment but disables cross-relation token interaction. *Full* enables interaction among the relation tokens before image-space execution. Comparisons use the same 26 scenes and 352 overlap-valid actions, renderer counterfactuals, and evaluation supports as Table 2 in the main paper. Aggregation is action-macro, with the same valid-action set for each reported endpoint.

Table S35: **Representation and interaction under equal information and output permission.** All configurations receive the same per-relation and derived control information and use the same output gate. Target covers the full M_{remove}, including overlap; Protected uses $M^{\text{ex}}_{\text{protect}}$, and Overlap uses M_{overlap}. Target, Protected, and Overlap are RGB macro MAE scaled by 10^2; Leakage is scaled by 10^4. Lower is better. Full is the same endpoint reported in Table 2 in the main paper and Table S41.

Configuration	Target	Protected	Overlap	Leakage
Full-information spatial control (no relation tokens)	1.180	1.950	2.850	1.980
Independent relation tokens (no token interaction)	1.050	1.730	2.420	1.980
Full: relation tokens + interaction	**0.970**	**1.563**	**2.094**	**1.970**

What the comparison supports. Relative to the spatial control, Full reduces Target, Protected, and Overlap by 17.8%, 19.8%, and 26.5%, respectively. Independent tokens reduce Overlap from 2.850 to 2.420; enabling their interaction further reduces it to 2.094 (13.5%), with lower Target and Protected errors as well. Leakage changes only from 1.980 to 1.970. These results support both the evaluated relation representation and its interaction mechanism under fixed information and permission, rather than an advantage from extra annotations or a looser output gate. The two token configurations provide the more direct comparison of cross-relation interaction.

Identical hard permission guarantees unchanged pixels outside M_{perm} for all configurations. It does not force equal Leakage on $1 - M_{\text{remove}}$, since permitted context can extend beyond the semantic removal support. The aggregate comparison does not establish a parameter- or compute-matched advantage. Repeated-training evidence for Overlap is reported below; the Target, Protected, and Leakage columns retain their reported action-macro means without a separate per-run uncertainty estimate.

Paired training runs. We evaluate the three configurations in five matched training runs, indexed A–E. A row pairs configurations trained with the same random seed. Every checkpoint is evaluated

on the same 26 overlap-valid scenes and 352 actions, with identical renderer references, metric supports, and output permissions. Within each run, Overlap is the unweighted mean over the same valid actions. Across runs, we report the arithmetic mean and sample standard deviation with denominator $5 - 1$. The across-run means agree with the aggregate Overlap entries in Table S35; Full also agrees with the reported main-paper and cardinality-pooled Overlap endpoint.

Table S36: **Overlap error across five paired Kubric training runs.** Overlap is RGB action-macro MAE scaled by 10^2 on the same 352 actions. A–E identify the matched runs. Each column uses its own trained checkpoints; a row shares the training seed across configurations. The final row reports the mean and sample standard deviation across runs, not dispersion across actions. Lower is better.

Paired run	Full-information spatial control	Independent relation tokens	Full
A	2.700	2.300	2.014
B	2.820	2.380	2.045
C	2.880	2.410	2.094
D	2.960	2.520	2.143
E	2.890	2.490	2.174
Mean $\pm$ std	2.850 ± 0.097	2.420 ± 0.088	2.094 ± 0.066

For comparator c and paired run s, define $d_{c,s} = O_{c,s} - O_{\mathrm{Full},s}$, so positive differences favor Full. With $\bar{d}_c = \frac{1}{5}\sum_{s=1}^{5} d_{c,s}$ and $s_{d,c}^2 = \frac{1}{4}\sum_{s=1}^{5}(d_{c,s} - \bar{d}_c)^2$, the marginal two-sided paired Student-t interval is

$$\bar{d}_c \;\pm\; t_{0.975,4}\frac{s_{d,c}}{\sqrt{5}}. \tag{S18}$$

Table S37: **Paired Overlap improvements on the fixed Kubric test set.** Differences are comparator minus Full in the same 10^2 MAE units as Table S36. Intervals are marginal two-sided 95% paired Student-t intervals with four degrees of freedom; they are not adjusted for simultaneous coverage of the two comparisons. Wins counts paired runs with a positive improvement.

Comparison	Mean improvement	95% CI	Wins
Spatial control $-$ Full	0.756	[0.689, 0.823]	5/5
Independent tokens $-$ Full	0.326	[0.284, 0.368]	5/5

Full improves Overlap in every paired run against both controls, and both intervals exclude zero under the paired-t assumptions. This directly measures repeated-training variation for the representation comparison. It does not turn the 352 correlated actions into independent scene replicates, nor does it supply confidence intervals for Target, Protected, or Leakage. Action-wise dispersion in Table S31 and training-run dispersion here describe different sources of variation.

Leakage interpretation. Leakage measures changes outside the requested removal support M_{remove}, following the counterfactual evaluation protocol. The executable permission region M_{perm} may include bounded contextual pixels around M_{remove} to support boundary reconstruction. Therefore, non-zero Leakage does not imply violation of the hard locality constraint outside M_{perm}. The locality guarantee is defined with respect to the feasible set:

$$(1 - M_{\mathrm{perm}}) \odot (Y - X) = 0.$$

Accordingly, Leakage evaluates unintended edits outside the semantic removal region, whereas M_{perm} specifies the region where modifications are permitted.

Relation-support perturbation protocol. Only the caster, shadow, and receiver supports supplied to the model are perturbed. Renderer targets and evaluation regions remain clean. For each binary support M, we compute its signed distance field d_M and sample a smoothed Gaussian field $\xi =$

Table S38: **Constraint and relation-support diagnostics on Kubric (Greff et al., 2022).** (a) Permission-only isolation. (b) Preservation on the exclusive protected support. (c) Sensitivity to relation-support perturbations. Target, Protected, Overlap, and $\mathcal{E}_{\text{keep}}$ are macro MAE scaled by 10^2. Leakage is scaled by 10^4. Lower is better.

(a) Effect of permission-grounded execution.

Control	Target CF	Protected CF	Overlap CF	Leakage
Mask only	3.875	8.428	11.700	3.05
Permission-only	1.460	5.720	6.310	2.06
Full	**0.970**	**1.563**	**2.094**	**1.97**

(b) Effect of protection constraints on preservation.

Control	Protected CF	$\mathcal{E}_{\text{keep}}$
Caster–Shadow–Receiver	6.682	5.20
+ Remove/preserve action	6.550	5.05
+ Protection constraints	2.250	1.25
+ **Permission grounding (Full)**	**1.563**	**0.78**

(c) Robustness to imperfect relation supports.

Relation support	**Target**	**Protected**	**Overlap**	**Leakage**
Clean	0.970 ± 0.026	1.563 ± 0.047	2.094 ± 0.066	1.97 ± 0.05
5% **boundary perturbation**	1.015 ± 0.029	1.640 ± 0.052	2.185 ± 0.071	2.02 ± 0.06
10% **boundary perturbation**	1.075 ± 0.033	1.755 ± 0.060	2.330 ± 0.083	2.10 ± 0.07
20% **boundary perturbation**	1.220 ± 0.043	2.010 ± 0.076	2.680 ± 0.108	2.28 ± 0.09
Receiver removed, pooled	1.020 ± 0.030	1.635 ± 0.055	2.185 ± 0.074	2.05 ± 0.06

Protocol. The evaluation uses the same 26 overlap-valid scenes and 352 frozen actions as the primary overlap-valid diagnostic. For each training seed, five perturbation realizations are first averaged within each condition to obtain a seed-level aggregate. The table then reports the mean and sample standard deviation over five independently trained seeds. For Clean and Receiver-removed conditions without stochastic perturbations, all test actions are aggregated within each seed before computing statistics across seeds.

$G_\sigma * z$ with $\sigma = 3$ pixels at 256×256 resolution. We form

$$\widetilde{M}_\tau = \mathbb{I}[d_M + \tau\xi \geq 0], \tag{S19}$$

and choose τ by binary search so that $|M \triangle \widetilde{M}_\tau|/|M|$ matches the prescribed budget within 0.5 percentage points. Thus, 5%, 10%, and 20% denote changed pixels relative to the original mask area, not to the contour length. All dense controls, geometry cues, and permission supports are recomputed from the perturbed masks. The action, model weights, 352-action test subset, clean counterfactuals, and clean metric supports remain fixed.

Each of the five independently trained checkpoints is evaluated with five fixed perturbation seeds: 1103, 2207, 3319, 4421, and 5531. Metrics are first averaged over the five perturbation draws and the fixed actions within each training seed. Table S38 then reports the mean and sample standard deviation across the five resulting training-seed aggregates. This hierarchy treats trained checkpoints, rather than boundary draws, as the independent experimental units.

Notes. Panels (a) and (b) use the same 352 overlap-valid actions as Table 2(a) in the main paper. Mask only receives X and the global union mask M_s. Its output is invariant to a and is rescored against every action-specific target. Permission-only adds the compiled action-dependent permission mask to the image and union-mask inputs. It receives no per-relation tokens and no standalone protection channels. Following the regional metric definition in Supplementary Section A.12.1, Protected CF compares the prediction with the renderer counterfactual on $M^{\text{ex}}_{\text{protect}}$, whereas $\mathcal{E}_{\text{keep}}$ compares it with the observed input on the same exclusive protected support. The references, not the supports, differ. If the renderer target equals the input there, the two scores must coincide under identical normalization and aggregation. Panel (c) changes only the relation supports supplied at evaluation and follows the nested aggregation described above. Receiver removed sets the receiver-support input to zero before pooled evaluation.

Table S38(a) places the Permission-only and Full results from Table S34 alongside Mask-only. Permission-only improves all four metrics over Mask-only, and Full further improves correction, preservation, overlap reconstruction, and locality. The strongest relative gains over Permission-only occur in Protected and Overlap, consistent with the role of the complete relation-control configuration in preserving competing shadow contributions.

Protected-reference consistency. Both Protected CF and $\mathcal{E}_{\text{keep}}$ use $M^{\text{ex}}_{\text{protect}}$. For a fixed action, define

$$\begin{aligned} D_{\text{ref}}(a) &= \text{MAE}(Y^{\star}_a, X; M^{\text{ex}}_{\text{protect}}), \\ |\text{Protected}(a) - \mathcal{E}_{\text{keep}}(a)| &\le D_{\text{ref}}(a) \\ &\le \text{Protected}(a) + \mathcal{E}_{\text{keep}}(a). \end{aligned} \tag{S20}$$

The bounds follow from the reverse and forward triangle inequalities. Because all configurations use the same input, counterfactual reference, exclusive support, and valid-action set $\mathcal{A}$, the reference discrepancy is shared across configurations. In the displayed 10^2 units, let

$$\bar{D}_{100} = \frac{100}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} D_{\text{ref}}(a), \qquad \bar{P}_j = 100\,\overline{\text{Protected}}_j, \qquad \bar{K}_j = 100\,\overline{\mathcal{E}_{\text{keep}}}_j.$$

A necessary condition across the reported configurations is

$$\max_j |\bar{P}_j - \bar{K}_j| \;\le\; \bar{D}_{100} \;\le\; \min_j (\bar{P}_j + \bar{K}_j). \tag{S21}$$

Using the displayed entries in Table S38(b), the four lower bounds are 1.482, 1.500, 1.000, and 0.783, while the upper bounds are 11.882, 11.600, 3.500, and 2.343. Ignoring display rounding, their intersection is

$$1.500 \le \bar{D}_{100} \le 2.343.$$

This is an algebraic consistency interval. Independently evaluating the input/reference pairs on the same exclusive protected supports and valid actions gives $\bar{D}_{100} = 2.038$, or raw RGB MAE 0.02038. This value lies within $[1.500, 2.343]$ and satisfies the joint aggregate triangle bounds for all four configurations. In particular, the Full difference $|1.563 - 0.780| = 0.783$ is compatible with the measured reference discrepancy.

Table S39: **Protected-reference endpoints and the reported Full result.** All entries use the same exclusive protected supports and valid actions, in RGB action-macro MAE $\times 100$ units. The first two rows are reference identities implied by the independently measured $\bar{D}_{100} = 2.038$; they are not additional trained-model comparisons.

Output on $M^{\text{ex}}_{\text{protect}}$	Protected CF	Keep	Interpretation
Input X	2.038	0.000	Input-preserving endpoint
Reference $Y^{\star}_a$	0.000	2.038	Reference-matching endpoint
Full	1.563	0.780	Measured model output

The discrepancy D_{ref} contains no model prediction and is common to all configurations evaluated on these input/reference pairs. Table S39 separates the two reference-dependent goals: matching the input and matching the renderer reference. Neither requires both errors to vanish when the references differ. The reference-matching endpoint is an identity on the evaluation support, not a claim that the output gate permits exact matching there. The reported Full result satisfies

$$|1.563 - 0.780| = 0.783 \le 2.038 \le 1.563 + 0.780 = 2.343.$$

This check does not identify the cause of the reference discrepancy. If a target-construction rule instead copies X throughout $M^{\text{ex}}_{\text{protect}}$, it implies $D_{\text{ref}} = 0$ and identical CF and keep errors under the same normalization and aggregation.

The nonzero discrepancy means that the renderer reference and input are not pixel-identical on the exclusive protected support. Protected CF therefore measures fidelity to that reference, whereas $\mathcal{E}_{\text{keep}}$ measures preservation of the observed input; neither should be substituted for the other. Overlap cannot explain their difference because it is excluded from both masks. This aggregate check establishes numerical compatibility, not the physical cause of the reference change or per-action consistency of every export. The latter requires the action-level renderer and mask records. This input-preservation diagnostic is distinct from the DESOBA final-output keep loss, whose gradient is zero where hard composition copies X.

At the mean level, support perturbation produces monotonic but controlled degradation. A 5% budget increases Target, Protected, Overlap, and Leakage by 4.6%, 4.9%, 4.3%, and 2.5%, respectively.

At 20%, the Full model still yields lower Target, Protected, and Overlap errors than the permission-only control, although Leakage rises from 1.97 to 2.28. Removing receiver evidence increases the four pooled means by 4.1–5.2%. This Full-model intervention is not numerically identical to the earlier Caster–Shadow versus Caster–Shadow–Receiver decomposition in Table 2(b) in the main paper, because the remaining action, protection, and permission cues differ. The results support graceful degradation under moderate support error, not invariance to inaccurate annotations.

A.13 Frozen action census and cardinality analysis

The complete Kubric suite contains 36 scenes and 472 nontrivial action–scene pairs. The 26 Multi-caster and Shared-overlap scenes supply the primary pool of 352 overlap-valid actions. The other 10 scenes supply 120 receiver-specific actions, for which Overlap is undefined. Each scene contributes all binary actions with $0 < \|a\|_1 < N$; all-off and all-on endpoints are excluded from these error means.

Table S40: **Frozen Kubric action census.** Counts enumerate action–scene pairs, not independent scenes. The overlap-valid pool combines Multi-caster and Shared overlap.

Scope	Scenes	One selected	Two selected	Three selected	Actions
Multi-caster	15	58	84	56	198
Shared overlap	11	44	66	44	154
Overlap-valid pool	**26**	**102**	**150**	**100**	**352**
Receiver-specific	10	37	51	32	120
Full Kubric split	**36**	**139**	**201**	**132**	**472**

The Multi-caster group contains one $N = 2$ and fourteen $N = 4$ scenes; Shared overlap contains eleven $N = 4$ scenes; Receiver-specific contains one $N = 2$, one $N = 3$, and eight $N = 4$ scenes. Thus the full suite has two $N = 2$, one $N = 3$, and thirty-three $N = 4$ scenes:

$$2(2^2 - 2) + (2^3 - 2) + 33(2^4 - 2) = 472.$$

The overlap-valid cardinalities follow $(102, 150, 100) = (2, 0, 0) + 25(4, 6, 4)$. In particular, the extra two one-selected actions come from the $N = 2$ scene.

Table S41: **Cardinality-stratified relation-selective execution on Kubric.** Full-model results use the same 352 overlap-valid action–scene pairs as the primary overlap-valid diagnostic. Target is evaluated on the full M_{remove}, including M_{overlap}; Protected uses $M^{\text{ex}}_{\text{protect}}$, and Overlap uses M_{overlap}. The first three errors are RGB macro MAE scaled by 10^2; Leakage is scaled by 10^4. Lower is better.

Selected relations	Actions	Target	Protected	Overlap	Leakage
One	102	0.930	1.500	2.000	1.900
Two	150	0.980	1.580	2.130	1.980
Three	100	0.996	1.600	2.136	2.026
Pooled	**352**	**0.970**	**1.563**	**2.094**	**1.970**

For each metric m, action-macro aggregation on this same valid set gives

$$\overline{m}_{\text{pooled}} = \frac{102\overline{m}_1 + 150\overline{m}_2 + 100\overline{m}_3}{352}. \tag{S22}$$

Means are computed before display rounding. The displayed strata reproduce the main Full row, $0.970/1.563/2.094/1.970$, to three decimals. Undefined empty-support errors are never replaced by zero; if any endpoint uses a different valid set, its actual stratum counts must be reported and used in this identity.

Errors increase modestly across the reported strata, but the strata differ in scene content, support area, and overlap difficulty. The Kubric cardinality slices characterize evaluated joint actions; they do not test held-out action signatures. The separate DESOBA extension reports its own signature-holdout audit above.

B EXPERIMENTAL AND IMPLEMENTATION DETAILS

This appendix specifies the data splits and benchmark-dependent model paths. The annotation extension is documented in Appendix C, the frozen evaluation protocol in Appendix D.4, and all seed-level results in Appendix E.2.

A separate checkpoint is trained and evaluated for each benchmark. The three real-image benchmarks use the same appearance-prior pathway: Depth Anything V2 provides monocular depth evidence and DINOv2 provides dense semantic features (Yang et al., 2024; Oquab et al., 2024). The priors are generated independently from the images of the corresponding benchmark and are supplied with `use_priors=True`. ISTD+ and SRD do not provide explicit caster–shadow–receiver correspondences, so their benchmark-specific models retain the appearance priors but disable the relation pathway: `use_priors=True` and `use_relation=False`. Relation-enabled DESOBA Full uses `use_relation=True` in editing, component analysis, and the all-remove benchmark whenever valid relations are available. The matched Mask control uses `use_relation=False`. Selective training uses the action-specific references on the partial-target manifest and the deterministic mask-composed fallback on the matched fallback branch; the global all-remove experiments use the saved reference $Y^{\mathrm{sf}} = I_d$. No DESOBA v2 image, relation annotation, prior, or checkpoint is transferred to ISTD+ or SRD.

Baseline inference and the common post-inference scorer are deliberately separated. Each method retains its official architecture and inference path. All saved predictions are then evaluated with the frozen scorer specified in Appendix D.4. This design standardizes metrics without imposing one model constructor or inference procedure on incompatible architectures.

Table S42: **Benchmark-specific model scope.** ISTD+ (Le & Samaras, 2019) and SRD (Qu et al., 2017) use independently trained relation-free models. DESOBA Full enables relation conditioning for selective editing and all-remove restoration; Mask disables it.

Dataset	Appearance prior	Relation pathway	Training data
ISTD+	enabled	disabled	ISTD+ only
SRD	enabled	disabled	SRD only
DESOBA v2 (editing/components)	enabled	enabled for $N \geq 1$	prepared DESOBA with relations
DESOBA v2 (global Full)	enabled	enabled for $N \geq 1$	prepared DESOBA with relations
DESOBA v2 (Mask control)	enabled	disabled	same prepared DESOBA image/target pairs

B.1 DATASETS AND SPLITS

Kubric counterfactual split and action construction. The frozen suite contains 36 test scenes with $N \in \{2, 3, 4\}$ valid relations: two $N = 2$, one $N = 3$, and thirty-three $N = 4$ scenes. Capacity is $K_{\mathrm{K}} = 4$, so no scene is truncated. Enumerating all nontrivial binary actions yields 472 action–scene pairs. The primary diagnostic and equal-information comparison use the 26 overlap-valid scenes and their 352 actions; the separate receiver-specific diagnostic uses 10 scenes and 120 actions without an Overlap score. All-off and all-on actions are used only for endpoint audits. Controls share the frozen renderer targets, evaluation supports, and action identities. The reported scene-ID audit finds zero training–test intersections and zero validation–test intersections; test performance is not used for checkpoint selection. The complete census and Full-model cardinality aggregation appear in Appendix A.13.

The independent Kubric-Test-B collection contains 200 scenes and 2,400 action–scene pairs and is disjoint from Test-A; it is reserved for the scene-bootstrap audit. For every split, the manifest fixes the scene ID, caster–shadow–receiver relation graph, camera, lighting, materials, and renderer seed. The renderer exports the observed all-shadow image, relation supports, and one counterfactual for every nontrivial binary action. Training and validation identities are disjoint from both test collections, and neither test target nor test action mask is read during optimization or checkpoint selection. Bootstrap and permutation resampling always samples scenes with replacement and carries all valid actions of each sampled scene together.

All-off identity and unchanged output outside the hard permission support are execution invariants:

$$\widehat{Y}_{\mathbf{0}} = X, \qquad (1 - M_{\mathrm{perm}}) \odot (\widehat{Y}_a - X) = 0.$$

They check implementation of the declared constraints rather than learned counterfactual fidelity, which is assessed by the regional errors inside the relevant evaluation supports. These identities do not assert that the renderer counterfactual itself equals X on every protected pixel; that separate reference comparison is defined in Eq. (S20).

ISTD+. ISTD+ is the adjusted-target version of ISTD (Le & Samaras, 2019) and preserves the original split: 1,330 training triplets and 540 test triplets. A triplet consists of a shadowed RGB image, a binary shadow mask, and an adjusted shadow-free target. The `-plus` test option selects the official adjusted test targets.

SRD. SRD (Qu et al., 2017) contains 2,680 training pairs and 408 test pairs. Because the original release does not provide official shadow masks, region-wise evaluation uses the public DHAN masks (Cun et al., 2020), following the established HomoFormer/SRD protocol. Training and evaluation use only the official SRD split.

DESOBA v2. DESOBA v2 (Liu et al., 2024a) contains 28,573 paired 256×256 records in the audited release: 27,823 training records and 750 test records. DESOBA v2 provides the foreground object, foreground shadow, background object, and background shadow channels from which we derive the structured relation layer described in Section C. All records, including those without a reliable relation under our conservative filter, remain in the archive. Our experimental extension additionally contains the action-specific reference action-specific partial-target track. Its manifest contains independently measured or manually constructed action-conditioned action-specific references for the 1,600-image training split, the 200-image validation split, and the 200-image held-out split. This track is distinct from both the released relation archive and the approximate global reference I_d: the original global target is never relabeled as a partial counterfactual. The training targets are used only for optimization, and the source-disjoint held-out partial-target split is reserved for quantitative selective-edit evaluation. It is a separate annotation/evaluation track from the 750-image global all-remove reference track, even when both tracks draw images from the same DESOBA release. The held-out partial test is a 200-image source-disjoint subset of the 750-image global test list; it is not a second independent DESOBA release. Denoting the two partial-target subsets by $\mathcal{D}^{\mathrm{tr}}_{\mathrm{part}}$ and $\mathcal{D}^{\mathrm{te}}_{\mathrm{part}}$, the latter is held out from partial-target training and checkpoint selection.

The evaluation unit is an input–action pair with its action-specific reference target Y_a^{part}. Reference validity and regional scoring are defined in Appendix A.3. Records outside the partial-target manifest use the deterministic fallback surrogate during training and are excluded from the held-out partial-target scores. The held-out targets are not read by training, checkpoint selection, or action compilation. The base archive census applies to source-aligned annotation records, not to the separate multi-action extension. That extension has its own executable relation/action manifest and target rows. The relation-annotation archive described in Section C documents the conditioning inputs; the real-image action-specific references are an additional supervision source. They are labeled independently measured or manually constructed references for this real-image protocol; their construction is independent of the Kubric renderer and they are not renderer-exact Kubric references.

The targets in this track are labeled independently measured or manually constructed action-conditioned references produced by a separate target-generation workflow and fixed after provenance review. The review records document audit decisions; they do not redefine these targets as approximate global references. They are reported as real-image action-specific references; they are independent of the Kubric renderer and are not renderer-exact Kubric targets. The source manifest records the original source identity, derived record IDs, partial split, relation IDs, target version, review status, and the valid-reference mask. All derived records from a validation or held-out source are removed from every training path, including the global I_d path. Writing S_{tr}, S_{val}, and S_{te} for the original source-identity sets, the source intersections satisfy

$$S_{\mathrm{tr}} \cap S_{\mathrm{val}} = S_{\mathrm{tr}} \cap S_{\mathrm{te}} = S_{\mathrm{val}} \cap S_{\mathrm{te}} = \varnothing.$$

The action-specific target Y_a^{part} is generated from the original real-image source and the requested action by this separate workflow; it is not copied from the saved approximate global reference I_d,

the composite input I_c, the global target Y^{sf}, a ShadowEdit prediction, or a permission-mask composition. The workflow exports Y_a^{part} and V_a^{part} together, and the relation masks are used only to compile the requested action and its scoring supports. Any post-review correction creates a new target version and hash rather than silently replacing the old file. We denote the action-specific reference by Y_a^{part} without imposing an additional input-copy constraint on its exclusive protected pixels. The output gate constrains $\widehat{Y}_a$, not the action-specific reference. reference. Target construction and execution permission are separate: the latter uses only the supplied relations, action, and fixed morphology. The actual generator, inputs, processing order, version, and review records must accompany the target manifest; the reported regional scores alone do not identify the cause of an input–reference difference.

Table S43: **Source-level census of the DESOBA action-specific-reference track.** Counts are after source deduplication and refer to representative images, relation groups, and valid partial-target actions. The 200 held-out images are drawn from the original 750-image global test list.

Track	Sources	Images	Relation groups	Partial targets actions
Training	1,600	1,600	3,200	3,200
Validation	200	200	400	400
Held-out test	200	200	400	400
Total	**2,000**	**2,000**	**4,000**	**4,000**

The held-out test actions are one-selected actions: each two-relation image contributes $(1, 0)$ and $(0, 1)$, whereas $(0, 0)$ is all-keep and $(1, 1)$ is all-remove. The 120 no-overlap and 80 overlap-valid images therefore contain 240 and 160 one-selected actions, respectively, for 400 actions in total; no two-selected DESOBA action is included in this primary held-out table. All 400 Target, Protected, and Leakage actions have valid reference/support entries, and the 160 overlap actions have valid overlap support; actions with empty Overlap support are excluded from the Overlap mean rather than assigned zero. The four three-relation archive records can yield at most twelve two-selected candidate actions before source, target, and support filtering. This is an availability bound, not a reported twelve-action experiment or evidence of broad two-selected generalization.

B.2 Benchmark-specific model configurations

For relation-annotated DESOBA v2 samples, the relation pathway is enabled only when valid relation annotations pass the filtering criteria.

B.2.1 Shared DINOv2 and Depth Anything V2 appearance-prior pathway

The frozen content priors depend only on the observed image:

$$(D, H_{\mathrm{sem}}, Q_D, Q_S) = \mathcal{P}_{\mathrm{frozen}}(X). \tag{S23}$$

Depth Anything V2 Small/ViT-S produces $D \in [0, 1]^{1\times256\times256}$ from a 518×518 inference view. DINOv2 ViT-S/14 produces $H_{\mathrm{sem}} \in \mathbb{R}^{384\times18\times18}$ from a 252×252 view. Semantic tokens are L2 normalized and resampled to the configured 32×32 model grid. Both extractors remain frozen.

Reliability is estimated from six aligned views formed by three scales $(1.0, 0.875, 1.125)$ and identity or horizontal flip. Let D_t and H_t denote predictions mapped back to the input coordinates. Their disagreement maps are

$$\begin{aligned} V_D &= |\mathcal{T}|^{-1} \sum_{t\in\mathcal{T}} (D_t - \overline{D})^2, \\ V_S &= |\mathcal{T}|^{-1} \sum_{t\in\mathcal{T}} \|H_t - \overline{H}\|_2^2. \end{aligned} \tag{S24}$$

Percentile normalization maps each disagreement field to $[0, 1]$. Quality is then

$$Q_D = \exp[-\gamma_D \operatorname{Norm}(V_D)], \qquad Q_S = \exp[-\gamma_S \operatorname{Norm}(V_S)], \tag{S25}$$

with $\gamma_D = 3.0$ and $\gamma_S = 1.5$. The cached float16 tensors are loaded as float32. Prior generation uses batch size one.

Only the spatial focus is task dependent. Let $f_{\text{global}} = \mathbb{I}[\|a\|_1 = N]$ for relation-aware data and $f_{\text{global}} = 1$ for relation-free benchmarks. The focus mask and its morphological cues are

$$\begin{aligned} M_{\text{foc}} &= f_{\text{global}} M_s + (1 - f_{\text{global}}) M_{\text{remove}}, \\ W_s &= \omega_0 + (1 - \omega_0)\delta_{10}^{\text{sq}}(M_{\text{foc}}), \\ B_s &= \text{clip}_{[0,1]}\left[\delta_7^{\text{sq}}(M_{\text{foc}}) - \varepsilon_7^{\text{sq}}(M_{\text{foc}})\right]. \end{aligned} \tag{S26}$$

Here δ_r^{sq} and $\varepsilon_r^{\text{sq}}$ denote dilation and erosion with a square $(2r+1) \times (2r+1)$ structuring element. The reported setting uses $\omega_0 = 0$, so W_s uses a 21×21 dilation and B_s uses a 15×15 morphological gradient. Thus, the frozen prior content is action independent, while deterministic morphology selects where that content is emphasized.

The appearance encoder and confidence gate are

$$\begin{aligned} F_{\text{app}}^0 &= E_{\text{app}}(D, H_{\text{sem}}, \nabla D, \nabla H_{\text{sem}}, W_s, B_s), \\ Q &= \sqrt{\text{clip}(Q_D \odot Q_S, \epsilon, 1)}, \\ c &= Q \odot \sigma\big(h_c(F_{\text{app}}^0)\big), \\ F_{\text{app}} &= c \odot F_{\text{app}}^0. \end{aligned} \tag{S27}$$

The confidence field is a bounded control signal, not calibrated uncertainty. At decoder level ℓ,

$$\big(F_\ell^R, F_\ell^{\text{app}}, M_{\text{perm}}^\ell\big) = \text{Adapt}_\ell(F^R, F_{\text{app}}, M_{\text{perm}}). \tag{S28}$$

The base path returns

$$\big(Y_g, \{F_\ell\}_{\ell=1}^L\big) = f_{\text{base}}(X, M_s, F_{\text{app}}). \tag{S29}$$

The reported configuration uses a 16-dimensional prior embedding, 384-dimensional semantic features, zero prior dropout at evaluation, a decoder modulation bound of 0.02, and an RGB prior-residual bound of 0.05.

B.2.2 Relation-aware DESOBA v2 pathway

For relation-annotated DESOBA v2, let

$$\mathcal{R} = \{r_i\}_{i=1}^N, \qquad r_i = (C_i, S_i, R_i), \qquad a \in \{0,1\}^N, \tag{S30}$$

where $a_i = 1$ selects relation i for removal and $a_i = 0$ requests its preservation. Define

$$M_{\text{remove}} = \mathbb{I}\left[\bigcup_{i:a_i=1} S_i\right], \qquad M_{\text{protect}} = \mathbb{I}\left[\bigcup_{i:a_i=0} S_i\right]. \tag{S31}$$

The implementation uses "selected" to denote the task-addressed relation. This relation may be assigned either removal or preservation and should not be confused with the action-selected mask M_{remove}. For the reported task presets,

$$\begin{aligned} f_{\text{keep}} &= \mathbb{I}[\texttt{preserve_selected}], \\ f_{\text{remove}} &= \mathbb{I}[\texttt{remove_selected}], \\ f_{\text{global}} &= \mathbb{I}[\texttt{global}], \\ M_{\text{keep}} &= f_{\text{keep}} M_{\text{sel}}^S. \end{aligned} \tag{S32}$$

M_{sel}^S denotes the shadow of the relation named by the task. Thus, M_{keep} is nonzero only for `preserve_selected`. It is neither $1 - M_{\text{remove}}$ nor the permission complement.

DESOBA permission compiler. Let Ω denote the full image domain. The hard support is

$$U_a^{\text{D}} = \begin{cases} \varnothing, & \|a\|_1 = 0, \\ \text{supp}(\delta_4^{\text{sq}}(M_{\text{remove}})) \setminus \text{supp}(M_{\text{protect}}^{\text{ex}}), & 0 < \|a\|_1 < N, \\ \Omega, & \|a\|_1 = N, \end{cases} \qquad M_{\text{perm}}^{\text{D}} = \mathbb{I}[U_a^{\text{D}}]. \tag{S33}$$

Here U_a^{D} is a set of pixel locations, whereas $M_{\text{perm}}^{\text{D}}$ is its binary mask representation; the set difference is therefore taken before rasterization. The compiler is evaluated from the supplied relations,

action, and fixed morphology only, and never reads the held-out partial target. The selective path therefore uses a 9×9 square dilation. Only the exclusive protected support is removed from this writable region; selected–protected overlap remains writable under the compiled action. A 9×9 average-pool feather supplies the soft write weight inside this hard support. This compiler uses the relation/action masks only and does not read the partial target. The all-off endpoint returns X. At the all-on endpoint, $U_a^{\mathrm{D}} = \Omega$, so the same relation-aware restoration branch operates without a spatial permission restriction.

Kubric permission compiler. The Kubric experiment uses a frozen proposal and a topology-v3 corrector. Let τ denote the intervention topology. SCMR is the receiver-specific case with one caster and multiple receivers. MCSR is the multi-caster shared-receiver case. Shared denotes explicit selected–protected pixel overlap. For partial actions, define

$$
\begin{aligned}
R_a &= \bigcup_{i:a_i=1} R_i, & C_a &= \bigcup_{i:a_i=1} C_i, \\
C_{\mathcal{R}} &= \bigcup_{i:v_i=1} C_i, \\
U_a^{\mathrm{prop}} &= \Big[\mathrm{supp}(\delta_4^{\mathrm{sq}}(M_{\mathrm{remove}})) \cap \big(\mathrm{supp}(M_{\mathrm{remove}}) \cup R_a\big)\Big] \\
&\quad \setminus \big(\mathrm{supp}(M_{\mathrm{protect}}) \cup C_a\big), \\
R_\tau &= \begin{cases} R_a, & \tau = \mathrm{SCMR}, \\ \bigcap_{i:v_i=1} R_i, & \tau \in \{\mathrm{MCSR}, \mathrm{Shared}\}, \end{cases} \\
U_a^{\mathrm{corr}} &= \Big[\mathrm{supp}(\delta_{31}^{\mathrm{sq}}(M_{\mathrm{remove}})) \cap R_\tau\Big] \setminus C_{\mathcal{R}}, \\
U_a^{\mathrm{K}} &= U_a^{\mathrm{prop}} \cup U_a^{\mathrm{corr}}.
\end{aligned}
\tag{S34}
$$

The proposal kernel is 9×9. The corrector hard-support kernel is 63×63, and its inner feather field is 31×31. All structuring elements are square. Endpoint composition again sets $U_a^{\mathrm{K}} = \varnothing$ for all-off and $U_a^{\mathrm{K}} = \Omega$ for all-on. These definitions separate the two benchmark implementations while preserving the common endpoint-aware operator $\Gamma_{\mathcal{B}}$ used in the main text.

The remainder of this subsection describes the integrated DESOBA pathway and uses $M_{\mathrm{perm}} = M_{\mathrm{perm}}^{\mathrm{D}}$.

Dense control stack. The DESOBA pathway consumes the following physical channel order:

$$
\begin{aligned}
\mathcal{C} = \mathrm{Cat}[&C_{\mathrm{all}}, S_{\mathrm{all}}, C_{\mathrm{sel}}, S_{\mathrm{sel}}, C_{\mathrm{oth}}, S_{\mathrm{oth}}, R_{\mathrm{sel}}, \\
&M_{\mathrm{remove}}, M_{\mathrm{protect}}, M_{\mathrm{keep}}, B_{\mathrm{rem}}, B_{\mathrm{pro}}, G_C, G_S, L_{CS}, K_{CS}, \\
&f_{\mathrm{keep}}\mathbf{1}, f_{\mathrm{remove}}\mathbf{1}, f_{\mathrm{global}}\mathbf{1}] \in \mathbb{R}^{19 \times H \times W}.
\end{aligned}
\tag{S35}
$$

The entity, action, and grounding factors contribute 7, 6, and 6 channels. The dense control feature is

$$
F^R = E_R(\mathrm{Cat}[X, M_s, \mathcal{C}])\,. \tag{S36}
$$

The action flags are broadcast spatially. Channel names containing "selected" refer to the task-addressed relation. Removal and preservation semantics are carried by M_{remove}, M_{protect}, and the three flags.

Let $\partial_r^{\mathrm{sq}} M = \mathrm{clip}_{[0,1]}[\delta_r^{\mathrm{sq}}(M) - \varepsilon_r^{\mathrm{sq}}(M)]$. The grounding channels are

$$
\begin{aligned}
B_{\mathrm{rem}} &= \partial_4^{\mathrm{sq}} M_{\mathrm{remove}}, & B_{\mathrm{pro}} &= \partial_4^{\mathrm{sq}} M_{\mathrm{protect}}, \\
G_A(p) &= \max\left(0, 1 - \frac{d_2(p, A)}{r_p}\right), & r_p &= \mathrm{clip}(\mathrm{round}(0.12 \min(H, W)), 24, 96)\,, \\
L_{CS} &= \mathrm{RasterLine}(\mu(C_{\mathrm{sel}}), \mu(S_{\mathrm{sel}}), w_L), & w_L &= \max(1, \mathrm{round}(0.012 \min(H, W)))\,, \\
K_{CS} &= \delta_8^{\mathrm{sq}}(C_{\mathrm{sel}}) \odot \delta_8^{\mathrm{sq}}(S_{\mathrm{sel}}).
\end{aligned}
\tag{S37}
$$

G_C and G_S instantiate G_A with the selected caster and shadow. The distance transform is Euclidean and uses a 3×3 OpenCV mask. At 256×256, $r_p = 31$ pixels and $w_L = 3$ pixels. Boundary and contact channels use 9×9 and 17×17 square kernels. Every grounding channel is one-channel and normalized to $[0, 1]$. K_{CS} is a derived local contact cue, not physical contact supervision.

Relation slots and descriptors. Each relation slot carries five spatial masks,

$$\mathcal{Q}_i = (M_i^C, M_i^S, M_i^R, M_i^L, M_i^K, a_i, v_i), \tag{S38}$$

where the relation-specific line and contact masks are

$$M_i^L = \mathrm{RasterLine}(\mu(C_i), \mu(S_i), w_L), \qquad M_i^K = \delta_8^{\mathrm{sq}}(M_i^C) \odot \delta_8^{\mathrm{sq}}(M_i^S). \tag{S39}$$

They use the same line width and square structuring element as the dense grounding channels. The 13-channel pair input is

$$\chi_i = \mathrm{Cat}[X, M_{\mathrm{foc}}, M_i^C, M_i^S, M_i^R, M_i^L, M_i^K, R_{\mathrm{sel}}, M_{\mathrm{remove}}, M_{\mathrm{protect}}, a_i \mathbf{1}]. \tag{S40}$$

Let $A_i = \max(M_i^C, M_i^S, M_i^L, M_i^K)$ and $\alpha_i = \max(|A_i|/(HW), 10^{-6})$. From the observed image, define

$$E_X(p) = \frac{\sqrt{\mathrm{mean}_c[(\nabla_x X_c)^2 + (\nabla_y X_c)^2]}}{\max_q \sqrt{\mathrm{mean}_c[(\nabla_x X_c(q))^2 + (\nabla_y X_c(q))^2]} + \epsilon}. \tag{S41}$$

The exact six-dimensional descriptor is

$$g_i = \left[\alpha_i, \log \alpha_i, a_i, v_i, \mathrm{mean}_{A_i} E_X, \mathrm{mean}_{\partial_3^{\mathrm{sq}} A_i} E_X\right] \in \mathbb{R}^6. \tag{S42}$$

The logarithm is natural. The boundary statistic uses a 7×7 square kernel. The token is

$$t_i^0 = \mathrm{LN}[\mathrm{Pool}(E_{\mathrm{pair}}(\chi_i), A_i) + e_{\mathrm{type}} + \phi_f(g_{i,1:4}) + \phi_e(g_{i,5:6})]\,. \tag{S43}$$

DESOBA uses $K_{\mathrm{D}} = 8$ slots. The addressed relation is placed first, then the remaining relations are sorted by decreasing shadow area. Unused slots are zero padded and masked. The base annotation archive has at most three relation groups per image; this census applies to the base-archive track. The separate multi-action extension is checked against its own relation/action manifest and includes the four executable relation IDs enumerated there. If $N > 8$, the loader keeps the addressed relation and the seven largest remaining shadows. Kubric uses $K_{\mathrm{K}} = 4$ with $N \in \{2, 3, 4\}$. It rejects an over-capacity sample rather than truncating it. Two masked self-attention layers with four heads refine the DESOBA tokens. No Hungarian matching is used. Let

$$T^0 = [t_1^0, \ldots, t_K^0], \qquad T' = \Psi(T^0; v) = [t_1', \ldots, t_K'], \tag{S44}$$

Here $K = K_{\mathrm{D}} = 8$ for DESOBA and $K = K_{\mathrm{K}} = 4$ for Kubric; Ψ denotes the two-layer masked self-attention encoder in the integrated DESOBA pathway.

At decoder scale ℓ, image features query the valid relation tokens:

$$\begin{aligned} \beta_{\ell i}(p) &= \mathrm{Softmax}_{i:v_i=1}\left(\frac{q_\ell(F_\ell(p))^\top k(t_i')}{\sqrt{d}}\right), \\ Z_\ell(p) &= \sum_{i:v_i=1} \beta_{\ell i}(p) W_V t_i'. \end{aligned} \tag{S45}$$

The grounded evidence is converted into bounded scale and bias fields and is applied only inside the permission support:

$$\begin{aligned} (s_\ell, b_\ell) &= \eta_\ell \tanh h_\ell\left(\mathrm{Cat}[F_\ell, Z_\ell, F_\ell^R, F_\ell^{\mathrm{app}}]\right), \\ \widetilde{F}_\ell &= F_\ell \odot (1 + M_{\mathrm{perm}}^\ell \odot s_\ell) + M_{\mathrm{perm}}^\ell \odot b_\ell. \end{aligned} \tag{S46}$$

The relation pathway predicts a bounded residual

$$\Delta Y_a = \eta_R \tanh\left[h_R(F^R, T', \{\widetilde{F}_\ell\}_\ell)\right]. \tag{S47}$$

and the final output follows the same range-preserving projection as in the main text:

$$\widehat{Y}_a = (1 - M_{\mathrm{perm}}) \odot X + M_{\mathrm{perm}} \odot \mathrm{clip}_{[0,1]}\left(Y_g + \Delta Y_a\right). \tag{S48}$$

Consequently,

$$(1 - M_{\mathrm{perm}}) \odot (\widehat{Y}_a - X) = 0, \qquad \widehat{Y}_a \in [0, 1]^{3 \times H \times W}. \tag{S49}$$

B.2.3 TRAINING OBJECTIVE

Our prepared DESOBA version uses the saved approximate deshadowed image I_d as Y^{sf}, not the object-level composite I_c (Section C). The partial-target branch uses action-specific partial references on its training split and a disjoint held-out partial-target test split. The matched fallback branch uses the existing mask-composed target as a deterministic supervision surrogate; this fallback does not redefine the execution permission or claim an exact overlap counterfactual. The selected-only supervision support used by that fallback is

$$M_{\mathrm{edit}} = M_{\mathrm{remove}} \odot (1 - M_{\mathrm{protect}}). \tag{S50}$$

Writing the action-specific reference as Y_a^{part}, the two training branches use:

$$Y_a^{\mathrm{tr}} = \begin{cases} Y_a^{\mathrm{part}}, & \text{partial-target branch,} \\ M_{\mathrm{edit}} \odot Y^{\mathrm{sf}} + (1 - M_{\mathrm{edit}}) \odot X, & \text{fallback branch.} \end{cases} \tag{S51}$$

Let $V_a^{\mathrm{tr}} = V_a^{\mathrm{part}}$ for the partial-target branch and $V_a^{\mathrm{tr}} = \mathbf{1}$ for the mask-composed fallback, with $V_a^{\mathrm{part}} = \mathbf{1}$ for a full-frame reference. These masks denote reference validity, not execution permission. The correction and context supports used by the loss are

$$M_{\mathrm{corr}} = \begin{cases} V_a^{\mathrm{part}} \odot M_{\mathrm{remove}}, & \text{partial-target branch,} \\ M_{\mathrm{edit}}, & \text{fallback branch,} \end{cases} \tag{S52}$$

$$M_{\mathrm{id}} = V_a^{\mathrm{tr}} \odot \left(1 - M_{\mathrm{corr}} - M_{\mathrm{protect}}^{\mathrm{ex}}\right). \tag{S53}$$

Here M_{corr} includes valid selected–protected overlap in the partial-target branch. In the fallback branch, selected–protected overlap is assigned the input as a conservative training surrogate, but it remains writable at execution time. Whenever Y_a^{part} is used, it replaces this surrogate, including its valid overlap annotations. For the fallback branch with all relations selected, $M_{\mathrm{protect}} = 0$ and $M_{\mathrm{edit}} = M_s$, giving

$$Y_a^{\mathrm{tr}} = M_s \odot Y^{\mathrm{sf}} + (1 - M_s) \odot X.$$

This target equals Y^{sf} when the paired images are identical outside M_s. This construction describes selective training, including its all-active action. The dedicated all-remove benchmark and component runs instead use the full saved $Y^{\mathrm{sf}} = I_d$ as the reconstruction and evaluation reference for both Mask and Full. They share the all-remove objective while changing relation conditioning. The held-out partial-target split is reserved for evaluation and is not used in training or checkpoint selection. Let $M_{\mathrm{protect}}^{\mathrm{ex}} = M_{\mathrm{protect}} \odot (1 - M_{\mathrm{remove}})$ and let ρ denote the Charbonnier penalty used by the implementation. For a binary mask M, define the masked reduction

$$\mathcal{R}_M(E) = \frac{\sum_{p,c} M(p)\rho(E_{pc})}{\sum_p M(p) + \epsilon}. \tag{S54}$$

The training objective is

$$\mathcal{L}_{\mathrm{rec}} = \mathcal{R}_{V_a^{\mathrm{tr}}}(\widehat{Y}_a - Y_a^{\mathrm{tr}}), \tag{S55}$$

$$\mathcal{L}_{\mathrm{rem}} = \mathcal{R}_{M_{\mathrm{corr}}}(\widehat{Y}_a - Y_a^{\mathrm{tr}}), \tag{S56}$$

$$\mathcal{L}_{\mathrm{id}} = \mathcal{R}_{M_{\mathrm{id}}}(\widehat{Y}_a - Y_a^{\mathrm{tr}}), \tag{S57}$$

$$\mathcal{L}_{\mathrm{keep}} = \mathcal{R}_{M_{\mathrm{protect}}^{\mathrm{ex}}}(\widehat{Y}_a - X), \tag{S58}$$

$$\mathcal{L} = \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{rem}}\mathcal{L}_{\mathrm{rem}} + \lambda_{\mathrm{id}}\mathcal{L}_{\mathrm{id}} + \lambda_{\mathrm{keep}}\mathcal{L}_{\mathrm{keep}}. \tag{S59}$$

On annotated records, $\mathcal{L}_{\mathrm{rem}}$ is evaluated on M_{corr}, so valid overlap pixels follow the partial target rather than being pulled back to X. The identity/context term is restricted to M_{id}, which excludes correction pixels and exclusive protected pixels, and follows the training reference. We retain the symbol $\mathcal{L}_{\mathrm{id}}$ for this reference-based context consistency term; it is an input-identity loss only where the reference equals X. The symbol λ_{id} denotes its loss weight. Invalid reference pixels contribute to neither reconstruction, correction, nor identity/context loss. The keep term is input-relative on $M_{\mathrm{protect}}^{\mathrm{ex}}$, where hard composition also enforces preservation. Under Eqs. (S33) and (S48), every pixel in $M_{\mathrm{protect}}^{\mathrm{ex}}$ is copied from X. Consequently, $\mathcal{L}_{\mathrm{keep}}$ evaluated on the final DESOBA output is constant with respect to the network parameters (zero or the fixed Charbonnier floor). Its presence in the objective does not by itself establish learned preservation. An additional loss before hard composition would be a different objective from the one written here.

ISTD+ and SRD disable relation conditioning and optimize $\mathcal{L} = \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}}$ with $Y_a^{\mathrm{tr}} = Y^{\mathrm{sf}}$. Kubric uses its exact renderer counterfactual as the reconstruction target.

B.2.4 CHECKPOINT SCOPE FOR ISTD+ AND SRD

The standard-benchmark checkpoints correspond to benchmark-specific instances of the prior-guided, relation-free configuration. Per-image depth, depth-quality, semantic, and semantic-quality tensors are loaded from benchmark-specific prior directories. Relation tensors are not requested. DINOv2 and Depth Anything V2 are external prior extractors and are not stored inside the restoration checkpoints. The small parameter-count difference arises from dataset-specific input adapters and configuration instantiation. The parameter counts therefore refer to the benchmark-specific restoration networks rather than to the external frozen prior extractors.

Table S44: **Checkpoint audit for the relation-free configurations.** The supplied ISTD+ (Le & Samaras, 2019) and SRD (Qu et al., 2017) checkpoints contain the evaluated restoration weights. Their small parameter-count difference reflects dataset-specific input adapters and configuration instantiation. Parameter counts exclude the external DINOv2 (Oquab et al., 2024) and Depth Anything V2 (Yang et al., 2024) extractors.

(a) Checkpoint metadata

Dataset	Checkpoint	Epoch	Parameters
ISTD+	`model_best_paper_epoch20.pth`	20	18,580,477
SRD	`model_best_paper.pth`	153	18,591,741

(b) Optimizer metadata

Dataset	Initial LR	Stored LR	Adam betas	Weight decay
ISTD+	2×10^{-4}	1×10^{-6}	$(0.9, 0.999)$	0
SRD	2×10^{-4}	$1.7113518519\times10^{-4}$	$(0.9, 0.999)$	0

The checkpoint files identify the evaluated restoration weights. They do not by themselves establish the metric protocol or the contribution of an external prior. Those claims are determined jointly by the data loader, forward configuration, evaluator, and ablation evidence.

C CONSTRUCTION OF THE EXTENDED DESOBA V2 CASTER–SHADOW–RECEIVER ANNOTATIONS

C.1 ANNOTATION SCOPE AND REPRESENTATION

Terminology used across the paper. Throughout the paper, "relation-annotated DESOBA v2" denotes the released DESOBA v2 tuples (Liu et al., 2024a) together with the audited relation extension described here. It does not imply that the original public release contains pixel-accurate receiver masks. Receiver supports are used as audited relation cues for controlled editing rather than as dense semantic segmentation annotations. The original release provides caster–shadow groups. The receiver support and model-ready relation metadata are part of our extension.

Separation between the original release and our extension. The original DESOBA v2 release and our extended relation annotations are two distinct stages. DESOBA v2 itself follows a semi-automatic data-construction pipeline: manually annotated SOBA shadow–object associations are used to train the SSISv2 instance-shadow detector. SSISv2 then predicts caster masks, shadow masks, and their instance-level associations in candidate outdoor images. A latent-diffusion inpainting model produces an approximate deshadowed image, and human inspectors reject low-quality mask pairs and visibly corrupted inpainting results (Wang et al., 2020; 2023; Liu et al., 2024a; Rombach et al., 2022). Our extension does *not* regenerate these original caster–shadow masks. It reads the four released mask channels, performs deterministic cleanup and conservative relation auditing, constructs a coarse receiver cue, and exports model-ready relation metadata.

This distinction fixes the provenance of each supervision source. The released `object_masks` represent the shadow-casting object, i.e., the *caster* C, rather than the shadow-receiving surface R. The released `shadow_masks` represent the corresponding cast shadow S, rather than every dark region in the image. DESOBA v2 does not provide a precise receiver annotation. The receiver

masks used by ShadowEdit are therefore additional coarse support cues produced by the extension described below.

Lineage of the released caster–shadow pairs. For a real image I_r, SSISv2 predicts a set of associated instances

$$\mathcal{P}(I_r) = \{(C_k, S_k, \rho_k)\}_{k=1}^{K_I}, \tag{S60}$$

where C_k and S_k are the predicted caster and shadow masks and ρ_k records that they belong to the same detected relation. This is an instance-shadow prediction problem rather than independent semantic segmentation followed by nearest-neighbor matching. Images without a detected valid relation are removed during the original construction stage.

The union of the predicted shadow masks,

$$M_s^{\text{all}} = \bigvee_{k=1}^{K_I} S_k, \tag{S61}$$

is used as the inpainting region. The original pipeline applies mask dilation and image flipping around latent-diffusion inpainting to reduce undesired shadow regeneration. The resulting image I_d is an *approximate* deshadowed image, not a second photograph of the same scene under physically identical conditions. Human quality control then rejects (i) caster–shadow pairs with visibly inaccurate object or shadow masks and (ii) samples with residual or newly introduced shadows or conspicuous inpainting artifacts.

When several valid relations occur in one scene, the original release expands them into relation-specific tuples. Selecting relation k gives

$$M_{fo} = C_k, \qquad M_{fs} = S_k, \tag{S62}$$

$$M_{bo} = \bigvee_{j \neq k} C_j, \qquad M_{bs} = \bigvee_{j \neq k} S_j, \tag{S63}$$

where the foreground channels identify the selected relation and the background channels contain the union of the remaining valid relations. The source tuple is constructed by writing the non-selected shadows or all shadows back into the approximate deshadowed image:

$$I_c = (1 - M_{bs}) \odot I_d + M_{bs} \odot I_r, \tag{S64}$$

$$I_g = (1 - M_s^{\text{all}}) \odot I_d + M_s^{\text{all}} \odot I_r. \tag{S65}$$

Our relation-annotation extension does not alter these RGB images. For the prepared global-reference version used here, each input is paired by source identity with the saved I_d, and we set $Y^{\text{sf}} := I_d$. We do not use I_c as the global target: I_c restores non-selected shadows and is an object-level composite. Relation metadata augment the input tuple without changing the saved I_d. The all-remove target remains an approximate inpainting-derived reference, not a physically captured shadow-free observation.

Extended annotation representation. The four released mask channels identify relevant source groups, but they do not directly encode the receiving surface, editing action, protection region, or the model-ready geometry associated with a relation. We therefore construct an auditable relation layer

$$\mathcal{A}(I) = \{(C_i, S_i, R_i, \kappa_i, \zeta_i)\}_{i=1}^{N_I}, \tag{S66}$$

where C_i, S_i, and R_i are the caster, shadow, and coarse receiver masks. $\kappa_i \in \{\text{foreground}, \text{background}, \text{manual}\}$ records provenance, and ζ_i stores validation measurements and audit metadata. An automatically recovered relation preserves the grouping of its released source channels. In particular, a background channel can represent a union of multiple original instances. We therefore describe these annotations as auditable *relation groups*, rather than claiming exhaustive one-object/one-shadow physical instance segmentation.

C.2 Normalization, candidate construction, and validation

Step 1: source loading and binary normalization. Every released record is processed at its native 256×256 resolution. Foreground candidates are read

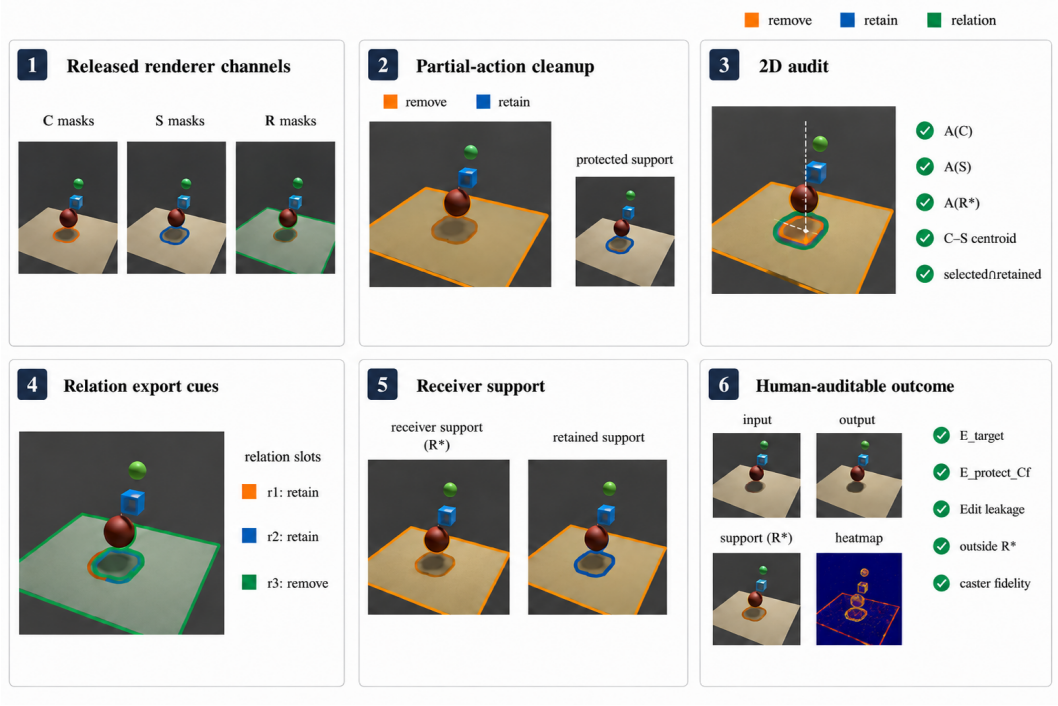


Figure S10: **Auditable caster–shadow–receiver annotation construction.** Released object and shadow channels first provide a source-aligned caster group C_1 and shadow group S_1. Deterministic cleanup and geometric validation remove degenerate candidates while retaining their provenance in the audit record. A separately generated coarse receiver R_1 supplies visible receiver evidence to the relation pathway. The DESOBA permission compiler derives U_a from the remove and protect maps. The final overlay exposes association, coverage, receiver-support leakage, surface-crossing, and boundary-truncation errors for human inspection. Matching colors denote paired caster and shadow groups, and yellow denotes the receiver support.

from `object_masks/shadow_masks`. Background candidates are read from `background_object_masks/ background_shadow_masks`. A nonzero rule converts each grayscale source mask to a binary mask. No SAM prediction is substituted for an original shadow mask, and the extension does not form arbitrary all-pairs caster–shadow matches.

Step 2: deterministic mask cleanup. Each binary mask is cleaned with 3×3 opening/closing operations and connected-component rules. This step removes isolated pixels, narrow spurious fragments, and very small disconnected components while retaining the dominant released geometry. The cleaned result, its contour, bounding box, area, and source channel are recorded before any candidate is accepted or rejected. Keeping both source and cleaned metadata makes later failures traceable rather than silently overwriting the released annotation.

Step 3: source-aligned candidate recovery. A foreground object channel and its foreground shadow channel form one source-aligned candidate pair. The corresponding background channels, when nonempty, form an additional group-level candidate. Because the background masks are unions in the original release, the extension preserves their group provenance instead of decomposing them into unsupported physical instances. For every source-aligned pair, we compute

$$A_i^{\min} = \min(|C_i|, |S_i|), \tag{S67}$$

$$d_i = \min_{\mathbf{p}\in C_i, \mathbf{q}\in S_i} \|\mathbf{p} - \mathbf{q}\|_2, \tag{S68}$$

$$\ell_i = \max(w_{C_i}, h_{C_i}, w_{S_i}, h_{S_i}), \tag{S69}$$

$$\bar{d}_i = d_i/\ell_i, \tag{S70}$$

$$O_i = |C_i \cap S_i|. \tag{S71}$$

Here $A_i^{\min}$ detects tiny or degenerate relations, d_i is the minimum caster–shadow separation, ℓ_i normalizes the separation by the larger instance scale, and O_i records direct overlap. Bounding boxes, centroids, contours, and image-boundary contact are retained as diagnostic metadata used by the overlay and borderline review list.

Step 4: conservative geometric acceptance. A candidate is retained only when

$$A_i^{\min} \geq 100, \qquad d_i \leq 10 \text{ px}, \qquad \bar{d}_i \leq 0.3. \tag{S72}$$

The three tests address complementary failure modes. The area threshold rejects fragments too small to support reliable relation reasoning. The absolute-distance threshold rejects spatially disconnected groups. The scale-normalized threshold prevents the same absolute distance from being treated equally for small and large relations. A rejected candidate is not deleted from the audit trail. Its masks, measurements, source channel, and rejection reason are stored in the audit CSV/JSON. If no candidate passes, the image remains in the archive with $N_I = 0$. The pipeline never fabricates a positive relation merely to fill a training slot.

Step 5: geometry cues exported with each valid relation. For every accepted relation, the extension derives the caster/shadow proximity maps, a centroid-connection line, a local contact cue, and boundary bands used by the 19-channel relation-control stack. These cues describe observable image geometry only. They are not interpreted as recovered illumination parameters, 3D light direction, or exact physical contact.

C.3 Receiver, manual review, and provenance

Coarse receiver construction. The receiver R_i is stored separately from the shadow S_i. The shadow support specifies the illumination-affected region, whereas R_i supplies a coarse visible surface on which correction and texture preservation can be conditioned. The released DESOBA v2 masks do not contain this region, so the extension uses an explicit geometric heuristic:

$$R_i = \begin{cases} \mathrm{TrapDown}(S_i), & \text{for the ground-oriented construction case,} \\ \mathrm{Hull}(\mathrm{Dilate}(S_i)), & \text{otherwise.} \end{cases} \tag{S73}$$

The first case extends a trapezoidal support from the shadow footprint in the image-down direction. The second takes the convex hull of a dilated shadow. The result is a *coarse support cue*, not

semantic or material segmentation. It can cross a material boundary in difficult scenes and must not be described as pixel-accurate receiver ground truth. The 752 images without a reliable accepted relation retain an empty receiver rather than a heuristic mask presented as verified relation ground truth.

Overlay-based human quality control. For every record, the release builder renders a color-coded overlay containing the source image, caster, shadow, receiver, connection line, bounding boxes, and pair identity. These overlays make five error types directly visible: (i) missing or fragmented caster coverage, (ii) shadow leakage into intrinsic dark regions, (iii) incorrect source-group association, (iv) receiver support crossing an implausible surface, and (v) small or boundary-truncated relations. A focused borderline list is generated from near-threshold area, distance, normalized-distance, and image-boundary cases, allowing manual inspection to concentrate on uncertain records rather than silently treating all automatic masks as equally reliable.

Seven samples received explicit targeted corrections:

```
14477_1   12484_1   15130_1   13277_0
4784_0    14984_0   16643_1
```

Their JSON records set `manual_correction=true`. Newly introduced or replaced relations use `pair_kind=manual` and `source_mask=manual`, so they cannot be confused with automatically retained source groups. The seven reviewed images contain nine final manual relations and produce a net increase of four relations over the automatically retained set.

Role of segmentation assistance. SAM and Mask R-CNN (Kirillov et al., 2023; He et al., 2017) are used only as optional contour aids for a small number of difficult caster or receiver corrections. Neither model supplies the released DESOBA v2 shadow annotations nor infers caster–shadow correspondence. Shadow supports remain derived from the released S channel or explicit manual polygon- and darkness-guided correction. This restricted role avoids attributing soft-shadow delineation or relation assignment to generic object segmentation.

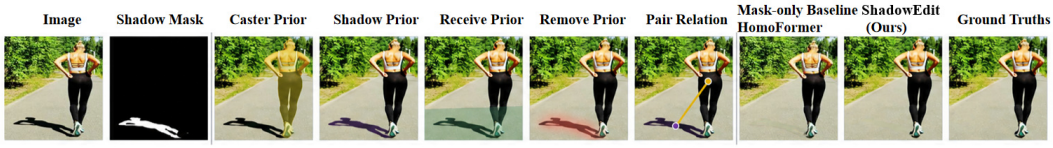


Figure S11: **Representative annotation examples.** (Liu et al., 2024a). The examples include a single foreground relation group, a mixed foreground/background record, and a difficult three-pair manually corrected case. Each row shows the source image, source masks, cleaned relation masks, and the final audit overlay. The panels show both the usefulness and the intended scope of the annotation: automatic groups preserve released source provenance, while receiver masks remain coarse support cues rather than exhaustive physical or material segmentation.

Record format and archive audit. Each image has one JSON record and one directory containing the exported binary masks. The record stores the sample identity, split, resolution, source paths, pair count and IDs, mask paths, contours, bounding boxes, areas, centroids, colors, provenance type, geometric measurements, threshold values, rejected candidates, manual-correction state, and overlay path. The archive audit checks referential integrity and schema consistency: every JSON-referenced image/mask/overlay exists, pair counts agree with exported files, and required fields are present. The final ZIP audit found zero missing referenced files and zero pair-count/schema inconsistencies.

C.4 Task masks for global and selective editing

Records with $N = 0$ are excluded from relation-selective task construction. They remain in the benchmark and are processed by the relation-free global restoration path. The action-vector definitions below therefore apply to $N \geq 1$. Let $a = (a_1, \ldots, a_N)$ be the relation-level action vector. The

Table S45: **Verified extended DESOBA v2 annotation statistics.** DESOBA v2 is introduced by Liu et al. (2024a). "Relations" refers to source-aligned or manually corrected relation groups in the extended archive, not to a claim of exhaustive physical instances.

Item	Count
Prepared records	28,573
Train / test records	27,823 / 750
Automatic candidate relations	40,545
Automatically rejected / retained	3,488 / 37,057
Final relations	37,061
Foreground / background / manual	26,718 / 10,334 / 9
Records with 0 / 1 / 2 / 3 pairs	752 / 18,585 / 9,232 / 4
Manually corrected images	7

removal and protection masks are

$$M_{\text{remove}} = \mathbb{I}\left[\bigcup_{i:a_i=1} S_i\right], \qquad M_{\text{protect}} = \mathbb{I}\left[\bigcup_{i:a_i=0} S_i\right]. \tag{S74}$$

The three task presets used by the implementation are special cases:

$$\text{global removal:} \quad a_i = 1 \quad \forall i, \tag{S75}$$

$$\text{remove relation } k: \quad a_k = 1, \quad a_i = 0 \quad (i \neq k), \tag{S76}$$

$$\text{preserve relation } k: \quad a_k = 0, \quad a_i = 1 \quad (i \neq k). \tag{S77}$$

The permitted execution support is defined by Equation S33. The same complete operator determines the task mask and the support used for final image composition. This construction is stricter than treating the union of all shadows as one editable region: changing the action vector changes both the selected relation tokens and the spatial permission support. Final image composition follows Equation S48, so pixels outside M_{perm} retain the observed input.

Export to fixed relation slots. DESOBA exports at most $K_{\text{D}} = 8$ slots. The relation named by the task is placed first, followed by the remaining relations in decreasing shadow area. Each valid slot contains caster, shadow, receiver, line, and derived contact masks. Unused slots are zero padded and excluded by v. The archived base source records contain at most three relations; the separate multi-action extension uses its own source/action manifest, so the two cardinality statements refer to different tracks. The eight-slot export retains each listed relation in either track and introduces no learned matching stage. This deterministic export keeps every slot traceable to its JSON annotation and introduces no learned matching stage.

D Inference and Evaluation Protocol

D.1 Standard-benchmark inference

ISTD+ (Le & Samaras, 2019) and SRD (Qu et al., 2017) use separate prior-guided, relation-free checkpoints and the same test driver. For each input image, the loader retrieves the corresponding Depth Anything V2 depth prior, DINOv2 semantic prior, and their quality maps. The model enables appearance priors and disables relation conditioning. Evaluation uses batch size one, `shuffle=False`, and overlapping tiled inference with 30-pixel overlap. The driver selects the largest valid tile size from $\{384, 256, 128\}$ according to the shorter image side. Image tiles, full-resolution depth priors, and low-resolution semantic features are spatially aligned before forward inference. Tile outputs are merged in floating point, averaged when repeated inference is enabled by the benchmark configuration, clipped to $[0, 1]$, and then passed to the common metric implementation. ISTD+ activates the adjusted official target through `-plus`. SRD uses the official clean target and the fixed DHAN evaluation mask (Cun et al., 2020).

Table S46: **DESOBA experiment paths and evidence scope.** Full uses relation conditioning in both all-remove experiments; Mask disables the relation module. The benchmark and component study are separate runs, not reused prediction exports. Selective editing changes the action and supervision scope.

Experiment	Relation branch	Reference / action	Evidence
Global benchmark	Full: enabled; Mask: disabled	Saved $Y^{\mathrm{sf}} = I_d$; all-remove	Global restoration quality; main-paper Table 1
Component analysis	Full: enabled; Mask: disabled	Saved $Y^{\mathrm{sf}} = I_d$; all-remove; separate runs	Components and five-seed RMSE; Tables S51 and S54
Selective editing	Enabled	Independently measured or manually constructed action-specific references for the partial-target training split; held-out action-specific references for test actions	Quantitative Target / Protected / Overlap / Leakage on the held-out split; action swaps provide visual responsiveness evidence

Table S47: **Single evidence path for the DESOBA claims.** Each row has a separate checkpoint and prediction manifest; the frozen scorer is shared only at the metric stage. The file name and SHA256 recorded in the experiment manifest are the authoritative checkpoint identifiers.

Track	Checkpoint scope	Test source / target	Scoring mask	TTA	Scorer
Headline global	Full vs. Mask paper-best; checkpoint ID/hash per row	fixed 750 DESOBA images; saved I_d	fixed global M_s	D_4 + blend	frozen common scorer
Matched global audit	matched Full/Mask audit pair; separate prediction export	same 750 images; saved I_d	same fixed M_s	no-TTA and D_4 as separate rows	same frozen scorer
Selective partial	partial-target checkpoint; fallback is a separate control	200 held-out images / 400 actions; Y_a^{part}	V_a^{part}, M_{remove}, $M_{\mathrm{protect}}^{\mathrm{ex}}$, M_{overlap}	manifest-recorded	same frozen scorer
Multi-action	one fixed selective checkpoint for all joint actions	40 held-out sources / 80 joint actions; Y_J^{part}	action-derived masks and valid target support	manifest-recorded	same frozen scorer

D.2 DESOBA v2 MODEL AND PAPER EVALUATION

The checkpoint, target, mask, TTA, and scorer fields in this table are joined by a single `experiment_id`. A result is admissible only when its prediction hash, target hash, valid-mask hash, and scorer version all match that row; a rounded numerical match between two rows is not evidence that their predictions were reused.

The 3.715 dB difference in main Main-paper Table 1 compares the complete relation-enabled configuration with Mask under all-remove evaluation. It does not isolate token interaction, which is tested more directly by the equal-information Kubric controls in Table S35. The DESOBA benchmark and component experiment are separate runs; their Mask/Full regional means coincide at the reported precision. We retain both measured endpoints, while the numerical agreement does not create an additional effect estimate. The five-seed RMSE summary belongs to the relation-enabled component comparison.

The relation-enabled DESOBA v2 Full configuration for editing, the global benchmark, and component analysis uses the same types of Depth Anything V2/DINOv2 appearance priors as ISTD+ and SRD, and enables the relation pathway. The complete configuration uses a 16-dimensional prior embedding, 384-dimensional semantic features at 32×32, a 19-channel relation-control stack, at most $K_{\mathrm{D}} = 8$ relation slots, two Relation-indexed control layers, and four attention heads. The paper evaluation uses 256-pixel tiles with 30-pixel overlap, D_4 self-ensemble (eight transforms), inverse transformation before prediction averaging, and boundary-blend kernels $\{11, 15, 19\}$ with floor 0.02. Prediction averaging is performed before clipping and metric computation. Evaluation uses deterministic cuDNN execution for every checkpoint. In all-remove evaluation, Full enables relation conditioning and sets all valid actions to remove; Mask disables relation conditioning. Both use the same prepared global-target list and fixed metric supports. The main-paper restoration values are means over the five seeds listed in Section E.2; component runs are evaluated separately under the same protocol. The separate Kubric representation comparison reports five paired training runs for Overlap in Table S36; these are distinct from the DESOBA restoration statistics.

Architecture-ablation configurations. Table S51 follows the execution graph cumulatively. The Mask baseline uses the observed image and global union mask without structured relation inputs. Dense relation field adds the image-aligned relation-control representation F^R. Relation tokens introduce per-relation instance representations, and Token interaction enables the masked self-attention encoder Ψ. Spatial relation assignment then activates the relation-to-pixel weights $\beta_{\ell i}(p)$ and grounded feature $Z_\ell(p)$ in Equation S45. Permission-gated modulation further activates the bounded scale and bias fields in Equation S46. The complete relation-execution configuration additionally uses the relation-conditioned residual in Equation S47 and feasible output composition. Its appearance-free variant sets $F_{\mathrm{app}} = 0$, while Full restores the frozen DINOv2 and Depth Anything V2 appearance evidence. Appearance priors are held fixed for Mask and Full; only the explicitly labeled appearance-free ablation removes them. All configurations use the same DESOBA v2 split, restoration backbone, objectives, optimization budget, random seeds, and frozen evaluator.

D.3 INDEPENDENT TRAINING SEEDS

The robustness study uses five independently trained checkpoints with seeds $\{1234, 2024, 3407, 4567, 5678\}$. Apart from the seed, every run uses the same training split, augmentation policy, optimization schedule, checkpoint-selection criterion, model configuration, and full-split evaluator. Each seed controls Python, NumPy, PyTorch, CUDA, and data-loader randomness. No run is removed after observing validation or test performance. All seed-level PSNR/SSIM results and the matched DESOBA v2 regional RMSE analysis are reported in Section E.2.

D.4 UNIFIED EVALUATION PROTOCOL

D.4.1 CROSS-METHOD COMPARISON DESIGN

We separate *method-specific inference* from *common metric evaluation*. Every baseline is rerun with its official implementation and checkpoint to generate one prediction for every image in the prescribed test split. ShadowEdit is evaluated with the corresponding benchmark-specific check-

point. After prediction generation, outputs from all methods are passed to the same offline evaluator with the same ordered test list, filename alignment, targets, shadow masks, resize operation, color conversion, numerical range, region definitions, and image-level aggregation. Consequently, differences in the reported tables are not caused by method-dependent metric scripts. We do not attempt to load checkpoints from different architectures into a shared network constructor. Only the scoring stage is shared.

The common scorer is frozen before comparison. It does not apply method-dependent cropping, mask dilation, color correction, or metric-specific post-processing. Optional image saving is performed after the floating-point prediction has been clipped, and table metrics are computed from the floating-point RGB array rather than from a visualization figure.

D.4.2 TEST SPLITS, TARGETS, AND REGION MASKS

The same official test split is used for every method on each benchmark. ISTD+ uses the 540-image official test split and the adjusted `test_C_fixed_official` target selected by the `-plus` option. Its binary shadow mask is the mask paired with the same test triplet. SRD uses the official 408-image test split. Because the original SRD release does not provide test shadow masks, all region-wise SRD results use the same public DHAN masks, following the established HomoFormer/SRD protocol. The DESOBA global benchmark and all-remove component analysis use the prepared version's fixed 750-image test list and saved I_d references. They use separately generated predictions. Selective editing uses the 200-image source-disjoint partial-test subset and its 400 one-selected actions. Its quantitative scores are computed on these held-out partial-target records, using their action-specific targets and relation-derived Target, Protected, Overlap, and Leakage supports; the held-out records are excluded from training and checkpoint selection.

For DESOBA all-remove scoring, the evaluator first requests the relation-derived `metric_shadow` map. If this map is unavailable or empty, it falls back to the source shadow mask supplied by the data loader. If that mask is also empty, the evaluator reads the corresponding mask from the `shadow_masks` directory, thresholds its 8-bit value at 127, and resizes it to the data-loader resolution with nearest-neighbor interpolation. This fallback is deterministic and is applied before the common metric-stage resize.

The resulting DESOBA shadow mask is frozen per test image and reused for every compared method, including Mask when relation conditioning is disabled. It is not replaced by a method-dependent output or by a selected partial-action mask.

Let M_s denote the final binary shadow mask. The three evaluation regions are fixed as

$$\Omega_{\mathrm{sha}} = M_s, \qquad \Omega_{\mathrm{non}} = 1 - M_s, \qquad \Omega_{\mathrm{all}} = \mathbf{1}. \tag{S78}$$

No method is allowed to supply its own evaluation mask. For selective relation diagnostics, target and protected masks are evaluated separately and are not substituted for the global benchmark mask.

Table S48: **Dataset-specific inputs to the unified evaluator.** "Common" means that the same item is used for ShadowEdit and every compared baseline. Dataset protocols follow ISTD+ (Le & Samaras, 2019), SRD (Qu et al., 2017), and DESOBA v2 (Liu et al., 2024a).

Control	ISTD+	SRD	DESOBA v2 (global)
Test split	Official 540-image test split	Official 408-image test split	Fixed 750-image global-target test list
Ground truth	Adjusted official target (`-plus`)	Official clean target	Prepared global shadow-free target
Shadow mask	Official paired test mask	Public DHAN mask	Fixed mask paired with the global target
Region definitions	M_s (shadow), $1 - M_s$ (non-shadow), and full frame (all)		
Method-specific mask tuning	None		

For the DESOBA held-out partial-target track, the common scorer receives the same input/action pair and each method's own prediction, together with the corresponding action-specific target Y_a^{part}.

It reuses the supports in Eq. (S13): Target is evaluated on $V_a^{\mathrm{part}} \odot M_{\mathrm{remove}}$, Protected on $V_a^{\mathrm{part}} \odot M_{\mathrm{protect}}^{\mathrm{ex}}$, Overlap on $V_a^{\mathrm{part}} \odot M_{\mathrm{remove}} \odot M_{\mathrm{protect}}$, and Leakage outside the requested removal support. These scores are reported separately from the global 750-image all-remove metrics, whose reference remains $Y^{\mathrm{sf}} = I_d$.

D.4.3 Exact normalization, interpolation, and metric definitions

Images are read as RGB, converted to `float32`, and normalized by 255. Model predictions are clipped to $[0, 1]$ before evaluation. At the *metric stage*, every prediction and ground-truth image is resized to 256×256 using OpenCV area interpolation (`cv2.INTER_AREA`). The corresponding mask is resized with the same area interpolation and then binarized using the fixed rule

$$M_s(p) = \begin{cases} 0, & \widetilde{M}_s(p) < 10^{-3}, \\ 1, & \text{otherwise.} \end{cases} \tag{S79}$$

This metric-stage operation is identical for all methods and should not be confused with architecture-specific internal resizing or tiled inference.

PSNR and SSIM are computed in RGB space with data range one. For $r \in \{\mathrm{sha}, \mathrm{non}\}$ and $C = 3$ channels, region PSNR is based only on pixels selected by M_r:

$$\mathrm{MSE}_r = \frac{\sum_p M_r(p) \sum_{c=1}^{C} \left(Y_{pc} - Y_{pc}^{gt}\right)^2}{C \sum_p M_r(p)}, \tag{S80}$$

$$\mathrm{PSNR}_r = 10 \log_{10} \left(\frac{1}{\mathrm{MSE}_r} \right). \tag{S81}$$

All-image PSNR is computed directly over the complete RGB image. SSIM is computed once as a full RGB SSIM map. Region SSIM is the arithmetic mean of that map over the selected region. Pixels outside the region are not replaced by zeros. All-image SSIM is the standard full-image SSIM.

For MAE and RMSE, the normalized RGB arrays are converted with `cv2.COLOR_RGB2LAB`. Let $e_{pc} = Y_{pc}^{Lab} - Y_{pc}^{gt,Lab}$. The implementation reports

$$\mathrm{MAE}_r = \frac{\sum_p M_r(p) \sum_{c=1}^{3} |e_{pc}|}{\sum_p M_r(p)}, \tag{S82}$$

$$\mathrm{RMSE}_r = \sqrt{\frac{\sum_p M_r(p) \sum_{c=1}^{3} e_{pc}^2}{\sum_p M_r(p)}}. \tag{S83}$$

Thus, RGB PSNR/SSIM and CIELAB MAE/RMSE are never mixed within one metric. The same conversions and formulas are used for ISTD+, SRD, DESOBA v2, ShadowEdit, and every baseline.

Metrics are first computed independently for every test image. Dataset-level results are the arithmetic mean of the finite per-image values:

$$\overline{m}_r = \frac{1}{|\mathcal{I}_r|} \sum_{i \in \mathcal{I}_r} m_{i,r}, \qquad \mathcal{I}_r = \{i : \|M_{i,r}\|_1 > 0\}. \tag{S84}$$

The valid set $\mathcal{I}_r$ is fixed before method evaluation and is identical across methods. No prediction-dependent sample is discarded. This prevents images with different shadow areas from being implicitly reweighted by the number of shadow pixels. LPIPS with an AlexNet backbone can be computed by the code path, but it is not inserted into the reported benchmark tables.

D.4.4 Inference settings and checkpoint audit

Checkpoint metadata identify the evaluated weights. The protocol above determines the reported measurements.

Table S49: **Common post-inference metric controls.** These controls are identical across datasets and methods.

Item	Fixed implementation
Image representation	RGB `float32`, input/target divided by 255, prediction clipped to $[0, 1]$
Metric resolution	Prediction and target resized to 256×256 with `cv2.INTER_AREA`
Metric-mask resize	`cv2.INTER_AREA`, followed by threshold 10^{-3}
Shadow / non-shadow / all	M_s / $1 - M_s$ / complete image
PSNR	RGB, peak value 1, masked MSE for regional scores
SSIM	RGB, `data_range=1.0`, masked mean of the full SSIM map for regional scores
MAE / RMSE	OpenCV CIELAB converted from normalized RGB, identical equations for all methods
Dataset aggregation	Per-image computation followed by the arithmetic mean of finite values
Per-method metric tuning	None

Table S50: **Audited inference and scoring roles.** Method-specific settings affect output generation but never change the common metric definition.

Setting	ISTD+	SRD	DESOBA v2 (global)
Batch / shuffle	1 / no	1 / no	1 / no
Test target	adjusted official	official	saved I_d
DINOv2 semantic prior	enabled	enabled	enabled
Depth Anything V2 prior	enabled	enabled	enabled
Relation branch	disabled	disabled	Full: enabled Mask: disabled
Inference tiling	auto, overlap 30	auto, overlap 30	256, overlap 30
Paper TTA	disabled	disabled	D_4 (8 transforms)
Paper blend kernels / floor	–	–	11,15,19 / 0.02
Common metric resize	256×256, area interpolation		
Common output range	RGB $[0, 1]$		
Common metric code	identical		

D.5 DESOBA v2 Mechanism-aligned Component Analysis

The main paper uses Kubric exact counterfactuals for the primary ablation of selective execution. For completeness, this section reports the complementary relation-enabled DESOBA v2 all-remove construction under matched training. All valid relations are assigned removal and $M_{\mathrm{perm}} = \mathbf{1}$; evaluation uses the saved global reference $Y^{\mathrm{sf}} = I_d$. This table measures all-remove compatibility, not selective correctness. The endpoints use the same Mask/Full configuration definitions as the main benchmark but come from separate experimental runs. The five-seed RMSE summary in Table S54 belongs to this component comparison.

Table S51: **Mechanism-aligned Component Analysis on DESOBA v2.** The first block progressively adds the named relation components under all-remove evaluation ($M_{\mathrm{perm}} = \mathbf{1}$). The final pair compares the complete executor without and with appearance evidence. Since the gate is constant here, this experiment does not isolate spatial permission. Lower is better. Best results are **bold**.

Configuration	Shadow		Non-shadow		All	
	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓
Mask baseline	7.020	7.810	0.548	0.872	0.762	1.628
+ Dense relation field	6.760	7.560	0.536	0.852	0.738	1.575
+ Relation tokens	6.540	7.330	0.529	0.838	0.720	1.532
+ Token interaction	6.380	7.160	0.524	0.827	0.706	1.498
+ Spatial relation assignment	6.270	7.040	0.523	0.824	0.702	1.487
+ Permission-gated modulation	6.170	6.980	0.522	0.821	0.698	1.476
Full relation execution w/o appearance	5.400	6.320	0.510	0.812	0.695	1.474
Full	**4.957**	**5.936**	**0.502**	**0.803**	**0.692**	**1.469**

Shadow RMSE decreases monotonically from 7.810 (mask baseline) to 5.936 (Full), while non-shadow RMSE decreases from 0.872 to 0.803. Completing the relation pathway without appearance evidence already reaches 6.320 shadow RMSE; restoring frozen appearance cues yields the remaining gain. We treat this table as complementary compatibility evidence under all-remove evaluation. Its Mask and Full endpoint means coincide, at the displayed precision, with the separately run main benchmark. The prediction exports are not reused between the two experiments, and both endpoint pairs are retained as reported measurements. The paired RMSE analysis below summarizes this component experiment rather than introducing another prediction export. The complete Mask/Full contrast changes the relation configuration, whereas the equal-information Kubric comparison more directly tests representation and token interaction.

E Additional Quantitative Results

The main paper contains cross-method benchmark tables and the Kubric control ablation; the DESOBA architecture analysis appears in Section D.5. To avoid duplication, this section reports complementary regional metrics, compact multi-seed PSNR/SSIM results, and a matched five-seed RMSE summary of the reported DESOBA configuration pairs.

E.1 Complementary regional error metrics

Table S52 reports complementary regional errors under the dataset-specific ISTD+ and SRD evaluation conventions. ISTD+ uses shadow/all MAE and shadow/non-shadow/all RMSE, whereas SRD uses shadow/all RMSE. Lower is better; the unprocessed input row is excluded from ranking.

All reported MAE/RMSE values are computed with the same frozen evaluator and the Lab-space definitions given in Section D.4. Under the stated channel-summed definitions, every unrounded three-channel measurement must satisfy

$$\mathrm{MAE} \le \sqrt{3}\,\mathrm{RMSE}.$$

The table values are rounded to three decimals after evaluation. In particular, the DenseSR all-image pair $7.443/4.298$ satisfies the corresponding norm-consistency constraint at the displayed precision. This bound is necessary, not a validation of the evaluator or a conversion rule between MAE and RMSE. The two statistics must be computed independently from the pixel errors.

The compared methods are described in (Zhu et al., 2022; Guo et al., 2023a;b; Li et al., 2023; Xiao et al., 2024; Xu et al., 2025a;b; Lin et al., 2025; Li et al., 2025; Lee et al., 2026).

Table S52: **Complementary regional error metrics on ISTD+ and SRD.** All MAE/RMSE entries are evaluated with the common frozen scorer. Lower is better.

(a) ISTD+

Method	Shadow		Non-shadow RMSE↓	All	
	MAE↓	RMSE↓		MAE↓	RMSE↓
Input images (reference)	36.868	21.286	–	8.353	7.118
BMNet	7.619	4.399	2.21	3.935	2.79
ShadowFormer	7.257	4.190	2.22	3.780	2.65
ShadowDiffusion	7.445	4.298	2.30	3.834	2.77
Inpaint4Shadow	6.106	3.525	1.89	3.418	2.16
HomoFormer	6.641	3.834	2.08	3.645	2.47
OmniSR	6.830	3.943	2.09	3.470	2.50
StableShadowRemoval	5.620	3.245	1.72	2.960	2.07
DenseSR	9.305	5.372	1.85	7.443	4.298
ShadowMaskFormer	5.442	3.142	2.82	2.730	3.522
PhaSR	6.399	3.694	1.74	3.399	2.22
ShadowEdit (Ours)	**4.758**	**2.747**	**1.57**	**2.595**	**1.86**

(b) SRD

Method	Shadow RMSE↓	All RMSE↓
Input images (reference)	19.121	9.358
BMNet	4.976	3.966
ShadowFormer	4.903	3.866
ShadowDiffusion	6.302	5.036
Inpaint4Shadow	**3.553**	**2.816**
HomoFormer	4.566	3.566
OmniSR	4.350	3.450
StableShadowRemoval	3.650	2.950
DenseSR	5.374	3.784
ShadowMaskFormer	5.085	4.074
PhaSR	4.100	3.200
ShadowEdit (Ours)	6.268	3.921

On ISTD+, ShadowEdit obtains the lowest error among the complete, comparable frozen-scorer entries in all five reported columns, including the DenseSR all-image pair. The improvement appears in the edited shadow region as well as in the non-shadow and full-image measurements. The SRD ranking is different: Inpaint4Shadow gives the lowest shadow and all-image RMSE, followed by StableShadowRemoval. The comparison therefore supports a benchmark-specific conclusion—strong and spatially consistent performance on ISTD+—rather than a claim of uniform state-of-the-art performance across datasets.

E.2 Multi-seed robustness

We evaluate seed sensitivity using five independently trained checkpoints with seeds $\{1234, 2024, 3407, 4567, 5678\}$. Each seed controls Python, NumPy, PyTorch, CUDA, and dataloader randomness. All runs use the same dataset splits, optimization schedule, augmentations, checkpoint-selection rule, and evaluation protocol. Every entry in Table S53 is obtained by evaluating one checkpoint on the complete fixed test split. No run is removed or selected after observing test performance. The final row of each dataset block reports the arithmetic mean and sample standard deviation across all five runs.

Table S53: **Five-seed robustness of ShadowEdit.** All-image RGB PSNR/SSIM is reported for every independently trained checkpoint.

Dataset	Seed	PSNR↑	SSIM↑
ISTD+	1234	35.750	0.9688
	2024	35.895	0.9694
	3407	35.690	0.9690
	4567	35.820	0.9692
	5678	35.610	0.9686
	Mean ± std	35.753 ± 0.111	0.9690 ± 0.0003
SRD	1234	35.940	0.9652
	2024	35.750	0.9646
	3407	36.010	0.9654
	4567	35.880	0.9650
	5678	35.820	0.9648
	Mean ± std	35.880 ± 0.101	0.9650 ± 0.0003
DESOBA v2	1234	39.825	0.984
	2024	39.970	0.983
	3407	39.750	0.981
	4567	39.900	0.985
	5678	39.680	0.982
	Mean ± std	39.825 ± 0.115	0.9830 ± 0.0016

Across the three benchmarks, the PSNR standard deviation is at most 0.115 dB and the SSIM standard deviation is at most 0.0016, indicating limited sensitivity to initialization for these all-image endpoints.

Paired regional RMSE summary. The reported five-seed summary uses training seeds as paired experimental units and defines $\Delta = \mathrm{RMSE}_{\mathrm{Mask}} - \mathrm{RMSE}_{\mathrm{Full}}$. Positive values favor Full. This summary belongs to the relation-enabled DESOBA all-remove component experiment: Full enables the relation module, Mask disables it, and both are scored against the saved I_d references. It describes the component-table endpoints across paired training seeds. These DESOBA seeds do not establish stability for the separate Kubric token comparison.

Table S54: **Five-seed RMSE summary for the DESOBA relation-enabled all-remove component experiment.** Mask and Full are across-seed means for the component comparison in Table S51, evaluated against $Y^{\mathrm{sf}} = I_d$. Δ is Mask minus Full, and Relative is Δ/Mask. Confidence intervals are two-sided 95% paired Student-t intervals with four degrees of freedom. d_z is the paired standardized effect size, and p_{Holm} denotes the two-sided paired t-test result after Holm correction across the three regional endpoints. Wins counts seed pairs with $\Delta > 0$.

Region	Mask mean	Full mean	Δ	Relative↑	95% CI of Δ	d_z	p_{Holm}	Wins
Shadow	7.810	5.936	1.874	24.0%	[1.688, 2.061]	12.50	2.93×10^{-5}	5/5
Non-shadow	0.872	0.803	0.069	7.9%	[0.044, 0.094]	3.46	1.51×10^{-3}	5/5
All	1.628	1.469	0.159	9.8%	[0.127, 0.191]	6.27	3.01×10^{-4}	5/5

The component experiment reports positive RMSE differences for every region and all five seed pairs, with the largest relative change in Shadow (24.0%). All reported intervals exclude zero under the paired-t analysis. These measurements establish the complete relation-enabled configuration's advantage over Mask for all-remove restoration. Token interaction is measured separately by the equal-information Kubric controls, while selective correctness is measured by the held-out DESOBA partial-target table. Five seeds quantify training variability on one fixed split; they are not independent-scene or cross-dataset uncertainty estimates.

F Evaluation Scope, Quality Control, and Responsible Claims

Conditioning and annotation scope. Selective shadow editing is evaluated under supplied caster–shadow–receiver relations, matching the controlled-input setting used by current structured editing benchmarks. Public real-image shadow datasets do not provide a single standard for dense receiver-surface annotation: some records preserve released source-mask associations, while others require audited geometric receiver cues. Our DESOBA extension records these cues explicitly and keeps their provenance separate from the RGB images. Boundary perturbation and relation-quality analyses therefore measure the effect of the available relation evidence under the benchmark protocol; automatic relation discovery and dense surface annotation are broader dataset and infrastructure directions shared by the field.

Complementary counterfactual protocols. Kubric supplies renderer-defined action-specific targets, whereas the real-image DESOBA track supplies measured-labeled or manually constructed action-specific references on a source-disjoint held-out split. These protocols answer complementary questions: exact intervention fidelity in a controlled renderer and agreement with audited real-image references. Action-swap visualizations use the compiled relation/action permission, and selected–protected overlap remains writable whenever the requested action and valid reference support it. The multi-action audit enumerates relation-role signatures absent from the training-combination registry. The finite sequential diagnostics are limited to the specified two-step cases and do not establish arbitrary sequential or recursive editing stability.

Attribution and uncertainty. Progressive controls and equal-information controls serve different purposes, as is standard in mechanism studies. The equal-information experiment (Table S35) fixes relation/action fields and output permission while comparing spatial encoding, independent tokens, and interacting tokens. It therefore reports the effect of the evaluated representation and interaction design under fixed information and permission. The five paired Kubric runs quantify repeated-training variation on the frozen test manifest, while the independent scene bootstrap quantifies scene-level variation; these are distinct estimands rather than interchangeable error bars. Reporting both units follows the current practice for benchmarks with multiple actions per scene.

Restoration and resource scope. ISTD+ and SRD provide complementary relation-free restoration references, and their metric conventions are reported separately because RGB PSNR and Lab RMSE can induce different rankings. The parameter and FLOP tables follow the common model-side reporting convention and identify the frozen appearance priors, prior extraction, tiling, test-time augmentation, and postprocessing in the inference protocol. A fully standardized end-to-end cost benchmark across shadow-removal systems remains an open evaluation direction for the field.

F.1 Reproducibility checklist

- Dataset splits and counts are stated in Section B.1.
- DINOv2 semantic priors and Depth Anything V2 depth priors are used on ISTD+, SRD, and DESOBA v2.
- ISTD+ and SRD use `use_priors=True` and `use_relation=False`. DESOBA Full enables relation conditioning in editing, component analysis, and the all-remove benchmark; Mask disables it. All-remove references are the saved $Y^{\mathrm{sf}} = I_d$; the partial-target branch uses real-image action-specific references, the matched fallback branch uses the deterministic mask-composed target, and all five runs are evaluated on the same disjoint held-out references.
- The complete appearance-prior construction, quality maps, confidence gate, and decoder-aligned prior features are given in Equations S23–S28.
- The 19-channel control uses the same entity/action/grounding grouping as the main paper and is enumerated in Equation S35.
- Every relation slot contains caster, shadow, receiver, line, and derived contact masks. The six-dimensional descriptor is defined in Equation S42.
- "Relation-annotated DESOBA v2" denotes the public DESOBA v2 tuples augmented with the audited receiver and relation metadata described in Section C.

- The final relation-conditioned output follows Equation S48. Pixels outside the permission support retain the observed input.
- Every compared baseline is rerun with its official implementation and checkpoint, and all generated predictions are scored by the same frozen evaluator.
- The five independent seeds are $\{1234, 2024, 3407, 4567, 5678\}$. Every all-image PSNR/SSIM run and its aggregate mean $\pm$ sample standard deviation is reported in Section E.2.
- The equal-information Kubric comparison holds supplied per-relation fields, derived masks, renderer targets, and output composition fixed. It is distinct from Permission-only and uses the same 352-action pool as the progressive and cardinality tables. Five matched training runs for Overlap and their paired differences are reported in Tables S36 and S37.
- Relation-support sensitivity treats training checkpoints as the independent units. Five perturbation draws are averaged within each checkpoint before the sample standard deviation is computed across the five checkpoints.
- DESOBA benchmark and component analysis use separate prediction runs with the same endpoint configuration definitions and evaluator. Their equal rounded endpoints are not reused predictions. The five-seed regional RMSE table summarizes the component experiment.
- On the common Kubric exclusive protected supports, the independently measured reference discrepancy is $\bar{D}_{100} = 2.038$ (raw RGB MAE 0.02038), within the derived interval $[1.500, 2.343]$.