SCARP: JOINT RISK-AWARE CORRECTION OF HIGH-RISK SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantic segmentation errors with comparable mean Intersection over Union (mIoU) can induce markedly different downstream consequences. We study a failure mode that exposes this mismatch: a *ghost*, defined as a connected prediction of a high-risk class with no ground-truth support for that class. Because ghost identity is unobservable at test time, correction is inherently joint: the value of rerouting one high-risk component depends on what the scene retains elsewhere. We characterize exactly when such context induces a Keep–Reroute preference reversal and introduce SCARP (**S**cene-level **C**onsequence-**A**ware **R**erouting with **P**reservation), a training-free post-hoc decision layer that jointly optimizes high-risk component interventions under prediction-only constraints on removed high-risk evidence, without test-time ground truth. Against a matched additive program differing only in the scene term, coupling changes decisions on 7–21% of test images across four paired benchmarks and improves supported-component retention across all five domains. Strict preference reversals occur in 17 of 96 eligible cases and vanish without coupling. At matched retention on ACDC, SCARP further lowers Ghost-FDR, demonstrating an improved preservation–risk frontier.

1 INTRODUCTION

Semantic segmentation is evaluated by aggregate overlap, yet its outputs increasingly inform decisions with asymmetric consequences. A small false positive on a critical class can alter autonomous planning (Philion et al., 2020), redirect a UAV operation, or bias monitoring in extreme environments while barely affecting mIoU (Figure 3). Under domain shift, an incorrect class can dominate the posterior over a contiguous region, and argmax decoding turns that region into a coherent object with no counterpart in the scene. We call such a component a *ghost*: a connected predicted region with no ground-truth support for its predicted class (Figure 1). When that class is high-risk, a single ghost can trigger a false alarm, whereas removing a supported high-risk component creates a missed hazard.

Post-hoc reliability methods for semantic segmentation largely address two questions. One asks *whether a prediction should be trusted*, using calibration, selective prediction, or conformal sets (Guo et al., 2017; Geifman & El-Yaniv, 2017; Angelopoulos et al., 2024). The other asks *which predicted regions are likely to be erroneous*, using component-level meta-classifiers (Rottmann et al., 2020). These approaches provide uncertainty, rejection, set-valued outputs, or region-wise reliability scores, but they do not directly solve the intervention problem considered here: selecting a corrected configuration of high-risk components under asymmetric consequences and shared preservation constraints. The missing abstraction is therefore not another component-wise confidence score, but a joint intervention rule whose action on one component can change with the configuration retained elsewhere in the scene.

This joint structure arises because ghost identity is unobservable at test time: every predicted high-risk component may represent either an unsupported alarm or genuine hazard evidence. The value of removing one component therefore depends on what its context retains. Several ambiguous components may jointly provide the only evidence of a real hazard, whereas several weak components may jointly sustain an unsupported alarm. High-risk correction therefore requires reasoning over configurations rather than components in isolation.

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.

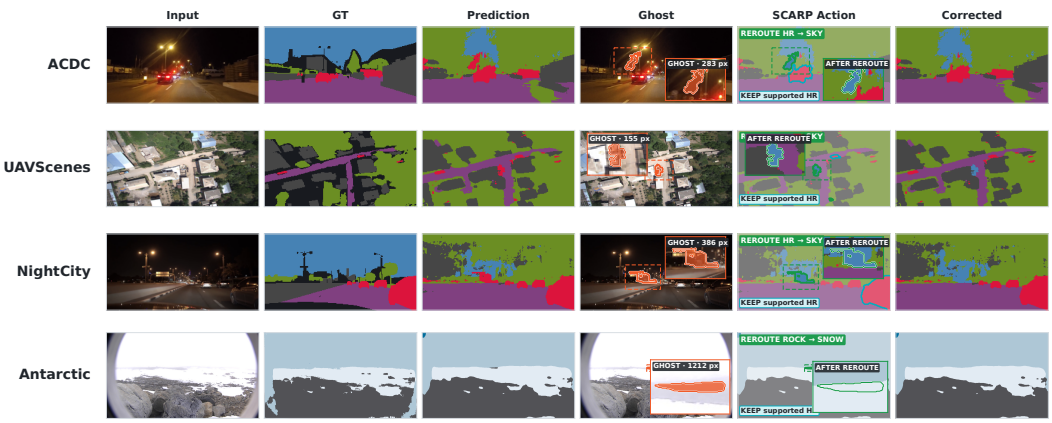


Figure 1: **Ghost failures and SCARP correction.** Rows show examples from ACDC, UAVScenes, NightCity, and Antarctic. Columns show the input image, ground truth, frozen segmentation, the ghost component, the SCARP decision, and the corrected prediction. Ground truth is used only to identify ghosts retrospectively and is never available to SCARP.

In this paper, we ask whether unsupported high-risk predictions can be suppressed post hoc, without retraining or test-time supervision, while preserving genuine high-risk evidence. We first characterize exactly when context reverses a component's Keep–Reroute preference (Proposition 3.1); within this regime, no configuration-invariant local score reproduces the preference. We then introduce **SCARP**, a post-hoc decision layer that either keeps each predicted high-risk component of a frozen segmenter (KEEP) or reroutes it to a non-high-risk class (REROUTE). SCARP prices each action from the frozen posterior and an application-specified consequence matrix, couples actions through a scene potential on the high-risk mass retained in each context, and bounds removed evidence with prediction-only budgets. The problem is solved exactly as a 0–1 MILP, while a matched additive program that omits only the scene term isolates the effect of coupling.

Across five test domains and four backbones, coupling is decision-relevant: it changes selected actions on 7–21% of test images across four paired benchmarks, vanishes when its weight is zero, and persists across contexts, consequence matrices, and recalibration. It acts through the reversals predicted by Proposition 3.1 (17 of 96 eligible cases, none at zero coupling) and through reallocation of a shared budget. It also improves the preservation–risk trade-off: at a fixed budget, SCARP retains more supported high-risk components than the additive program across all five domains, and at matched retention on ACDC it attains lower Ghost-FDR than additive and unary rules. Our contributions are as follows:

- We introduce the *ghost*, a component-level failure mode for unsupported predictions of an application-designated high-risk class, together with Ghost-FDR, a corresponding false-discovery metric for quantifying such errors.
- We cast post-hoc correction of high-risk predictions as a consequence-aware structured decision problem, derive the exact condition for context-induced Keep–Reroute preference reversal, and solve the resulting constrained objective to global optimality with SCARP.
- We evaluate SCARP through a matched Joint–Additive study across five domains, including a new Antarctic field domain, and perform targeted audits of the coupling mechanism, robustness, and failure modes.

2 RELATED WORK

Component-level reliability. Pixel-averaged overlap obscures localized failures, which motivates boundary-, topology-, and instance-aware evaluation (Cheng et al., 2021; Hu et al., 2019; Stucki et al., 2023; Valverde et al., 2026; Kirillov et al., 2019; Maier-Hein et al., 2024). Meta-classification estimates the quality of individual predicted components (Rottmann et al., 2020; Heinert et al.,

2024), and anomaly and out-of-distribution segmentation detect unsupported regions, increasingly at the object or mask level (Chan et al., 2021; Rai et al., 2023; Grcić et al., 2023; Zhu et al., 2025; Nayal et al., 2025; Shoeb et al., 2025). These methods estimate *whether* a component is reliable; SCARP decides whether to *change* it.

Post-hoc reliability and risk control. Calibration, selective prediction, and conformal prediction quantify when an output should be trusted (Guo et al., 2017; Chow, 1970; Geifman & El-Yaniv, 2017; Kahl et al., 2024; Heng & Soh, 2026). Segmentation variants provide pixel-, region-, or image-conditioned sets and risk guarantees (Mossina et al., 2024; Viti et al., 2026; Luo et al., 2026; Guarino et al., 2026; Bates et al., 2021; Angelopoulos et al., 2024; Yeh et al., 2025), including false-positive, false-discovery, and decision-level control (Mossina & Friedrich, 2025; Dai et al., 2025; Zhou et al., 2026; Hu et al., 2026; Zhang et al., 2025). These methods accept, reject, or represent predictions as sets; SCARP commits to a single corrected configuration.

Cost-sensitive and decision-focused learning. Cost-sensitive segmentation encodes asymmetric consequences in prediction or post-processing (Chan et al., 2019; 2020), and Bayes decision rules select the minimum-cost label for each prediction independently (Elkan, 2001). Decision-focused learning aligns training with a downstream decision problem (Wilder et al., 2019; Elmachtoub & Grigas, 2022; Mandi et al., 2024; Sadana et al., 2025; Ovalle et al., 2025; Pan et al., 2025); SCARP keeps the segmenter frozen and optimizes only the post-hoc interventions.

Structured inference. Dense CRFs (Krähenbühl & Koltun, 2011), higher-order and label-cost potentials (Kohli et al., 2009; Delong et al., 2012), and combinatorial panoptic grouping (Abbas & Swoboda, 2021) couple decisions to enforce labeling consistency or grouping; SCARP couples decisions through their downstream consequence. Its scene potential is supermodular in the retained set (Boros & Hammer, 2002; Kolmogorov & Zabih, 2004) and budget-constrained (Appendix C.6), so graph cuts do not apply; we solve it exactly as a 0–1 MILP.

3 METHOD

Section 3.1 prices each action of a single component, Section 3.2 adds a scene term that couples components, and Section 3.3 bounds the removed evidence and solves the joint problem exactly. Appendix A lists all symbols.

3.1 COMPONENT-LEVEL BAYESIAN DECISIONS

Let Ω be the pixels of an image, $\mathcal{Y} = \{c_0, \dots, c_{K-1}\}$ the label space, and $c_h \in \mathcal{Y}$ the high-risk class. A frozen segmenter assigns each $u \in \Omega$ a probability vector $p_u = (p_u(c))_{c\in\mathcal{Y}}$. Let $C_1, \dots, C_n$ be the connected components of $\{u \in \Omega : \arg\max_{c\in\mathcal{Y}} p_u(c) = c_h\}$; if $n = 0$, the prediction is returned unchanged. Each component C_i is one decision unit, summarized by its normalized area, its mean probability of each $c_q \in \mathcal{Y}$, and its mean high-risk probability:

$$M_i = \frac{|C_i|}{|\Omega|}, \qquad P_i(c_q) = \frac{1}{|C_i|}\sum_{u\in C_i} p_u(c_q), \qquad \pi_i = P_i(c_h). \tag{1}$$

The application specifies a consequence matrix $R = [R_{\ell q}] \in \mathbb{R}^{K\times K}_{\geq 0}$, where $R_{\ell q}$ is the cost of assigning c_ℓ to a pixel of true class c_q. Assigning c_ℓ to all pixels of C_i costs

$$L_i(c_\ell) = M_i \sum_{q=0}^{K-1} R_{\ell q} P_i(c_q), \tag{2}$$

the posterior expected cost when pixel probabilities are calibrated. Each component either keeps c_h or is rerouted to a non-high-risk class:

$$\mathcal{A}_i = \{\textsc{Keep}\} \cup \{\textsc{Reroute}(c) : c \in \mathcal{D}_i\}, \qquad \mathcal{D}_i \subseteq \mathcal{Y} \setminus \{c_h\}, \tag{3}$$

where $\mathcal{D}_i$ holds the destinations that survive exact Pareto pruning (Section 3.3). We write $L_{i,a}$ for the cost of $a \in \mathcal{A}_i$, so $L_{i,\textsc{Keep}} = L_i(c_h)$ and $L_{i,\textsc{Reroute}(c)} = L_i(c)$. Deciding each component alone gives the Bayes rule

$$a_i^{\text{Bayes}} \in \arg\min_{a\in\mathcal{A}_i} L_{i,a}, \tag{4}$$

which is jointly optimal only if no decision changes the cost or the available actions of another component. A joint configuration $z = (z_{i,a})$ selects one action per component, and x_i marks whether C_i is kept:

$$z_{i,a} \in \{0,1\}, \qquad \sum_{a \in \mathcal{A}_i} z_{i,a} = 1 \quad \forall i, \qquad x_i = z_{i,\textsc{Keep}}. \tag{5}$$

3.2 From Independent Decisions to Scene Coupling

Why independent decisions are insufficient. Summing the unary costs gives the additive objective

$$J_{\mathrm{add}}(z) = \sum_{i=1}^{n} \sum_{a \in \mathcal{A}_i} z_{i,a} L_{i,a}, \tag{6}$$

which splits into the n problems (4): the gain of replacing KEEP by a reroute a,

$$g_{i,a} = L_{i,\textsc{Keep}} - L_{i,a}, \tag{7}$$

depends only on C_i and a; $g_{i,a} > 0$ favors rerouting. Removing a weak component thus costs the same whether or not it is the last high-risk evidence in its region. SCARP adds this dependence in three steps.

Step 1: contexts. A fixed partition of $\{1, \dots, n\}$ into nonempty contexts $\mathcal{N}_1, \dots, \mathcal{N}_m$ groups components whose retention is assessed jointly; $r(i)$ is the context containing C_i. It is built from predicted geometry, never from z or ground truth: here, the nonempty cells of a 3×3 grid over centroids (Appendix B.8).

Step 2: confidence and retention. For $i \in \mathcal{N}_r$, ω_{ri} is C_i's share of the context's predicted high-risk area, e_r the context's area-weighted high-risk confidence, and $s_r(z)$ the fraction of its area that z retains:

$$\omega_{ri} = \frac{M_i}{\sum_{j \in \mathcal{N}_r} M_j}, \qquad e_r = \sum_{i \in \mathcal{N}_r} \omega_{ri} \pi_i, \qquad s_r(z) = \sum_{i \in \mathcal{N}_r} \omega_{ri} x_i. \tag{8}$$

Step 3: the scene potential. Each context is penalized when retention disagrees with confidence:

$$\Phi_r(z) = A_r \, s_r(z)^2 + B_r \big(1 - s_r(z)\big)^2, \qquad A_r = \lambda_r^{\mathrm{fa}}(1 - e_r), \qquad B_r = \lambda_r^{\mathrm{miss}} e_r, \tag{9}$$

where $\lambda_r^{\mathrm{fa}}, \lambda_r^{\mathrm{miss}} \geq 0$ weight false alarms and misses. Equivalently, $\Phi_r = \mathbb{E}_{Y \sim \mathrm{Bern}(e_r)}[w(Y)(s_r(z) - Y)^2]$ with $w(0) = \lambda_r^{\mathrm{fa}}$, $w(1) = \lambda_r^{\mathrm{miss}}$: a cost-weighted Brier score. We adopt this quadratic as a modeling choice; a penalty linear in s_r would split over components and couple nothing. If $S_r = A_r + B_r > 0$,

$$\Phi_r(z) = S_r \big(s_r(z) - s_r^\star\big)^2 + \frac{A_r B_r}{S_r}, \qquad s_r^\star = \frac{B_r}{S_r} \in [0,1], \tag{10}$$

so Φ_r pulls retention toward the target $s_r^\star$ with strength S_r; if $S_r = 0$, then $\Phi_r \equiv 0$. Because s_r is a decision rather than a forecast, the target stays below one even when every component is confident; the unary costs can counterbalance this effect when they favor KEEP.

Relation to the consequence matrix. We set $\lambda_r^{\mathrm{fa}} = \rho_{\mathrm{fa}} \lambda_{\mathrm{base}}/m$ and $\lambda_r^{\mathrm{miss}} = \rho_{\mathrm{miss}} \lambda_{\mathrm{base}}/m$ with dimensionless $\rho_{\mathrm{fa}}, \rho_{\mathrm{miss}} \geq 0$, where λ_{base} is the median of the nonzero $|\gamma_i|$, $\gamma_i = \max_{a \neq \textsc{Keep}} g_{i,a}$ (and 0 if none exist). R thus enters directly only through $L_{i,a}$, and indirectly through λ_{base}, which scales Φ_r; rescaling R leaves every decision unchanged. The ratio $\rho_{\mathrm{fa}}/\rho_{\mathrm{miss}}$ plays at scene level the role of the off-diagonal entries of R; we set $\rho_{\mathrm{fa}} = \rho_{\mathrm{miss}} = \rho$, so $s_r^\star = e_r$, and $\rho = 0$ gives $\Phi_r \equiv 0$. SCARP minimizes

$$J_{\mathrm{SCARP}}(z) = \sum_{i=1}^{n} \sum_{a \in \mathcal{A}_i} z_{i,a} L_{i,a} + \sum_{r=1}^{m} \Phi_r(z). \tag{11}$$

When does context reverse a local preference? Write $(z_{i=b}, z_{-i})$ for z with the action of C_i set to b. All reroutes of C_i share one scene term, so it suffices to compare KEEP with each reroute a. The conditional gain $\Delta_i(a \mid z_{-i}) = J_{\mathrm{SCARP}}(z_{i=\textsc{Keep}}, z_{-i}) - J_{\mathrm{SCARP}}(z_{i=a}, z_{-i})$ is positive when a is preferred. With $r = r(i)$ and $s_{r,-i} = \sum_{j \in \mathcal{N}_r \setminus \{i\}} \omega_{rj} x_j$, substitution gives (Appendix C)

$$\Delta_i(a \mid z_{-i}) = g_{i,a} + \omega_{ri}\big[S_r(2s_{r,-i} + \omega_{ri}) - 2B_r\big] = g_{i,a} + 2\omega_{ri}S_r\Big(s_{r,-i} + \tfrac{\omega_{ri}}{2} - s_r^\star\Big), \quad (12)$$

the second form requiring $S_r > 0$. The scene correction favors rerouting when the context, with C_i counted as half-kept, exceeds its target, so the preference depends on the neighbors. For two equal-area components ($\omega_{ri} = \frac{1}{2}$) with $s_r^\star = \frac{1}{2}$, the correction is $+S_r/4$ if the neighbor is kept and $-S_r/4$ otherwise; for $|g_{i,a}| < S_r/4$, each prefers the opposite of its neighbor, which no configuration-invariant local score can represent.

Proposition 3.1 (Configuration-dependent Keep–Reroute preference)**.** *Let $r = r(i)$, $\omega = \omega_{ri}$, $S = S_r > 0$, and $\mathcal{Z}_{-i}(a) = \{z_{-i} : (z_{i=a}, z_{-i}) \in \mathcal{F}\}$, with $\mathcal{F}$ from Section 3.3; since $\mathbf{v}_{i,\textsc{Keep}} = \mathbf{0}$, also $(z_{i=\textsc{Keep}}, z_{-i}) \in \mathcal{F}$. Let $\underline{s} = \min_{\mathcal{Z}_{-i}(a)} s_{r,-i}$ and $\overline{s} = \max_{\mathcal{Z}_{-i}(a)} s_{r,-i}$. The strict preference between* KEEP *and a differs between two configurations in $\mathcal{Z}_{-i}(a)$ if and only if*

$$S\omega\big(2s_r^\star - \omega - 2\overline{s}\big) \;<\; g_{i,a} \;<\; S\omega\big(2s_r^\star - \omega - 2\underline{s}\big). \quad (13)$$

The interval has width $2\omega S(\overline{s} - \underline{s})$ and is nonempty exactly when $\overline{s} > \underline{s}$.

Corollary 3.2. *If every keep pattern of the other components of $\mathcal{N}_r$ is feasible, the interval is $S\omega(\omega + 2s_r^\star - 2) < g_{i,a} < S\omega(2s_r^\star - \omega)$, of width $2\omega(1-\omega)S$, positive iff $\omega < 1$. Inside it, no score depending only on (i, a) reproduces the preference in every configuration.*

The example is the case $\omega = s_r^\star = \frac{1}{2}$. Reversals cannot occur for singleton contexts or when $S_r = 0$. Proofs are in Appendix C.

3.3 PRESERVATION-CONSTRAINED JOINT OPTIMIZATION

Why interventions must be bounded. Nothing so far limits how many components are rerouted. Under a false-alarm-sensitive R, rerouting is cheap, and unconstrained Bayes retains only 29% of supported components on ACDC (Section 4.3). We therefore add prediction-only budgets. A reroute $a = \textsc{Reroute}(c)$ has protection multiplier $\eta_{i,a} = 1$ only when c is the extractor-selected destination $\hat{c}_i$ and its prediction-only dominance margin over KEEP is strictly positive; otherwise $\eta_{i,a} = 2$ (Appendix B.7). Each action consumes a resource vector $\mathbf{v}_{i,a}$ from the image budget $\boldsymbol{\beta}$:

$$\mathbf{v}_{i,a} = \begin{bmatrix} \eta_{i,a}\pi_i \\ \eta_{i,a}M_i\pi_i \\ 1 \end{bmatrix}, \quad \mathbf{v}_{i,\textsc{Keep}} = \mathbf{0}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \delta_{\mathrm{comp}} \sum_i \pi_i \\ \delta_{\mathrm{mass}} \sum_i M_i\pi_i \\ \kappa \end{bmatrix}. \quad (14)$$

The rows bound removed confidence, removed area-weighted confidence, and the number of edits; $\delta_{\mathrm{comp}}, \delta_{\mathrm{mass}} \in (0, 1]$ are preservation fractions and $\kappa = \lceil 0.2n \rceil$. The feasible set $\mathcal{F}$ contains every z satisfying (5) and $\sum_{i,a} \mathbf{v}_{i,a} z_{i,a} \preceq \boldsymbol{\beta}$ elementwise.

The SCARP program. Writing out Φ_r from (9), SCARP solves

$$\boxed{z^\star \in \arg\min_{z \in \mathcal{F}} \left[\sum_{i=1}^{n} \sum_{a \in \mathcal{A}_i} z_{i,a} L_{i,a} + \sum_{r=1}^{m} \Big(A_r\, s_r(z)^2 + B_r\big(1 - s_r(z)\big)^2 \Big) \right]}. \quad (15)$$

Budgets couple decisions through $\mathcal{F}$, the scene term through the objective. Deleting only $\sum_r \Phi_r$ gives SCARP-ADDITIVE, so any difference from SCARP-JOINT is caused by objective coupling; deleting the budgets too recovers BAYES-INDEPENDENT.

Exact solution. A destination is pruned if another has no larger $L_{i,a}$ and $\eta_{i,a}$, one strictly smaller. Since Φ_r depends only on keep states and every budget row is nondecreasing in $\eta_{i,a}$, this preserves the optimal value. Since x_i is binary,

$$s_r(z)^2 = \sum_{i \in \mathcal{N}_r} \omega_{ri}^2 x_i + 2 \sum_{\substack{i<j \\ i,j \in \mathcal{N}_r}} \omega_{ri}\omega_{rj} x_i x_j, \quad (16)$$

and replacing each $x_i x_j$ by $\mu_{rij} \in [0, 1]$ with $\mu_{rij} \le x_i$, $\mu_{rij} \le x_j$, and $\mu_{rij} \ge x_i + x_j - 1$ (McCormick, 1976) yields an equivalent 0–1 MILP, solved to proven optimality (Appendix C).

4 EXPERIMENTS

We ask whether coupling improves the preservation–risk trade-off (Q1), how SCARP compares with other rules (Q2), what drives the difference (Q3), and where SCARP falls short (Q4).

4.1 SETUP

Data and models. We use five test domains (19,183 test images): ACDC (Sakaridis et al., 2021), CityscapesFoggy (Sakaridis et al., 2018), and NightCity (Tan et al., 2021) for urban domain shift; the camera subset of UAVScenes (Wang et al., 2025); and Antarctic, 752 test images captured hourly in the 2024–25 austral summer by two AIoT monitoring platforms in East Antarctica (Anonymous, 2025), released anonymously at `https://anonymous.4open.science/r/SCARP-Antarctic-3311/` (Appendix B.4). The main predictor is a frozen SegMAN-B (Fu et al., 2025), with its Cityscapes-trained (Cordts et al., 2016) checkpoint on urban benchmarks and in-domain checkpoints on UAVScenes and Antarctic. Mechanism audits add SegFormer-B5 (Xie et al., 2021), Mask2Former-Swin-L (Cheng et al., 2022; Liu et al., 2021), and DeepLabV3+-R101 (Chen et al., 2018; He et al., 2016). All datasets map to a six-slot interface with one high-risk slot (Appendix B.3).

Protocol. Components are 8-connected with $\alpha_{\min} = 10$ pixels. Development/test splits are 122/284, 150/350, and 390/909 images for ACDC, CityscapesFoggy, and NightCity; UAVScenes uses 5,790/1,448/16,888 train/validation/test images. SCARP fits no parameters, and every action vector is frozen before test ground truth is loaded. The consequence matrix is fixed a priori and false-alarm-sensitive (Appendix B.5); protection of real hazards comes from the budgets. Unless stated otherwise, budgeted methods use $\delta_{\text{comp}} = \delta_{\text{mass}} = \delta = 0.5$, $\rho_{\text{fa}} = \rho_{\text{miss}} = \rho = 4$, and the 20% edit cap, abbreviated $(\delta, \rho) = (0.5, 4)$.

Metrics. A predicted high-risk component is *supported* if it intersects the ground-truth support G_h of c_h and a *ghost* otherwise; G_h is used only for evaluation. For a decision z with retained set $\mathcal{J}(z) = \{i : x_i = 1\}$,

$$\text{Ghost-FDP}(z) = \frac{\sum_{i \in \mathcal{J}(z)} \mathbf{1}[C_i \cap G_h = \varnothing]}{\max(1, |\mathcal{J}(z)|)}, \tag{17}$$

and Ghost-FDR (G-FDR), its image-macro test average (Benjamini & Hochberg, 1995), measures risk. Supported-component retention (Ret.), ground-truth component recall (C-Rec), and high-risk pixel recall (HR-Rec) measure preservation; we also report mIoU and CostRisk (Appendix E). *Binding* is the fraction of images on which Joint and Additive select different actions.

4.2 Q1: COUPLING IMPROVES THE PRESERVATION–RISK TRADE-OFF

In short, coupling preserves more at a fixed budget and lowers risk at matched preservation. On the urban and UAV benchmarks, both budgeted programs trade supported components for lower G-FDR relative to NO-EDIT (Table 1).

Joint preserves more than Additive under the same budget. Joint achieves higher C-Rec, HR-Rec, retention, and mIoU than Additive on all five domains, by 1.22–2.11, 0.93–2.00, 0.78–5.00, and 0.15–0.42 points on the urban and UAV benchmarks. On ACDC, paired bootstrap 95% intervals (5,000 resamples) exclude zero for retention ($+5.00$, $[+3.50, +6.40]$), C-Rec ($+2.00$, $[+0.80, +3.30]$), HR-Rec ($+2.00$, $[+1.00, +3.00]$), and mIoU ($+0.20$, $[+0.05, +0.40]$), and retention and recall point estimates are positive on all four benchmarks with paired outputs (Figure 2). The gap is largest on Antarctic, where Additive lowers C-Rec from 0.91 to 0.66 but Joint keeps 0.84 and HR-Rec at the uncorrected 0.97, with no worse risk: Additive spends its budget on supported components that the scene term protects.

The gain is not bought with risk. At the fixed point, risk differences are modest and of mixed sign: Joint lowers G-FDR relative to Additive on ACDC, NightCity, and Antarctic and raises it slightly elsewhere; its CostRisk is higher by 2.6×10^{-6} to 2.1×10^{-4} except on Antarctic; on ACDC

Table 1: **Matched Joint–Additive comparison on five test domains** at $(\delta, \rho) = (0.5, 4)$ with a 20% edit cap. Joint and Additive share actions and the feasible set and differ only in the scene term. Bold marks the better of the two, determined from unrounded values. Entries are test means; CostRisk differences between Joint and Additive lie below the displayed precision and are reported in Section 4.2.

Dataset	Method	G-FDR↓	C-Rec↑	HR-Rec↑	Ret.↑	mIoU↑	CostRisk↓
ACDC	NO-EDIT	0.13	0.95	0.93	1.00	0.84	0.03
	SCARP-ADDITIVE	0.09	0.77	0.90	0.80	0.84	**0.02**
	SCARP-JOINT	**0.06**	**0.79**	**0.92**	**0.85**	**0.84**	0.02
CityscapesFoggy	NO-EDIT	0.01	0.90	0.96	1.00	0.88	0.02
	SCARP-ADDITIVE	**0.00**	0.75	0.93	0.84	0.87	**0.02**
	SCARP-JOINT	0.01	**0.76**	**0.93**	**0.85**	**0.87**	0.02
NightCity	NO-EDIT	0.40	0.86	0.85	1.00	0.49	0.12
	SCARP-ADDITIVE	0.37	0.65	0.77	0.77	0.48	**0.12**
	SCARP-JOINT	**0.37**	**0.67**	**0.78**	**0.79**	**0.48**	0.12
UAVScenes	NO-EDIT	0.06	1.00	0.98	1.00	0.92	0.00
	SCARP-ADDITIVE	**0.05**	0.83	0.84	0.83	0.72	**0.00**
	SCARP-JOINT	0.05	**0.84**	**0.85**	**0.84**	**0.72**	0.00
Antarctic	NO-EDIT	0.03	0.91	0.97	1.00	0.80	0.02
	SCARP-ADDITIVE	0.04	0.66	0.92	0.67	0.76	0.02
	SCARP-JOINT	**0.03**	**0.84**	**0.97**	**0.89**	**0.79**	**0.02**

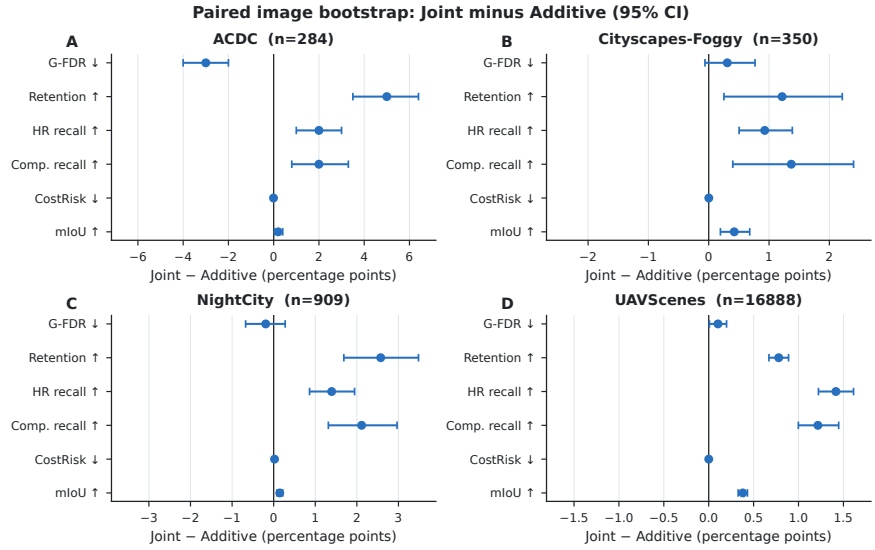


Figure 2: **Joint minus Additive: paired bootstrap 95% intervals** (5,000 image-level resamples) on the four benchmarks with paired per-image outputs. Retention and recall point estimates are positive on all four, whereas G-FDR differences at this fixed operating point have mixed sign and CostRisk differences are negligible.

only the G-FDR risk interval excludes zero. Because the programs sit at different retention levels, this comparison mixes the two axes. At *equal* retention, Joint attains lower G-FDR (Table 3), so coupling shifts the frontier. We treat $(0.5, 4)$ as a diagnostic point, not a deployment optimum (Appendix H.1).

Table 2: **Post-hoc correction rules on ACDC.** Budgeted methods share $(\delta, \rho) = (0.5, 4)$, the edit cap, and the feasible set; BAYES-INDEPENDENT removes the budgets. Bold marks the better of Joint and Additive (unrounded values).

Method	G-FDR↓	Ret.↑	HR-Rec↑	CostRisk↓	mIoU↑
NO-EDIT	0.13	1.00	0.93	0.03	0.84
SCARP-JOINT	**0.06**	**0.85**	**0.92**	0.02	**0.84**
SCARP-ADDITIVE	0.09	0.80	0.90	**0.02**	0.84
META@K	0.08	0.84	0.92	0.03	0.84
META-BUDGET	0.09	0.97	0.93	0.03	0.84
CRC-FDR	0.10	0.99	0.93	0.03	0.84
TOPB	0.09	0.84	0.92	0.03	0.84
Entropy	0.09	0.83	0.92	0.03	0.84
RANDOM@K	0.13	0.76	0.90	0.03	0.84
BAYES-INDEPENDENT	0.01	0.29	0.74	0.03	0.81

Table 3: **Matched-retention comparison on ACDC test.** Comparator operating points are matched to Joint within ± 0.01 supported-component retention; the nearest TOPB point lies at 0.84. Entries are Joint minus comparator (paired bootstrap, 5,000 resamples); negative values favor Joint.

Comparator	Matched Ret.	ΔG-FDR	95% CI for ΔG-FDR	ΔCostRisk
SCARP-ADDITIVE	0.85	-0.028	$[-0.041, -0.015]$	-1.1×10^{-4}
META@K	0.85	-0.024	$[-0.036, -0.012]$	-8.3×10^{-5}
TOPB	0.84	-0.030	$[-0.043, -0.017]$	-9.5×10^{-5}
Entropy	0.85	-0.022	$[-0.034, -0.010]$	-7.1×10^{-5}

4.3 Q2: COMPARISON WITH POST-HOC CORRECTION RULES

In short, Joint has the lowest G-FDR at matched retention on ACDC. We compare with unary rankings that reroute as many components as Joint (TOPB, entropy (Kendall & Gal, 2017), RANDOM@K); component meta-classification (META@K, META-BUDGET (Rottmann et al., 2020)); conformal risk control (CRC-FDR (Angelopoulos et al., 2024; Mossina & Friedrich, 2025)); and unconstrained local Bayes (Table 2).

A fixed operating point cannot rank the methods. At the reported budgeted operating point, Joint attains the lowest G-FDR (0.06) while retaining 0.85 of supported components, compared with 0.09/0.80 for Additive, 0.08/0.84 for META@K, and 0.09/0.84 for TOPB. Because these methods operate at different retention levels, the fixed operating point does not by itself establish a preservation–risk ordering. At operating points matched within ± 0.01 retention (Table 3), Joint lowers G-FDR by 2.2–3.0 points relative to Additive, META@K, TOPB, and entropy, with all four paired 95% intervals excluding zero, and also lowers CostRisk (Appendix H.8).

The other rules show what matters. RANDOM@K leaves G-FDR unchanged, so *which* components are rerouted matters; META-BUDGET and CRC-FDR remove fewer ghosts; BAYES-INDEPENDENT retains only 29% of supported components. As these controls are interface-matched proxies (Appendix D), mechanistic claims rest on the Joint–Additive ablation.

4.4 Q3: WHAT DRIVES THE DIFFERENCE, AND IS IT ROBUST?

In short, the scene term causes the difference through two mechanisms, robustly.

Cause. Joint and Additive differ on 7.4–21.1% of test images (Table 4), and only $\sum_r \Phi_r$ separates them. Over a 7×5 grid of (δ, ρ), binding is nonzero in 19 of 35 cells, exactly zero whenever $\rho = 0$, and increases monotonically with ρ at $\delta = 0.5$, all solved to optimality (Appendix H.4).

Table 4: **Scene-coupling audit.** Binding: fraction of images on which Joint and Additive select different action vectors. Development rows use SegMAN-B on ACDC ($\delta = 0.5$); strict reversals (Proposition 3.1) pool four backbones.

Test binding at $(0.5, 4)$	
ACDC / CityscapesFoggy	16.6% / 21.1%
NightCity / UAVScenes	18.0% / 7.4%
Coupling strength (dev.)	
$\rho = 0$ / 1 / 2	0.0% / 6.6% / 9.8%
$\rho = 4$ / 8	21.3% / 32.0%

Context grid (dev., $\rho = 4$*)*	
1×1 / 2×2	2.5% / 18.0%
3×3 / 4×4	21.3% / 24.6%
Temperature-scaled (ACDC test)	16.6% → 19.0%
Strict reversals at $(0.5, 4)$	17 / 96 cases

Mechanism 1: local preference reversal. For every (component, action, neighbor, neighbor action) case at $(0.5, 4)$ with four feasible joint states, we test whether the preferred action flips with the neighbor's action (Proposition 3.1). It does in 17 of 96 cases, on all four backbones and well above solver tolerance; reversals vanish at $\rho = 0$ and reach 23.91% at $\rho = 8$ (Appendix H.7).

Mechanism 2: budget reallocation. Because reroutes compete for shared budgets, the scene term also changes which interventions receive them. With $\rho = 4$ and every $g_{i,a}$ fixed, relaxing δ from 0.05 to 1.00 raises the share of feasible positive-gain reroutes from 0.33% to 100% and binding from 0% to 24.0% (Appendix F, Figure 4): tight budgets leave nothing to choose between.

Robustness. Binding persists for context grids from 1×1 to 4×4 (growing with resolution), for minimum component areas from 5 to 50 pixels (21.3–23.8%), after temperature scaling, and under three consequence-matrix families on UAVScenes (10.6% to 4.4%), retaining more in each (Appendices H.3, H.5, H.6, H.10).

4.5 Q4: Where SCARP Falls Short

Rerouting destinations are not semantically reliable. Among 123 components rerouted on the ACDC development split, the destination matches the majority ground-truth class in only 3.25% of cases, and within-component pixel accuracy improves for 5.69% of them and worsens for 61.79% (Appendix H.9). SCARP's decisions should therefore be read as *whether to retain* a high-risk label; the destination is a consequence-minimizing placeholder, not a semantic prediction. This also bounds dense accuracy: on UAVScenes, both budgeted programs lower mIoU from 0.92 to 0.72. The drop barely differs between Joint and Additive, so it stems from the budget and destinations, not coupling; an explicit *unknown* destination is a remedy.

Operating behavior depends on calibration and scale. Temperature scaling lowers ACDC test ECE from 0.058 to 0.014 but, at the same budgets, it lowers retention and raises G-FDR (Appendix H.10), and absolute G-FDR falls from 0.227 to 0.068 as $\alpha_{\min}$ grows from 5 to 50 pixels (Appendix H.6); budgets and comparisons must be fixed for the deployed calibration and extraction scale. Exact solving is fast on real instances (all 488 MILPs optimal, median 1.7 ms, at most 16 components) and on synthetic instances with up to 128 components (99.8% optimal, median 151 ms; Appendix H.11), but much larger scenes may require decomposition or approximation. Finally, matched-frontier comparisons are currently reported on ACDC only.

5 Conclusion

We introduced SCARP, a training-free decision layer that corrects high-risk predictions of frozen segmenters by jointly deciding which components to keep. Its central observation is that, when ghost identity is unobservable, the value of an intervention depends on what the scene retains; we characterize exactly when this induces preference reversals and solve the resulting budget-constrained problem to optimality. Empirically, coupling changes decisions on four paired benchmarks, preserves more supported components across all five domains, and lowers Ghost-FDR at matched retention on ACDC. Post-hoc reliability, we argue, should be optimized where consequences interact.

AI USE STATEMENT

Large language models assisted with drafting and language editing. The authors verified all technical claims, derivations, code, and reported results, and take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Appendix B specifies the label interface, consequence matrix, action policy, normalization, budgets, solver, and evaluation aggregation; the anonymized Antarctic data, split files, probability caches, and code are available at the anonymous repository `https://anonymous.4open.science/r/SCARP-Antarctic-3311/` (Appendix B.4); Appendix C contains complete proofs; Appendix F gives the mechanism protocols; and Appendix H reports every supplementary audit together with its exact configuration.

REFERENCES

Ahmed Abbas and Paul Swoboda. Combinatorial optimization for panoptic segmentation: A fully differentiable approach. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 15635–15649, 2021.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024.

Anonymous. Antarctic AIoT environmental-monitoring field dataset. National Antarctic data repository, 2025. Identifying details withheld for double-blind review.

Stephen Bates, Anastasios N. Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):43:1–43:34, 2021. doi: 10.1145/3478535.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Endre Boros and Peter L. Hammer. Pseudo-Boolean optimization. *Discrete Applied Mathematics*, 123(1–3):155–225, 2002. doi: 10.1016/S0166-218X(01)00341-9.

Robin Chan, Matthias Rottmann, Radin Dardashti, Fabian Hüger, Peter Schlicht, and Hanno Gottschalk. The ethical dilemma when (not) setting up cost-based decision rules in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1395–1403, 2019. doi: 10.1109/CVPRW.2019.00180.

Robin Chan, Matthias Rottmann, Fabian Hüger, Peter Schlicht, and Hanno Gottschalk. Controlled false negative reduction of minority classes in semantic segmentation. In *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2020.

Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. SegmentMeIfYouCan: A benchmark for anomaly segmentation. In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 801–818, 2018. doi: 10.1007/978-3-030-01234-2_49.

Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C. Berg, and Alexander Kirillov. Boundary IoU: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15334–15342, 2021.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1290–1299, 2022. doi: 10.1109/CVPR52688.2022.00135.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, 2016. doi: 10.1109/CVPR.2016.350.

Mengxia Dai, Wenqian Luo, and Tianyang Li. Statistical guarantees of false discovery rate in medical instance segmentation tasks based on conformal risk control. *arXiv preprint arXiv:2504.04482*, 2025.

Andrew Delong, Anton Osokin, Hossam N. Isack, and Yuri Boykov. Fast approximate energy minimization with label costs. *International Journal of Computer Vision*, 96(1):1–27, 2012. doi: 10.1007/s11263-011-0437-z.

Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 973–978, 2001.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68 (1):9–26, 2022. doi: 10.1287/mnsc.2020.3922.

Yunxiang Fu, Meng Lou, and Yizhou Yu. SegMAN: Omni-scale context modeling with state space models and local attention for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19077–19087, 2025. doi: 10.1109/CVPR52734.2025.01777.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 4878–4887, 2017.

Matej Grcić, Josip Šarić, and Siniša Šegvić. On advantages of mask-level recognition for outlier-aware segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2937–2947, 2023.

Vanessa Emanuela Guarino, Claudia Winklmayr, Jannik Franzen, Josef Lorenz Rumberger, Manuel Pfeuffer, Sonja Greven, Klaus Maier-Hein, Dagmar Kainmueller, Christoph Karg, and Carsten T. Lüth. Better than average: Spatially-aware aggregation of segmentation uncertainty improves downstream performance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13145–13156, 2026.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Edgar Heinert, Stephan Tilgner, Timo Palm, and Matthias Rottmann. Uncertainty and prediction quality estimation for semantic segmentation via graph neural network. In *British Machine Vision Conference (BMVC) Workshops*, 2024. Robust Recognition in the Open World (RROW) Workshop.

Alvin Heng and Harold Soh. Know when to abstain: Optimal selective classification with likelihood ratios. In *International Conference on Learning Representations (ICLR)*, 2026.

Xiaoling Hu, Fuxin Li, Dimitris Samaras, and Chao Chen. Topology-preserving deep image segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 5657–5668, 2019.

Yang Hu, Jieren Tan, Changliang Zou, Yajie Bao, and Haojie Ren. Conformal robustness control: A new strategy for robust decision. In *International Conference on Learning Representations (ICLR)*, 2026.

Qi Huangfu and J. A. Julian Hall. Parallelizing the dual revised simplex method. *Mathematical Programming Computation*, 10(1):119–142, 2018. doi: 10.1007/s12532-017-0130-5.

Kim-Celine Kahl, Carsten Lüth, Maximilian Zenk, Klaus Maier-Hein, and Paul F. Jaeger. ValUES: A framework for systematic validation of uncertainty estimation in semantic segmentation. In *International Conference on Learning Representations (ICLR)*, 2024.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 5574–5584, 2017.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9404–9413, 2019.

Pushmeet Kohli, L'ubor Ladický, and Philip H. S. Torr. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 82(3):302–324, 2009. doi: 10.1007/s11263-008-0202-0.

Vladimir Kolmogorov and Ramin Zabih. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, 2004. doi: 10.1109/TPAMI.2004.1262177.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 24, pp. 109–117, 2011.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

Rui Luo, Jie Bao, Xiaoyi Su, Wen Jung Li, and Suqun Cao. Enhancing image-conditional coverage in segmentation: Adaptive thresholding via differentiable miscoverage loss. In *International Conference on Learning Representations (ICLR)*, 2026.

Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Buettner, et al. Metrics reloaded: Recommendations for image analysis validation. *Nature Methods*, 21(2):195–212, 2024. doi: 10.1038/s41592-023-02151-z.

Jayanta Mandi, James Kotary, Senne Berden, Maxime Mulamba, Victor Bucarey, Tias Guns, and Ferdinando Fioretto. Decision-focused learning: Foundations, state of the art, benchmark and future opportunities. *Journal of Artificial Intelligence Research*, 80:1623–1701, 2024. doi: 10.1613/JAIR.1.15320.

Garth P. McCormick. Computability of global solutions to factorable nonconvex programs: Part I—convex underestimating problems. *Mathematical Programming*, 10(1):147–175, 1976. doi: 10.1007/BF01580665.

Luca Mossina and Corentin Friedrich. Controlling false positives in image segmentation via conformal prediction. *arXiv preprint arXiv:2511.15406*, 2025.

Luca Mossina, Joseba Dalmau, and Léo Andéol. Conformal semantic image segmentation: Post-hoc quantification of predictive uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 3574–3584, 2024. doi: 10.1109/CVPRW63382.2024.00361.

Nazir Nayal, Youssef Shoeb, and Fatma Güney. A likelihood ratio-based approach to segmenting unknown objects. *International Journal of Computer Vision*, 133(10):6860–6872, 2025. doi: 10.1007/s11263-025-02509-0.

Daniel Ovalle, Lorenz T. Biegler, Ignacio E. Grossmann, Carl D. Laird, and Mateo Dulce Rubio. Conformal mixed-integer constraint learning with feasibility guarantees. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Yiyuan Pan, Yunzhe Xu, Zhe Liu, and Hesheng Wang. Seeing through uncertainty: Robust task-oriented optimization in visual navigation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Jonah Philion, Amlan Kar, and Sanja Fidler. Learning to evaluate perception models using planner-centric metrics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14055–14064, 2020.

Shyam Nandan Rai, Fabio Cermelli, Dario Fontanel, Carlo Masone, and Barbara Caputo. Unmasking anomalies in road-scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4037–4046, 2023.

Matthias Rottmann, Pascal Colling, Thomas-Paul Hack, Robin Chan, Fabian Hüger, Peter Schlicht, and Hanno Gottschalk. Prediction error meta classification in semantic segmentation: Detection via aggregated dispersion measures of softmax probabilities. In *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9, 2020. doi: 10.1109/IJCNN48605.2020.9206659.

Utsav Sadana, Abhilash Chenreddy, Erick Delage, Alexandre Forel, Emma Frejinger, and Thibaut Vidal. A survey of contextual optimization methods for decision-making under uncertainty. *European Journal of Operational Research*, 320(2):271–289, 2025. doi: 10.1016/j.ejor.2024.03.020.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018. doi: 10.1007/s11263-018-1072-8.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10765–10775, 2021. doi: 10.1109/ICCV48922.2021.01059.

Youssef Shoeb, Azarm Nowzad, and Hanno Gottschalk. Out-of-distribution segmentation in autonomous driving: Problems and state of the art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4349–4359, 2025.

Nico Stucki, Johannes C. Paetzold, Suprosanna Shit, Bjoern Menze, and Ulrich Bauer. Topologically faithful image segmentation via induced matching of persistence barcodes. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32698–32727. PMLR, 2023.

Xin Tan, Ke Xu, Ying Cao, Yiheng Zhang, Lizhuang Ma, and Rynson W. H. Lau. Night-time scene parsing with a large real dataset. *IEEE Transactions on Image Processing*, 30:9085–9098, 2021. doi: 10.1109/TIP.2021.3122004.

Juan Miguel Valverde, Dim P. Papadopoulos, Rasmus Larsen, and Anders Bjorholm Dahl. Towards high-quality image segmentation: Improving topology accuracy by penalizing neighbor pixels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13123–13133, 2026.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Bruno Viti, Elias Karabelas, and Martin Holler. CONSIGN: Conformal segmentation informed by spatial groupings via decomposition. In *International Conference on Learning Representations (ICLR)*, 2026.

Sijie Wang, Siqi Li, Yawei Zhang, Shangshu Yu, Shenghai Yuan, Rui She, Quanjiang Guo, JinXuan Zheng, Ong Kang Howe, Leonrich Chandra, Shrivarshann Srijeyan, Aditya Sivadas, Toshan Aggarwal, Heyuan Liu, Hongming Zhang, Chujie Chen, Junyu Jiang, Lihua Xie, and Wee Peng Tay. UAVScenes: A multi-modal dataset for UAVs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28946–28958, 2025.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1658–1665, 2019. doi: 10.1609/aaai.v33i01.33011658.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 12077–12090, 2021.

Christopher Yeh, Nicolas Christianson, Adam Wierman, and Yisong Yue. Conformal risk training: End-to-end optimization of conformal risk control. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Botong Zhang, Shuo Li, and Osbert Bastani. Conformal structured prediction. In *International Conference on Learning Representations (ICLR)*, 2025.

Wenbin Zhou, Agni Orfanoudaki, and Shixiang Zhu. Conformalized decision risk assessment. In *International Conference on Learning Representations (ICLR)*, 2026.

Huachao Zhu, Zelong Liu, Zhichao Sun, Yuda Zou, Gui-Song Xia, and Yongchao Xu. Beyond pixel uncertainty: Bounding the OoD objects in road scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8472–8481, 2025.

A Notation

B Reproducibility Details

This appendix fixes the implementation contract used in the main paper. Unless stated otherwise, every quantity below is computed from the frozen probability field, the registered label adapter, and the application-specified consequence matrix before any ground truth is loaded.

B.1 Implementation Details

Compute environment. Evaluation was conducted on a server equipped with four NVIDIA A100 80 GB PCIe GPUs, two AMD EPYC 7402 CPU sockets (36 physical CPU cores visible to the virtual machine), and 334 GiB of usable RAM. The main experiments used Python 3.13.2, PyTorch 2.9.1, CUDA 12.8, and SciPy 1.16.3.

Optimization. SCARP's mixed-integer programs were solved with `scipy.optimize.milp` using HiGHS 1.8.0, with a relative MIP-gap tolerance of 10^{-8}. Gurobi was not used. Reported MILP runtimes measure the wall-clock duration of the complete SCARP `solve_milp` call, including problem construction and optimization, but excluding component extraction and segmentation inference.

Training and hyperparameters. SCARP itself fits no parameters. All segmentation backbones were pretrained separately and frozen before post-hoc correction. Unless otherwise stated, SCARP hyperparameters were fixed using development data prior to test evaluation and then held fixed across test domains.

B.2 Configurations at a Glance

Three configurations appear in the paper and should not be conflated.

Table 5: Notation used in Section 3.

Symbol	Meaning
Ω, $\mathcal{Y}$, K, c_h	Image pixels, label space, number of classes, high-risk class
p_u	Frozen predictive distribution at pixel u
C_i, n	i-th predicted high-risk connected component; number of components
M_i, $P_i(c_q)$, π_i	Normalized area, mean class probability, mean high-risk probability of C_i
G_h	Ground-truth support of c_h (evaluation only)
R, W	Consequence matrix R[action, true] and its input form $W = R^\top$
$L_i(c_\ell)$, $L_{i,a}$	Cost of assigning c_ℓ to C_i; cost of action a
$\mathcal{A}_i$, $\mathcal{D}_i$	Action set of C_i; its non-high-risk destinations after pruning
z, $z_{i,a}$, x_i	Configuration; action indicator; keep indicator $x_i = z_{i,\text{KEEP}}$
$(z_{i=b}, z_{-i})$	Configuration z with the action of C_i set to b
J_{add}, J_{SCARP}	Additive and SCARP objectives
$g_{i,a}$, γ_i	Unary gain of rerouting C_i with a; largest unary gain of C_i
$\mathcal{N}_r$, m, $r(i)$	r-th context of the partition; number of contexts; context of C_i
ω_{ri}, e_r, $s_r(z)$	Area share of C_i in context r; area-weighted high-risk confidence; retained fraction
Φ_r	Scene potential of context r, Eq. (9)
A_r, B_r, S_r, $s_r^\star$	False-alarm and miss weights; $S_r = A_r + B_r$; target B_r/S_r (if $S_r > 0$)
λ_r^{fa}, λ_r^{miss}	False-alarm and miss scales of context r
ρ_{fa}, ρ_{miss}, ρ, λ_{base}	Dimensionless coupling strengths (ρ when equal); per-image scale tied to R
$\Delta_i(a \mid z_{-i})$, $s_{r,-i}$	Conditional gain; retained fraction of the other components of context r
$\mathcal{Z}_{-i}(a)$, $\underline{s}$, $\overline{s}$	Feasible surrounding configurations; min and max of $s_{r,-i}$ over them
$\Gamma_i(c)$, $\hat{c}_i$	Prediction-only margin of destination c over KEEP; proposed destination
$\eta_{i,a}$, $\mathbf{v}_{i,a}$, $\boldsymbol{\beta}$	Protection multiplier; resource vector; budget vector
δ_{comp}, δ_{mass}, δ, κ	Preservation fractions (δ when equal); edit cap
$\mathcal{F}$, μ_{rij}	Feasible set; McCormick product variable

- **Main operating point**: $(\delta, \rho) = (0.50, 4)$, 3×3 centroid grid, 20% edit cap, and $\eta_{i,a} \in \{1, 2\}$ for certified and uncertified reroutes.
- **Controlled preservation sweep**: $\rho = 4$ fixed and $\delta \in \{0.05, 0.075, 0.10, 0.15, 0.20, 0.30, 0.50, 0.75, 1.00\}$.
- **Coupling-strength sweep**: $\delta = 0.5$ fixed and $\rho \in \{0, 1, 2, 4, 8\}$.

B.3 Label Interface and Dataset Adapters

Urban six-slot interface. For the Cityscapes-family experiments (Cordts et al., 2016), we use

$$(c_0, c_1, c_2, c_3, c_4, c_5) = (\text{reserved}, \text{dynamic/high-risk}, \text{structure}, \text{surface}, \text{nature}, \text{sky}),$$

with $c_h = c_1$. The Cityscapes trainId-to-slot vector is

$$[3, 3, 2, 2, 2, 255, 255, 255, 4, 4, 5, 1, 1, 1, 1, 1, 1, 1, 1],$$

where 255 denotes ignored probability mass. Road and sidewalk map to surface; building, wall, and fence to structure; vegetation and terrain to nature; sky to sky; and person, rider, car, truck, bus, train, motorcycle, and bicycle to the high-risk slot. Probabilities of source classes assigned to one slot are summed and renormalized over the registered slots.

For labelId-form annotations, the evaluation mapping is

slot	Cityscapes labelIds
c_1	$24, 25, 26, 27, 28, 29, 30, 31, 32, 33$
c_2	$9, 10, 11, 12, 13, 14$
c_3	$7, 8$
c_4	$21, 22$
c_5	23

and all remaining raw IDs are ignored.

UAVScenes. Only categories with explicit registered counterparts are evaluated:

1	$\mapsto$	c_2 (roof)	2	$\mapsto$	c_3 (dirt motor road)
3	$\mapsto$	c_3 (paved motor road)	6	$\mapsto$	c_2 (bridge)
9	$\mapsto$	c_2 (container)	10	$\mapsto$	c_3 (airstrip)
13	$\mapsto$	c_4 (green field)	14	$\mapsto$	c_4 (wild field)
17	$\mapsto$	c_2 (transparent roof)	18	$\mapsto$	c_3 (car park)
19	$\mapsto$	c_3 (paved walk)	20	$\mapsto$	c_1 (sedan)
24	$\mapsto$	c_1 (truck)			

Unregistered native IDs are ignored rather than semantically guessed.

Antarctic. Antarctic is processed through its native frozen six-channel adapter. The deployment-designated hazard is mapped to the high-risk slot before optimization. SCARP operates on registered slots and the consequence matrix rather than on an inferred cross-domain semantic taxonomy.

B.4 Antarctic Data Access

The Antarctic imagery was collected by two AIoT environmental-monitoring platforms deployed at a protected site in East Antarctica, logging sensor measurements and images hourly during the 2024–25 austral summer (Anonymous, 2025). The raw record contains approximately 2.8k files and is archived in a national Antarctic repository. For double-blind review, identifying metadata are withheld. The de-identified evaluation images, annotations, split files, and frozen probability caches used in the experiments are available at the anonymous repository `https://anonymous.4open.science/r/SCARP-Antarctic-3311/`; a public link to the archived record will be added in the camera-ready version.

B.5 Consequence Matrix and Orientation

The external convention is $W[\text{true}, \text{action}]$, which is converted to $R[\text{action}, \text{true}] = W^\top$. The main experiments use

$$W = \frac{1}{3}\begin{bmatrix} 0 & 3 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & .8 & .3 & .1 \\ 1 & 3 & 0 & 1 & 1 & 1 \\ 1 & 3 & 1 & 0 & 1 & 1 \\ 1 & 3 & 1 & 1 & 0 & 1 \\ 1 & 3 & 1 & 1 & 1 & 0 \end{bmatrix}.$$

Normalization does not change any decision: it rescales every $L_{i,a}$ and, through λ_{base}, every Φ_r by the same constant. The matrix is deliberately false-alarm-sensitive; protection against wholesale removal of genuine hazards is supplied separately by the preservation constraints.

B.6 Component Extraction and Local Posterior Summaries

The frozen predictor is decoded by standard argmax. We extract 8-connected components of the predicted high-risk slot and discard components containing fewer than $\alpha_{\min} = 10$ pixels. With uniform pixel weights,

$$M_i = \frac{|C_i|}{|\Omega|}, \qquad P_i(c_q) = \frac{1}{|C_i|}\sum_{u \in C_i} p_u(c_q), \qquad \pi_i = P_i(c_h).$$

Ground truth is not loaded during construction of these quantities.

B.7 Action Space, Certificates, and Exact Pareto Pruning

Let $\mathcal{D}_i^\circ \subseteq \mathcal{Y} \setminus \{c_h\}$ be the registered non-high-risk destinations, from which Pareto pruning yields $\mathcal{D}_i$.

Prediction-only destination certificate. Let $R_{\max} = \max_{\ell,q} R_{\ell q}$ and let $\bar{R} = R/R_{\max}$ be the normalized consequence matrix, with $\bar{R} = \mathbf{0}$ if $R_{\max} = 0$. For a candidate destination $c \in \mathcal{D}_i^\circ$, the implementation defines the prediction-only dominance margin

$$\Gamma_i(c) = \min_{\ell \in \mathcal{K}_i(c)} \min_{\nu \in \mathcal{Q}_i} \sum_{q=0}^{K-1} \nu(c_q) \Big[\max\{\bar{R}_{\ell q} - \tau, 0\} - \min\{\bar{R}_{cq} + \tau, 1\} \Big],$$

where $\mathcal{Q}_i$ is a prediction-only set of beliefs ν over $\mathcal{Y}$, $\tau \geq 0$ is a consequence-radius parameter, and $\mathcal{K}_i(c)$ is a set of competing classes whose assignment c must dominate. In all experiments, $\mathcal{Q}_i = \{P_i\}$, $\tau = 0$, and $\mathcal{K}_i(c) = \{c_h\}$, so the margin reduces exactly to

$$\Gamma_i(c) = \sum_{q=0}^{K-1} P_i(c_q) \big(\bar{R}_{hq} - \bar{R}_{cq} \big),$$

where $\bar{R}_{hq}$ is the row of $\bar{R}$ that assigns c_h. Thus $\Gamma_i(c) > 0$ means that, under the component posterior P_i, rerouting to c has lower expected normalized cost than keeping the high-risk label. By Eq. (2), $\Gamma_i(c)$ is the dimensionless form of the unary gain in Eq. (7):

$$g_{i,\textsc{Reroute}(c)} = L_{i,\textsc{Keep}} - L_{i,\textsc{Reroute}(c)} = M_i R_{\max} \Gamma_i(c).$$

Certification and protection multipliers. The extractor selects

$$\hat{c}_i \in \arg\max_{c \in \mathcal{D}_i^\circ} \Gamma_i(c),$$

breaking ties first by lower expected cost $L_i(c)$ and then by class index. Certification is

$$\mathrm{cert}_i(c) = \mathbf{1}[c = \hat{c}_i]\, \mathbf{1}[\Gamma_i(\hat{c}_i) > 0], \qquad \eta_{i,\textsc{Reroute}(c)} = \begin{cases} 1, & \mathrm{cert}_i(c) = 1, \\ 2, & \mathrm{cert}_i(c) = 0. \end{cases}$$

Only $\hat{c}_i$ can be certified. Uncertified destinations remain admissible but consume twice the preservation resources in the first two rows of Eq. (14).

Pareto pruning. A candidate c is Pareto-pruned if another candidate c' satisfies $L_i(c') \leq L_i(c)$ and $\eta_{i,\textsc{Reroute}(c')} \leq \eta_{i,\textsc{Reroute}(c)}$, with at least one strict inequality. Because the scene potential depends only on keep states and every budget row is nondecreasing in $\eta_{i,a}$, replacing c by c' in any feasible configuration keeps it feasible and does not increase the objective, so pruning preserves the optimal value.

B.8 Spatial Contexts and Scene Normalization

Let $(\xi_i^x, \xi_i^y) \in [0,1]^2$ be the normalized centroid of C_i. The main configuration assigns components to nonempty cells of a 3×3 grid. For each context,

$$\omega_{ri} = \frac{M_i}{\sum_{j \in \mathcal{N}_r} M_j}, \qquad e_r = \sum_{i \in \mathcal{N}_r} \omega_{ri} \pi_i.$$

The scene scale is

$$\lambda_{\mathrm{base}} = \mathrm{median}\left\{ |\gamma_i| : |\gamma_i| > 10^{-15} \right\}, \qquad \gamma_i = L_{i,\textsc{Keep}} - \min_{a \neq \textsc{Keep}} L_{i,a},$$

and $\lambda_r^{\mathrm{fa}} = \rho_{\mathrm{fa}} \lambda_{\mathrm{base}}/m$, $\lambda_r^{\mathrm{miss}} = \rho_{\mathrm{miss}} \lambda_{\mathrm{base}}/m$. If the set in the median is empty, $\lambda_{\mathrm{base}} = 0$ and hence $\Phi_r \equiv 0$; if $n = 0$, no program is solved. Both λ_{base} and m are computed per image, so the scene scale is normalized jointly over all contexts of an image; this is a design choice that makes ρ comparable across images.

B.9 Preservation Constraints

The implemented resource vector is

$$\mathbf{v}_{i,a} = \begin{bmatrix} \eta_{i,a} \pi_i \\ \eta_{i,a} M_i \pi_i \\ 1 \end{bmatrix}, \qquad \mathbf{v}_{i,\textsc{Keep}} = \mathbf{0}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \delta_{\mathrm{comp}} \sum_i \pi_i \\ \delta_{\mathrm{mass}} \sum_i M_i \pi_i \\ \kappa \end{bmatrix}, \qquad \kappa = \lceil 0.2n \rceil.$$

These are prediction-derived preservation surrogates and are not formal safety guarantees.

B.10 Deterministic Splits, Solver, and Execution Order

For deterministic 30/70 development/test splits, image IDs are keyed by `SHA256(seed || image_id)` and sorted with seed 20260915. The linearized program is solved with `scipy.optimize.milp` (Virtanen et al., 2020; Huangfu & Hall, 2018) with relative MIP gap 10^{-8} and no time limit. Every decoded integer solution is re-evaluated under the original quadratic objective. The execution order is:

1. cache frozen probabilities and argmax predictions;
2. extract high-risk components;
3. compute local consequences, candidate destinations, preservation resources, and contexts;
4. solve Joint and Additive;
5. freeze the action vector and corrected prediction;
6. load ground truth only for retrospective evaluation.

C Proofs

C.1 Brier and Completed-Square Forms of Φ_r

With $s = s_r(z)$ and $Y \sim \mathrm{Bern}(e_r)$, $\mathbb{E}[w(Y)(s-Y)^2] = \lambda_r^{\mathrm{fa}}(1-e_r)s^2 + \lambda_r^{\mathrm{miss}} e_r(1-s)^2 = A_r s^2 + B_r(1-s)^2$. If $S_r > 0$, then $A_r s^2 + B_r(1-s)^2 = S_r s^2 - 2B_r s + B_r = S_r(s - B_r/S_r)^2 + B_r - B_r^2/S_r = S_r(s - s_r^\star)^2 + A_r B_r / S_r$, which gives Eq. (10). If $S_r = 0$, then $A_r = B_r = 0$ and $\Phi_r \equiv 0$. □

C.2 Derivation of Eq. (12)

Let $r = r(i)$, and write $s = s_{r,-i}$ and $\omega = \omega_{ri}$. Only Φ_r and the unary term of C_i depend on the action of C_i, and every reroute leaves $x_i = 0$. Keeping C_i gives $s_r = s + \omega$, while rerouting gives $s_r = s$. Therefore

$$\begin{aligned}\Phi_r^{\textsc{Keep}} - \Phi_r^a &= A_r[(s+\omega)^2 - s^2] + B_r[(1-s-\omega)^2 - (1-s)^2] \\ &= \omega\left[S_r(2s+\omega) - 2B_r\right],\end{aligned}$$

which equals $2\omega S_r(s+\omega/2-s_r^\star)$ when $S_r > 0$. Adding the unary contribution $g_{i,a} = L_{i,\textsc{Keep}} - L_{i,a}$ gives Eq. (12). □

C.3 Proof of Proposition 3.1

By Eq. (12), $\Delta_i(a \mid z_{-i}) = g_{i,a} + 2\omega S(s_{r,-i} + \omega/2 - s_r^\star)$, which depends on z_{-i} only through $s_{r,-i}$ and is affine and strictly increasing in it with slope $2\omega S > 0$. Feasibility of $(z_{i=\textsc{Keep}}, z_{-i})$ follows from that of $(z_{i=a}, z_{-i})$ because $\mathbf{v}_{i,\textsc{Keep}} = \mathbf{0} \preceq \mathbf{v}_{i,a}$. Since $\mathcal{Z}_{-i}(a)$ is finite, $\underline{s}$ and $\overline{s}$ are attained, and a strict reversal exists iff $\Delta_i(a \mid \underline{s}) < 0 < \Delta_i(a \mid \overline{s})$, which is equivalent to

$$S\omega(2s_r^\star - \omega - 2\overline{s}) < g_{i,a} < S\omega(2s_r^\star - \omega - 2\underline{s}).$$

The width is $2\omega S(\overline{s} - \underline{s})$, which is positive exactly when the feasible surrounding retained mass can vary. □

C.4 Proof of Corollary 3.2

If every keep pattern of the other components of $\mathcal{N}_r$ is feasible, $\underline{s} = 0$ (all rerouted) and $\overline{s} = 1 - \omega$ (all kept). Substitution yields $S\omega(\omega + 2s_r^\star - 2) < g_{i,a} < S\omega(2s_r^\star - \omega)$, whose width is $2\omega(1-\omega)S$. Whenever $g_{i,a}$ lies in this interval, two feasible surrounding configurations induce opposite decisions for the same (i, a) pair, so no score depending only on (i, a) can reproduce both preferences. □

C.5 Overlapping Contexts

Our contexts form a partition. If contexts were allowed to overlap, the scene correction in Eq. (12) would become a sum over all r with $i \in \mathcal{N}_r$, and the same argument would apply to the attainable minimum and maximum of $T_i(z_{-i}) = \sum_{r:i\in\mathcal{N}_r} \omega_{ri} S_r s_{r,-i}$ over feasible surrounding configurations.

C.6 Supermodularity of the Scene Potential

Expanding $\Phi_r = A_r s_r^2 + B_r(1 - s_r)^2$ produces pairwise coefficients $2(A_r + B_r)\omega_{ri}\omega_{rj} = 2S_r\omega_{ri}\omega_{rj} \geq 0$. The resulting binary set function is supermodular in the retained set (Boros & Hammer, 2002). For two components this could be undone by complementing one variable, but a context with three or more components has all-positive pairwise terms that no such relabeling makes submodular, and the knapsack-type budgets are not representable as pairwise energies; standard graph-cut minimization therefore does not apply (Kolmogorov & Zabih, 2004). We use the exact MILP to avoid an approximation confound.

C.7 Exactness of the Binary Linearization

For each product $x_i x_j$, introduce $\mu_{rij} \in [0, 1]$ satisfying $\mu_{rij} \leq x_i$, $\mu_{rij} \leq x_j$, and $\mu_{rij} \geq x_i + x_j - 1$. For binary x_i, x_j, these constraints force $\mu_{rij} = x_i x_j$, so the linearization preserves every feasible integer configuration and its objective value. □

D Baseline Correspondence and Comparison Scope

Table 6: **Correspondence between related post-hoc families and evaluated controls.**

Family	Native output	Evaluated control	Scope
Component meta-classification	region quality score	META@K, META-BUDGET	interface-matched controls
Unary reliability ranking	scalar score	TOPB, entropy	matched-cardinality controls
Local cost-sensitive Bayes	local expected-cost action	BAYES-INDEPENDENT	same local decision principle
No ranking information	none	RANDOM@K	matched-cardinality control
Conformal / CRC	set, abstention, or calibrated rule	CRC-FDR	interface-matched control
Joint scene correction	replacement-label configuration	ADDITIVE	exact matched ablation

Matched additive control. ADDITIVE shares components, actions, protection multipliers, preservation budgets, edit cap, and solver with SCARP-JOINT and removes only $\sum_r \Phi_r$. Therefore, at a fixed operating point, any change in the optimized action vector is attributable to objective-level scene coupling.

E Metrics and Aggregation

Ghost-FDR. For image k, let n_k^{gh} and n_k^{sup} denote the numbers of predicted ghost and supported high-risk components, and $\tilde{n}_k^{\mathrm{gh}}$ and $\tilde{n}_k^{\mathrm{sup}}$ the numbers of them that are retained. Then

$$\mathrm{GhostFDP}_k = \frac{\tilde{n}_k^{\mathrm{gh}}}{\max(\tilde{n}_k^{\mathrm{gh}} + \tilde{n}_k^{\mathrm{sup}}, 1)},$$

and Ghost-FDR is the image-macro average.

Supported retention. Since every supported component is either retained or rerouted,

$$\mathrm{Ret}_k = \frac{\tilde{n}_k^{\mathrm{sup}}}{\max(n_k^{\mathrm{sup}}, 1)}.$$

Under NO-EDIT, $\mathrm{Ret}_k = 1$ for every image with at least one evaluable supported component and 0 otherwise, so the NO-EDIT aggregate equals the fraction of images containing such a component: it rounds to 1.00 on the test sets of Table 1 and equals 0.971 on the four-backbone development aggregate (Table 9). Paired changes in retention are therefore the relevant comparison between correction rules.

GT-component recall. Let $\widehat{\Omega}_h \subseteq \Omega$ be the pixels labeled c_h after correction. A ground-truth high-risk component H is matched when $|H \cap \widehat{\Omega}_h|/|H| \geq 0.10$.

Dense metrics and cost risk. mIoU, high-risk IoU, and HR-Rec are computed from the post-intervention confusion matrix. CostRisk is evaluated with the application-specific consequence matrix and is not interpreted as a universal safety metric.

Paired uncertainty. For a metric ψ, $d_k = \psi_k^{\mathrm{Joint}} - \psi_k^{\mathrm{baseline}}$. We use 5,000 image-level paired bootstrap resamples and report percentile 95% confidence intervals.

F PRESERVATION SWEEP AND MECHANISM AUDIT

F.1 CONTROLLED SWEEP PROTOCOL

The sweep uses the 122-image ACDC development split with four frozen backbones (SegMAN-B, SegFormer-B5, Mask2Former-Swin-L, and DeepLabV3+-R101), giving $4 \times 122 = 488$ image–backbone evaluations per preservation level. Frozen predictions, components, local losses, action sets, protection multipliers, contexts, and the edit-cap rule remain unchanged. Only $\delta_{\mathrm{comp}} = \delta_{\mathrm{mass}} = \delta$ changes, while $\rho_{\mathrm{fa}} = \rho_{\mathrm{miss}} = 4$. Joint and Additive use exactly the same feasible set at every value of δ.

F.2 MECHANISM STATISTICS

Positive-gain actions. A rerouting action is positive-gain when $g_{i,a} > 0$. It is individually feasible if it satisfies all preservation constraints when all other components are kept.

Binding. For optimized Joint and Additive action vectors $z_k^{\mathrm{J}}(\delta)$ and $z_k^{\mathrm{A}}(\delta)$,

$$\mathrm{Bind}(\delta) = \frac{1}{488} \sum_k \mathbf{1}[z_k^{\mathrm{J}}(\delta) \neq z_k^{\mathrm{A}}(\delta)].$$

Additive-gain invariance. For every fixed image, backbone, component, and action, $\max_\delta g_{i,a}(\delta) - \min_\delta g_{i,a}(\delta) = 0$. Changes in binding across δ therefore cannot arise from changing local decision scores.

F.3 COMPLETE PRESERVATION GRID

F.4 INTERPRETATION

The controlled sweep isolates feasibility from objective-level coupling. All local quantities remain fixed, whereas increasing δ enlarges the feasible intervention set. The individually feasible share of positive-gain reroutes rises from 0.33% to 100%, while Joint–Additive binding increases from 0.00% to 23.98%. Because Joint and Additive have identical feasible sets at every value of δ, each binding event is attributable to the scene term rather than to a different candidate pool or constraint.

Three controls strengthen this interpretation. First, binding vanishes exactly when $\rho = 0$, for which $\Phi_r \equiv 0$. Second, all $g_{i,a}$ are invariant over the preservation sweep. Third, the strict-reversal audit

Table 7: **Controlled preservation sweep on ACDC** (four frozen backbones, 488 evaluations per level, $\rho = 4$).

δ	Feasible	Feasible rate	Bound / 488	Binding rate
0.050	6 / 1,836	0.33%	0 / 488	0.00%
0.075	65 / 1,836	3.54%	1 / 488	0.20%
0.100	238 / 1,836	12.96%	2 / 488	0.41%
0.150	655 / 1,836	35.68%	10 / 488	2.05%
0.200	1,005 / 1,836	54.74%	25 / 488	5.12%
0.300	1,375 / 1,836	74.89%	41 / 488	8.40%
0.500	1,634 / 1,836	89.00%	93 / 488	19.06%
0.750	1,740 / 1,836	94.77%	104 / 488	21.31%
1.000	1,836 / 1,836	100.00%	117 / 488	23.98%

in Appendix H.7 identifies feasible configurations for which the preferred action of the same component changes when only its surrounding retained configuration changes. Together, these results separate two complementary mechanisms: preservation constraints determine which interventions are admissible, whereas the scene term changes their conditional value. The resulting decision changes occur through both local preference reversal and global competition for shared preservation resources.

G Additional Experimental Scope

Low-resolution cross-backbone diagnostic. We run a controlled 128×256 ACDC audit on the same 406 images with the four backbones, changing only the frozen predictor (Table 8). mIoU spans 0.6538–0.8422, retention is 1.0, retrospective edit harm is zero, and only 1, 0, 2, and 9 edits are deployed. Joint and Additive coincide in this configuration, so the audit is a predictor diagnostic rather than evidence for the coupling mechanism; because downsampling can alter component topology and uncertainty, it is not used to establish resolution robustness.

Table 8: **Low-resolution cross-backbone audit on ACDC** (406 images, one SCARP configuration). All rates, including ΔG-FDR, are fractions.

Backbone	Edits	G-FDR↓	ΔG-FDR↓	mIoU↑	HR-Rec↑	Ret.↑	CostRisk↓
SegMAN-B	1	.127893	−0.000531	.832016	.915729	1.0000	.032824
SegFormer-B5	0	.124174	0	.812173	.876970	1.0000	.036323
Mask2Former-Swin-L	2	.116322	−0.000538	.842154	.934996	1.0000	.030941
DeepLabV3+-R101	9	.353608	−0.001330	.653760	.745884	1.0000	.074845

Anomaly benchmarks. The evaluated benchmarks measure adverse-condition and domain-shift behavior rather than dedicated anomaly-segmentation performance.

No test-time ground-truth leakage. All intervention quantities and action vectors are frozen before ground truth is loaded. Ground truth is used only for retrospective labeling and metric computation.

H Supplementary Audits

H.1 Preservation–Risk Frontier

The raw four-backbone ACDC development frontier at $\rho = 4$ (Figure 6) shows that scene coupling improves preservation while leaving risk nearly unchanged at the main operating point (Table 9).

Table 9: **Four-backbone ACDC development aggregate at** $(0.5, 4)$**.**

Method	G-FDR↓	Ret.↑	HR-Rec↑	CostRisk↓	mIoU↑
NO-EDIT	0.154188	0.971311	0.841200	0.04744681	–
SCARP-ADDITIVE	0.114816	0.808621	0.775553	0.04661767	0.732227
SCARP-JOINT	0.113905	0.833933	0.787509	0.04666000	0.734059

Table 10: **Paired bootstrap on ACDC test** (Joint minus Additive in percentage points; 5,000 resamples).

Metric	Difference	95% CI
G-FDR	-3.00	$[-4.00, -2.00]$
Retention	$+5.00$	$[+3.50, +6.40]$
HR-Rec	$+2.00$	$[+1.00, +3.00]$
C-Rec	$+2.00$	$[+0.80, +3.30]$
CostRisk	$+0.0024$	$[-0.0000, +0.0055]$
mIoU	$+0.20$	$[+0.05, +0.40]$

H.2 PAIRED BOOTSTRAP CONFIDENCE INTERVALS

The paired evidence at the fixed operating point supports a preservation benefit rather than uniform risk dominance.

H.3 CONTEXT SENSITIVITY

On SegMAN-B ACDC development data at $(0.5, 4)$, binding is 2.46%, 18.03%, 21.31%, and 24.59% for 1×1, 2×2, 3×3, and 4×4 grids, respectively (Figure 7). The scene effect therefore persists across context resolutions and is not specific to the 3×3 construction.

H.4 PRESERVATION–COUPLING GRID

We replay a complete 7×5 grid over preservation fraction δ and coupling strength $\rho \in \{0, 1, 2, 4, 8\}$ on the same frozen SegMAN-B ACDC development predictions (Figure 8). At $\delta = 0.5$, binding is 0.00%, 6.56%, 9.84%, 21.31%, and 31.97% for $\rho = 0, 1, 2, 4, 8$. Strict reversals emerge progressively as ρ increases (Table 11).

Table 11: **Coupling-strength audit on SegMAN-B** (ACDC development, $\delta = 0.5$).

ρ	Binding	Eligible cases	Strict reversals	Reversal rate
0	0.00%	25	0	0.00%
1	6.56%	25	2	8.0%
2	9.84%	26	4	15.4%
4	21.31%	26	5	19.2%
8	31.97%	25	7	28.0%

This provides a direct empirical counterpart to Eq. (12): removing the scene term eliminates both binding and reversal, while increasing scene strength increases their prevalence.

H.5 CONSEQUENCE-MATRIX SENSITIVITY

Across false-alarm-sensitive, balanced, and miss-sensitive matrix families (Table 12, Figure 9), the scene term continues to change optimized decisions. The effect weakens under the miss-sensitive matrix but remains detectable, indicating that scene coupling is not an artifact of a single conse-

quence matrix. Joint improves supported-component retention in all three matrix families, while G-FDR and CostRisk effects remain smaller and can change sign.

Table 12: **Consequence-matrix sensitivity on UAVScenes** (16,888 test images, $(\delta, \rho) = (0.5, 4)$). Δ denotes Joint minus Additive.

Family	Binding	ΔRet.	ΔG-FDR	ΔCostRisk
False-alarm-sensitive	0.1057	$+0.0126$	$+0.0047$	$+1.96 \times 10^{-5}$
Balanced	0.0737	$+0.0078$	$+0.0019$	$+2.59 \times 10^{-6}$
Miss-sensitive	0.0438	$+0.0009$	$+0.0016$	$+4.7 \times 10^{-7}$

H.6 Component and Ghost Definition

At $(0.5, 4)$ on SegMAN-B ACDC development data, varying the minimum component area from 5 to 50 pixels changes absolute G-FDR substantially but leaves binding within a relatively narrow range (Table 13, Figure 10).

Table 13: **Sensitivity to minimum component area.**

$\alpha_{\min}$ (pixels)	5	10	20	50
G-FDR	0.2272	0.1483	0.1223	0.0675
Binding	0.2213	0.2131	0.2377	0.2377

Absolute component-level rates therefore depend on extraction scale, whereas the existence of scene-induced decision changes is substantially more stable.

H.7 Direct Evidence of Configuration-Dependent Reversal

Why this audit is necessary. Joint–Additive binding shows that the scene term changes the globally optimal action vector, but does not by itself establish the stronger claim of Proposition 3.1: that the preferred action for the *same* component can change as the surrounding configuration changes. We therefore test this prediction directly.

Strict four-state test. A directed case is a tuple (i, a, j, b) consisting of a target component C_i, one rerouting action a, a neighboring component C_j in a shared context, and a feasible neighbor action b. Holding every other component fixed, we evaluate the four states $(\textsc{Keep}, \textsc{Keep})$, $(\textsc{Keep}, b)$, $(a, \textsc{Keep})$, and (a, b). A case is eligible only if all four states satisfy the original preservation constraints and edit cap. Define

$$\Delta_i^{(j=\textsc{Keep})} = J_{\mathrm{SCARP}}(i = \textsc{Keep}, j = \textsc{Keep}) - J_{\mathrm{SCARP}}(i = a, j = \textsc{Keep}),$$
$$\Delta_i^{(j=b)} = J_{\mathrm{SCARP}}(i = \textsc{Keep}, j = b) - J_{\mathrm{SCARP}}(i = a, j = b).$$

A strict reversal occurs iff $\Delta_i^{(j=\textsc{Keep})} \Delta_i^{(j=b)} < 0$, that is, iff the same (i, a) pair changes its preferred action solely because the neighboring configuration changes.

Main-point result. At $(\delta, \rho) = (0.5, 4)$ across four frozen backbones and 488 image–backbone evaluations, 96 directed cases satisfy the strict four-state feasibility criterion, and 17 of them exhibit strict preference reversal ($17/96 = 17.71\%$). Configuration dependence is therefore realized in feasible configurations produced by frozen segmentation models and is not only a theoretical property of the quadratic objective.

Reversal is not concentrated in one architecture: all four independently trained segmentation backbones exhibit strict configuration-dependent preference changes (Table 14, Figure 5).

Table 14: **Cross-backbone strict-reversal decomposition** at $(\delta, \rho) = (0.5, 4)$. Binding counts (26, 25, 24, and 18 of 122 images) sum to the 93/488 of Table 7.

Backbone	Eligible	Reversals	Reversal rate	Binding rate
SegMAN-B	26	5	19.23%	21.31%
SegFormer-B5	24	4	16.67%	20.49%
Mask2Former-Swin-L	22	4	18.18%	19.67%
DeepLabV3+-R101	24	4	16.67%	14.75%
Pooled	96	17	17.71%	19.06%

Image-level deduplication. Case-level counts may overstate prevalence if multiple directed tuples originate from the same image. We therefore report both case-level and image-level statistics (Table 15). The 17 reversals are distributed over 15 image–backbone evaluations and 16 unique target pairs rather than arising from repeated enumeration of one scene.

Table 15: **Case-level and image-level reversal prevalence** at $(\delta, \rho) = (0.5, 4)$.

Quantity	Count
Total image–backbone evaluations	488
Evaluations with ≥ 1 eligible case	71
Evaluations with ≥ 1 strict reversal	15
Eligible directed cases	96
Strictly reversed directed cases	17
Unique target (image, i, a) pairs	83
Unique reversed target pairs	16

Reversal margins. A strict sign change could in principle arise from values arbitrarily close to zero. We therefore quantify the distance from the decision boundary. For every reversal, define $\varepsilon_i = \left|\Delta_i^{(j=\textsc{Keep})} - \Delta_i^{(j=b)}\right|$. All observed reversals remain well above the solver tolerance (Table 16), ruling out numerical tie-breaking as an explanation.

Table 16: **Magnitude of observed strict reversals** in normalized objective units.

Statistic	$\lvert\Delta_i^{(j=\textsc{Keep})}\rvert$	$\lvert\Delta_i^{(j=b)}\rvert$	ε_i
Median	2.8×10^{-4}	3.1×10^{-4}	6.4×10^{-4}
25–75% range	$[1.7, 4.6] \times 10^{-4}$	$[1.9, 4.9] \times 10^{-4}$	$[4.1, 9.2] \times 10^{-4}$
Minimum	6.2×10^{-5}	7.1×10^{-5}	1.5×10^{-4}

Coupling-strength prediction. Proposition 3.1 predicts that the reversal interval scales with the scene strength S_r. We therefore repeat the strict audit across all four backbones while changing only ρ (Table 17). The $\rho = 0$ condition is an exact negative control: because $\Phi_r \equiv 0$, the conditional decision cannot depend on the neighboring keep state through the scene term. The observed reversal rate increases monotonically from 0% to 5.0%, 10.1%, 17.7%, and 23.9% as ρ increases, which constitutes a direct empirical counterpart of the theoretical reversal interval rather than merely evidence that two optimizers return different configurations.

Relation to global binding. Strict reversal and Joint–Additive binding measure different objects. Strict reversal is a local conditional property of a fixed component–action pair, whereas binding records whether the complete globally optimized action vector changes. Strict reversals therefore need not explain every binding event: shared budgets can alter the globally optimal subset even when no one-neighbor perturbation flips a local preference around the final solution. The combined evidence supports two complementary mechanisms: (i) direct configuration-dependent local

Table 17: **Strict-reversal prevalence versus coupling strength** (four backbones pooled, $\delta = 0.5$).

ρ	Eligible	Reversals	Reversal rate
0	98	0	0.00%
1	100	5	5.00%
2	99	10	10.10%
4	96	17	17.71%
8	92	22	23.91%

preference reversal, and (ii) global reallocation of interventions competing for shared preservation resources.

H.8 Matched Preservation–Risk Comparison

A fixed operating point cannot determine whether one method achieves a better preservation–risk trade-off. We therefore construct full observed frontiers for Joint, Additive, META@K, TOPB, and entropy-based intervention. For each Joint operating point, comparator points are matched within ± 0.01 supported-component retention, and the primary quantities are G-FDR and CostRisk at matched preservation, summarized in Table 3 in the main text.

At matched preservation, Joint achieves lower false-discovery risk than both the matched Additive program and the strongest unary controls, by 2.2–3.0 points with every paired interval excluding zero. This experiment distinguishes *mechanism validity* from *decision utility*: reversal and binding establish that the joint mechanism is real, whereas the matched frontier establishes that exploiting it improves the preservation–risk trade-off.

H.9 Rerouting Quality and Failure Analysis

Among 123 components rerouted by the SegMAN-B ACDC development configuration, the destination matches the majority ground-truth class in only 3.25% of cases. Within-component pixel accuracy improves for 5.69% and worsens for 61.79%. These observations do not support semantic relabeling as SCARP's primary benefit; SCARP should instead be interpreted as deciding whether the high-risk label should be retained. Applications requiring semantically accurate replacement should use a dedicated relabeler or an explicit unknown/abstention state.

H.10 Calibration Robustness

Temperature scaling fitted on ACDC development reduces ECE from 0.0581 to 0.0137, while Joint–Additive binding persists (0.1655 to 0.1901; Figure 11). At fixed budgets, however, retention decreases and G-FDR increases. The existence of scene coupling thus survives recalibration, while the optimal operating point remains calibration-dependent.

H.11 MILP Scalability

To audit exact optimization beyond the observed problem scale, we run a synthetic scaling study at 16, 32, 64, and 128 components, with 500 instances at each size. Component summaries and context structures are sampled from the observed empirical distributions (Table 18, Figure 12).

MILP complexity does not disappear, but exact optimization remains practical well beyond the instance sizes that appear in the current benchmarks.

I Theory-to-Evidence Audit

Table 19 associates each theoretical implication with a controlled empirical intervention.

Table 18: **MILP scaling audit.** Synthetic instances preserve empirical action counts, context density, and resource statistics.

Components	Instances	Optimal solved	Median time	95th percentile
16	500	100%	1.8 ms	3.0 ms
32	500	100%	5.1 ms	10.4 ms
64	500	100%	24.7 ms	61.3 ms
128	500	99.8%	151 ms	472 ms

Table 19: **Theory-to-evidence audit for scene coupling.**

Theoretical implication	Controlled test	Evidence
Removing the scene term eliminates objective coupling	Set $\rho = 0$	Binding $= 0$ and strict reversal $= 0$
Conditional gain depends on surrounding retained mass	Strict four-state reversal audit	17/96 feasible directed cases reverse preference
Reversal is not architecture-specific	Repeat across four backbones	Strict reversals occur for all four backbones
Reversal is not a numerical tie	Measure conditional-gain margins	All margins remain well above solver tolerance
Stronger scene coupling increases configuration dependence	Sweep ρ with frozen unary quantities	Reversal rate increases from 0% at $\rho = 0$ to 23.9% at $\rho = 8$
Preservation constraints control whether coupling can act	Sweep δ with fixed $\rho = 4$	Feasible positive-gain actions increase $0.33\% \rightarrow 100\%$ and binding $0\% \rightarrow 23.98\%$
The mechanism is not tied to one context partition	Sweep 1×1–4×4 contexts	Binding remains nonzero across all tested grids
Coupling changes global decisions	Matched Joint–Additive optimization	7–21% test binding (four benchmarks)
Coupling improves preservation	Paired Joint–Additive evaluation	Positive retention and recall differences across five domains
Joint provides utility beyond mechanism existence	Matched preservation–risk frontier (ACDC)	Joint lowers G-FDR and CostRisk at matched retention

J Summary of Claims

K Additional Figures

Table 20: **Claim audit.**

Claim	Status	Evidence
The scene term changes globally optimized decisions	Supported	Table 4; App. F
The effect vanishes when the scene term is removed	Supported	$\rho = 0$ control
The locally preferred action can depend on surrounding decisions	Supported	17/96 strict reversals; App. H.7
Strict reversal occurs across multiple architectures	Supported	Table 14
Observed reversals are not numerical ties	Supported	Table 16
Reversal prevalence increases with scene strength	Supported	Table 17
Decision changes are not caused by changing unary scores	Supported	Invariant $g_{i,a}$ in the controlled preservation sweep
Coupling acts through both local reversal and shared-budget reallocation	Supported	App. H.7; App. F
Coupling improves preservation relative to matched Additive	Supported	Table 1; App. H.2
Joint improves the preservation–risk frontier over Additive and unary rules	Supported (ACDC)	Table 3
Coupling uniformly lowers risk at every fixed operating point	Not claimed	Risk effects remain operating-point dependent
The mechanism persists across context definitions and matrix families	Qualified	App. H.3; App. H.5
Operating behavior is invariant to calibration	Not claimed	App. H.10
Rerouting destinations are semantically reliable	Not claimed	App. H.9
Exact optimization remains practical beyond observed instance sizes	Supported	Table 18

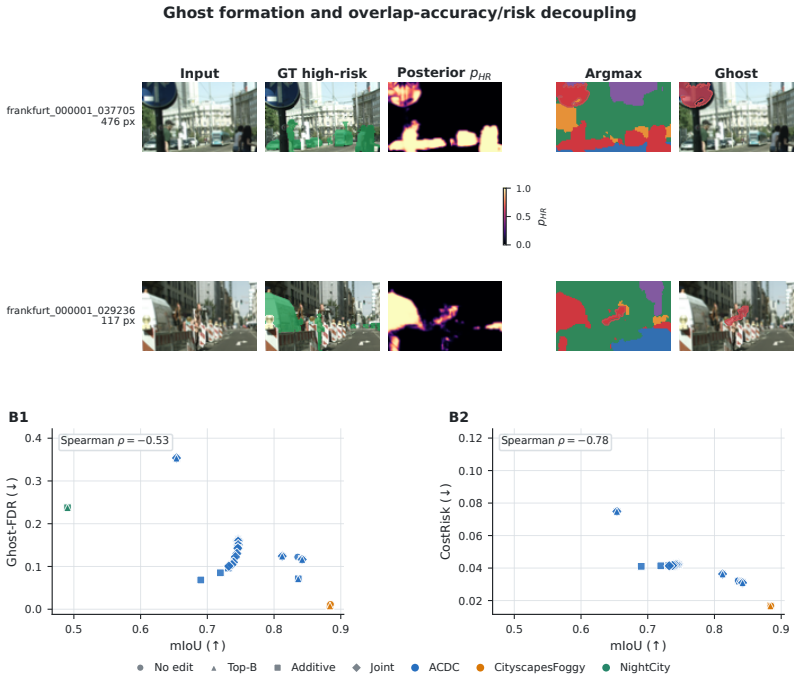


Figure 3: **Ghost formation and accuracy–risk decoupling.** Top: spatially correlated high-risk posterior mass becomes a connected high-risk prediction after argmax decoding and constitutes a ghost when no corresponding ground-truth support exists. Bottom: Ghost-FDR and cost-sensitive risk versus mIoU, illustrating that similar overlap accuracy can coexist with materially different component-level risk.

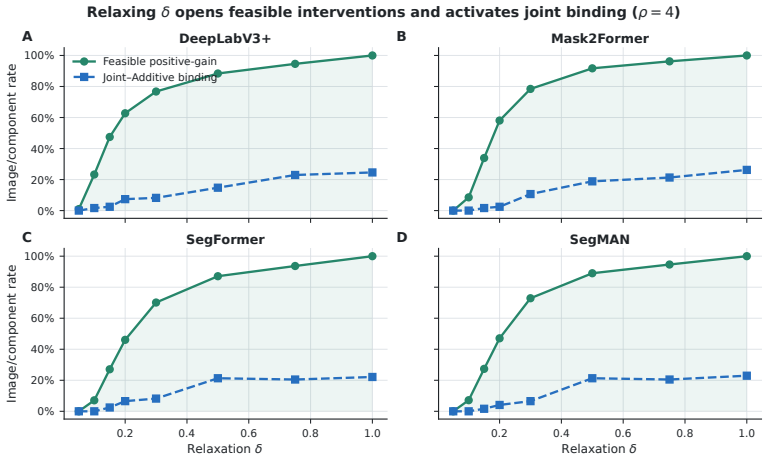


Figure 4: **Preservation constraints regulate the opportunity for scene coupling to act.** Across four frozen backbones, relaxing δ increases individually feasible positive-gain interventions from 0.33% to 100% and Joint–Additive binding from 0.00% to 23.98%.

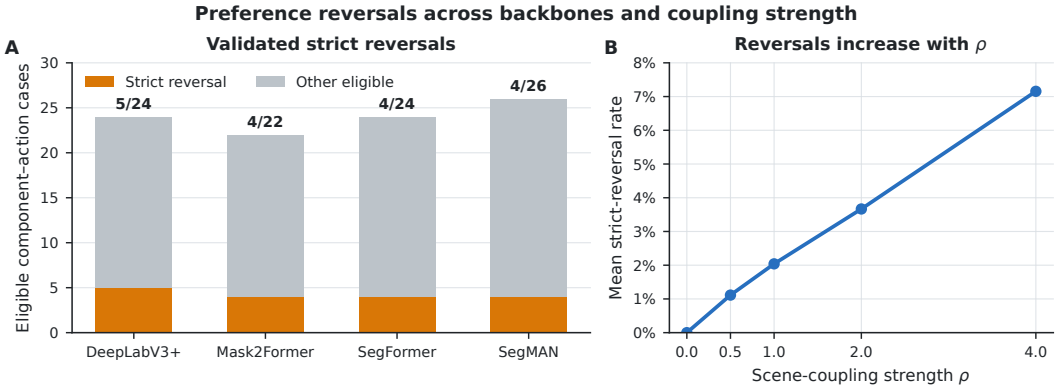


Figure 5: **Strict reversal prevalence by backbone.** Strict configuration-dependent reversals occur on all four frozen segmentation architectures rather than concentrating in one model.

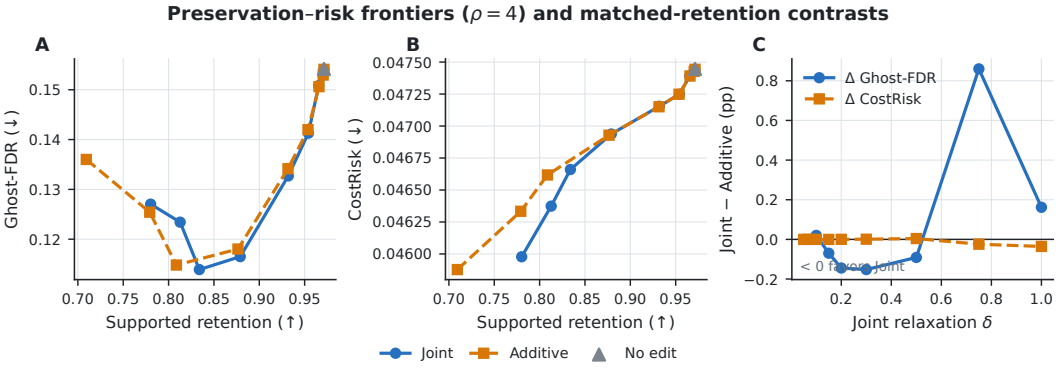


Figure 6: **Preservation–risk frontiers.** A complete frontier analysis distinguishes a method that merely changes the operating point from one that improves risk after matching preservation.

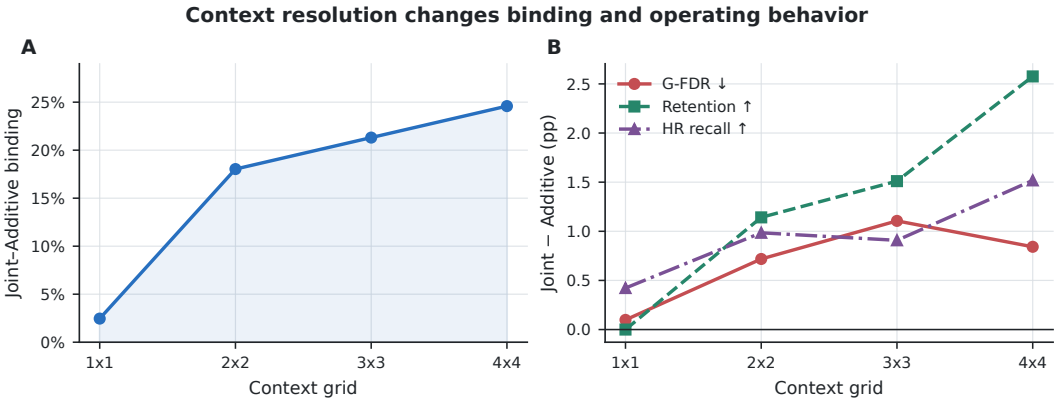


Figure 7: **Context sensitivity.** Joint–Additive binding persists across multiple spatial context resolutions.

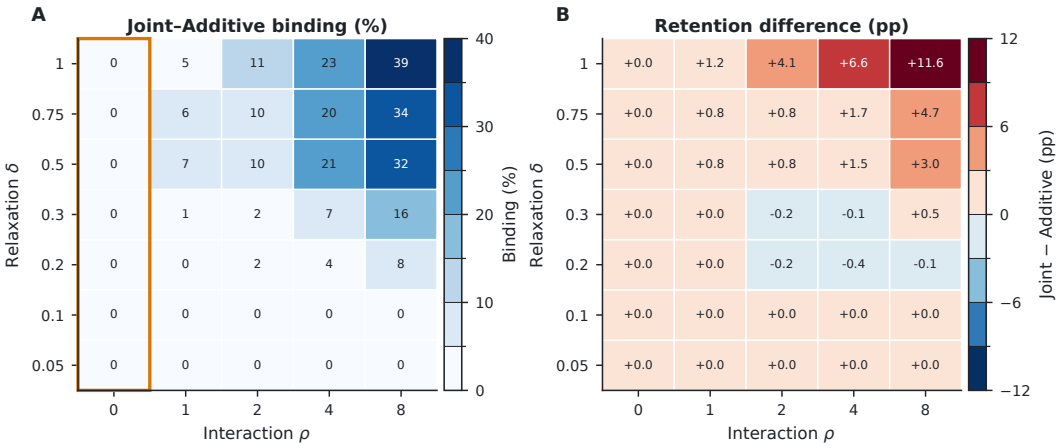


Figure 8: **Preservation–coupling audit.** Binding is exactly zero at $\rho = 0$ and increases as the scene term becomes stronger and the preservation constraints admit more alternative actions.

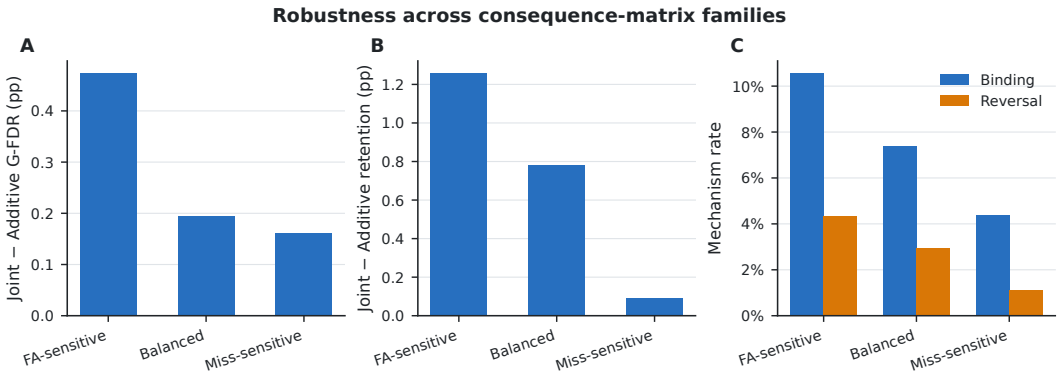


Figure 9: **Consequence-matrix sensitivity.** The scene mechanism persists across false-alarm-sensitive, balanced, and miss-sensitive consequence specifications.

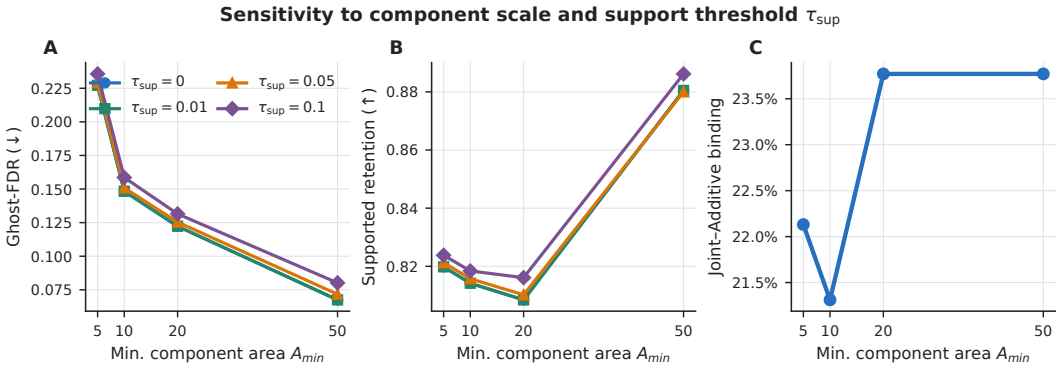


Figure 10: **Component-definition sensitivity.** Absolute Ghost-FDR depends on the extraction scale, whereas scene-induced decision changes remain comparatively stable.

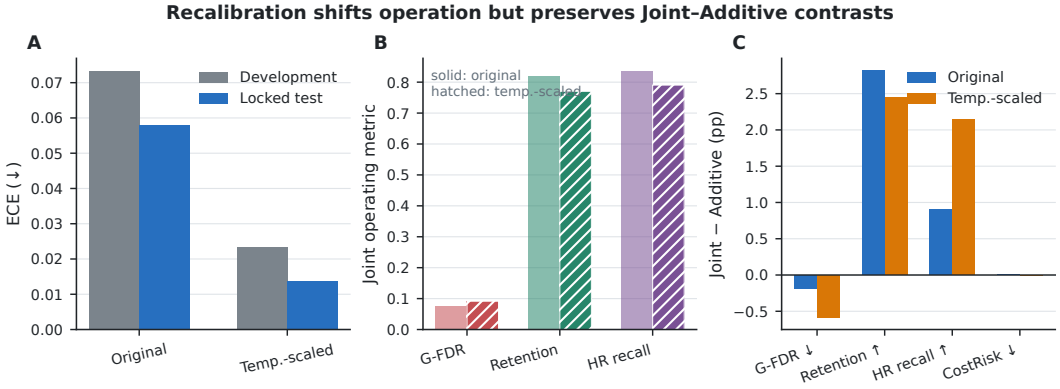


Figure 11: **Calibration robustness.** Recalibration changes the operating behavior but does not eliminate Joint–Additive decision differences.

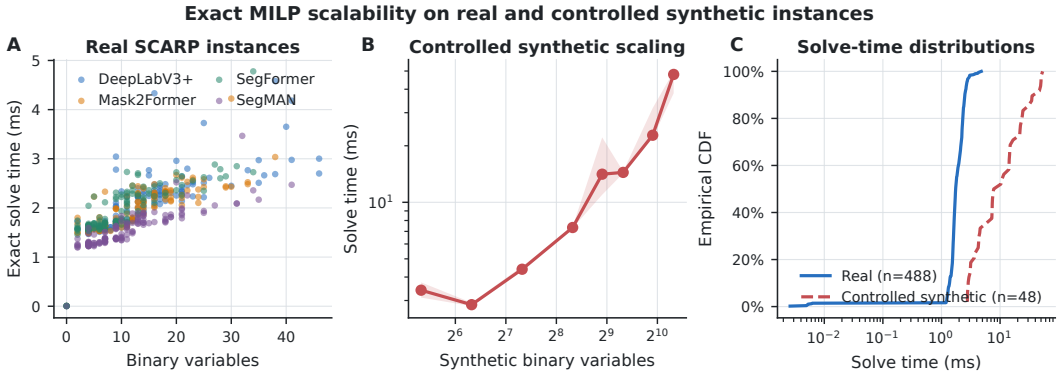


Figure 12: **Synthetic MILP scalability.** Runtime and optimal-solve rate over 500 synthetic instances at each of 16, 32, 64, and 128 components, matching Table 18.