TRAINING-FREE CROSS-DOMAIN FEW-SHOT SEGMENTATION VIA PRESERVE-RELATIVE COUNTERFACTUAL REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-domain few-shot segmentation aims to segment unseen classes in new domains using only a few annotated examples. In training-free settings, frozen semantic representations guide initial predictions, while promptable segmentation models generate structurally coherent candidate masks. However, under domain shift, structurally coherent masks can still misidentify target regions, undermining the semantic reliability of refinement. When the initial prediction also guides candidate selection, refinement can preserve or reinforce its semantic errors. We therefore propose a training-free framework that uses an audited semantic reference to improve the semantic reliability of mask refinement under domain shift. To this end, Cycle-Consistent Semantic Cover (CSC) uses semantic matching to select reliable prompts covering complementary target regions. Building on these prompts, Preserve-Relative Counterfactual Mask Selection (PRCMS) compares mask responses across prompt subsets and combines this evidence with support semantics to audit the reference and select refinements. Across eight datasets, our framework achieves mean mIoU of 69.33% in one-shot and 75.20% in five-shot segmentation, exceeding the evaluated state-of-the-art baselines. Ablation studies further show that complementary prompt coverage and reference auditing both improve cross-domain segmentation. Code will be released upon publication.

1 INTRODUCTION

Cross-domain few-shot segmentation (CD-FSS) aims to segment unseen target classes from only a few annotated examples under domain shift (Shaban et al., 2017; Nguyen & Todorovic, 2019; Lei et al., 2022). Recent training-free approaches use frozen foundation models to transfer target information from annotated support examples to query images without task-specific parameter updates (Liu et al., 2024; Zhang et al., 2024a; Zhu & Zhang, 2025; Sun et al., 2026). Although semantic matching helps locate target regions, the resulting masks can remain incomplete or spatially fragmented. This gap between locating a target and delineating its full extent motivates the use of promptable segmentation models to generate more coherent mask candidates. Yet structural coherence alone does not establish whether a candidate proposes a semantically valid change.

Determining which changes should be accepted under domain shift is therefore the central challenge. This challenge becomes harder when candidate changes are evaluated against the initial semantic prediction, which may itself be imperfect. For example, an incomplete reference may disagree with a candidate that correctly recovers missing target regions, yet agree with one that preserves its errors. Moreover, a single candidate can recover missing foreground while removing regions that were already correct. Its revision value therefore depends on the current prediction and the changes it introduces, rather than structural quality alone. We refer to this mismatch between candidate quality and revision value as *state transition ambiguity*. This motivates a reference-first strategy that audits what should be preserved before deciding which candidate changes to accept.

To implement this strategy, we propose a training-free framework that constructs an audited semantic reference and evaluates candidate changes relative to it. The framework integrates two complementary modules for prompt construction and mask revision. First, Cycle-Consistent Semantic Cover (CSC) selects informative prompts by balancing correspondence reliability with semantic coverage. Reliable

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.

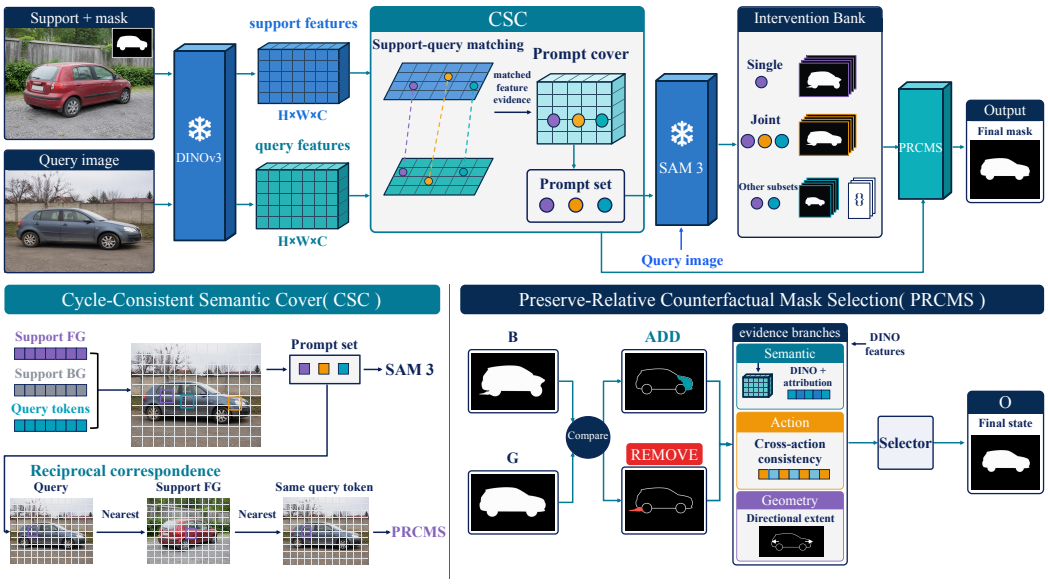

Figure 1: **Method overview with illustrative mask states.**

matches keep the prompts grounded in the support semantics, while complementary coverage discourages multiple prompts from concentrating on the same target region. Together, these criteria guide SAM 3 (Carion et al., 2025) to generate diverse mask candidates for subsequent refinement. Building on these prompts, Preserve-Relative Counterfactual Mask Selection (PRCMS) compares SAM 3 responses across prompt subsets to determine how each prompt affects the predicted regions. These contributions, together with support semantics, guide an audit of a structural proposal against the initial prediction, yielding a semantic preserve state. PRCMS then evaluates candidate foreground additions and removals relative to this audited state using direction-specific semantic and structural evidence. All component models remain frozen throughout inference.

We evaluate the proposed framework on eight datasets spanning natural images, underwater scenes, medical images, and aerial imagery. The method achieves 69.33% mean mIoU in the one-shot setting and 75.20% in the five-shot setting, improving matched RSRM (Sun et al., 2026) predictions by 4.47 and 3.18 percentage points, respectively. Controlled experiments further show that complementary prompt coverage and reference auditing both contribute substantially to the final performance.

Our main contributions are summarized as follows:

- We identify state transition ambiguity in training-free mask refinement: a structurally plausible candidate may not constitute a beneficial revision of an imperfect semantic prediction. This motivates a reference-first refinement principle that separates what should be preserved from what should be changed.
- We propose CSC and PRCMS, which respectively construct reliable and complementary semantic prompts and use prompt-induced responses with support semantics to audit the reference and evaluate subsequent revisions, without task-specific training.
- Extensive experiments across eight datasets demonstrate consistent cross-domain improvements, including mean gains of 4.47 and 3.18 points over matched RSRM in one-shot and five-shot settings, with ablations supporting the major design choices.

2 Related Work

2.1 Cross-Domain Few-Shot Segmentation

Few-shot segmentation transfers target information from annotated support examples to a query image (Shaban et al., 2017; Nguyen & Todorovic, 2019). Methods model this support–query relation

through prototypes, feature proxies, or dense correlations (Zhang et al., 2022; Fan et al., 2022; Peng et al., 2023). In-context systems extend this support–query interface through general segmentation models (Wang et al., 2023; Meng et al., 2024; Zhu et al., 2024). Under domain shift, CD-FSS methods use domain-agnostic transformation, test-time adaptation, or learned rectification (Lei et al., 2022; Herzog, 2024; Su et al., 2024). RSRM instead performs semantic matching with frozen representations (Sun et al., 2026). Its support-grounded prediction provides our initialization and the matched baseline for performance comparison.

2.2 Training-Free Segmentation with Foundation Models

Training-free methods use pretrained representations to guide segmentation from annotated supports. PerSAM personalizes SAM from one reference, while Matcher combines dense matching with SAM proposals (Zhang et al., 2024b; Liu et al., 2024). GF-SAM, MAUP, and PR-MaGIC study prompting and mask prediction (Zhang et al., 2024a; Zhu & Zhang, 2025; Lee et al., 2026), while INSID3 uses DINOv3 features for in-context segmentation (Cuttano et al., 2026). Retrieval and concept search provide other routes to prediction (Boudiaf et al., 2026; Chen et al., 2026).

Our focus is on evaluating mask revisions relative to a reliable reference. When candidate selection relies on an imperfect reference, agreement alone cannot establish whether a revision corrects or preserves its errors. We therefore audit the reference using complementary prompt responses and support semantics. Relative to this audited reference, we assess foreground additions and removals with direction-specific semantic and structural evidence.

3 Methodology

3.1 Overview

An episode contains an annotated support set $\mathcal{S} = \{(I_s^{(r)}, M_s^{(r)})\}_{r=1}^{K}$, with $K \in \{1, 5\}$, and a query image I_q. Frozen RSRM supplies the initial prediction B_0. As shown in Figure 1, CSC selects complementary point prompts whose SAM 3 responses form the evidence bank for PRCMS. PRCMS constructs a structural proposal G_0 from this bank and audits its disagreements with B_0 to obtain the *semantic preserve state* B. This audited state serves as the reference for evaluating subsequent candidate changes. All masks share a pixel grid Ω. Relative to this audited reference, a candidate G proposes foreground additions and removals defined by

$$E^+(G, B) = G \setminus B, \qquad E^-(G, B) = B \setminus G. \tag{1}$$

Because both sets are defined relative to B, changing the reference changes the revisions attributed to the same candidate. All component models remain frozen throughout inference. Algorithm 1 and Appendix A.1 give the inference flow and supporting definitions.

3.2 Cycle-Consistent Semantic Cover (CSC)

Reliable correspondence alone can concentrate prompts on the same object part. CSC therefore balances prompt reliability with coverage of semantically supported query regions. Foreground–background margins from DINOv3 (Siméoni et al., 2025) support features define semantic demand w_j on the query. Foreground matching and agreement across support–layer fields determine anchor reliability r_a. For normalized query features $\widehat{f}_j$ and token coordinates p_j, these quantities define the coverage kernel and objective

$$\kappa(a, j) = r_a \left[(1 - \alpha)[\cos(\widehat{f}_a, \widehat{f}_j)]_+ + \alpha e^{-\|p_a - p_j\|_2^2/(2\tau_p^2)} \right], \tag{2}$$

$$F(A) = \sum_j w_j \max_{a \in A} \kappa(a, j). \tag{3}$$

Here $[x]_+ = \max(x, 0)$, α balances feature and spatial similarity, and τ_p sets the spatial scale. Because each token contributes only through its strongest anchor, the objective discounts redundant coverage and is monotone submodular (Nemhauser et al., 1978). Starting from the strongest semantic match, CSC greedily adds anchors from a pool screened for agreement and foreground correspondence.

After selection, reciprocal query–support–query matching tests whether each prompt returns to its original query token through annotated support foreground. One-shot eligibility requires a reciprocal return at one or more layers. Five-shot eligibility requires returns through a majority of supports at the same layer. Eligible prompts contribute to attribution and coalition evidence in PRCMS, while response generation and total-effect estimation use the full selected set. The same DINOv3 features also provide support–query semantic fields $D^{\ell,r}$ for reference auditing in PRCMS.

3.3 Preserve-Relative Counterfactual Mask Selection (PRCMS)

Using the CSC prompts and semantic fields, PRCMS constructs an audited reference B and evaluates candidate changes relative to this fixed state. It evaluates every subset of the selected prompts $N = \{1, \dots, k\}$, $k \leq 5$, through SAM 3, keeping the support-conditioned query state and negative prompt fixed. The resulting bank contains at most 32 soft and binary responses (h_S, M_S), including the negative-only response. Exact pixelwise Shapley attribution aggregates each prompt's marginal effect across subsets (Shapley, 1953). Summing eligible contributions gives $\Phi_{\mathcal{E}}$, while the total effect $T = h_N - h_\varnothing$ includes every selected prompt. Support semantics complement these response-derived fields through $D^{\ell} = K^{-1} \sum_r D^{\ell,r}$.

Component replacements from singleton and joint responses in this bank form the structural proposal G_0. PRCMS audits G_0 against B_0 by comparing initial assignments X with proposed assignments Y within each connected disagreement group. For a signed evidence field F, the restoration utility is

$$u_F(X, Y) = \frac{\sum_{x \in X \setminus Y} \widehat{F}(x) - \sum_{x \in Y \setminus X} \widehat{F}(x)}{|X \triangle Y|}, \tag{4}$$

where $\widehat{F} = F / \max_x |F(x)|$. Zero fields remain zero, and $u_F = 0$ when $X = Y$. Positive utility favors restoring the initial assignments. The audit restores the initial assignments when supported jointly by eligible attribution and all semantic layers, or by the total-effect test for large disagreements. Otherwise, it retains the proposed assignments. These decisions construct B while preserving assignments shared by B_0 and G_0.

Once B is fixed, each proposal ξ specifies additions E^+_ξ, removals E^-_ξ, and their response provenance. Response consistency measures agreement across prompt subsets, while support semantics and edit geometry provide complementary evidence for evaluating each direction. These quantities jointly determine admissibility under the corresponding shot setting. Let $\mathcal{C}_K(B)$ be the proposal set and $\mathrm{Adm}_K(\xi \mid B)$ indicate whether a proposal meets the branch-specific requirements for its shot setting. The admissible set is

$$\mathcal{A}_K(B) = \{\xi \in \mathcal{C}_K(B) : \mathrm{Adm}_K(\xi \mid B) = 1\}. \tag{5}$$

For one-shot inference, the branches assess individual-response and consensus candidates. We select one admissible complete mask according to branch priority and within-branch rank, preserving its overall structure. Five-shot inference instead uses support-averaged semantics to guide regional acceptance through coalition or correspondence evidence. All accepted, disjoint changes are composed simultaneously relative to B. Both settings yield accepted additions $U^+ \subseteq \Omega \setminus B$ and removals $U^- \subseteq B$, giving the final mask

$$O = (B \setminus U^-) \cup U^+. \tag{6}$$

If no proposal is admissible, both edit sets are empty and the output remains B. Appendix A.1.4 specifies the branch conditions and selection order.

4 Experiments

4.1 Experimental Setup

Datasets and metric We evaluate segmentation across underwater (SUIM), medical (ISIC, Chest X-ray), aerial (iSAID-5^i), and natural-image segmentation (FSS-1000, PASCAL-5^i, COCO-20^i, VOC2012) (Islam et al., 2020; Li et al., 2020; Codella et al., 2019; Tschandl et al., 2018; Candemir et al., 2014; Jaeger et al., 2014; Shaban et al., 2017; Nguyen & Todorovic, 2019; Everingham et al., 2010; Zamir et al., 2019). Within each seed, matched methods use the same episodes, with each query excluded from its support set (Table 7 and Appendix A.2.1). We report class-macro foreground mIoU

Table 1: **One-shot comparison with existing methods on eight datasets.** Dataset columns report three-seed mean mIoU (%). Mean reports the seed-level macro mean and sample standard deviation. Bold marks column maxima.

Method	SUIM	FSS	ISIC	CXR	PASCAL	COCO	VOC	iSAID	Mean
PerSAM (ICLR'24) (Zhang et al., 2024b)	32.31	71.05	22.23	30.89	49.10	21.99	43.88	20.17	36.45±0.26
MAUP (MICCAI'25) (Zhu & Zhang, 2025)	24.36	54.23	36.98	53.90	36.67	22.16	33.74	25.88	35.99±0.29
INSID3-B (CVPR'26) (Cuttano et al., 2026)	52.87	83.12	39.17	31.17	67.69	52.20	64.09	36.25	53.32±0.25
INSID3-L (CVPR'26) (Cuttano et al., 2026)	55.37	83.57	53.68	77.15	67.52	57.00	64.64	**51.26**	63.77±0.31
GF-SAM (NeurIPS'24) (Zhang et al., 2024a)	60.33	87.90	50.01	50.14	71.91	**58.39**	69.82	48.20	62.09±0.49
Matcher (ICLR'24) (Liu et al., 2024)	46.23	86.97	36.13	69.63	68.23	52.55	63.57	35.71	57.38±0.89
RSRM (ECCV'26) (Sun et al., 2026)	57.58	82.48	54.87	85.45	72.70	52.24	67.90	45.71	64.87±0.92
Ours	**60.85**	**90.23**	59.05	**90.47**	77.37	56.06	73.03	47.59	**69.33±0.88**
GF-SAM DINOv3-B/SAM-B	57.48	86.58	48.71	42.46	72.94	51.60	67.44	42.65	58.73±0.46
Matcher+PR-MaGIC (Lee et al., 2026)	43.70	87.15	39.69	74.28	65.05	44.81	60.55	32.82	56.00±0.38
SAM 3 propagation	51.35	79.26	52.07	89.85	48.61	25.77	47.74	32.73	53.42±0.36
SAM 3 concatenation	41.94	88.10	**68.14**	90.09	**78.54**	53.78	**78.69**	28.71	66.00±0.10

Table 2: **Paired one-shot comparison with RSRM.** Three-seed mean mIoU (%) with standard deviations below in gray. Gains are in points and use unrounded matched results.

Method	SUIM	FSS	ISIC	CXR	PASCAL	COCO	VOC	iSAID	Mean
RSRM	57.58 ±3.98	82.48 ±0.09	54.87 ±2.72	85.45 ±0.20	72.70 ±0.28	52.24 ±0.71	67.90 ±1.07	45.71 ±0.09	64.87 ±0.92
Ours	**60.85** ±3.66	**90.23** ±0.12	**59.05** ±2.50	**90.47** ±0.04	**77.37** ±0.82	**56.06** ±1.20	**73.03** ±1.50	**47.59** ±0.34	**69.33** ±0.88
Paired gain	+3.27 ±1.64	+7.75 ±0.14	+4.18 ±0.93	+5.02 ±0.19	+4.68 ±0.55	+3.82 ±0.61	+5.13 ±0.52	+1.88 ±0.34	**+4.47** ±0.30

(%) and weight datasets equally when computing overall means. Unless stated otherwise, results are means and sample standard deviations over three seeds. Extended controls report three-seed mean mIoU. We compute paired RSRM gains from unrounded matched results, then average across datasets.

Models and inference The primary pipeline uses frozen RSRM, DINOv3 ViT-B/16 (Siméoni et al., 2025), and SAM 3 (Carion et al., 2025). RSRM provides both the initial prediction B_0 and the matched baseline (Sun et al., 2026). Predictions use a 480×480 grid, with DINOv3 margins bilinearly aligned from the native 30×30 token grid. For each shot setting, the inference configuration is fixed across all evaluation datasets and random seeds (Appendix A.1.1).

4.2 Comparison with Existing Methods

Comparison protocol Table 1 compares visual-matching and prompting pipelines on matched episodes and seeds. Backbone choices and inference budgets differ across the evaluated pipelines. The language-guided CG-ICS host is examined in Table 10. Table 6 separately tests selection with features, prompts, and the response bank fixed. The composition controls combine GF-SAM with DINOv3-B or Matcher with PR-MaGIC, or use SAM 3 directly, and appear in the lower block of Table 1. FSS/CXR abbreviate FSS-1000/Chest X-ray throughout the tables. SAM 3 propagation treats support and query images as consecutive frames, initialized with the support mask. Concatenation predicts the query half of side-by-side images using a box around the support foreground.

One-shot results Our method achieved an eight-dataset mean mIoU of 69.33%, exceeding SAM 3 concatenation, the strongest composition control, by 3.33 points (Table 1). Our method outperformed this control on five datasets, with the largest gains on SUIM and iSAID-5^i. Concatenation led on ISIC, PASCAL-5^i, and VOC2012 by 9.09, 1.17, and 5.66 points. Across all evaluated methods, GF-SAM led on COCO-20^i and INSID3-L on iSAID-5^i. In the paired RSRM comparison, the mean gain was 4.47 ± 0.30 points, with improvements on all eight datasets and all 24 dataset–seed pairs (Table 2).

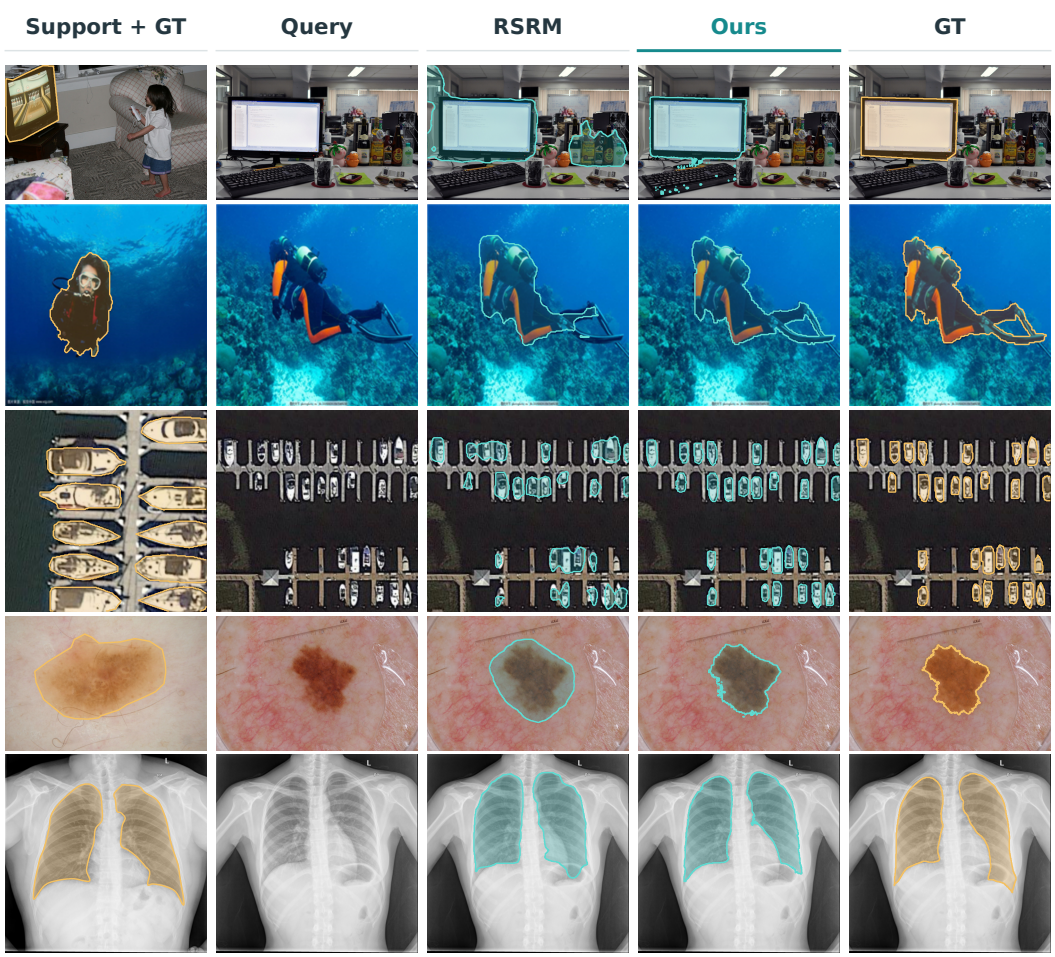

Figure 2: **Qualitative one-shot comparison with RSRM.** Amber marks annotations and teal marks predictions.

Table 3: **Five-shot comparison on eight datasets.** Three-seed mean mIoU (%) with standard deviations below in gray. Bold marks the highest method mean in each column. Paired gains (points) compare ours with RSRM.

Method	SUIM	FSS	ISIC	CXR	PASCAL	COCO	VOC	iSAID	Mean
RSRM	**68.25** ±1.00	84.01 ±0.04	60.73 ±2.96	86.95 ±0.82	81.01 ±0.19	60.92 ±1.49	78.90 ±0.76	55.34 ±1.18	72.01 ±0.43
INSID3-L	67.53 ±1.03	85.52 ±0.12	58.30 ±3.85	77.90 ±0.40	79.15 ±0.34	64.81 ±0.41	76.22 ±1.82	**58.77** ±1.11	71.03 ±0.41
GF-SAM	67.83 ±0.74	88.90 ±0.05	53.52 ±2.46	50.09 ±1.10	82.68 ±0.21	**67.45** ±0.70	81.15 ±0.96	55.52 ±0.30	68.39 ±0.44
Ours	67.75 ±0.57	**91.02** ±0.21	**63.09** ±1.66	**90.56** ±0.30	**84.28** ±0.71	65.82 ±1.22	**81.88** ±0.72	57.18 ±1.35	**75.20** ±0.21
Paired gain	−0.49 ±1.15	+7.00 ±0.25	+2.36 ±1.71	+3.61 ±0.52	+3.27 ±0.60	+4.90 ±0.37	+2.98 ±0.39	+1.84 ±0.99	**+3.18** ±0.36

Five-shot results Our method achieved 75.20% mean mIoU, exceeding INSID3-L and GF-SAM by 4.17 and 6.81 points, respectively (Table 3). Among these four methods, ours led on five datasets. RSRM, GF-SAM, and INSID3-L led on SUIM, COCO-20^i, and iSAID-5^i, respectively. The paired RSRM gain was 3.18 ± 0.36 points, with improvements on seven datasets and 22/24 dataset–seed pairs, although SUIM decreased by 0.49 points. Appendix A.2.2 gives the matched evaluation protocol.

Table 4: **One-shot design ablations on six datasets over three seeds.** Cells report class-macro foreground mIoU (%) with standard deviations below in gray. Mean weights datasets equally. Δ is the drop from the full method in percentage points. Bold marks column maxima.

Variant	SUIM	FSS	ISIC	CXR	PASCAL	COCO	Mean ↑	Δ
Full (CSC + PRCMS)	60.85 ±3.66	90.23 ±0.12	59.05 ±2.50	90.47 ±0.04	77.37 ±0.82	**56.06** ±1.20	**72.34** ±1.15	—
Reliability + spatial NMS prompts	56.96 ±3.38	88.82 ±0.11	**59.29** ±2.22	90.54 ±0.04	74.65 ±0.78	53.41 ±1.02	70.61 ±1.14	1.73
Reliability top-k prompts	**61.19** ±4.18	**90.39** ±0.14	53.09 ±2.43	90.03 ±0.04	72.92 ±0.91	52.63 ±1.10	70.04 ±1.24	2.30
Without preserve audit ($B = G_0$)	59.27 ±3.89	89.80 ±0.11	56.86 ±2.19	90.10 ±0.04	75.12 ±0.92	52.71 ±1.05	70.64 ±1.07	1.70
Initial prediction as reference ($B = B_0$)	56.97 ±3.30	88.65 ±0.13	54.57 ±2.32	89.88 ±0.05	73.68 ±0.73	50.81 ±1.33	69.09 ±1.27	3.25
No attribution-eligibility filter	59.84 ±3.42	89.92 ±0.13	57.50 ±2.30	90.64 ±0.05	76.63 ±0.88	54.45 ±1.21	71.50 ±1.10	0.84
Leave-one-out for Shapley	60.17 ±4.02	89.94 ±0.11	58.28 ±2.21	**90.65** ±0.04	**77.42** ±0.75	55.13 ±1.13	71.93 ±0.99	0.41

Table 5: **Direct composition and fixed-reference component controls.** (a) Mean mIoU (%). (b) Paired gains in points.

(a) Direct composition

Method	mIoU ↑	Δ RSRM
RSRM (B_0)	65.31	+0.00
Direct composition	63.85	−1.46
Component overlap	64.98	−0.33
Ours	**70.04**	**+4.73**

(b) Component contrasts

Comparison	Paired gain
CSC vs. top-anchor actions	+1.35±0.12
Selection vs. unconditional	+3.20±0.59
Component pipeline vs. RSRM	+4.41±0.39

Table 6: **PRCMS controls with CSC fixed.** One-shot three-seed mean mIoU (%) on eight datasets. Δ is the full-method mean (Table 1) minus the control mean, in points.

ID	Candidate pool	Selection rule	Mean mIoU ↑	Δ
N1	Raw	Semantic p	65.41	3.92
N2	Raw	Equal fusion $(p+q)/2$	66.73	2.60
C1	PRCMS	Semantic p	66.98	2.35
C2	PRCMS	Consistency q	66.63	2.70
C3	PRCMS	Equal fusion $(p+q)/2$	67.66	1.67
Full method (CSC + PRCMS)			**69.33**	—

4.3 Ablation Studies

The ablations test prompt and reference construction (Table 4), component selection (Table 5), and selection rules under shared responses (Table 6). Table 4 uses SUIM, FSS-1000, ISIC, Chest X-ray, PASCAL-5^i, and COCO-20^i over three seeds.

Prompt and reference construction With PRCMS retained, replacing CSC coverage with reliability plus spatial NMS or reliability ranking reduced mean mIoU by 1.73 and 2.30 points, respectively. The prompt budget was fixed, isolating the benefit of complementary coverage within the complete one-shot pipeline. Spatial separation alone did not match semantic coverage, supporting the allocation of prompts to complementary target regions. We next fix CSC to examine reference construction. Using B_0 or unaudited G_0 instead of the audited reference B reduced the six-dataset mean by 3.25 and 1.70 points (Table 4). Removing eligibility filtering or replacing Shapley attribution with leave-one-out effects reduced the mean by 0.84 and 0.41 points, respectively, with effects varying across datasets.

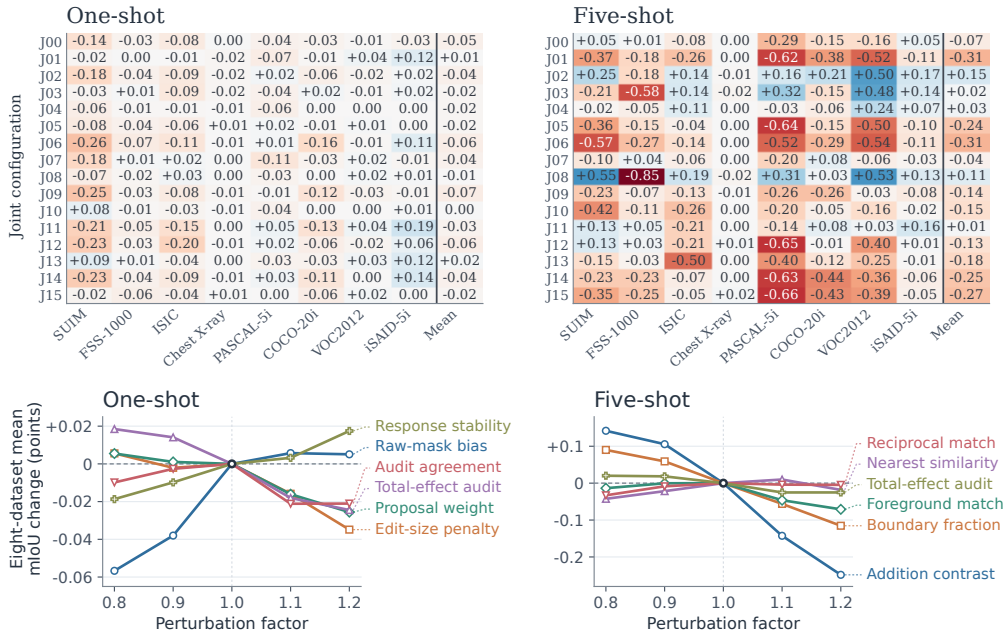

Figure 3: **Local parameter sensitivity.** Red indicates decreases and blue indicates increases. Values are three-seed mean paired mIoU changes (points) from the default.

Composition and component selection Table 5(a) compares direct model composition with the full refinement framework on eight datasets at seed 0. The controls return either the SAM 3 mask from one DINOv3 anchor or component replacements chosen by overlap with RSRM. Both underperformed RSRM, whereas the full method gained 4.73 points. Appendix A.3.1 specifies the replacement and retention rules. To separate action diversity from component selection, Table 5(b) fixes B_0 and retains the full evidence bank, excluding auditing and final mask selection. On six datasets over three seeds, CSC actions improved on the top-anchor action with component selection fixed. Conversely, selection improved over unconditional output with only the top-anchor action available. These contrasts separate action diversity from component selection (Appendix A.3.1).

Controlled PRCMS comparisons We then examine candidate scoring with CSC prompts, DINOv3 fields, B_0, and the SAM 3 response bank fixed on eight datasets over three seeds (Table 6). N1–N2 select from B_0 and raw subset masks, whereas C1–C3 use PRCMS-generated candidates, including the final output. Here p scores semantic agreement and q scores response consistency. The full method exceeded equal semantic/consistency score fusion by 2.60 points in the raw pool (N2) and 1.67 with PRCMS candidates (C3). Equal fusion benefited from PRCMS-generated candidates, improving from 66.73% to 67.66%.

We evaluated reference auditing under a common global-IoU selector. Changing the reference from B_0 to B raised mean mIoU from 66.50% to 67.49% in the raw pool and from 66.97% to 67.97% in the PRCMS pool (Table 8). The full method exceeded all extended controls, including calibrated weighted fusion, by at least 1.36 points. Together, these comparisons support both reference auditing and the subsequent decision about which changes to accept.

Contribution of revision admissibility Finally, we hold the response bank and audited B fixed to examine revision admissibility. Omitting the one-shot disagreement branch while retaining later branches yielded 68.75% mean mIoU, compared with 69.33% for the full procedure. In five-shot inference, composing edits improved the mean from 73.61% to 75.20%. Removing both coalition-eligibility and interaction-mass conditions reduced the mean by 1.47 points (Table 9).



Figure 4: **Reference auditing followed by mask revision.** Selected FSS-1000 (top) and PASCAL-5^i (bottom) episodes. Values are episode IoUs.

4.4 Further Analysis

To connect the aggregate controls to prediction behavior, Figure 4 follows two selected episodes through the recorded states B_0, G_0, B, and O. For the buckets, G_0 omits one target instance, and auditing restores both before final selection reduces excess foreground. For the train, auditing likewise restores B_0, while subsequent revision recovers missing foreground and increases IoU from 46.81% to 97.03%. These trajectories illustrate how PRCMS protects valid target regions during reference construction while allowing subsequent correction of remaining errors.

Local stability and host adaptation With CSC prompts and SAM 3 responses fixed, we perturbed PRCMS parameters individually and jointly by factors in $[0.8, 1.2]$ across eight datasets and three seeds (Appendix A.5.1). Figure 3 shows joint changes by dataset and mean above, and ranks individual parameters by their largest absolute mean change over four factors below. Maximum mean decreases from the unchanged defaults were 0.068 and 0.307 points for one-shot and five-shot, respectively, indicating local stability within this range. We also evaluated host-adapted selection in CG-ICS, which augments SAM 3 with Qwen3-VL concept reasoning and search (Chen et al., 2026). This adaptation improved all eight datasets at seed 0, from 76.60% to 77.68% mean mIoU (Table 10).

Inference cost Uniform response averaging and the full pipeline averaged 342.62 and 387.65 ms per episode, respectively, in the one-shot timing comparison. Both recomputed identical features, CSC prompts, and SAM 3 response banks. The full pipeline added 13.1% end-to-end latency relative to this full-bank control (Table 14 and Appendix A.5.3).

Qualitative results Across five selected episodes in Figure 2, refinement reduced spillover around the monitor and skin lesion and recovered diver, ship, and lung regions relative to RSRM. Rows correspond to COCO-20^i, FSS-1000, iSAID-5^i, ISIC, and Chest X-ray, with support images lightly cropped for alignment. Residual errors included lesion-boundary mismatches and missed small ship instances.

5 Conclusion

We introduced a training-free framework for cross-domain few-shot segmentation that audits a semantic reference before deciding which mask changes to accept. CSC selects reliable, complementary prompts, while PRCMS combines support semantics with counterfactual mask evidence to audit the reference and evaluate directional revisions. Experiments across eight datasets demonstrate state-of-the-art average performance in the primary one-shot and five-shot comparisons. However, exhaustive prompt-subset evaluation increases inference cost and limits scalability. Future work will therefore investigate counterfactual evidence estimation without exhaustive enumeration, aiming to improve efficiency while retaining the benefits of reference auditing and directional refinement.

USE OF GENERATIVE AI

Generative AI tools assisted with manuscript drafting and language revision, code development, experimental analysis, and figure preparation. The authors are responsible for verifying AI-assisted outputs and retain full responsibility for the methods, reported results, interpretations, and final content of this work.

REPRODUCIBILITY STATEMENT

The inference procedure and implementation details of CSC and PRCMS are described in Section 3 and Appendix A.1. Section 4.1 and Appendix A.2 specify the datasets, episode construction, preprocessing, and evaluation protocols. Additional ablations and robustness analyses are reported in Appendices A.3 and A.5. The anonymous supplementary materials include the source code, inference configurations, and logs for the reported results.

REFERENCES

Abderrahmene Boudiaf, Irfan Hussain, and Sajid Javed. SegRAG: Training-free retrieval-augmented semantic segmentation. arXiv preprint arXiv:2605.17630, 2026.

Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P. Musco, Rahul K. Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J. McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Transactions on Medical Imaging*, 33(2):577–590, February 2014.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719, 2025.

Zhigang Chen, Xiawu Zheng, and Rongrong Ji. Toward robust in-context segmentation via concept guidance. In *European Conference on Computer Vision (ECCV)*, 2026.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). arXiv preprint arXiv:1902.03368, 2019.

Claudia Cuttano, Gabriele Trivigno, Christoph Reich, Daniel Cremers, Carlo Masone, and Stefan Roth. INSID3: Training-free in-context segmentation with DINOv3. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21638–21648, June 2026.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88 (2):303–338, Jun 2010.

Qi Fan, Wenjie Pei, Yu-Wing Tai, and Chi-Keung Tang. Self-support few-shot semantic segmentation. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 701–719, Cham, 2022. Springer Nature Switzerland.

Jonas Herzog. Adapt before comparison: A new perspective on cross-domain few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23605–23615, June 2024.

Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic Segmentation of Underwater Imagery: Dataset and Benchmark. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1769–1776. IEEE/RSJ, 2020.

Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K. Singh, Sameer Antani, George Thoma, Yi-Xiang Wang, Pu-Xuan Lu, and Clement J. McDonald. Automatic tuberculosis screening using chest radiographs. *IEEE Transactions on Medical Imaging*, 33(2):233–245, February 2014.

Minjae Lee, Sungwoo Hur, Soojin Hwang, and Won Hwa Kim. PR-MaGIC: Prompt refinement via mask decoder gradient flow for in-context segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21659–21668, June 2026.

Shuo Lei, Xuchao Zhang, Jianfeng He, Fanglan Chen, Bowen Du, and Chang-Tien Lu. Cross-domain few-shot semantic segmentation. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 73–90, Cham, 2022. Springer Nature Switzerland.

Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. FSS-1000: A 1000-class dataset for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2869–2878, June 2020.

Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 17904–17925, 2024.

Lingchen Meng, Shiyi Lan, Hengduo Li, Jose M. Alvarez, Zuxuan Wu, and Yu-Gang Jiang. SegIC: Unleashing the emergent correspondence for in-context segmentation. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 203–220, Cham, 2024. Springer Nature Switzerland.

G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14(1):265–294, Dec 1978.

Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 622–631, October 2019.

Bohao Peng, Zhuotao Tian, Xiaoyang Wu, Chengyao Wang, Shu Liu, Jingyong Su, and Jiaya Jia. Hierarchical dense correlation distillation for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23641–23651, June 2023.

Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In *British Machine Vision Conference (BMVC)*, 2017.

L. S. Shapley. 17. a value for n-person games. In *Contributions to the Theory of Games (AM-28), Volume II*, pp. 307–318. Princeton University Press, December 1953.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. arXiv preprint arXiv:2508.10104, 2025.

Jiapeng Su, Qi Fan, Wenjie Pei, Guangming Lu, and Fanglin Chen. Domain-rectifying adapter for cross-domain few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24036–24045, June 2024.

Sujun Sun, Mingwu Ren, and Haofeng Zhang. Training-free cross-domain few-shot segmentation via robust semantic representation and matching. In *European Conference on Computer Vision (ECCV)*, 2026.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1): 180161, Aug 2018.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. SegGPT: Towards segmenting everything in context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1130–1140, October 2023.

Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. iSAID: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

Anqi Zhang, Guangyu Gao, Jianbo Jiao, Chi Harold Liu, and Yunchao Wei. Bridge the points: Graph-based few-shot segment anything semantically. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 33232–33261. Curran Associates, Inc., 2024a.

Jian-Wei Zhang, Yifan Sun, Yi Yang, and Wei Chen. Feature-proxy transformer for few-shot segmentation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 6575–6588. Curran Associates, Inc., 2022.

Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Yu Qiao, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 18250–18279, 2024b.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 633–641, July 2017.

Muzhi Zhu, Yang Liu, Zekai Luo, Chenchen Jing, Hao Chen, Guangkai Xu, Xinlong Wang, and Chunhua Shen. Unleashing the potential of the diffusion model in few-shot semantic segmentation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 42672–42695. Curran Associates, Inc., 2024.

Yazhou Zhu and Haofeng Zhang. MAUP: Training-free multi-center adaptive uncertainty-aware prompting for cross-domain few-shot medical image segmentation. In James C. Gee, Daniel C. Alexander, Jaesung Hong, Juan Eugenio Iglesias, Carole H. Sudre, Archana Venkataraman, Polina Golland, Jong Hyo Kim, and Jinah Park (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pp. 326–336, Cham, 2025. Springer Nature Switzerland.

A Appendix

A.1 Method Details

Algorithm 1 summarizes prompt construction, response comparison, reference auditing, and revision. The following sections detail one-shot mask selection and five-shot edit composition. Masks are foreground pixel sets on Ω; $\langle F \rangle_R$ denotes the mean of a field over region R.

A.1.1 Models, inputs, and implementation conventions

Frozen DINOv3 ViT-B/16 (LVD-1689M) supplies correspondence streams from blocks 9–12, with masks aligned to a 480×480 grid. RSRM provides the initial prediction, and the SAM 3 video tracker generates each response bank from a fixed tracker snapshot. SAM 3 uses the first support in both shot settings; five-shot DINOv3 evidence averages all five supports.

The one-shot and five-shot inference configurations are selected on ADE20K (Zhou et al., 2017). They specify numerical bounds, score coefficients, normalization scales, and numerical floors. The following equations and decision tables define the evidence statistics, acceptance conditions, and output order.

Algorithm 1: Preserve-relative counterfactual inference

Input: Query I_q, annotated supports $\mathcal{S}$, frozen models.
Output: Binary query mask O.

1. **Construct prompts.** Obtain B_0 and support-conditioned semantic fields. CSC selects complementary prompts N and identifies their reciprocal attribution eligibility $\mathcal{E}$.
2. **Collect response evidence.** Evaluate every prompt subset from the same SAM 3 state. Compute eligible attribution $\Phi_{\mathcal{E}}$ and total effect T from the shared response bank.
3. **Audit the reference.** Form a structural proposal G_0 from singleton and joint responses. Audit its disagreements with B_0 to construct the fixed preserve state B.
4. **Resolve admissible revisions.** Construct proposals and apply the admissibility conditions in Equation (5). Select one admissible whole-mask proposal for one-shot inference, or compose admissible disjoint groups for five-shot inference. Return $(B \setminus U^-) \cup U^+$; an empty admissible set returns B.

A.1.2 Prompt evidence and reciprocal eligibility

Semantic demand and reliability At each correspondence layer, annotated support tokens define foreground and background prototypes. For each normalized query token, the semantic margin is the difference between its inner products with the foreground and background prototypes. CSC uses normalized prototypes; the audit fields $D^{\ell,r}$ retain prototype magnitudes. Each support–layer margin field is first median-centered, then scaled by the larger of its normalized median absolute deviation and standard deviation. Averaging the standardized fields combines support evidence on a common scale.

Let $\bar{z}_j$ be the mean standardized CSC margin, g_j the fraction of fields placing token j in their upper quintile, and v_j its foreground matching frequency. With the soft initial-mask prior $\widetilde{B}_{0,j}$, the demand and reliability are

$$w_j = \mathrm{softplus}(\bar{z}_j)\frac{1+g_j}{2}(1+\widetilde{B}_{0,j}), \qquad r_j = g_j v_j \,\mathrm{sigmoid}(\bar{z}_j). \tag{A.1.1}$$

Demand identifies regions to cover; reliability weights the anchors. The prior increases demand inside B_0 without excluding regions outside it.

Complementary selection For selected anchors A, let $c_A(j) = \max_{a\in A} \kappa(a,j)$. The marginal gain in Equation (3) is

$$\Delta(a \mid A) = \sum_j w_j[\kappa(a,j) - c_A(j)]_+. \tag{A.1.2}$$

An anchor therefore contributes only where it improves existing coverage. The strongest fused semantic match seeds the selection, followed by greedy additions until coverage is complete, the prompt budget is reached, or the screened pool is exhausted.

Reciprocal eligibility Let $e_i^{\ell,r} = 1$ when prompt i matches a foreground support token whose nearest query match returns to the original prompt location, and zero otherwise. The eligible set is

$$\mathcal{E} = \left\{ i \in N : \max_{\ell} \sum_{r=1}^{K} e_i^{\ell,r} > K/2 \right\}. \tag{A.1.3}$$

One-shot inference requires a return at one layer; five-shot inference requires returns through at least three supports at the same layer. Eligibility governs attribution and coalitions. Response generation uses the full selected prompt set.

A.1.3 Prompt responses and reference auditing

A.1.3.1 Controlled response evidence For each $S \subseteq N$, SAM 3 receives the corresponding positive prompts while the query encoder state, support memory, and negative point stay fixed. The negative point is the unselected token with the lowest fused semantic margin. Accordingly, the empty subset is also evaluated with the same support memory and negative point. These controlled prompt-subset comparisons provide the counterfactual evidence. Because adding a prompt can decrease a response, attribution is signed.

The exact pixelwise Shapley contribution is

$$\phi_i(x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(k-|S|-1)!}{k!}[h_{S\cup\{i\}}(x) - h_S(x)]. \tag{A.1.4}$$

The two audit fields are

$$\Phi_{\mathcal{E}}(x) = \sum_{i\in\mathcal{E}} \phi_i(x), \qquad T(x) = h_N(x) - h_\varnothing(x) = \sum_{i\in N} \phi_i(x). \tag{A.1.5}$$

The eligible attribution is normalized after summation; an empty eligible set yields zero attribution without changing the response bank or total effect. In one-shot inference, attribution guides the construction of B, which defines the changes evaluated during final selection. Directional scores assess these changes using response and correspondence evidence.

A.1.3.2 Constructing and auditing the proposal Singleton and all-positive responses supply structural replacements for each 8-connected component C of B_0. A response contributes its components that overlap C and their union; arbitrary sub-unions are not enumerated. Identical candidates retain their first source. For a replacement Q, define

$$\begin{aligned} a &= |C \cap Q|/|C|, \qquad p = |B_0 \cap Q|/|Q|, \qquad I = \frac{2ap}{a+p}, \\ r_A &= |Q|/|C|, \qquad r_P = \frac{P(Q)/\sqrt{|Q|}}{P(C)/\sqrt{|C|}}. \end{aligned} \tag{A.1.6}$$

Here P counts foreground pixels removed by one zero-padded binary erosion with an eight-neighbor structuring element. Reproduction s_Q is the maximum Dice with a component from another response. For each alternative response, we take the union of components overlapping C. Averaging the resulting foreground indicators gives v_C. A response without an overlapping component contributes a zero indicator. Extent agreement is

$$e_Q = \frac{\sum_{x\in Q\setminus C} v_C(x) + \sum_{x\in C\setminus Q}[1 - v_C(x)]}{|C \triangle Q|}, \tag{A.1.7}$$

with $e_Q = 1$ when $C = Q$. The replacement score is an affine combination of these statistics,

$$J_0(Q; C) = I + \mathbf{w}_0^{\mathsf{T}}(-|\log r_A|, -[\log r_P]_+, s_Q, e_Q, 1). \tag{A.1.8}$$

The final coordinate absorbs fixed offsets. Geometry penalizes area change and excess boundary complexity; cross-response evidence rewards reproduction and agreement on changed pixels. If no alternative response exists, both cross-response terms, including their offsets, are omitted. The first score maximizer replaces C only when its score is positive; otherwise C is retained. Candidate order is singleton responses, the joint response, then component order followed by the within-response union. The union of selected or retained components forms G_0. Unmatched components are not introduced at this step, and an empty B_0 yields an empty G_0.

An overlap graph groups initial components and their proposed replacements into disjoint regions. In group t, let X_t and Y_t be the initial and proposed foreground assignments, and let $d_t = |X_t \triangle Y_t|/|\Omega|$. Using the restoration utility in Equation (4), the audit restores X_t when

$$r_t = \mathbf{1}\left\{ \min\left(u_{\Phi_{\mathcal{E}}}(X_t, Y_t), \min_{\ell} u_{D^\ell}(X_t, Y_t) \right) > \tau_{\mathrm{a}} \ \vee \ (d_t > \tau_{\mathrm{d}} \ \wedge \ u_T(X_t, Y_t) > \tau_{\mathrm{t}}) \right\}. \tag{A.1.9}$$

The first test combines prompt attribution and support semantics; the second uses total-effect evidence for larger disagreements. Keeping X_t when $r_t = 1$ and Y_t otherwise produces the preserve state

$$B(x) = \sum_t \left[r_t \mathbf{1}_{X_t}(x) + (1 - r_t)\mathbf{1}_{Y_t}(x) \right]. \tag{A.1.10}$$

Because the groups are disjoint, B stays binary and shared assignments remain unchanged. This state is then held fixed during terminal revision.

A.1.4 Evidence and resolution of revisions

All branches assess changes against fixed B using prompt responses, support semantics, and spatial evidence. One-shot inference selects a mask in branch order; five-shot inference accepts disjoint edits through direction-specific branches. Shared quantities are defined before the branches.

Shared notation Let $o_R(j)$ be the occupancy of region R in query-token cell j and

$$\mathcal{W}_R(q) = \frac{\sum_j o_R(j) q_j}{\max(\sum_j o_R(j), \epsilon)}. \tag{A.1.11}$$

ϵ denotes a positive numerical floor. Write $[z]_{[a,b]}$ for clipping, $\mathrm{HM}(a,b) = 2ab/\max(a+b,\epsilon)$ for the harmonic mean, and $T_F(z) = \tanh(z/a_F)$ for a fixed-scale transform. The harmonic mean is zero when both inputs vanish. Each shot setting uses a fixed configuration. Its lower and upper bounds are

$$\begin{aligned} L_v(x) &= [x > \ell_{v,x}], \qquad U_v(x) = [x < u_{v,x}], \\ \bar{L}_v(x) &= [x \geq \ell_{v,x}], \qquad \bar{U}_v(x) = [x \leq u_{v,x}]. \end{aligned} \tag{A.1.12}$$

Subscripts identify branch-specific bounds for each statistic; brackets denote Boolean indicators. Table conditions are joined by AND unless OR is shown. Exact numerical configurations accompany the implementation.

A.1.4.1 One-shot mask selection

Candidates from individual responses For a subset S, let $\Delta_S = h_S - h_\varnothing$, $A_S = M_S \setminus B$, and $R_S = B \setminus M_S$. Directional evidence defines

$$\begin{aligned} H^\sigma_{S,\mathrm{strict}} &= \{x : \sigma\Delta_S(x) > 0,\ \min_\ell \sigma D^\ell(x) > 0\}, \\ H^\sigma_{S,\mathrm{mean}} &= \{x : \sigma\Delta_S(x) > 0,\ \sigma\bar{D}(x) > 0\}, \end{aligned} \tag{A.1.13}$$

where $\bar{D}$ averages layers and $\sigma \in \{+1, -1\}$. For either semantic view v, the joint projection is

$$G_{S,v} = (B \cup [A_S \cap H^{+1}_{S,v}]) \setminus [R_S \cap H^{-1}_{S,v}]. \tag{A.1.14}$$

Addition-only and deletion-only projections keep just the corresponding edit. Together with M_S, they form the raw and projected candidate family. Pixel-identical candidates retain their source and construction identity because their response evidence can differ.

Response evidence The bank's foreground frequency is $v(x) = 2^{-k}\sum_T M_T(x)$. With $Z_S = M_S \triangle B$, all projections from S share stability

$$s_S = \langle M_S v + (1 - M_S)(1 - v)\rangle_{Z_S}. \tag{A.1.15}$$

All projections inherit the stability measured over their source response's changes relative to B. Define $m_S = \langle |h_S - 1/2|\rangle_{Z_S}$ and $p_S = \langle h_S\rangle_{M_S} - \langle h_S\rangle_{\Omega\setminus M_S}$. Response confidence is

$$q_S = \tfrac{1}{2}\left([T_m(m_S)]_{[0,1]} + [(1 + T_p(p_S))/2]_{[0,1]}\right). \tag{A.1.16}$$

Correspondence and edit geometry Dense nearest-neighbor matching records foreground membership and reciprocal return through foreground for each support and layer. Their averages give foreground matching frequency q_{fg} and reciprocal-foreground frequency q_{rf}. The field q_{sim} averages nearest-match cosine similarity over the same supports and layers. For these dense fields, if no support token reaches foreground occupancy one half, the maximum-occupancy token is retained as foreground. Prompt eligibility is computed separately using Equation (A.1.3). Relative fidelity is the geometric mean of normalized correspondence,

$$\begin{aligned} f(R;B) &= \left[\prod_{q\in\{q_{\mathrm{fg}},q_{\mathrm{rf}},q_{\mathrm{sim}}\}} \left[\frac{\mathcal{W}_R(q)}{\max(\mathcal{W}_B(q),\epsilon_f)}\right]_{[0,c_f]}\right]^{1/3}, \\ f^+ &= [f(E^+;B)]_{[0,1]}, \qquad f^- = [1 - f(E^-;B)]_{[0,1]}. \end{aligned} \tag{A.1.17}$$

c_f is the fixed clipping cap. For an empty B, only the normalization means use a foreground surrogate: rank tokens with positive q_{fg} by a fixed affine score of $(q_{\mathrm{rf}}, q_{\mathrm{fg}}, [q_{\mathrm{contrast}}]_+, q_{\mathrm{sim}})$ and retain the configured number of highest-ranked tokens. If none have positive foreground membership, rank all tokens. q_{contrast} is the layer-averaged difference between maximum foreground and background cosine similarities, or zero without background tokens. Stable sorting followed by taking the last entries resolves ties. The surrogate supplies normalization means, while B remains the binary reference.

A fixed occupancy cutoff defines the token core of B. Its exterior and interior distance transforms are d_{out} and d_{in}. For the fixed exterior and interior band widths, $c_0 = \mathcal{W}_{E^+}(\mathbf{1}\{d_{\text{out}} \le r_{\text{out}}\})$ and $b = \mathcal{W}_{E^-}(\mathbf{1}\{d_{\text{in}} \le r_{\text{in}}\})$. Scaled addition contact is $c = [c_0/a_c]_{[0,1]}$. The direction-specific scores are

$$C^+ = s_S[c + (1-c)f^+], \qquad C^- = s_S[b + (1-b)f^-]. \tag{A.1.18}$$

Detached additions therefore rely more on correspondence, whereas interior removals require weaker foreground correspondence to receive support.

Let $\mu_\sigma = \sum_j o_{E^\sigma}(j)$, $\rho_\sigma = \mu_\sigma / \max(\mu_+ + \mu_-, \epsilon)$, and $\bar{C} = \sum_\sigma \rho_\sigma C^\sigma$. Relative extent is $L_p = \log(1 + \ell_+ + \ell_-)$, with ℓ_σ the capped ratio $|E^\sigma|/\max(|B|, 1)$. For each present direction, proposal support is

$$P^\sigma = \min\left\{ T_\Delta(\sigma\langle \Delta_S \rangle_{E^\sigma}), \min_\ell \tanh(\sigma \mathcal{W}_{E^\sigma}(D^\ell)) \right\}, \qquad \bar{P} = \sum_\sigma \rho_\sigma P^\sigma. \tag{A.1.19}$$

Here semantic fields are block-averaged before occupancy weighting.

Response-derived selection (P) This branch assesses masks from individual subset responses and their projections. Response consistency, confidence, edit extent, and directional evidence determine admissibility. A candidate is admissible if and only if

$$\begin{aligned} |G \triangle B| > 0 \ \wedge L_{\text{P}}(s_S) \wedge U_{\text{P}}(L_p) \wedge L_{\text{P}}(q_S) \\ \wedge \bigwedge_{\sigma \in \{+,-\}} \big[\mu_\sigma = 0 \ \vee \ \bar{U}_{\text{P}}(\rho_\sigma) \ \vee \ L_{\text{P}}(C^\sigma)\big]. \end{aligned} \tag{A.1.20}$$

A direction whose mass share is below its configured floor does not independently veto the candidate. Admissible candidates receive the score

$$J_{\text{P}}(G) = s_S + \varepsilon_{\text{rank}}\, \mathbf{w}_{\text{P}}^{\mathsf{T}}(\bar{C}, -L_p, \bar{P}, \mathbf{1}_{\text{raw}}, \mathbf{1}_{\text{mean}}). \tag{A.1.21}$$

Within this branch, candidates are ranked by the scalar score $J_{\text{P}}(G)$. The mean indicator includes all mean-layer projection modes.

Candidates from response consensus If no response-derived candidate is admissible, PRCMS assesses aggregate evidence from response consensus. Binary frequency v and mean soft response $\bar{h} = 2^{-k}\sum_S h_S$ each produce a majority mask and two hysteresis masks. A hysteresis mask is

$$G_{q,l,u}(x) = \begin{cases} 0, & q(x) \le l, \\ B(x), & l < q(x) < u, \\ 1, & q(x) \ge u. \end{cases} \tag{A.1.22}$$

The two fixed intervals and majority construction give six raw consensus candidates. Their semantic projections use the same strict/mean signs as above, without a subset-specific Δ_S condition. Cross-view comparisons stay within a projection mode; only raw consensus masks can be selected, so projected masks do not enter raw-mask evidence averages.

Consensus stability s_G is the minimum of $\langle q \rangle_{E^+}$ and $\langle 1-q \rangle_{E^-}$ over directions present. Its semantic score d_G is the minimum signed semantic-layer average over those directions. Addition safety is $c_0 + (1-c_0)\,\text{density}$; deletion safety is $b + (1-b)\exp[-\mathcal{W}_{E^-}(d_{\text{in}})/a_b]$. Their changed-area-weighted mean is t_G. Density averages neighboring addition occupancies with eight-neighbor zero padding, then weights by addition occupancy. The divisor remains eight at borders. Absent directions are omitted from directional minima.

Agreement across views and calibration to the reference For consensus view v, let $Q_v = G_v \triangle B$. V_{all} and V_{cross} average edit indicators over Q_v from other views and from the opposite binary/soft family, respectively. I_{all} is mean other-view edit IoU; I_{cross} is maximum other-family edit IoU. Empty edit unions contribute zero. Then

$$\begin{aligned} \mathcal{P}_v &= \text{HM}(V_{\text{all}}, V_{\text{cross}}), \quad \mathcal{C}_v = \text{HM}(I_{\text{all}}, I_{\text{cross}}), \\ a_v &= (1-\eta)\mathcal{P}_v + \eta \mathcal{C}_v, \quad u_G = \text{HM}([\tanh d_G]_{[0,1]}, a_v). \end{aligned} \tag{A.1.23}$$

These averages use the other five raw views and the three opposite-family views, respectively. The mixing and deletion-depth scales can differ by assessment, as specified in the fixed configuration.

Reference calibration uses foreground/background weighted medians of each block-averaged D^ℓ, with weights o_B and $1 - o_B$. Let m_ℓ be their midpoint, t_ℓ their absolute difference and ξ_ℓ the sign of foreground minus background, taking $+1$ at equality. Define

$$a_\ell = \max\{t_\ell/2, \operatorname{med}_j |D_j^\ell - m_\ell|, \epsilon\}, \qquad \widetilde{D}_j^\ell = \xi_\ell (D_j^\ell - m_\ell)/a_\ell. \tag{A.1.24}$$

Weighted medians select the first sorted value reaching half the total weight. Calibration requires nonvanishing foreground and background masses; otherwise it uses the field median, positive orientation, and zero separation. The score γ_G is the hyperbolic tangent of the minimum signed regional average of $\widetilde{D}^\ell$ over layers and present directions. Reference separation z_B is the capped minimum of t_ℓ/a_ℓ over layers.

Consensus branches and shared checks The remaining branches assess the same consensus candidates, using disagreement (D), persistence (V), then calibration (R). Define $a_G = |G \triangle B|/|\Omega|$, $L_G = \log[1 + |G \triangle B|/\max(|B|, 1)]$, and $b_G = 2\min(|E^+|, |E^-|)/\max(|G \triangle B|, 1)$. All three apply the same spatial and response checks,

$$V_v(G) = [a_G > 0] \wedge \bar{L}_v(s_G) \wedge \bar{L}_v(t_G) \wedge \bar{U}_v(a_G) \wedge \bar{U}_v(L_G) \wedge \bar{U}_v(b_G). \tag{A.1.25}$$

Disagreement (D) This branch assesses response-supported changes under semantic disagreement, using the shared response and spatial checks.

Persistence (V) This branch combines semantic support and cross-view agreement through their harmonic mean u_G for admissibility and ranking.

Calibration (R) This branch uses semantics recalibrated to B, reference separation, and cross-view agreement. Its lower bound on γ_G need not be positive.

The table gives each branch's additional evidence and within-branch score.

Assessment	Additional condition beyond V_v	Ranking score J_v
Disagreement D	$\bar{L}_{\mathrm{D}}([-\tanh d_G]_{[0,1]})$	$s_G + \mathbf{w}_{\mathrm{D}}^{\mathsf{T}}(t_G, -L_G)$
Persistence V	$\bar{L}_{\mathrm{V}}(u_G)$	$s_G + \mathbf{w}_{\mathrm{V}}^{\mathsf{T}}(t_G, u_G, -L_G)$
Calibration R	$\bar{L}_{\mathrm{R}}(\gamma_G) \wedge \bar{L}_{\mathrm{R}}(z_B) \wedge \bar{L}_{\mathrm{R}}(a_v)$	$s_G + \mathbf{w}_{\mathrm{R}}^{\mathsf{T}}(t_G, \gamma_G, a_v, -L_G)$

Output choice Evaluate the families in the fixed order $\mathrm{P} \prec \mathrm{D} \prec \mathrm{V} \prec \mathrm{R}$. Stop at the first nonempty admissible family and maximize its stated score. Equivalently, if π_ξ is a candidate's first admitting assessment,

$$\xi^\star = \operatorname*{arg\,lex\,max}_{\xi \in \mathcal{A}_1(B)} (-\pi_\xi, J_{\pi_\xi}(\xi)). \tag{A.1.26}$$

Exact ties retain the first candidate, ordered by projection mode and then subset bitmask in CSC selection order. Consensus order is binary before soft, each followed in majority, narrow-hysteresis, wide-hysteresis order. The selected candidate supplies U^+ and U^-; an empty admissible set returns B.

A.1.4.2 Five-shot edit composition

From response disagreement to local proposals Unlike one-shot selection of a complete mask, five-shot inference composes local changes using support-averaged semantic evidence. An atom C is an 8-connected region with constant (B, B_0, G_0, M_N) signature where at least one of the latter states differs from B. Its proposed label is $y_C = 1 - B|_C$, with direction $\sigma_C = 2y_C - 1$. For eligible nonempty $S \subseteq \mathcal{E}$, define

$$a_C(S) = \frac{1}{|C|} \sum_{x \in C} \mathbf{1}\{M_S(x) = y_C\}, \qquad n_C(S) = \min_{i \in S} [a_C(S) - a_C(S \setminus \{i\})]. \tag{A.1.27}$$

A sufficient coalition has $a_C(S) > 1/2$ and $n_C(S) > 0$. Choose the smallest sufficient coalition, breaking ties by greatest necessity and then smallest subset identifier. Necessity tests immediate one-prompt removals, using the measured negative-only response for a singleton. Adjacent atoms merge only when direction and selected nonempty coalition match. Atoms without a coalition remain separate. The resulting disjoint groups form $\mathcal{R}$.

Regional response and semantic evidence For $\nu_C(S) = \langle h_S - h_\varnothing \rangle_C$, define

$$\delta_C(S) = \sum_{T \subseteq S} (-1)^{|S|-|T|} \nu_C(T), \tag{A.1.28}$$

$$m_R = \max_{C \subseteq R} \sum_{\varnothing \neq S \subseteq \mathcal{E}} [\sigma_C \delta_C(S)]_+. \tag{A.1.29}$$

The decomposition uses the full response lattice; the sum includes eligible singleton and higher-order contributions, without averaging or clipping. With image-wide normalized semantic fields $\widehat{D}^\ell$, set $d_C^\ell = \sigma_C \langle \widehat{D}^\ell \rangle_C$ and

$$d_{\min}(R) = \sum_{C \subseteq R} \frac{|C|}{|R|} \min_\ell d_C^\ell, \qquad d_{\max}(R) = \sum_{C \subseteq R} \frac{|C|}{|R|} \max_\ell d_C^\ell. \tag{A.1.30}$$

Layer-wise minima and maxima are computed within each atom before area-weighted aggregation over the group. For deletion, the reference percentile is

$$\pi_B(R) = \min_\ell \frac{1}{|B|} \sum_{x \in B} \mathbf{1}\{\widehat{D}^\ell(x) \leq \langle \widehat{D}^\ell \rangle_R\}. \tag{A.1.31}$$

It ranks the group using the foreground-positive orientation of $\widehat{D}^\ell$. Since a deletion group is a nonempty subset of B, this reference is nonempty. Background-relative contrast is

$$\chi_B(R) = \frac{1}{|\mathcal{L}|} \sum_\ell \left[\langle \widehat{D}^\ell \rangle_R - \mathrm{med}_{x \notin B}\, \widehat{D}^\ell(x) \right]. \tag{A.1.32}$$

An empty background uses the full-image median. Let f_R, n_R, r_R be pixel-averaged foreground matching, nearest similarity, and reciprocal matching, after support averaging and bilinear interpolation. Here, reciprocal matching does not require the matched support token to belong to the foreground. Let b_R be the fraction of group pixels on the inner pixel boundary of B under one zero-padded eight-neighbor erosion, and $\ell_R = |R| / \max(|B|, 1)$. This boundary statistic differs from the one-shot token-grid statistic.

Acceptance by direction and evidence source Let c_R indicate that all atoms in R have a sufficient coalition. Coalition branches are assessed first. Groups not accepted there proceed to correspondence deletion and then contrast addition, as applicable to their direction. The table gives the acceptance conditions, using the branch-specific bounds in Equation (A.1.12). A group is accepted if any applicable branch accepts it.

Coalition addition (CA) This branch checks coalition support for a proposed foreground region together with its minimum and maximum layer-wise semantic evidence.

Coalition deletion (CD) This branch combines coalition support with upper bounds on edit extent and interaction mass, and a lower bound on semantic percentile. Positive interaction contributions can accumulate even when the net response change is small. The upper bound on m_R constrains this accumulated effect, while the percentile expresses the group's relative semantic rank within B.

Correspondence deletion (FD) This branch uses weak foreground or nearest-match evidence to assess removal, together with reciprocal matching, boundary contact, and edit extent. It can accept groups that fail the coalition branch.

Contrast addition (BA) This branch assesses additions using foreground correspondence and semantic contrast against the background of B. Like correspondence deletion, it requires no coalition.

Evidence source	Conditions for accepting group R
Coalition addition CA	$\sigma_R = +1$, $c_R = 1$, $L_{\mathrm{CA}}(d_{\min})$, $L_{\mathrm{CA}}(d_{\max})$
Coalition deletion CD	$\sigma_R = -1$, $c_R = 1$, $U_{\mathrm{CD}}(\ell_R)$, $L_{\mathrm{CD}}(\pi_B)$, $U_{\mathrm{CD}}(m_R)$
Correspondence deletion FD	$\sigma_R = -1$, $(\bar{U}_{\mathrm{FD}}(f_R) \vee \bar{U}_{\mathrm{FD}}(n_R))$, $\bar{U}_{\mathrm{FD}}(r_R)$, $\bar{L}_{\mathrm{FD}}(b_R)$, $\bar{U}_{\mathrm{FD}}(\ell_R)$
Contrast addition BA	$\sigma_R = +1$, $\bar{L}_{\mathrm{BA}}(\chi_B)$, $\bar{L}_{\mathrm{BA}}(f_R)$

Simultaneous composition All groups use fixed B, without intermediate updates or group ranking. The accepted changes are

$$U^\pm = \bigcup_{\substack{R \in \mathcal{R}: \mathrm{Adm}_5(R|B)=1 \\ \sigma_R = \pm 1}} R, \qquad O = (B \setminus U^-) \cup U^+. \tag{A.1.33}$$

Disjointness makes the output independent of update order and preserves labels outside accepted groups. If no group is admissible, the output is B.

A.2 Experimental Protocols

Evaluation episodes withhold current-query annotations from prediction inputs and retain the one-shot query when extended to five shots. Their sampling rules are specified below.

A.2.1 Evaluation episodes

Episode roles and source splits Each episode pairs one query with either one or five distinct support images containing the same target class. The current query is excluded from its support set. Query and support roles can change across episodes. Matched comparisons share fixed episode lists for seeds 0, 1, and 2. Table 7 lists the one-shot source pools and query construction rules. Fold-based datasets retain their released evaluation class folds.

The suite combines natural-image generalization with underwater, medical, and aerial domain shifts. Its macro mean equally weights the eight datasets. Consequently, the episode counts in Table 7 describe coverage, not the relative contribution of each dataset to the reported macro mean.

In Table 7, counts are episodes per seed, summed over evaluation folds where applicable. The five-shot extension retains these episodes; images can recur across episodes.

Table 7: **One-shot episode sources and coverage.**

Dataset	Source pool and query construction	Classes	Episodes
SUIM	Test images; enumerate valid, nonempty class–image pairs	7	285
FSS-1000	Released test classes; each image used as a query	240	2,400
ISIC	RSRM evaluation pool from ISIC2018 challenge training data; 400 sampled queries per category	3	1,200
Chest X-ray	RSRM lung-segmentation pool; sampled queries	1	1,200
PASCAL-5^i	Four validation folds; cycle through each fold's query list for 1,000 episodes	20	4,000
COCO-20^i	Four COCO2014 validation folds; sample a class and a query for 1,000 episodes per fold	80	4,000
VOC2012	Validation queries and training supports; 100 episodes per class	20	2,000
iSAID-5^i	Three validation folds; enumerate class–patch entries	15	10,063

One-shot support sampling For each query, a support is sampled from images containing the target class after excluding the query identity. Both roles use the listed source pool, except that VOC2012 draws its first support from the training split. FSS-1000, ISIC, and Chest X-ray follow the RSRM test-loader sampling procedures. ISIC images retain their source partition name, "training", but serve as target evaluation images in this protocol. For iSAID, the sampling unit is an individual image patch. Query exclusion prevents reuse of that patch as support; separation of the original aerial scenes is not enforced.

Five-shot extension Five-shot episodes retain the one-shot query and first support, then add four same-class supports. The pool for additional supports contains images appearing as queries or supports in the one-shot lists for that seed. After excluding the current query and first support, a fixed hash ordering determines the additional images. The hash uses the seed-specific protocol, dataset, class, and image identity. Additional supports follow this shared within-class ordering.

For VOC2012, additional supports can come from both training and validation images represented in the one-shot lists. The first support remains training-only. An image can therefore be a query in one episode and an annotated support in another, but cannot fill both roles within one episode. The paired RSRM comparison uses these same support lists.

Masks and inference inputs Support masks encode the target class as binary foreground. Shared mask-only preprocessing minimally adjusts small or nearly full support regions to retain usable foreground and background on the feature grid. All matched methods receive these same processed supports. The query evaluation masks remain unchanged. Prediction lists contain query RGB images and annotated support RGB images. Current-query target masks and class identifiers are excluded from prediction inputs and stored separately for evaluation.

A.2.2 FIVE-SHOT COMPARISON PROTOCOL

Matched evaluation GF-SAM and INSID3-L were evaluated on all eight datasets with seeds 0, 1, and 2 (Table 3). Each query and its ordered set of five supports matched the full method's episodes. We used the official method implementations with frozen weights and adapted their episode inputs and evaluation to our common protocol. Predictions were finalized before query labels were accessed for evaluation. Masks were resized by nearest-neighbor interpolation to the 480×480 grid, with dataset-specific ignore regions handled consistently.

Aggregation and scope For each dataset and seed, foreground intersections and unions were accumulated by class before averaging class IoUs. The overall score first averaged the eight dataset scores within each seed, then reported the mean and sample standard deviation across seeds.

A.3 EXTENDED QUANTITATIVE RESULTS

The controls separate prompt and component effects from candidate scoring and terminal revision decisions. The six additional selectors and six decision variants share eight datasets, three seeds, episode lists, and metric aggregation. The comparisons are followed by analyses across support samplings, target classes, and individual episodes.

A.3.1 DIRECT COMPOSITION AND FIXED-REFERENCE COMPONENT CONTROLS

Direct composition The seed-0 comparison in Table 5(a) uses eight datasets. Direct composition returns the SAM 3 mask generated from one DINOv3 anchor. The component-overlap control replaces each 8-connected RSRM component with the SAM 3 component having the greatest intersection. Unmatched RSRM components are retained, and unselected SAM 3 components are discarded.

Fixed-reference comparisons The component contrasts in Table 5(b) use the six datasets in Table 4: SUIM, FSS-1000, ISIC, Chest X-ray, PASCAL-5^i, and COCO-20^i. They use three seeds, fixing B_0 and the full cross-action evidence bank. Auditing and final mask selection are excluded from these comparisons. Each action supplies overlapping components and their union within the response mask. The action contrast replaces the top-anchor action with CSC singleton and joint actions while keeping component selection fixed. Conversely, the selection contrast compares component selection with unconditional output while using only the top-anchor action. The combined component pipeline is compared separately with RSRM.

Evidence terms Geometry-based selection uses overlap, area change, and boundary complexity. It gained 2.91 ± 0.66 points over unconditional top-anchor output; reproduction and extent added 0.29 ± 0.08 points. Reproduction measures maximum cross-action component Dice, whereas extent measures alternative-action agreement on foreground additions and deletions. With all actions available, their individual gains over geometry were 0.03 ± 0.11 and 0.48 ± 0.24 points, respectively.

A.3.2 SELECTION FROM RAW AND PRCMS-GENERATED CANDIDATES

Shared inputs and candidate pools The controls in Table 6 share CSC prompts, DINOv3 fields, B_0, and the full SAM 3 subset-response bank. They use eight datasets over three seeds. N1–N2 select from B_0 and raw subset masks. C1–C3 select from the complete masks generated at PRCMS stages, including B_0 and the final output. Each candidate pool is deduplicated by pixel content before scoring.

Scoring and tie handling Let $\widehat{D}$ denote the fused DINOv3 foreground–background margin, normalized by its maximum absolute value; zero fields remain unchanged. For a candidate mask M, the semantic score is

$$p(M) = [1 + \mathrm{mean}_x((2M(x) - 1)\widehat{D}(x))]/2. \tag{A.3.1}$$

The consistency score $q(M)$ is mean Dice with all original subset masks, counting each prompt subset once. Two empty masks have Dice 1. Equal score fusion uses $(p + q)/2$. Exact ties prefer B_0, followed by a fixed mask identifier. Query labels are used only for evaluation.

Interpretation N1 and N2 test selection from raw responses at a shared SAM 3 response budget. C1–C3 test scoring rules with PRCMS-generated candidates also available. Within each pool, these comparisons hold candidates and responses fixed. Post-processing costs vary with candidate construction and selection. Uniform response averaging in Appendix A.5.3 instead combines probability maps to produce a new mask.

Global-overlap and calibrated-fusion controls Table 8 adds six selector configurations. For each candidate source, global-IoU selection maximizes mask IoU against either the initial prediction B_0 or the audited reference B. The explicitly inserted reference retention item is excluded from its own argmax, avoiding a trivial self-match. Candidate responses that naturally coincide with the reference remain subject to the preset mask-handling rules. Empty-mask, deduplication, tie, and fallback conventions are held fixed across the two reference choices. Thus the global-IoU candidate lists differ from N1–N2/C1–C3 in their handling of the explicit retention item. The calibrated-fusion controls use $wp(M) + (1 - w)q(M)$ in the corresponding N2/C3 pools, with w following the configuration-selection protocol in Section 4.1. The definitions of p and q remain unchanged.

Table 8. Eight-dataset mean mIoU (%) over three seeds, with CSC prompts, DINOv3 fields, and the SAM 3 response bank fixed. Columns distinguish candidate sources. Global-IoU rows omit the explicitly inserted reference retention item; score-fusion rows retain the pools defined in Appendix A.3.2. Equal fusion repeats N2/C3 from Table 6 for context.

Table 8: **Extended one-shot selection controls.**

Selection rule	Raw pool mIoU ↑	PRCMS pool mIoU ↑
Global IoU against B_0	66.50	66.97
Global IoU against audited B	67.49	67.97
Equal fusion $(p + q)/2$	66.73	67.66
Calibrated weighted fusion	66.96	67.94
Full method (CSC + PRCMS)	**69.33**	

Changing the global-IoU reference from B_0 to B raised mean mIoU from 66.50% to 67.49% in the raw pool and from 66.97% to 67.97% with PRCMS candidates. Reference auditing therefore improved selection even with a symmetric overlap criterion. The full method exceeded the strongest extended selector, global IoU against B with PRCMS candidates, by 1.36 points. Calibrated fusion reached 66.96% and 67.94%, respectively, compared with 66.73% and 67.66% for equal fusion. These shared-response comparisons support both reference auditing and PRCMS revision evidence. The two columns distinguish candidate availability from the rule used to rank candidates. Gains from the larger pool should therefore be read separately from gains obtained by changing the reference within a fixed candidate source.

A.3.3 Ablations of revision admissibility

Fixed state and interventions Table 9 tests the admissibility conditions in Appendix A.1.4 after reference auditing. The one-shot variant omits the disagreement branch in Appendix A.1.4.1. It then continues to the persistence and reference-calibration branches, with the usual fallback to B. The five-shot variants keep B, the response bank, edit atoms, and edit groups fixed. Direct output of B removes terminal composition. The remaining variants remove coalition eligibility, the deletion interaction-mass condition, both conditions, or the deletion-percentile condition described in Appendix A.1.4.2. All other acceptance conditions and fallback assessments are retained, and accepted edits are applied simultaneously.

Table 9. Eight-dataset mean mIoU (%) over three seeds. Each shot setting keeps the response bank and audited B fixed; five-shot edit atoms and groups are also fixed. Δ is the corresponding full-method mean minus the variant mean in points, computed from the displayed values.

Effect of terminal decisions Omitting the one-shot disagreement assessment reduced mean mIoU from 69.33% to 68.75%, with later branches retained. In five-shot inference, composing accepted edits improved the mean by 1.59 points over direct output of the audited reference. Removing coalition eligibility, the interaction-mass condition, or both reduced the mean by 1.10, 0.84, and 1.47 points, respectively. Removing the deletion-percentile condition produced a smaller 0.34-point decrease. These are net effects within the fixed composition procedure, with all other assessments retained. The decreases are conditional on the remaining procedure and should not be added together as independent component contributions. In particular, removing both conditions measures their joint absence rather than the sum of separate effects.

Table 9: **One-shot selection and five-shot edit-composition ablations.**

Setting	Output or decision rule	Mean mIoU ↑	Δ
One-shot	**Full method**	**69.33**	—
	Without disagreement assessment	68.75	0.58
Five-shot	**Full method**	**75.20**	—
	Output audited B	73.61	1.59
	Without coalition-eligibility condition	74.10	1.10
	Without interaction-mass condition	74.36	0.84
	Without both conditions	73.73	1.47
	Without deletion-percentile condition	74.86	0.34

A.3.4 Five-shot comparisons and host adaptation

The full method exceeded INSID3-L on seven datasets and GF-SAM on six. GF-SAM scored 67.83% on SUIM and 67.45% on COCO-20^i, versus 67.75% and 65.82% for ours; INSID3-L scored 58.77% on iSAID-5^i, versus 57.18%.

Selection within the CG-ICS host Table 10. One-shot mIoU (%) on matched host outputs and episodes. Both rows retain the host's Qwen3-VL concept reasoning and SAM 3 candidate generation. Gains are in points.

Table 10: **Selection within the CG-ICS host at seed 0.**

Method	SUIM	FSS	ISIC	CXR	PASCAL	COCO	VOC	iSAID	Mean
CG-ICS replay	64.74	90.60	68.54	88.21	88.60	71.54	88.20	52.33	76.60
With host-adapted selection	**68.81**	**90.66**	**70.04**	**88.36**	**88.81**	**72.74**	**88.54**	**53.45**	**77.68**
Paired gain	+4.07	+0.06	+1.50	+0.14	+0.21	+1.20	+0.34	+1.13	**+1.08**

CG-ICS uses Qwen3-VL to generate textual concepts and refines them through SAM 3-scored tree search (Chen et al., 2026). Table 10 evaluates a host-specific selection adaptation at seed 0. Both rows retain this language-guided candidate source and share the frozen host outputs and episode lists. Adapted selection improved all eight datasets, increasing mean mIoU from 76.60% to 77.68%. The comparison measures the selection gain within CG-ICS. The visual-support pipeline is evaluated separately in Table 1. Gains vary across datasets within this single-seed, language-guided host setting.

A.3.5 Evaluation coverage and class-resolved robustness

We examine two levels that dataset averages alone do not resolve: consistency across support-sampling seeds and the distribution of gains across target classes. These analyses reuse the final evaluations, without additional model fitting or inference. Table 11 separates the full benchmark coverage from the narrower scope of the available episode-level analysis. A paired comparison in Table 11 uses ours and RSRM under the same dataset, shot setting, and seed. Counts refer to evaluation conditions or episode records, not unique images or independent training runs.

Table 11: Coverage of the quantitative analyses.

Evaluation axis	Verified coverage
Target datasets in the final benchmark	8
Support-set sizes	1 and 5
Support-sampling seeds per setting	3
Paired dataset–shot–seed comparisons	$8 \times 2 \times 3 = 48$
Dataset-level scores in these paired comparisons	$48 \times 2 = 96$
One-shot datasets with class-resolved episode analysis	6
Paired episode records in the class-resolved analysis	39,255
Dataset-specific class entries in this analysis	351

Consistency across support samplings Figure 5 reports all 48 paired gains across eight datasets and three seeds. Improvements occur in 24/24 one-shot and 22/24 five-shot conditions; the five-shot

SUIM exceptions show that this consistency is not universal. Each cell gives ours minus RSRM in mIoU percentage points, not an episode win rate. Panels (a) and (b) show one-shot and five-shot results on a shared scale; teal and rust denote positive and negative gains, respectively.

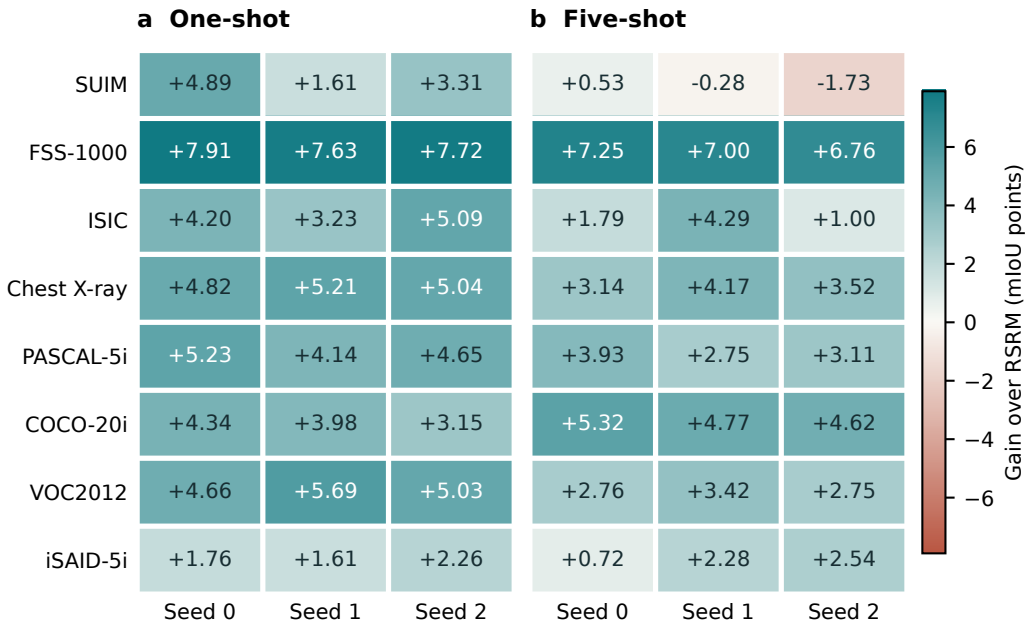

Figure 5: **Cross-domain gains across support-sampling seeds.**

Distribution across target classes We next reconstruct class IoU from the 39,255 paired episode records available for six one-shot datasets. For each dataset, seed, and class, intersection and union counts are summed over episodes before division. We then compute the paired class-IoU difference and average it over the three seeds. This follows the class-first benchmark aggregation; it does not average episode IoUs. Classes retain their dataset-specific identities and are not treated as a shared taxonomy.

Table 12 summarizes 351 class entries, of which 309 have a positive three-seed mean gain. Episode counts sum over three seeds. "Mean > 0" counts positive seed-averaged gains, whereas "All seeds > 0" requires a positive gain in every seed. The median summarizes class-level mean gains in IoU percentage points, not dataset mIoU.

Table 12: Class-resolved one-shot gains over RSRM.

Dataset	Episodes	Classes	Mean > 0	All seeds > 0	Median
SUIM	855	7	5/7	2/7	+2.28
FSS-1000	7,200	240	229/240	207/240	+7.35
ISIC	3,600	3	3/3	3/3	+4.71
Chest X-ray	3,600	1	1/1	1/1	+5.02
PASCAL-5i	12,000	20	15/20	15/20	+4.91
COCO-20i	12,000	80	56/80	42/80	+3.24

Figure 6 shows all 351 classes, each with one marker for its three-seed mean gain and a line spanning the seed-specific minimum and maximum, not a confidence interval. Classes are ordered by mean gain within each panel, with panel-specific axes. Teal and rust denote positive and non-positive means. Positive dataset averages therefore need not imply improvement for every class.

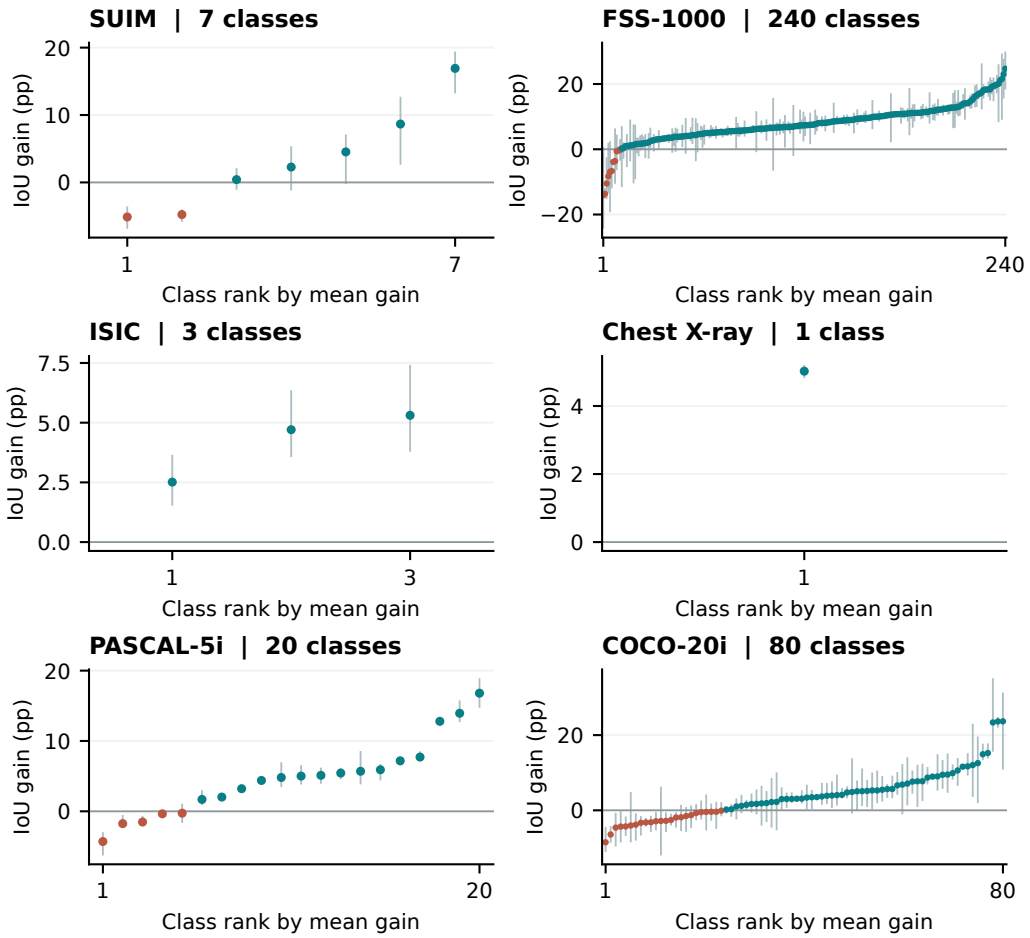

Figure 6: **Class-level gain profiles.**

A.3.6 Scope and distribution of improvements

We next examine how gains from the frozen one-shot pipeline vary across episodes. Aligned intermediate records are available for six datasets and three seeds, covering 39,255 episodes. The qualitative overview in Appendix A.6.1 uses one episode per dataset. These diagnostics reuse predictions from the fixed experimental configuration. Table 13 gives the population behind the six-domain summaries. Counts follow the episode protocols, while the benchmark mean weights datasets equally.

Table 13 reports episode records pooled across all three seeds.

Table 13: Episode populations used for the mechanism diagnostics.

	SUIM	FSS	ISIC	CXR	PASCAL	COCO
Per seed	285	2,400	1,200	1,200	4,000	4,000
Total	855	7,200	3,600	3,600	12,000	12,000

Two complementary estimands The class-macro benchmark sums foreground intersections and unions within each class before averaging class IoUs. An episode-level gain instead compares the two IoUs on a single query. Figure 7 separates these quantities: positive class-macro gains coexist with both beneficial and harmful episodes. Its intervals describe seed variation or episode quantiles, as specified, rather than confidence intervals. Queries may recur across seeds, so we do not treat the pooled episodes as independent samples for significance testing.

In Figure 7(a), diamonds and whiskers show mean class-macro mIoU gains over matched RSRM and sample standard deviations across three seeds. Open circles show individual seeds. Panel (b) uses all episode gains, with a Gaussian kernel and Scott bandwidth, evaluated over the observed range and scaled to equal peak height. Below each density, an open circle marks the median, the thick segment the interquartile range, and the thin segment the 10th–90th percentiles. Density heights do not represent sample sizes.

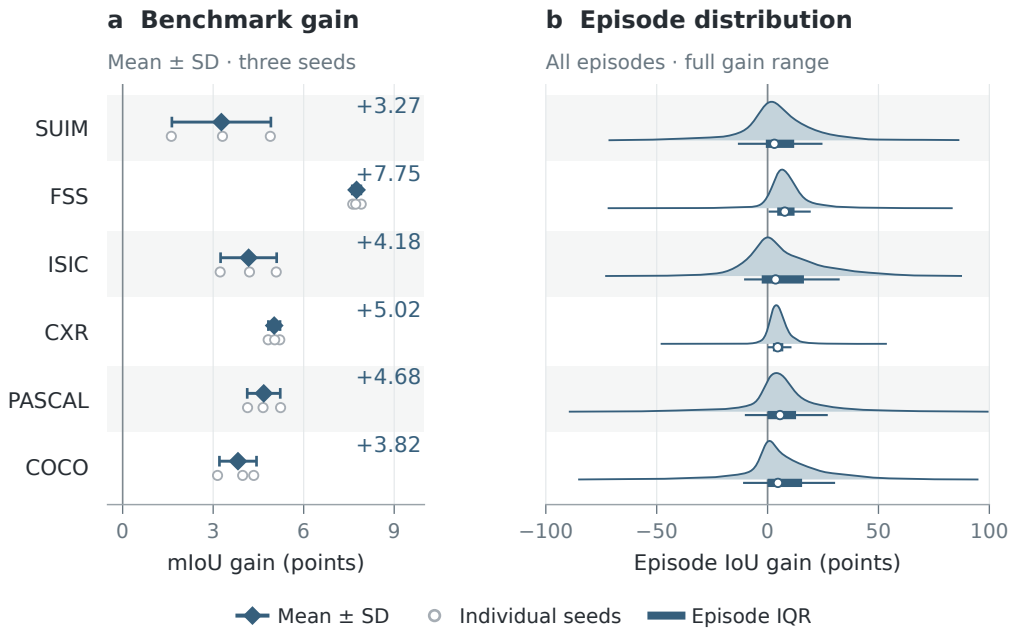

Figure 7: **Magnitude and heterogeneity of refinement gains.**

A.3.7 Per-dataset episode distributions

Episode frequencies throughout this section are distinct from class-macro benchmark scores.

Shared panel conventions All six domain figures use the same four-panel layout. Panel (a) shows all paired episode IoUs, with logarithmic color normalization of occupied-bin counts; the diagonal marks equality. Panel (b) shows the empirical cumulative distribution of gains: each ordinate is the fraction with gain at most the horizontal coordinate. Panel (c) reports the median and interquartile range within baseline-IoU bins, with labels giving episode counts. Panel (d) separates the gain distribution by seed. No bins or tails are trimmed. These strata describe the association between initial accuracy and revision gain.

A.3.7.1 SUIM The logged population contains 855 episodes across three seeds. Relative to RSRM, 63.5% improved in foreground IoU and 31.1% degraded; the remaining episodes tied. Figure 8 shows that improvements and degradations coexist over a broad range of initial accuracies. The positive median gains in the middle baseline-IoU bins contrast with the substantial negative tail. The aggregate improvement therefore does not imply uniformly reliable revision of underwater scenes.

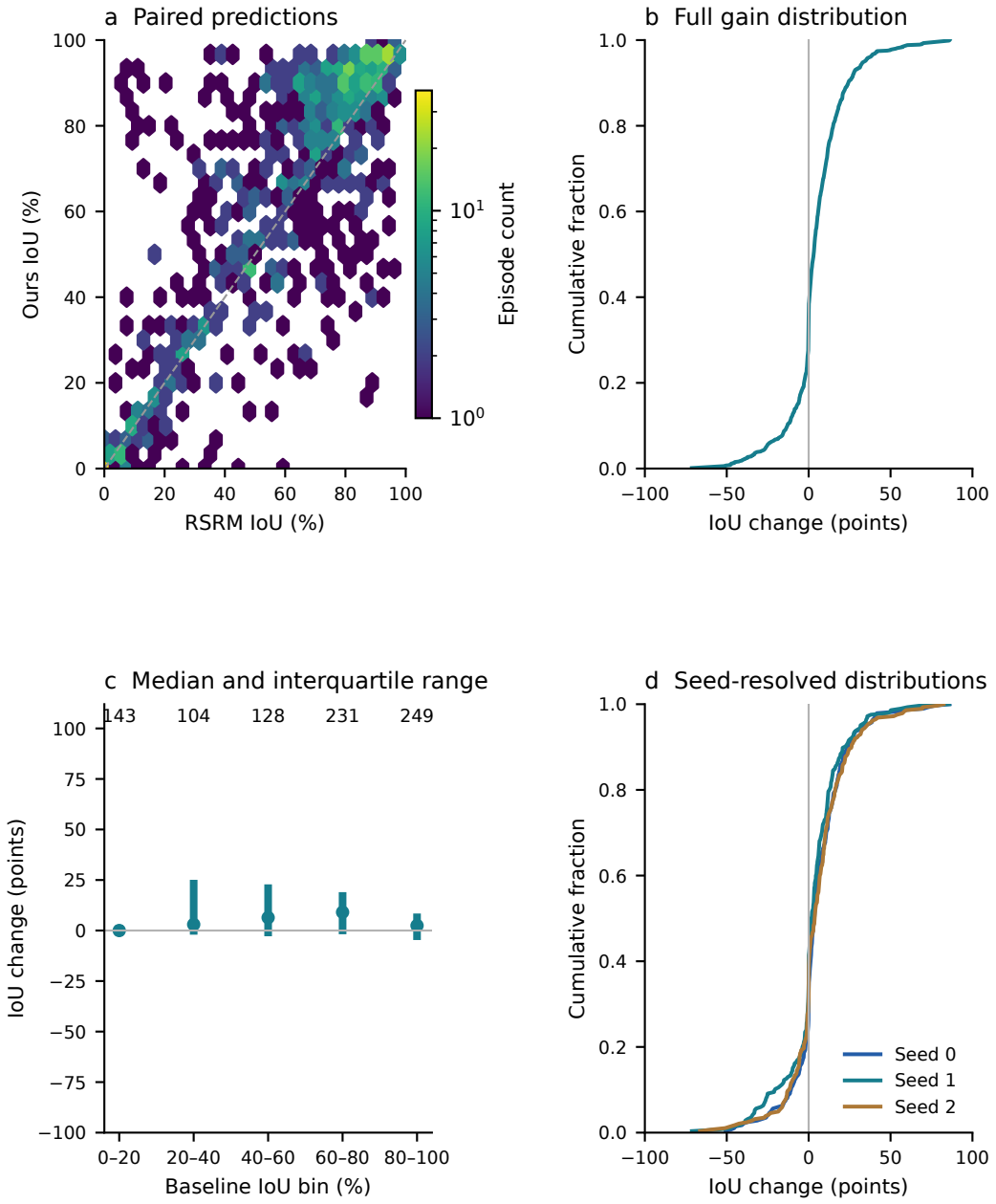

Figure 8: **Episode-level gain distribution on SUIM.**

A.3.7.2 FSS-1000 The logged population contains 7,200 episodes across three seeds. Relative to RSRM, 91.9% improved in foreground IoU and 7.8% degraded; the remaining episodes tied. The distribution in Figure 9 indicates that gains are widespread rather than confined to a small set of large improvements. The intermediate baseline-IoU bins show larger median gains than the highest bin, which contains most episodes and has less available headroom. The three seed-resolved curves are closely aligned, although individual degradations remain visible in the paired plot and negative tail.

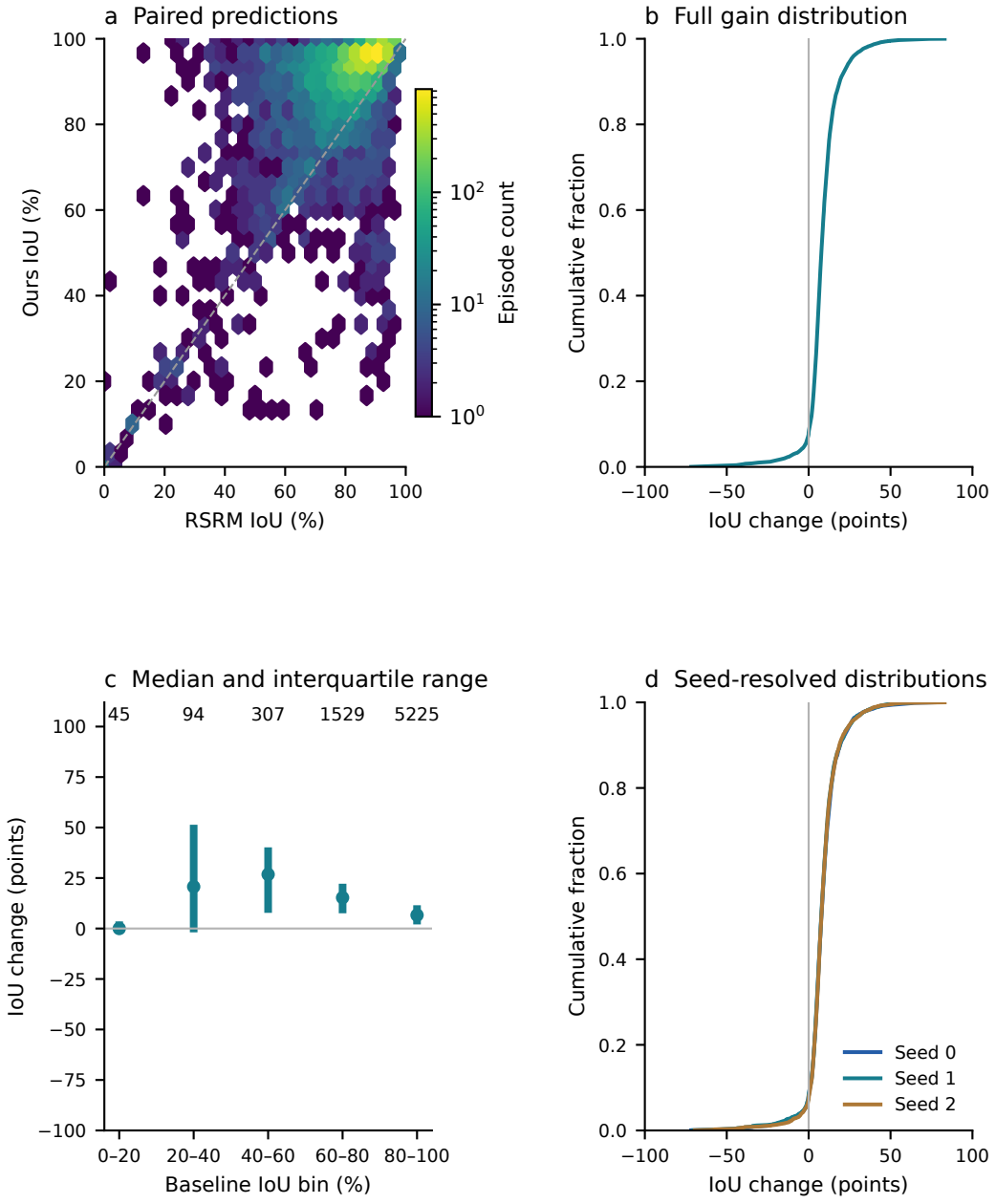

Figure 9: **Episode-level gain distribution on FSS-1000.**

A.3.7.3 ISIC The logged population contains 3,600 episodes across three seeds. Relative to RSRM, 62.3% improved in foreground IoU and 34.8% degraded; the remaining episodes tied. Figure 10 shows a more mixed response than FSS-1000 or Chest X-ray. Middle-range initial masks have positive median gains, whereas the highest-IoU bin is centered near zero and includes losses. This distinction separates improving incomplete predictions from preserving already accurate lesion masks.

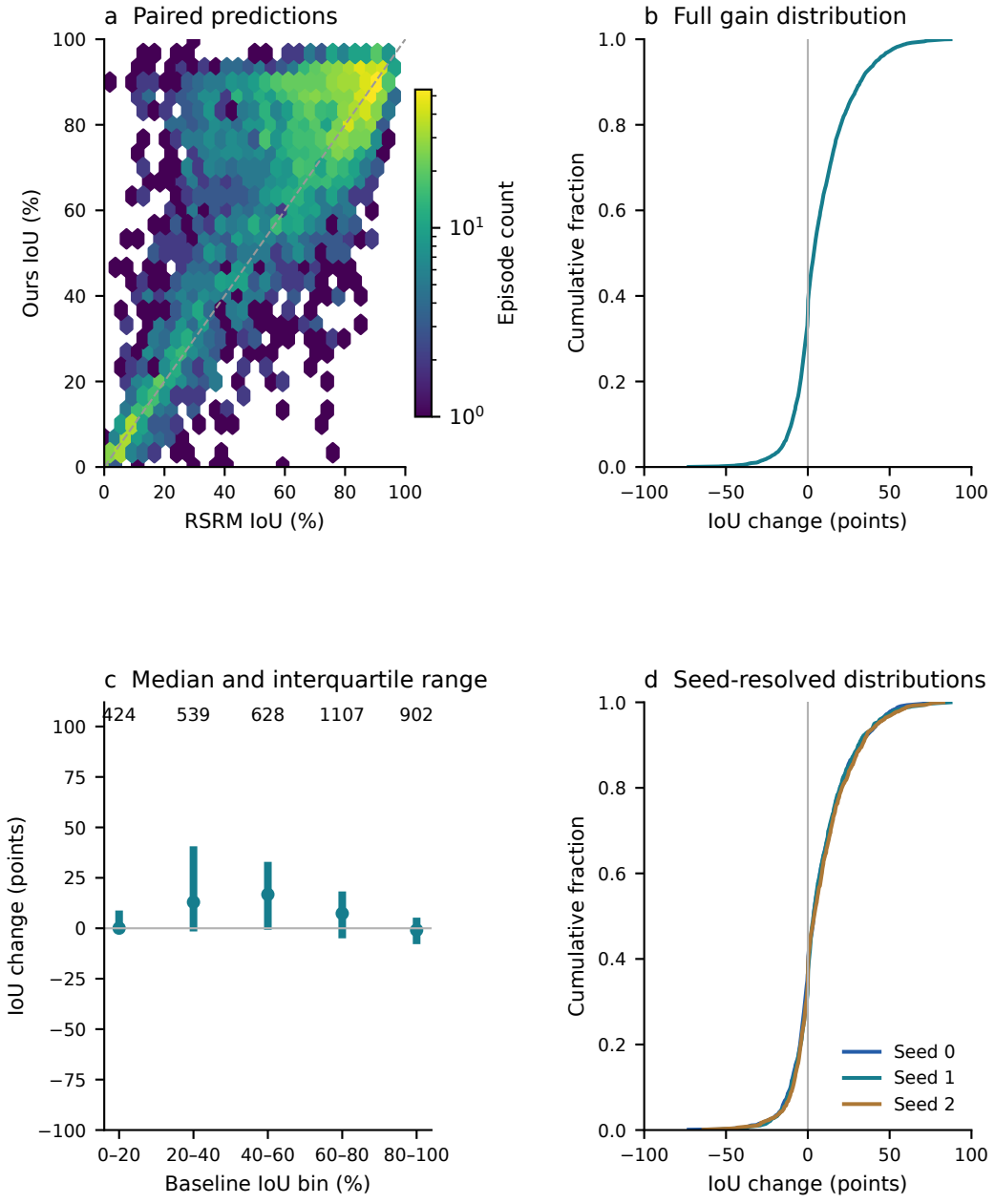

Figure 10: **Episode-level gain distribution on ISIC.**

A.3.7.4 Chest X-ray The logged population contains 3,600 episodes across three seeds. Relative to RSRM, 92.2% improved in foreground IoU and 7.7% degraded; the remaining episodes tied. Most observations in Figure 11 lie in the highest baseline-IoU bin, where the median gain remains positive. Thus, improvements also occur among initially accurate lung predictions. The two lowest bins each contain only one episode and cannot support general conclusions about severely inaccurate initial masks. The closely aligned seed curves summarize the much larger high-accuracy population.

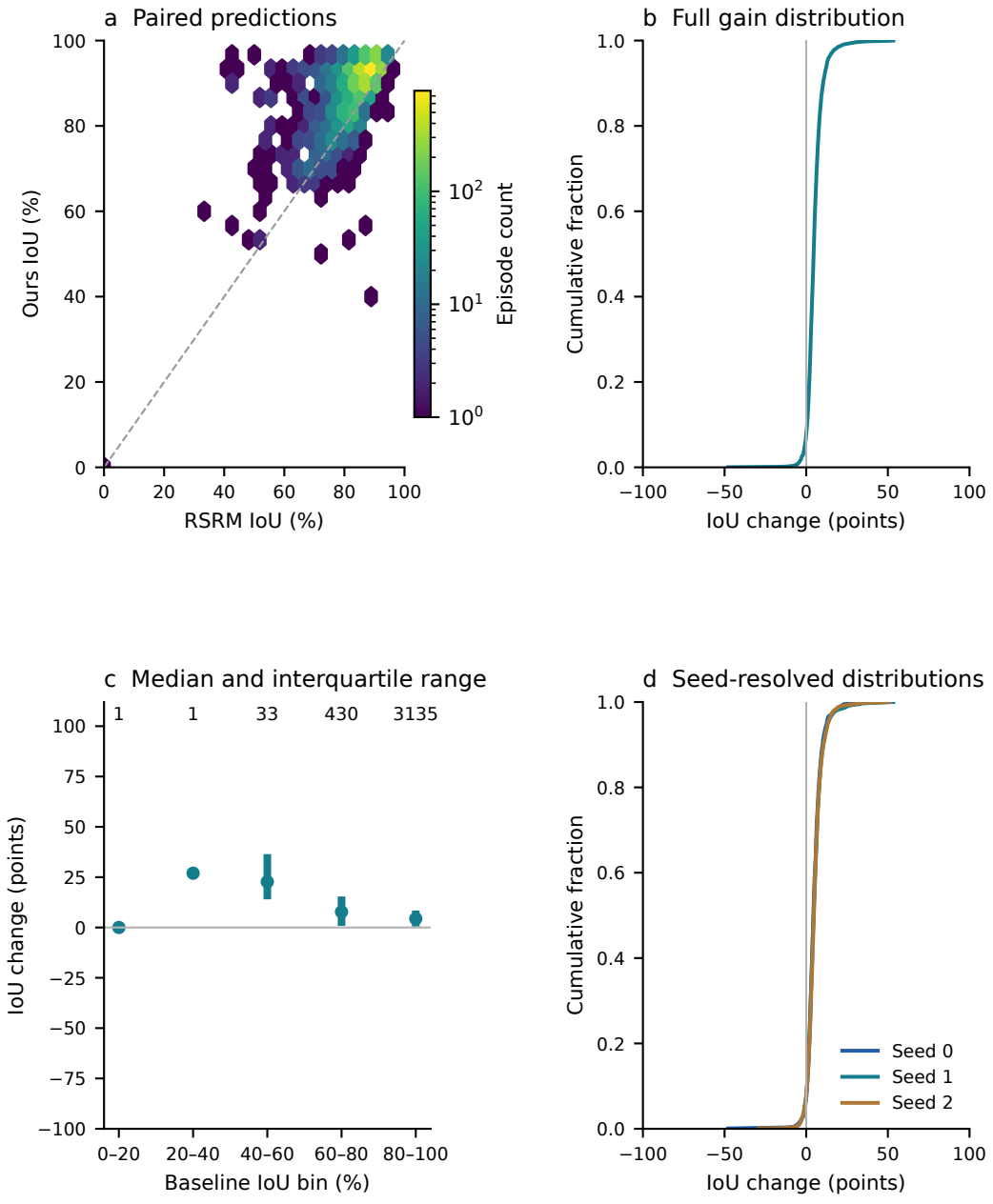

Figure 11: **Episode-level gain distribution on Chest X-ray.**

A.3.7.5 PASCAL-5i The logged population contains 12,000 episodes across three seeds. Relative to RSRM, 74.6% improved in foreground IoU and 21.2% degraded; the remaining episodes tied. In Figure 12, positive median gains extend across the intermediate and high baseline-IoU bins. Their magnitude is smaller in the highest bin, consistent with its limited room for improvement. The full paired distribution additionally retains large positive and negative changes that the binned medians alone would conceal.

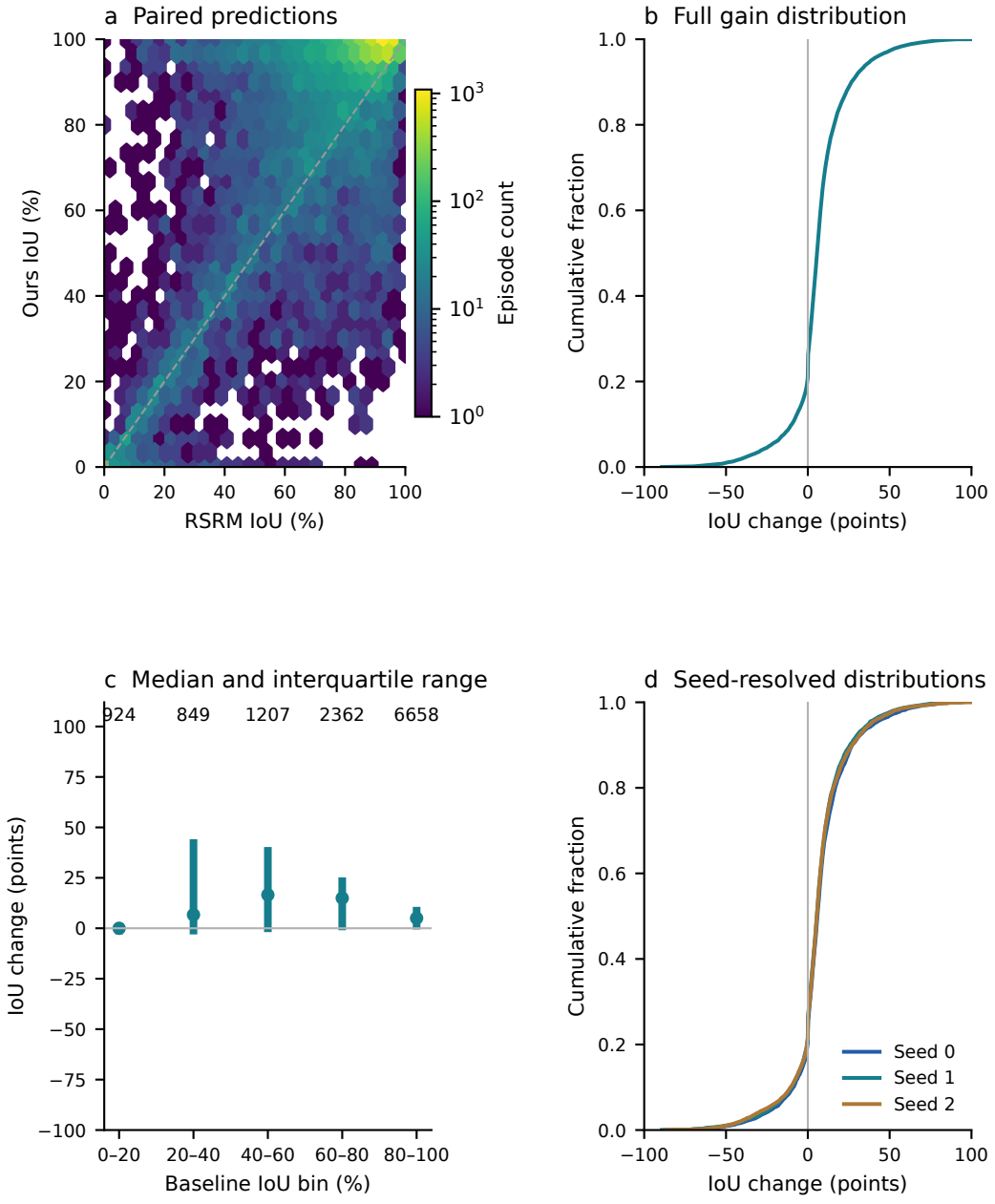

Figure 12: **Episode-level gain distribution on PASCAL-5i.**

A.3.7.6 COCO-20i The logged population contains 12,000 episodes across three seeds. Relative to RSRM, 66.0% improved in foreground IoU and 24.7% degraded; the remaining episodes tied. Figure 13 combines a majority of improved episodes with ties and a substantial degradation tail. The middle baseline-IoU bins show positive median gains, while the lowest bin remains concentrated near zero. The similar seed curves indicate comparable aggregate gain distributions across the three episode lists, not identical outcomes for individual queries.

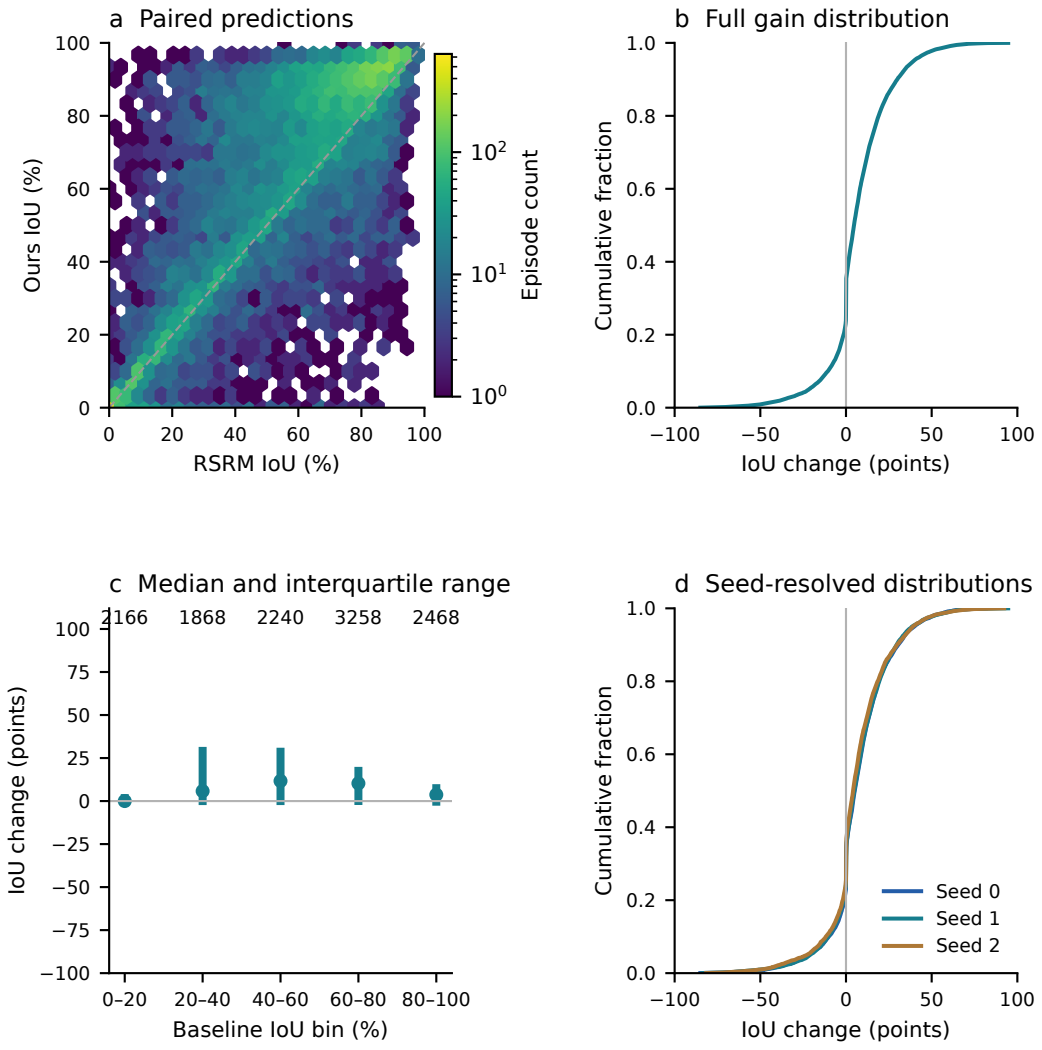

Figure 13: **Episode-level gain distribution on COCO-20i.**

A.4 Mechanism Analysis

The following analyses examine CSC coverage, reference construction, and reference-relative changes using the recorded responses and predictions. These descriptive diagnostics characterize recorded behavior rather than isolate the causal contribution of individual evidence channels.

A.4.1 Semantic coverage and spatial dispersion in CSC

We compare the logged CSC prompts with reliability TOPK and spatial NMS using the same candidate pool and per-episode prompt budget. Across the six domains, the median CSC-to-TOPK coverage-objective ratio ranges from 1.10 to 1.21. NMS often produces larger nearest-prompt distances, whereas CSC achieves higher median demand coverage. Thus, geometric separation and coverage of the weighted semantic demand are distinct selection criteria. The objective is the quantity CSC

optimizes; this diagnostic characterizes its behavior and is not an independent segmentation-quality test.

Figure 14(a) reports within-episode coverage-objective ratios relative to TOPK. Panel (b) shows the minimum Euclidean distance between anchors on the 30×30 token grid. Points and lines denote medians and interquartile ranges. Panel (c) stacks the percentages with zero to five eligible CSC prompts, from light to dark. Small segments remain visible without numeric labels.

Selection and eligibility have different roles Eligibility varies substantially across domains even under the same prompt budget. In particular, 77.7% of Chest X-ray episodes have four or five eligible prompts, whereas 22.1% of SUIM episodes have none. These are return-consistency outcomes, not estimates of prompt correctness. All selected prompts still participate in the response bank; eligibility determines which attributions enter the eligible evidence sum. Figure 15 illustrates the distinction on a separate FSS episode selected for clear prompt-placement differences.

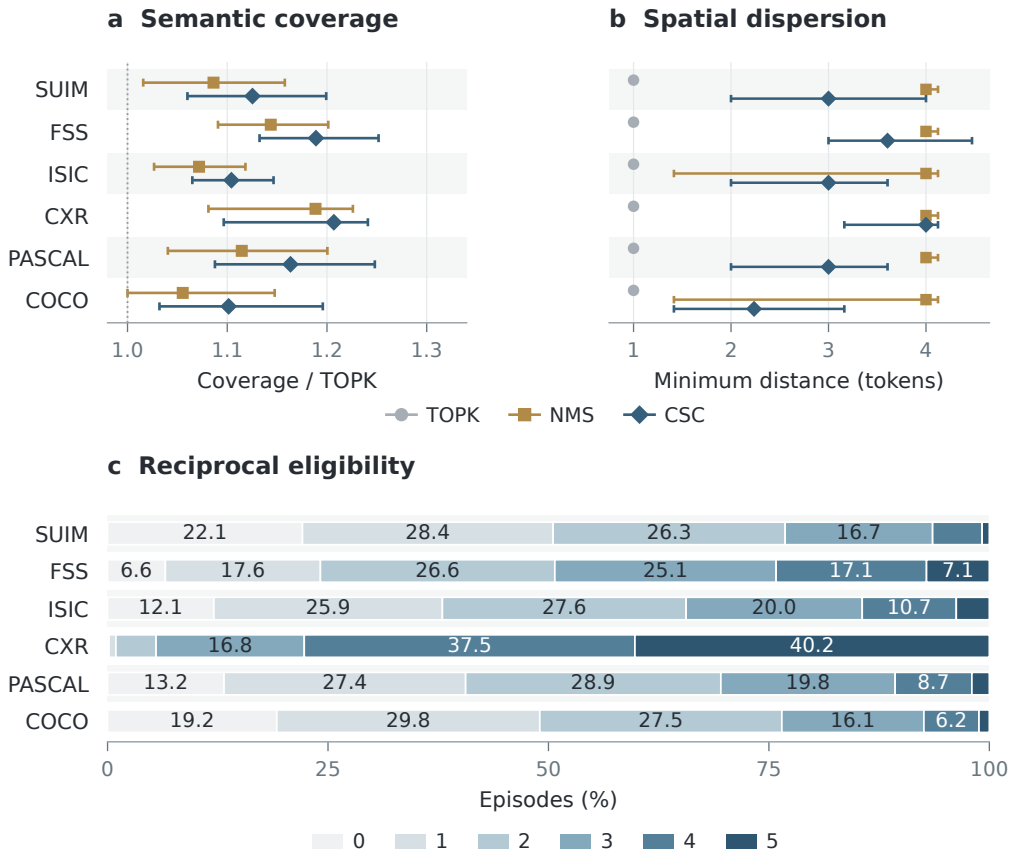

Figure 14: **Prompt geometry, coverage, and evidence eligibility.**

A.4.2 Reference construction and terminal revision are distinct

The sequence $B_0 \rightarrow G_0 \rightarrow B \rightarrow O$ distinguishes structural proposal, reference construction, and final revision. The following examples illustrate these roles rather than isolate their causal contributions. First, Figure 15 shows how different selection policies place their prompts on the same query. The NMS and CSC anchors need not coincide because their selection criteria differ. Gold and white cells additionally distinguish prompt placement from eligibility to supply attribution evidence.

Why preserving a prediction can be a successful outcome The ISIC trace in Figure 16 illustrates an audit that restores the initial assignment after the proposal loses foreground accuracy. Its final IoU equals the initial IoU, so an end-to-end gain plot alone would conceal the intermediate correction.

The FSS trace illustrates a different behavior: the audit retains the proposal, and terminal selection then improves it. These two episodes were selected retrospectively to expose distinct state transitions; they do not estimate their prevalence. Across domains, auditing and subsequent selection do not each yield uniformly positive mean increments, and an accepted revision can still reduce query IoU.

In Figure 15, this FSS eagle example uses five prompts under each policy. CSC distributes prompts across both wings and the body, whereas three TOPK prompts occupy adjacent cells on one wing. Each square occupies one token cell; gold denotes reciprocal eligibility and white denotes ineligibility. All five CSC prompts are ineligible in this example: spatial coverage and attribution eligibility are distinct. Guide lines are spaced every three tokens. Locations and eligibility flags are recorded values.



Figure 15: **Different prompt sets on the same query.**

Numbers in Figure 16 are episode foreground IoUs in percent. The same FSS and ISIC episodes appear in the response-bank diagnostics and eight-domain overview, linking the evidence fields to these trajectories. The eagle in Figure 15 is a separate episode.

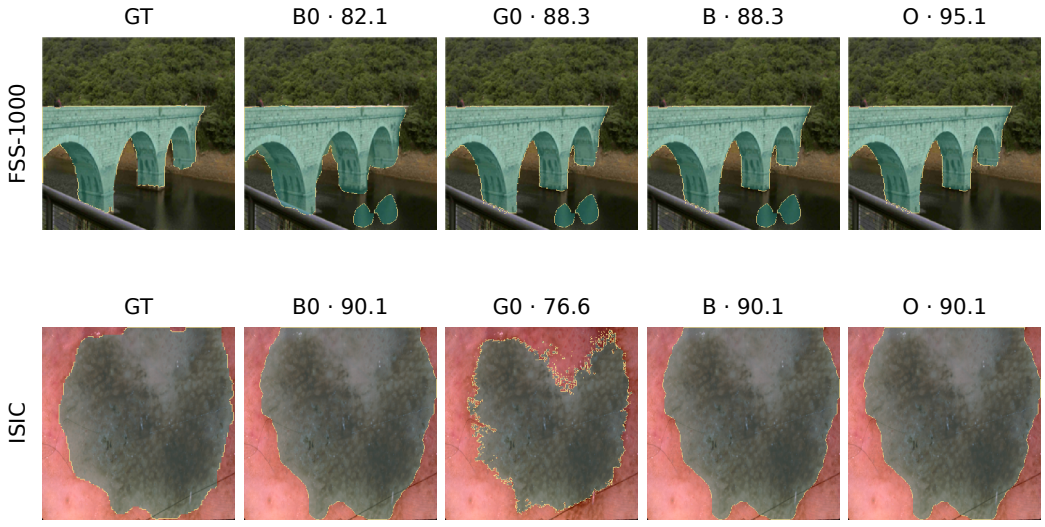

Figure 16: **Two different uses of the reference state.**

A.4.3 Recovering foreground versus removing false positives

Foreground IoU can improve through different error changes. Let I and U be a reference mask's intersection and union with a fixed target, and let $\Delta I = I_O - I$ and $\Delta U = U_O - U$ denote the output

differences. Since $I = \mathrm{TP}$ and $U = |Y| + \mathrm{FP}$, we obtain

$$\Delta\mathrm{TP} = \Delta I, \qquad -\Delta\mathrm{FP} = -\Delta U. \tag{A.4.1}$$

We sum each quantity within a dataset and seed and divide by the summed reference union, reporting pixels per 100 reference-union pixels. This separates net true foreground recovery from net false-positive removal without introducing another segmentation metric or changing the evaluation masks. Read Figure 17 by quadrant: the upper-right combines both forms of correction, while opposite-sign coordinates indicate a trade-off. The terminal panel isolates changes after reference construction; compare its smaller changes with the end-to-end panel using the axes, not visual distance.

Figure 17 compares full refinement relative to B_0 in panel (a) with terminal selection relative to B in panel (b). Each labeled point represents one dataset; horizontal and vertical whiskers show sample standard deviations across three seeds around the mean. Moving right indicates net false-positive removal; moving up indicates net true-foreground recovery. Negative coordinates indicate the reverse. Both axes use pixels per 100 reference-union pixels. The panels use different references and axis scales. These are net changes, not the precision of individually added or deleted pixels.

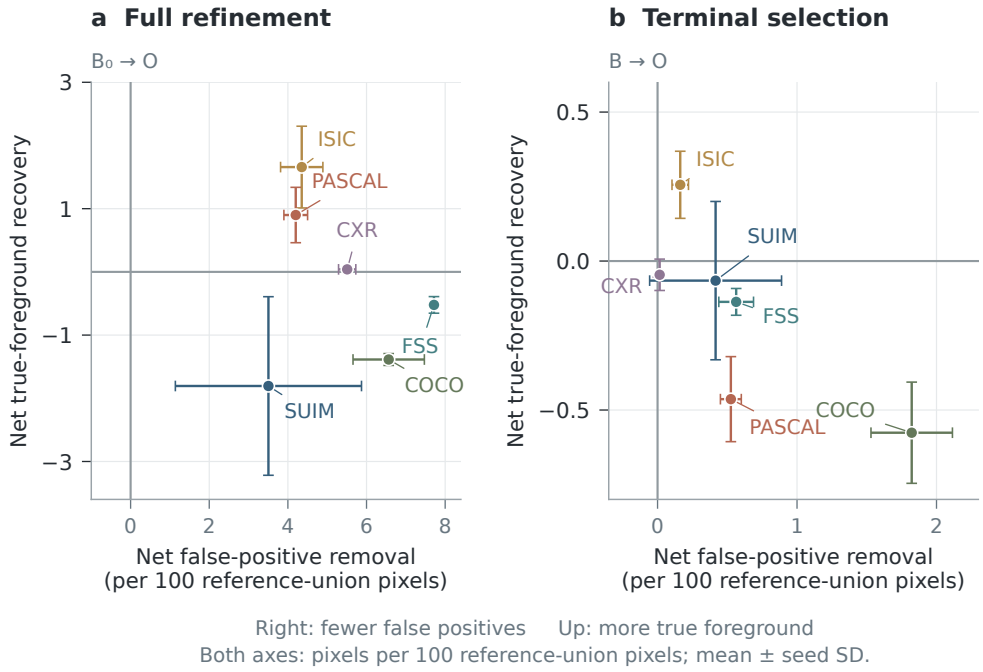

Figure 17: **The balance of foreground recovery and false-positive removal.**

An exact state-conditioned relation For positive reference and output unions, the IoU change on the unit scale is

$$\mathrm{IoU}(O) - \mathrm{IoU}(B) = \frac{U\Delta I - I\Delta U}{U(U + \Delta U)}. \tag{A.4.2}$$

Thus, the change is positive exactly when $\Delta I > \mathrm{IoU}(B)\Delta U$. This identity makes the role of the reference explicit: net foreground recovery and false-positive change are evaluated relative to its existing intersection and union. It also explains why neither larger foreground area nor fewer foreground pixels alone establishes a beneficial revision. This is a retrospective accounting identity, not an oracle used at inference; the method must estimate revision value without query labels.

A.4.4 What the response bank and attribution fields reveal

We inspect the recorded response banks of the FSS and ISIC traces. Each bank contains all 32 subsets of five selected prompts under the fixed query state. Figure 18 separates support-derived

correspondence from intervention-derived evidence: the mean DINO margin, response variability across subsets, the sum of eligible Shapley contributions, and the sum of the remaining contributions. The correspondence map describes support–query evidence, whereas the variability map identifies locations whose responses depend on the prompt subset. The signed attribution maps allocate response changes to the two prompt groups.

In Figure 18, columns show the mean cached DINO semantic margin, pixelwise standard deviation of the 32 soft responses, $\sum_{i\in\mathcal{E}} \phi_i$, and $\sum_{i\notin\mathcal{E}} \phi_i$. Signed fields use red for positive and blue for negative values. Each column uses a common scale across both episodes, with separate scales across columns and no per-image normalization. Response variability measures prompt-subset sensitivity rather than calibrated predictive uncertainty. All fields are computed without query ground truth.

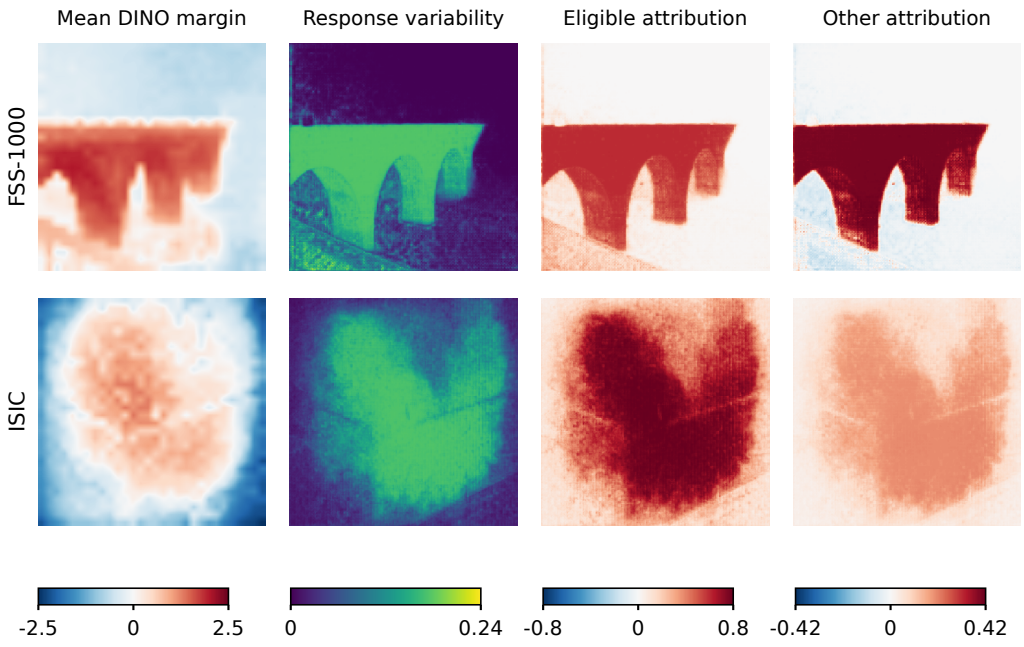

Figure 18: **Distinct evidence fields on two shared episodes.**

Response generation and attribution eligibility The eligible and remaining contributions together reconstruct $h_N - h_\varnothing$, but only the eligible sum enters the eligible-attribution channel. All selected prompts participate in response generation, while eligibility determines their inclusion in the audited attribution sum. In one-shot inference, these attributions affect terminal selection through the construction of B, not as an additional direct term in the final ranking score.

Numerical reconstruction check For the full logged population, we inspect the recorded residual

$$\epsilon_e = \max_x \left| \sum_{i\in N} \phi_i(x) - \big(h_N(x) - h_\varnothing(x)\big) \right|. \tag{A.4.3}$$

The maximum over all 39,255 episodes is 2.38×10^{-7}. This residual measures numerical consistency of the attribution allocation.

Prompt composition and response variation Figure 19 displays every subset response for the same two episodes, grouped by prompt count. Subsets of equal size can produce different IoUs, and adding a prompt need not improve IoU. The aligned membership matrices identify the prompts in each subset, separating composition from cardinality. Query ground truth is used for retrospective evaluation of these cached responses, not for subset selection.

In Figure 19, each column is one of the 32 cached binary subset responses, ordered first by cardinality and then by its binary subset identifier. Points and stems show IoU; the aligned five-row matrix

identifies the included prompts. Group labels above the chart give the prompt count; the zero-prompt response remains included. IoU is evaluated on the unchanged valid region. The two episodes are the same ones used in Figures 16 and 18.

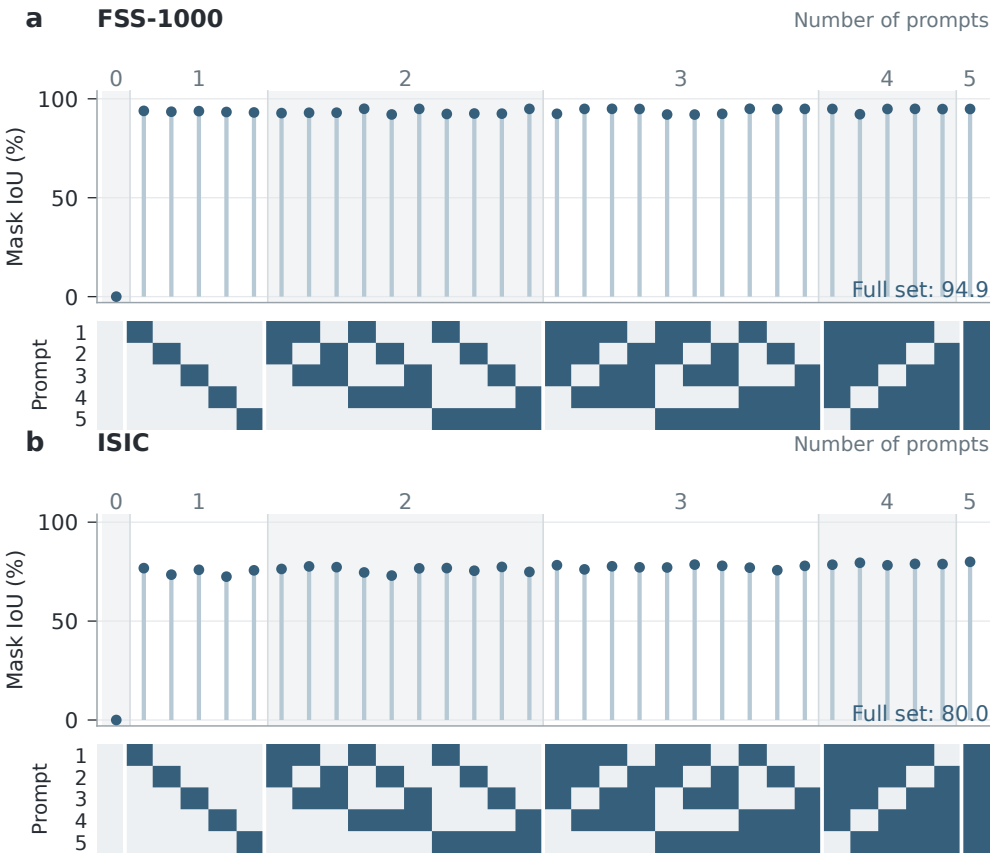

Figure 19: **Prompt-subset composition versus prompt count.**

An explicit boundary of reference auditing The groupwise audit chooses between B_0 and G_0 assignments on their disagreement regions while preserving their common assignments. Consequently,

$$B_0 \cap G_0 \subseteq B \subseteq B_0 \cup G_0. \tag{A.4.4}$$

Auditing cannot recover target pixels absent from both inputs. Recovering these shared omissions requires subsequent revision with additional candidate support. This is a structural boundary of reference auditing, whose occurrence frequency is not measured here.

A.4.5 Reference-relative revisions in selected examples

The following cached one-shot examples connect three questions: which regions change, why the reference matters, and how prompt context affects the responses. They were selected retrospectively to illustrate these behaviors rather than estimate their frequency. Query annotations are used only for this analysis, not for the recorded predictions.

A.4.5.1 Recovering the intended target extent Figure 20 separates locating the target from delimiting its extent. The support mask specifies a display, whereas the initial query prediction also includes the unit beneath it. The preserve state still contains much of this non-target region. The final revision removes 16,416 background pixels while losing 216 target pixels and recovering 15 target pixels on the evaluation grid; its IoU rises from 72.88% to 99.09%. The edit map identifies where

the mask changes; the error maps show which changes correct the target assignment. The benefit is predominantly removal of an attached non-target region rather than foreground expansion. Here $B = G_0$, so this example illustrates a successful final revision, not an improvement from reference auditing alone.

For the PASCAL-5^i example in Figure 20, panels (a–c) show the support, query, and query annotation. Panels (d–f) show the recorded initial, preserve, and final masks. Panels (g,i) distinguish correct foreground, false positives, and missed foreground; panel (h) shows additions and removals relative to B. Query panels share the same spatial extent and preserve the native image aspect ratio.

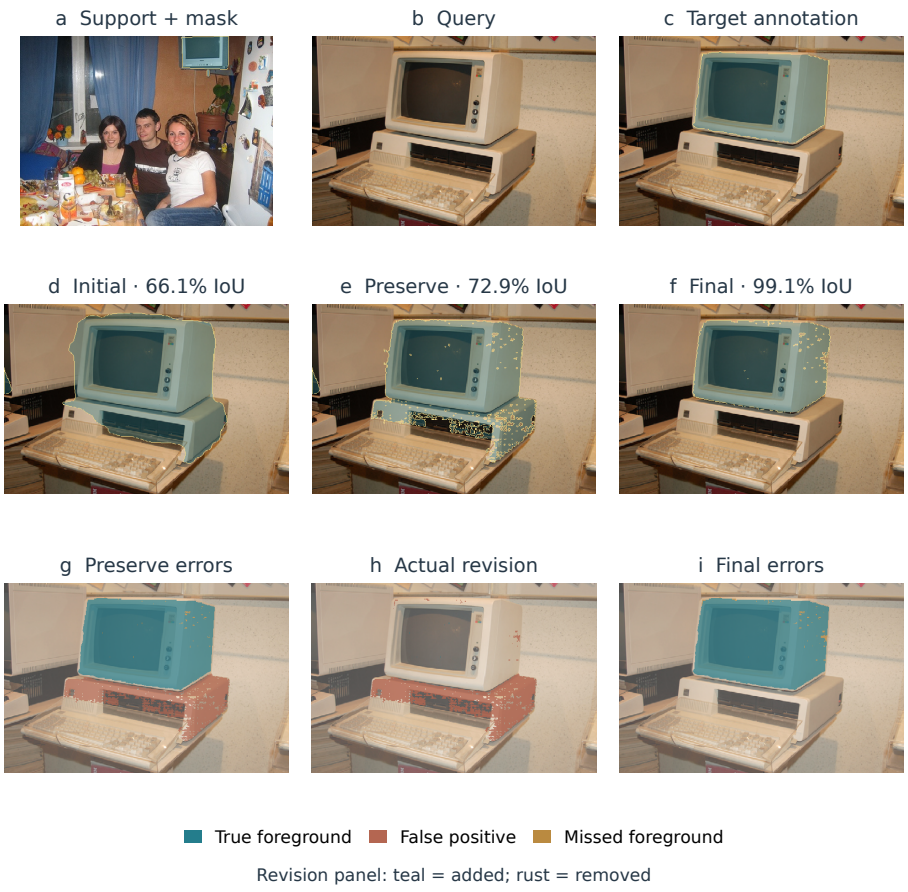

Figure 20: **Target extent in a cluttered scene.**

A.4.5.2 A fixed mask can imply opposite revision values The value of a candidate depends on what it replaces. In Figure 21, we fix one cached SAM 3 mask C and compare it with two recorded states from the same SUIM episode. Its IoU is 74.03% in both comparisons. Replacing B_0 with C would improve IoU by 9.96 points, whereas replacing B with the same C would reduce IoU by 14.47 points. The edit maps expose the difference: the replacement removes 1,859 target pixels relative to B_0, but 4,915 relative to B. A candidate-level score alone does not describe this change in the consequences of replacement.

This retrospective comparison uses the recorded response to prompts 2 and 3 and unmodified cached reference states. Both replacements are hypothetical, rather than executed selector decisions.

Figure 21(a–c) shows the initial state B_0, fixed candidate C, and preserve state B, with their individual IoUs. Panels (d,e) show additions and removals for the two hypothetical replacements. Panel (f) reports $\mathrm{IoU}(C) - \mathrm{IoU}(R)$ for $R \in \{B_0, B\}$. All panels use the common 480×480 inference grid. Only the reference changes; the candidate is identical.

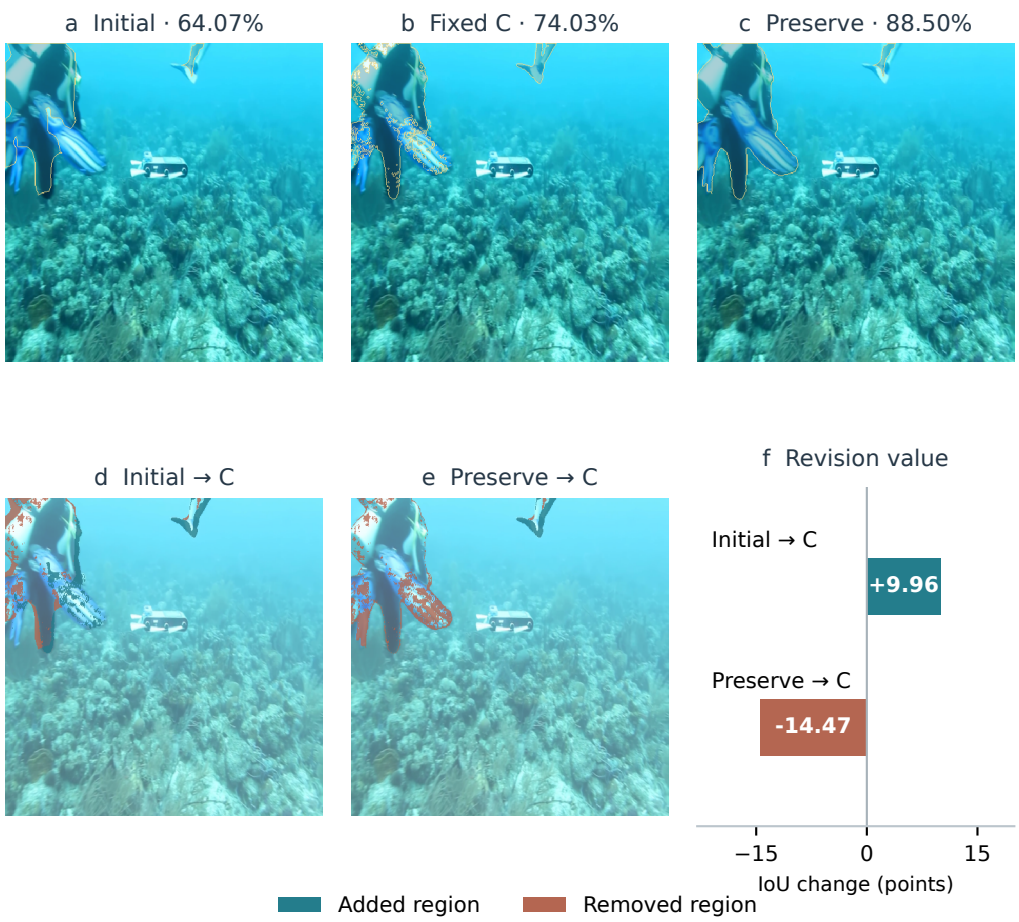

Figure 21: **Same candidate, different revision value.**

A.4.5.3 Prompt contributions depend on the other prompts Figure 22 revisits the display example to inspect response formation rather than the final mask. It compares the exact Shapley contribution ϕ_i with the leave-one-out difference $h_N - h_{N\setminus\{i\}}$ for every selected prompt. The former averages marginal effects over prompt contexts; the latter examines only the context containing all remaining prompts. Their spatial patterns differ, particularly for prompt 4: its Shapley map covers more of the display, while its leave-one-out effect is concentrated on the right side. This example makes the context dependence of prompt contributions visible. The common scale is essential: comparing independently normalized maps would obscure differences in effect magnitude as well as spatial extent.

All five prompts participate in the response bank, but only prompts 2 and 5 are eligible for the audited attribution sum. An ineligible prompt can therefore affect response generation without supplying eligible attribution. These maps describe response changes, not calibrated semantic confidence or direct terms in the final one-shot ranking; attribution influences that selection through the construction of B.

Figure 22(a) numbers CSC tokens in bank order. Panel (b) shows the all-prompts response, not the final output in Figure 20. Panel (c) shows exact Shapley contributions from all 32 cached soft responses; panel (d) shows full-coalition leave-one-out differences. All ten maps share one symmetric linear color scale on the 480×480 grid; teal and rust indicate positive and negative changes. No per-map rescaling is applied.

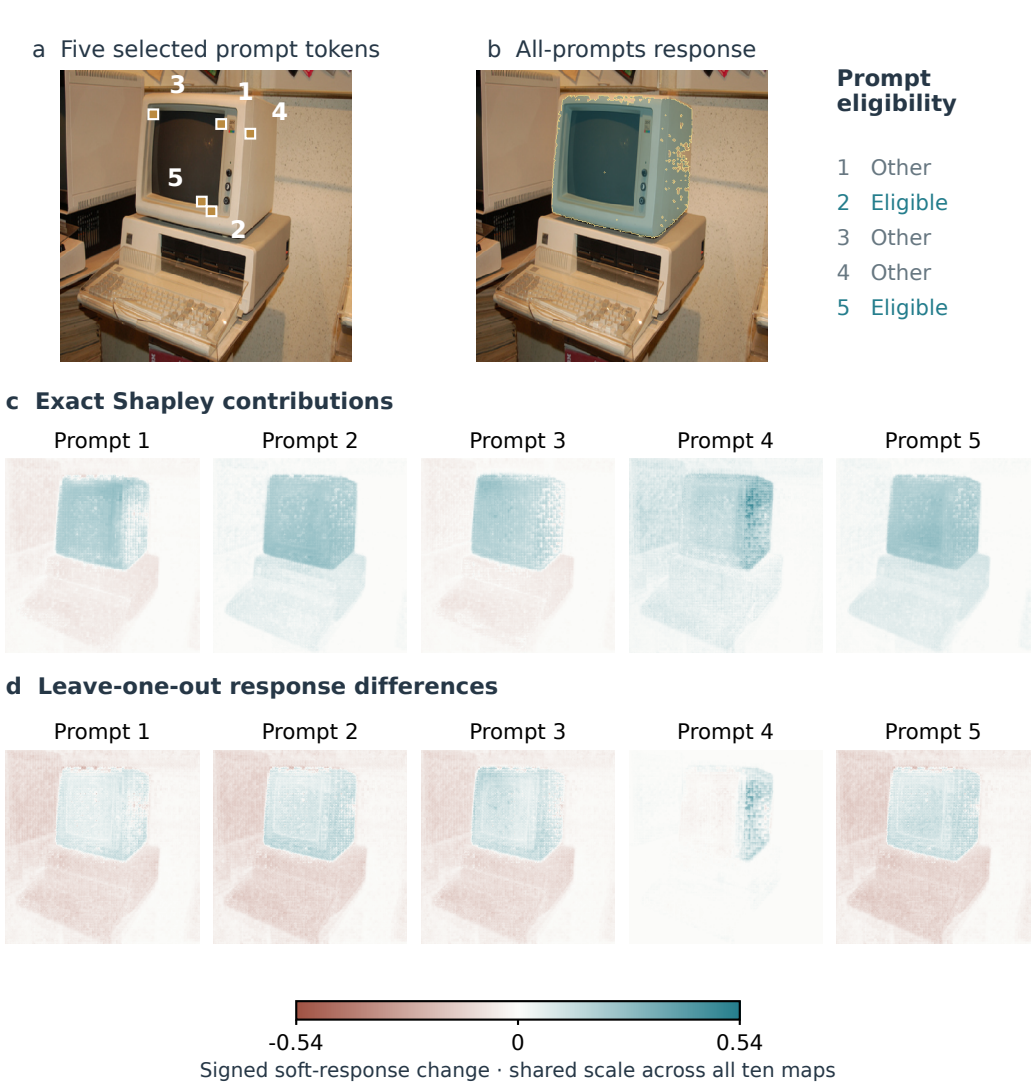

Figure 22: **Prompt context and pixelwise attribution.**

A.5 Robustness and Efficiency

A.5.1 Parameter perturbation and summary statistics

Perturbation design Individual perturbations used factors 0.8, 0.9, 1.1, and 1.2. Joint perturbations used fixed settings with log-uniform factors in $[0.8, 1.2]$ for each shot setting. Factors scaled parameters directly, except for bounded mask-revision parameters, whose odds were scaled after $[0, 1]$ normalization. All settings used eight datasets and three seeds, with CSC prompts and SAM 3 responses held fixed.

Dependencies and displayed effects Audit perturbations rebuilt B and all dependent quantities. One-shot selection and five-shot composition perturbations held B fixed. Changes in Figure 3 are paired mIoU differences from the default, averaged over three seeds and expressed in points. The upper panels show joint perturbations by dataset and their eight-dataset means. The lower panels rank parameters by their maximum absolute eight-dataset mean change across the four factors. They display the highest-ranked parameters for each shot setting, including the default at factor 1. The default configurations were retained throughout the analysis.

A.5.2 Response-bank reuse and computational scaling

PRCMS reuses prompt-subset responses for attribution, candidate construction, stability, consensus, and five-shot coalitions. The leave-one-out ablation retains this bank while changing the attribution estimator (Table 4). Response reuse therefore allows multiple forms of revision evidence to be computed from the same SAM 3 outputs. Bank construction scales as 2^k, with $k \leq 5$ in these experiments.

A.5.3 Inference cost

Comparison and measurement We compared the full pipeline with uniform response averaging under identical DINOv3 features, CSC prompts, and complete SAM 3 response banks. The control thresholds the FP32 average of nonempty-subset probability maps at 0.5. N2/C3 instead select a candidate by equal semantic/consistency score fusion (Table 6). Measurements used 12 seed-1 one-shot ISIC episodes, repeated across two fresh-process rounds for 96 measurements per method, on a shared NVIDIA A800 80GB PCIe GPU with at most 16 physical CPU cores. We alternated method order. Timing used warmed-up CUDA graph execution plans and ran from decoded inputs to the final 480×480 CPU mask. It included preprocessing, transfers, model inference, response generation, post-processing, and synchronization. Features and responses were recomputed for every episode.

In Table 14, latency summaries pool 96 measurements per method. Memory is the maximum over the same 12 episodes in separate, single-method processes.

Table 14: **Shared-response-bank inference cost.**

	End-to-end latency (ms)		Peak memory (GiB)	
Method	Mean	Median	Allocated	Reserved
Uniform response averaging	342.62	331.06	2.076	3.840
Full pipeline (CSC + PRCMS)	387.65	370.75	2.084	3.900

Latency and memory The full pipeline added 45.03 ms, or 13.1%, over uniform response averaging with the complete response bank (Table 14). Peak allocated and reserved memory report PyTorch live-tensor and allocator peaks, respectively, for each method. Because both methods construct the complete response bank, this comparison isolates the additional processing beyond shared response generation. It does not measure the cost difference from a baseline that builds a smaller bank or uses a different segmentation pipeline. The timing scope is the measured episode set and hardware configuration, rather than all evaluation datasets.

A.6 Qualitative Results and Error Analysis

A.6.1 One illustrative episode per dataset

Figures 23 and 24 retain eight distinct episodes, one per dataset, to compare the spatial forms of correction across domains. They include seven improvements and one unchanged final IoU following reference restoration (ISIC). The examples were selected retrospectively to illustrate different behaviors and are not a representative sample for estimating frequencies. The FSS and ISIC episodes also appear in Appendix A.4 to relate their internal evidence to their state transitions.

In Figure 23, columns show the annotated support, query, query GT, matched RSRM, and final prediction. Numbers are episode foreground IoUs in percent. Teal overlays and gold outlines mark foreground; gray marks ignored evaluation pixels. Query panels share the 480×480 evaluation frame, while support aspect ratios are preserved. Prediction masks are unchanged. SUIM shows improved target boundaries; ISIC shows preservation of the initial IoU after reference auditing.

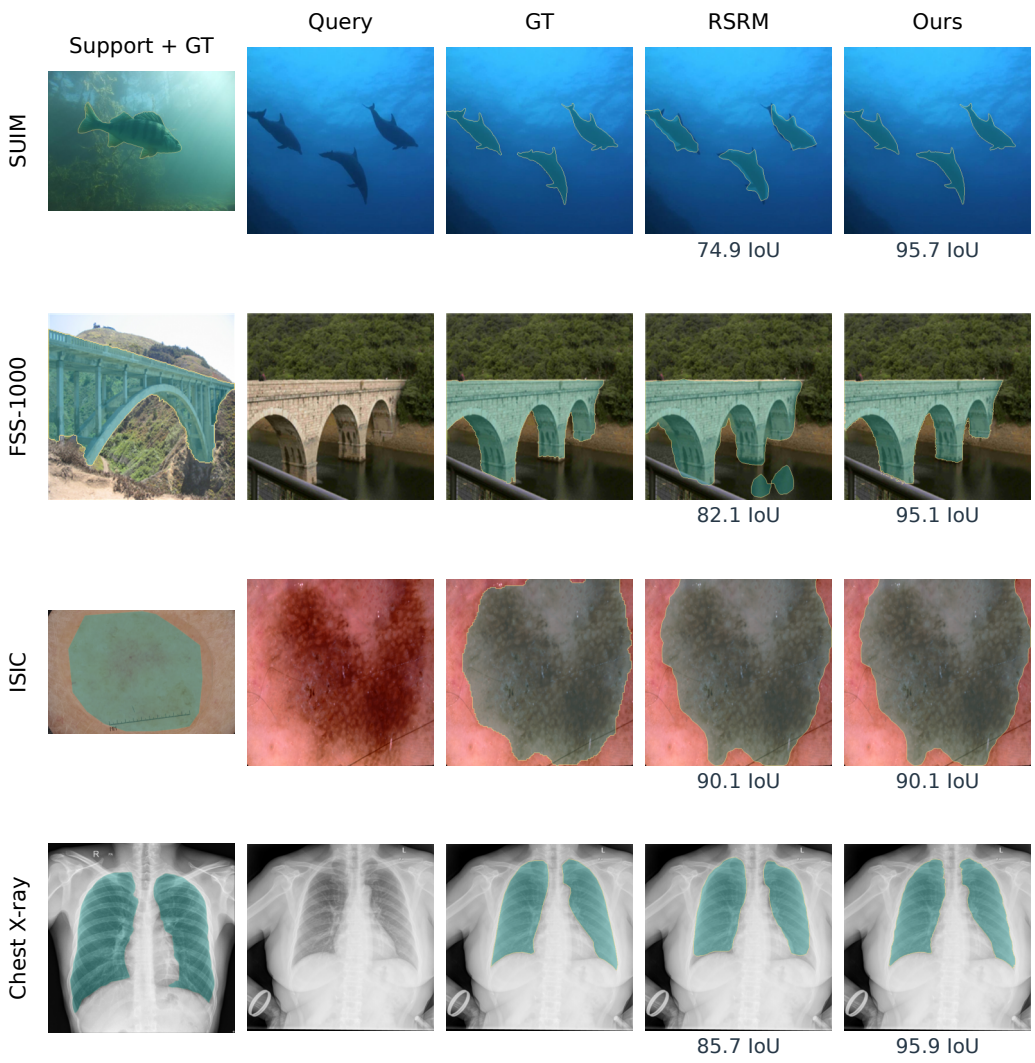

Figure 23: **Compact cross-domain examples: SUIM, FSS, ISIC, and Chest X-ray.**

In Figure 24, conventions match Figure 23. These examples illustrate different spatial corrections, including foreground recovery and removal of extraneous regions. They add no further episodes from a domain already shown in that figure. No cropping, contrast enhancement, or generative image processing is applied.

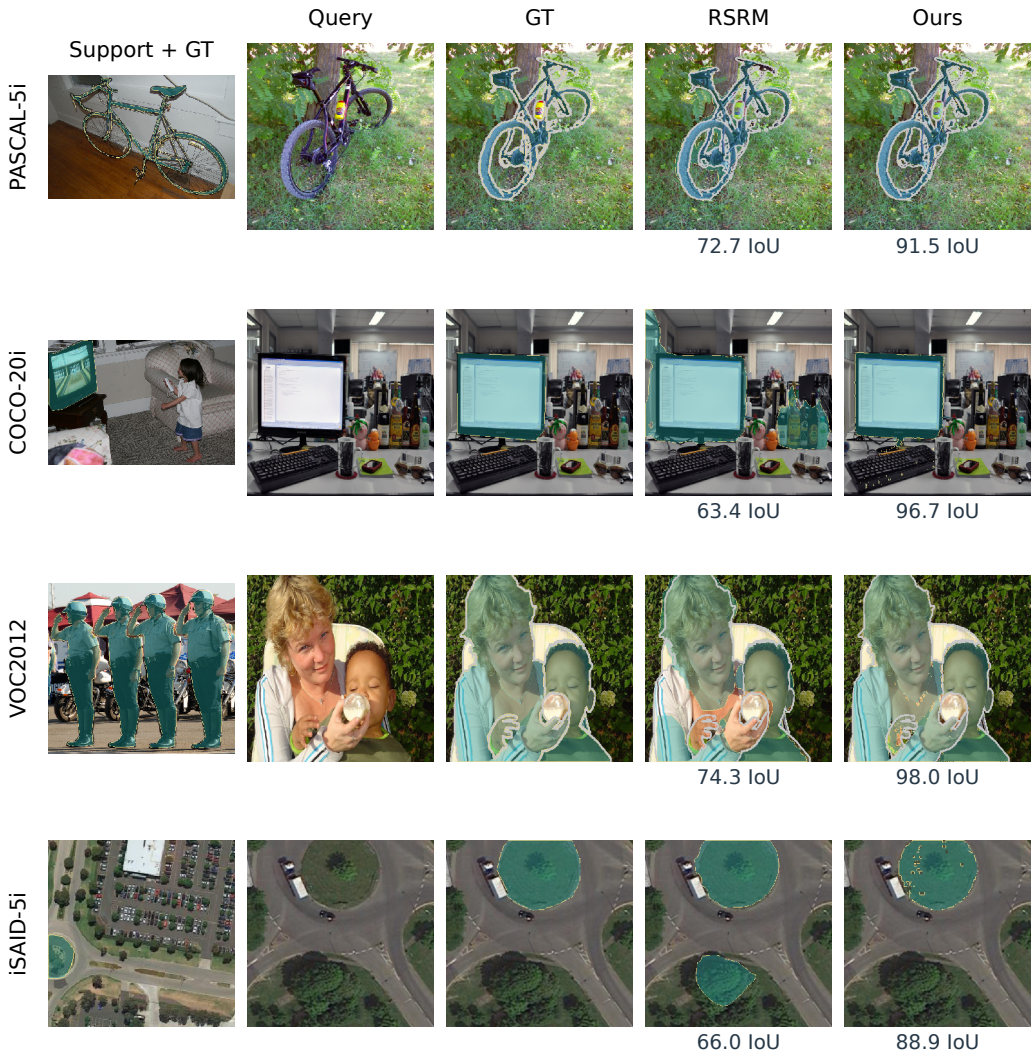

Figure 24: **Compact cross-domain examples: PASCAL, COCO, VOC2012, and iSAID.**

A.6.2 Spatial error diagnostics for the eight retained examples

The following panels expand the same eight episodes in the compact overview, with one example per dataset. Foreground overlays show the predicted extent, whereas categorical error maps distinguish missed targets from excess foreground. Orange indicates false positives, blue false negatives, pale gray-green correct foreground, white correct background, and gray ignored pixels. Prediction labels report episode IoU in percent. The support retains its aspect ratio; query panels share the unchanged 480×480 evaluation frame.

For each valid pixel, the local error-change map averages $\mathbf{1}[O \neq Y] - \mathbf{1}[B_0 \neq Y]$ over a 25×25 window, dividing by its valid-pixel count. Outside-image pixels contribute zero to both numerator and denominator; a zero denominator yields zero and ignored centers are masked. All panels use the same $[-1, 1]$ scale: blue indicates error reduction and red error increase. These maps use query ground truth for retrospective diagnosis; they are not model attention or confidence maps. Local cancellation is possible, so the categorical error maps remain necessary.

In Figure 25, the final mask follows the three target silhouettes more closely, reducing both excess foreground and missed target pixels. The smaller orange and blue error regions explain the episode

IoU improvement from 74.9% to 95.7%. This is the selected SUIM example in Figure 23. Colors and normalization follow Appendix A.6.2.

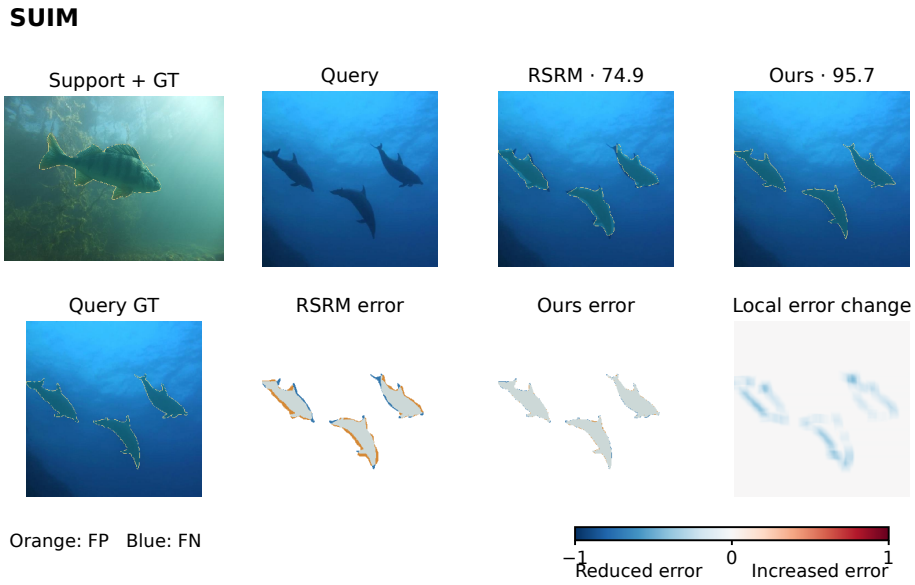

Figure 25: **SUIM: refining boundaries across multiple targets.**

In Figure 26, the final mask reduces excess foreground beneath the arches and around the lower supports. The corresponding orange false positives shrink, with blue local error changes marking the corrected areas. This is the bridge episode used in the response-bank and state-transition analyses.

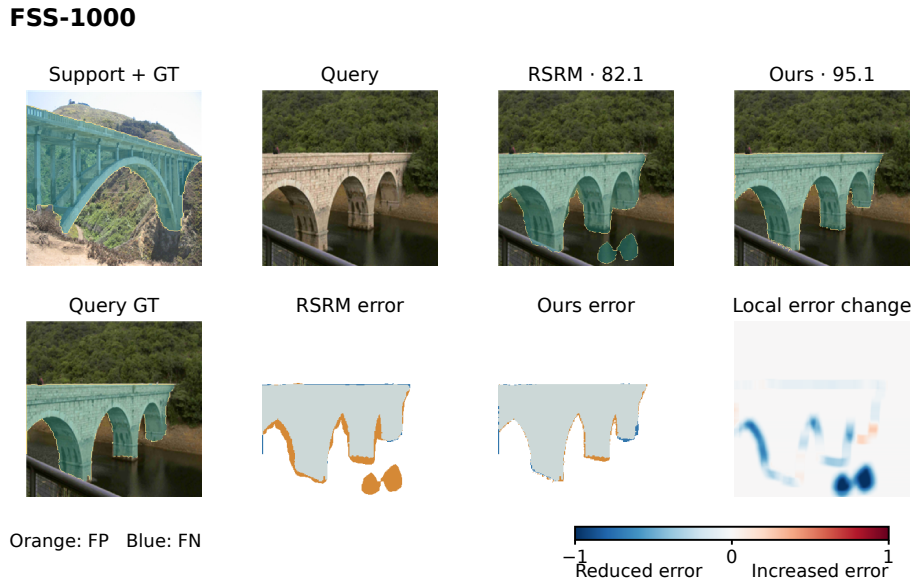

Figure 26: **FSS-1000: separating a bridge from surrounding regions.**

In Figure 27, the final and initial masks coincide after reference auditing, so their categorical errors are identical and the local error-change map is neutral. The remaining boundary errors show that preservation prevents the observed proposal degradation without making the retained prediction error-free.

In Figure 28, the final prediction reduces both excess foreground along the outer contours and missed foreground near the lower lung boundaries. The error maps distinguish these complementary

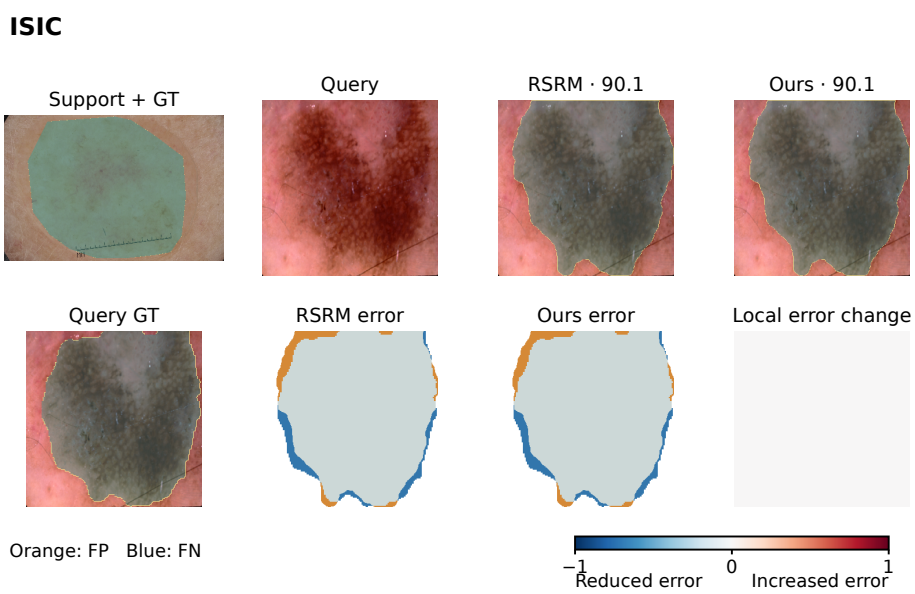

Figure 27: **ISIC: preserving the initial lesion boundary.**

corrections, which would be difficult to separate from foreground overlays alone. Thin residual errors remain along parts of the contour.

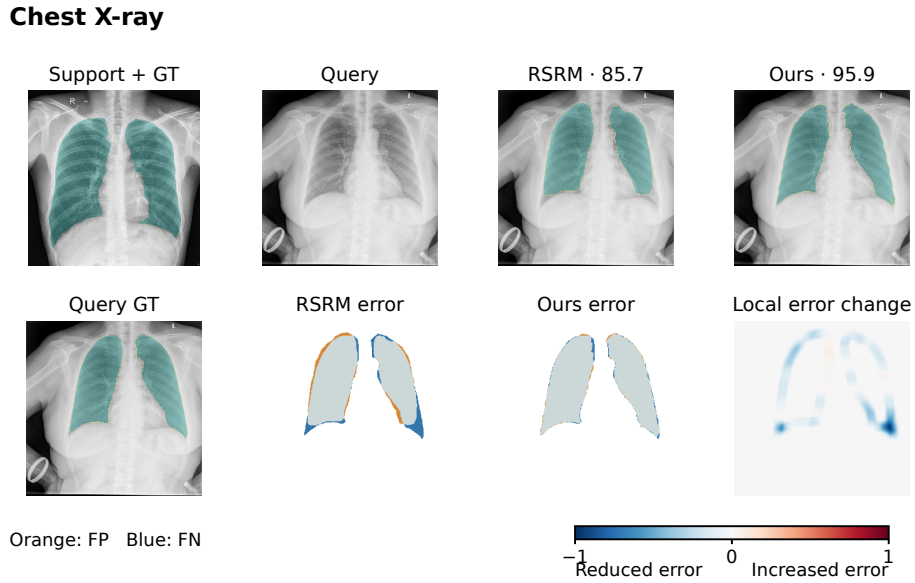

Figure 28: **Chest X-ray: correcting lung boundaries.**

In Figure 29, false-positive regions within and around the wheel openings are reduced while the main bicycle structure remains segmented. The localized blue error changes show where these corrections occur. This example illustrates that improving object extent can require removing interior background rather than filling it.

In Figure 30, the initial prediction includes a separate region to the right of the display and excess foreground near its boundary. Both are largely removed in the final mask. The categorical maps localize the false-positive reduction, while the local map distinguishes it from the small residual errors.



Figure 29: **PASCAL-5^i: resolving the bicycle's open structure.**

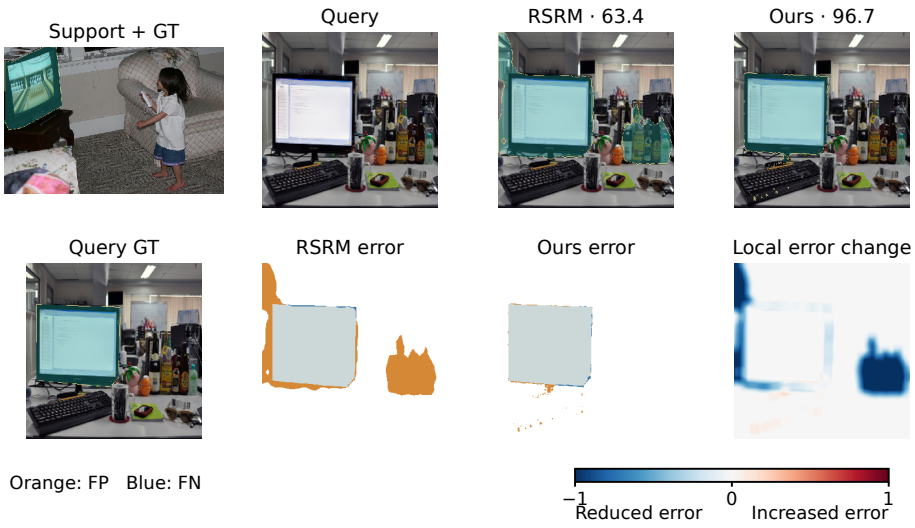

Figure 30: **COCO-20^i: excluding nearby non-target objects.**

In Figure 31, the initial mask omits substantial foreground in the lower part of the people. These false negatives shrink in the final prediction, producing concentrated blue local error changes. Unlike the nearby-object removal in Figure 30, this improvement primarily concerns target coverage.

In Figure 32, the final prediction removes the lower false-positive region while retaining the annotated circular target. Small false negatives remain within the target, as shown by the blue categorical errors. The paired maps reveal both the main correction and the residual omissions in this selected example.



Figure 31: **VOC2012: recovering omitted person regions.**

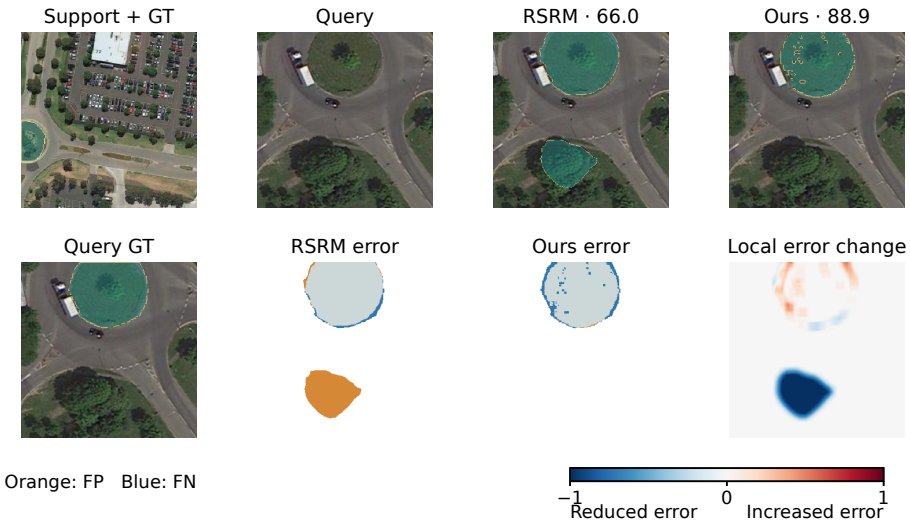

Figure 32: **iSAID-5^i: suppressing a separate non-target region.**