# SAMLENS: CANONICAL QUALITY CAN HIDE NEGATIVE-POINT REVERSAL IN SPECIALIZED SEGMENT ANYTHING MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Specialized versions of the Segment Anything Model (SAM) keep their point-prompt interface, yet release checks usually validate only the single positive prompt used in training. Among the prompts that remain available, a negative point has the least ambiguous meaning: the mask should not grow into the object it marks. We introduce SAMLens, a paired audit that gives reference and specialized endpoints identical images and prompts and reports each prompt relation both as an absolute response and as a change. On dense cell images, full mask-decoder tuning reverses this instruction: a negative point inside a neighboring object raises the predicted overlap with that object from 0.03 to 0.53, whereas endpoints of other SAM-family adaptations whose canonical Dice lies within 0.03 of this endpoint keep the object excluded or add only a small part of it. The reversal recurs in two further training seeds on cell images and in three on histopathology nuclei, and no other audited adaptation shows a reversal of comparable size. Changes in same-target stability and target switching do not separate the reversed endpoint from the others. A candidate decomposition places the reversal in the returned masks rather than in score-based selection, and shows that the relative decline seen after partial tuning is mostly response-range compression. Canonical Dice and change-only summaries of prompt behavior can therefore hide whether a specialized segmenter still follows negative points.

## 1 INTRODUCTION

Promptable segmentation models such as the Segment Anything Model (SAM; Kirillov et al., 2023) are often specialized to a cell type, scanner, or object class (Ma et al., 2024; Archit et al., 2025; Tejero et al., 2025) while retaining their point-conditioned interface. Yet a release check based on the canonical positive-point condition used for training does not establish whether the specialized model still preserves an object when the point moves, switches to another selected object, or suppresses a region marked by a negative point. These interactions are routine in interactive annotation, where users click off-center, move between neighboring objects, and add negative clicks to correct errors (Liu et al., 2023; Antonov et al., 2024; Archit et al., 2026).

A single robustness score cannot capture these interactions, because the three prompt changes call for opposite behaviors: same-target jitter calls for preservation, a point on a second instance calls for a target switch, and a negative point on an excluded object calls for suppression. Of the three, the negative point carries the least ambiguous instruction. A negative point placed inside a neighboring object, which we call the distractor, asks the model to keep that object out of the mask. Canonical-only training never presents such a point, so the training objective provides no direct supervision for this behavior.

SAMLens measures how each of these behaviors, which we call prompt relations, changes between the same model before and after specialization, not how robust a single model is to prompt changes. We call a model checkpoint with fixed inference settings an endpoint. Reference and specialized endpoints receive the same images, target pairs, and prompts, and every relation is reported as an absolute response and as a paired change, with source images as statistical units. For SAM, a candidate decomposition separates changes in the returned masks from changes in their selection.



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.

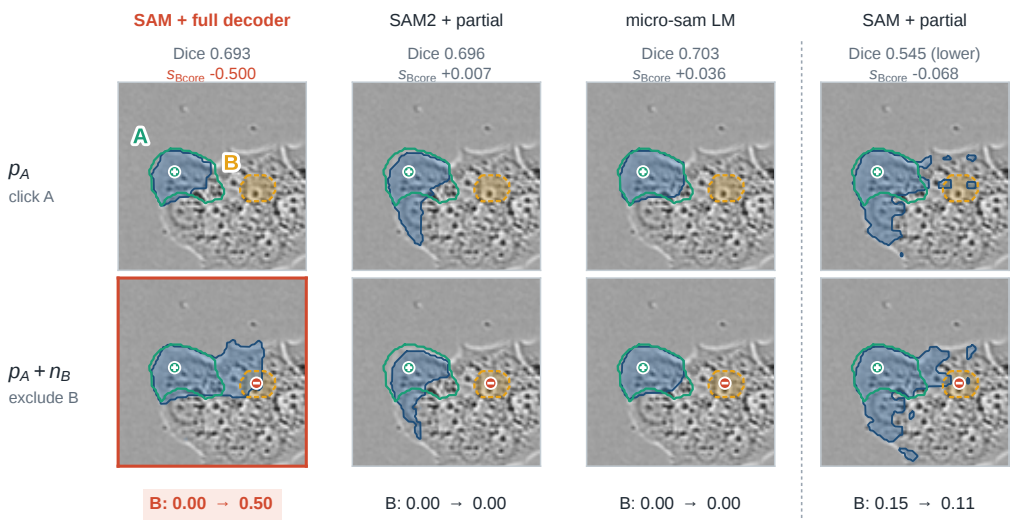



Figure 1: **Canonical quality does not determine negative-point response.** One LIVECell pair with a positive point $p_A$ on target $A$ (top) and an added negative point $n_B$ at the center of distractor $B$ (bottom). Blue: selected mask; green: $A$; dashed amber: $B$. Below: fraction of $B$ covered in this pair. Above: mean canonical Dice (400 pairs) and mean B-core response $s_{\mathrm{Bcore}}$, the drop in $B$ coverage caused by $n_B$ (387 pairs; negative: $n_B$ adds $B$). Full decoder: whole mask decoder tuned; partial: mask tokens, hypernetwork MLPs, and upscaling tuned (Section 4); micro-sam LM: released light-microscopy model. SAM2 partial and micro-sam LM have canonical Dice similar to the full decoder's but differ in provenance; SAM partial (right of the dashed rule) shares the full decoder's base and seed-42 training images but has lower Dice. Case selection: Appendix J.

Within the audited SAM family (Kirillov et al., 2023; Ravi et al., 2025), the absolute view reveals a failure that canonical Dice does not show (Figure 1). On dense LIVECell cell images (Edlund et al., 2021), full mask-decoder tuning reverses the instruction: a negative point inside the distractor raises the predicted overlap with the distractor from 0.03 to 0.53. Endpoints of other SAM-family adaptations whose canonical Dice lies within 0.03 of this endpoint behave differently: the released micro-sam light-microscopy model (micro-sam LM) and partially tuned SAM2 keep the distractor excluded, and two SAM seeds, tuned partially or with low-rank adaptation (LoRA; Hu et al., 2022), add a small part of it. The reversal recurs in two further LIVECell seeds and in all three training seeds on PanNuke histopathology nuclei (Gamper et al., 2019), and no other audited adaptation, including LoRA and a specialized SimpleClick model (Liu et al., 2023), shows a reversal of comparable size. In dense scenes, same-target stability falls and target switching improves after every audited adaptation, reversed or not, so an audit of those relations alone would not flag the reversal.

Our contributions are:

- A paired audit of prompt relations after specialization that reports absolute responses next to paired changes and uses a negative point inside the distractor as its most direct semantic test.
- Evidence that similar canonical quality can hide qualitatively different negative-point semantics, including a reversal after full mask-decoder tuning that recurs across seeds and domains and that no other audited adaptation matches in magnitude.
- An output-level diagnosis showing that a relative decline in negative-point response can mean either response-range compression or reversal, and that the reversal is already present in the masks SAM returns and is not created by score-based selection.

## 2 Related Work

### 2.1 SAM and its specialization

SAM couples prompt-conditioned mask generation with an intersection-over-union (IoU) prediction head that ranks candidates (Kirillov et al., 2023); SAM2 extends this interface to images and video

(Ravi et al., 2025). MedSAM, micro-sam, SAM-DA, SUM, and HQ-SAM specialize this design for medical, microscopy, domain-adaptive, or high-quality segmentation (Ma et al., 2024; Archit et al., 2025; Tejero et al., 2025; Liu et al., 2024; Ke et al., 2023). PP-SAM and Stable-SAM train under perturbed boxes or points, while HBF-BCER evaluates prompt-preserving medical adaptation under repeated prompt perturbations (Rahman et al., 2024; Fan et al., 2025; Ping et al., 2026). ProMISe is closest in motivation: it freezes SAM and adapts prompt patterns because fine-tuning may impair prompting capability (Wang et al., 2024). These methods improve accuracy, efficiency, or robustness under a chosen prompt distribution; we measure what specialization changes in the prompt interface.

### 2.2 Interactive segmentation

Interactive segmentation is commonly evaluated through click trajectories, prompt noise, and the number of corrective interactions (Liu et al., 2023; Chen et al., 2022; Antonov et al., 2024). TETRIS and BREPS test sensitivity to click placement and bounding-box prompts, respectively (Moskalenko et al., 2024; 2026). MedicoSAM (Archit et al., 2026), conSAMme (Myers-Dean et al., 2025), Kinalioglu (2026), and CAFE (Liang et al., 2026) study correction gains, consistency under nearby guidance, structure-preserving perturbations, and counterfactual grounding. These studies motivate testing beyond one mask score. We instead pair each specialized endpoint with its reference across separate preserve, switch, and suppress relations.

### 2.3 Behavioral testing and model auditing

Behavioral and metamorphic testing provide a methodological precedent. CheckList organizes tests around expected input–output capabilities (Ribeiro et al., 2020); behavioral-shift auditing compares model versions under controlled false-positive rates (Richter et al., 2025); and underspecification explains why similar held-out risk can conceal different off-condition behavior (D'Amour et al., 2022). We follow this practice but specify a different expected relation for each prompt change: preservation for nearby points, and a change for retargeting and negative points.

### 2.4 Candidate masks and reranking

Multiple plausible masks create a second distinction between what a model returns and what its scoring rule selects. PiClick, SAMRefiner, and UncertainSAM generate, refine, or rerank candidate masks (Yan et al., 2024; Lin et al., 2025; Kaiser et al., 2025); SAM itself reports oracle-mask results to expose candidate quality (Kirillov et al., 2023). We use a ground truth (GT) oracle over candidates only as a diagnostic, to separate changes in the returned masks from changes in their ranking.

### 2.5 Fine-tuning and retained capabilities

Fine-tuning a pretrained model can erode what it did well before: sequential training causes catastrophic forgetting (Kirkpatrick et al., 2017), full fine-tuning can distort pretrained features and hurt out-of-distribution accuracy (Kumar et al., 2022), and weight-space ensembling restores some of this robustness (Wortsman et al., 2022). These studies measure accuracy; we ask which prompt relations of a promptable model survive specialization.

## 3 SAMLens: A paired audit of prompt behaviors

### 3.1 Paired endpoints and prompt relations

We compare a reference endpoint $\theta_0$ with a specialized endpoint $\theta_1$ on identical images, target pairs, and prompts, with GT masks $G$. For every prompt relation $\mathcal{H}$ we report both levels, $\mathcal{H}(\theta_0)$ and $\mathcal{H}(\theta_1)$, next to the change $\Delta\mathcal{H} = \mathcal{H}(\theta_1) - \mathcal{H}(\theta_0)$, because two very different behaviors can produce the same change (Section 5.4).

A prompt $p$ is a set of signed (positive or negative) points. A SAM endpoint maps an image $x$ and a prompt $p$ to $K$ mask–score pairs and selects the highest-scoring mask,

$$f_\theta(x, p) = \{(m_k, q_k)\}_{k=1}^{K}, \qquad M_\theta(x, p) = m_{\arg\max_k q_k}. \tag{1}$$

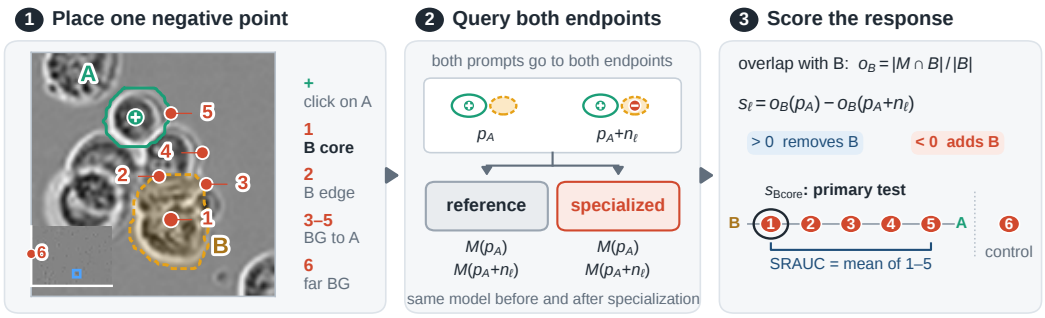



Figure 2: **Paired audit of six negative-point locations.** (1) A real LIVECell pair with fixed $p_A$; each query adds one negative point: 1 B core, 2 B edge, 3–5 background (BG) from near $B$ to near $A$, or 6 far BG (inset). (2) Both prompts query both endpoints. (3) $s_\ell$ is the drop in $B$ overlap: positive removes $B$, negative adds it. Location 1 gives the primary B-core response; SRAUC averages 1–5. Steps 2–3 are schematic.

Single-output models return one mask. Here $p_A$ denotes a positive point inside target $A$ and $n_\ell$ a negative point at location $\ell$; $M_\theta(x, p_A, n_\ell)$ abbreviates $M_\theta(x, \{p_A, n_\ell\})$, one compound prompt. A controlled adaptation minimizes the canonical risk $\mathcal{R}_{\text{can}} = \mathbb{E}[\mathcal{L}(M_\theta(x, p_A), G_A)]$, which evaluates only $(x, p_A)$. For each annotated pair $(A, B)$ of a target and a nearby distractor, we measure four relations.

**Canonical quality ($D_A$).** The Dice similarity coefficient at the training prompt, $D_A(\theta) = \text{Dice}(M_\theta(x, p_A), G_A)$.

**Same-target stability (STS).** A second positive point $p'_A$ inside $A$ should preserve the mask: $\text{STS}(\theta) = \text{IoU}(M_\theta(x, p_A), M_\theta(x, p'_A))$. Because a stable but wrong mask also scores high, we report target-conditioned stability $\text{STS}_{target}(\theta) = \min\{\text{IoU}(M_\theta(x, p_A), G_A), \text{IoU}(M_\theta(x, p'_A), G_A)\}$ alongside it.

**Target-switch obedience (TSO).** Moving the point to $p_B$ should switch the output to $B$: $\text{TSO}(\theta) = \text{IoU}(M_\theta(x, p_B), G_B) - \text{IoU}(M_\theta(x, p_B), G_A)$.

**Negative-point response ($s_\ell$).** We keep $p_A$ and add $n_\ell$. With $o_B(m) = |m \cap G_B|/|G_B|$ the fraction of $B$ covered by mask $m$,

$$s_\ell(\theta) = o_B(M_\theta(x, p_A)) - o_B(M_\theta(x, p_A, n_\ell)). \tag{2}$$

Positive $s_\ell$ means that the negative point removes part of $B$; negative $s_\ell$ means that it adds part of $B$.

### 3.2 Where the negative point is placed

Negative points occupy six fixed locations (Figure 2): the core and edge of $B$, three BG points from near $B$ to near $A$, and a far-BG control.

**B-core response ($s_{\text{Bcore}}$).** Our primary test is $s_\ell$ with the negative point $n_B$ at the center of $B$. It audits a relation that we specify at the object level: a negative point inside an object should not increase the predicted overlap with that object, so $s_{\text{Bcore}} < 0$ violates it. We report the two overlaps that define $s_{\text{Bcore}}$, $o_B(p_A)$ and $o_B(p_A{+}n_B)$, and target retention, the Dice with $A$ before and after the negative point. SAM's interface marks only the clicked pixel as background, so Section 5.1 also checks that pixel.

**Spatial-response area under the curve (SRAUC).** SRAUC averages $s_\ell$ over the five local locations, $\text{SRAUC}(\theta) = \frac{1}{5}\sum_{\ell=1}^{5} s_\ell(\theta)$, a normalized area under the response curve from $B$ toward $A$. The far-BG point is excluded from this mean and serves as a location-independent control.

### 3.3 Returned candidates versus selection

For SAM, a change in the selected mask can come from the candidates the model returns or from how its IoU head ranks them. Let $\mathcal{C}^0_\theta = \{m^0_k\}$ and $\mathcal{C}^-_\theta = \{m^-_k\}$ be the candidates returned without

and with the negative point, $t^0$ and $t^-$ the score-selected indices, and $r^0$ and $r^-$ the candidates with the lowest overlap with $B$ in each set. At each location,

$$S_{\text{total}} = o_B(m^0_{t^0}) - o_B(m^-_{t^-}) = \underbrace{o_B(m^0_{r^0}) - o_B(m^-_{r^-})}_{S_{\text{set}}} + \underbrace{\left[o_B(m^0_{t^0}) - o_B(m^0_{r^0})\right] - \left[o_B(m^-_{t^-}) - o_B(m^-_{r^-})\right]}_{S_{\text{selection}}}. \quad (3)$$

The identity is exact and needs no matching of candidates across prompts. The best-candidate term $S_{\text{set}}$ compares the best returned candidate under each prompt; the score-selection term measures how the gap between the selected and the best candidate changes. We apply it both to absolute responses and to specialized-minus-reference changes, and repeat it among candidates that reach at least 80%, 90%, or 95% of the best target Dice (Appendix D).

### 3.4 Prompt bank and statistics

Within each image we pair annotated instances whose mask boundaries lie within a dataset-specific distance (8–120 pixels on LIVECell), keeping the four closest pairs, so $B$ is a neighbor that a user could confuse with $A$. Positive points are distance-transform centers; random-interior and boundary-interior replications test the center-point policy. Pair effects are averaged within each source image, and images (or seeds and images) are bootstrapped with 10,000 draws. Effects are confirmed when their 95% confidence intervals (CIs) exclude zero and paired sign-flip tests pass Holm correction (Holm, 1979) within a fixed family. The B-core analyses in Section 5.1 are post-hoc analyses of frozen endpoints: endpoints, metrics, and test families were fixed in a dated plan before the final recomputation (Appendix J).

## 4 Experimental setup

### 4.1 Datasets

LIVECell (Edlund et al., 2021) provides dense phase-contrast cell images; we audit 100 official test images (387 pairs with all six negative-point locations). PanNuke (Gamper et al., 2019) provides dense histopathology nuclei; we train on fold 1 and audit 93 fold-3 images (349 pairs). COCO (Lin et al., 2014) (123 val2017 images) serves as a natural-image comparison for the preservation and switching relations, and Kvasir-SEG (Jha et al., 2020) supports a single-object background-point analysis (Appendix F). For our controlled adaptations, training and audit images never overlap (Appendix Table S2); released endpoints keep the provenance limits of Section 7 and Appendix J.

### 4.2 Endpoints

We audit SAM with a ViT-B backbone (Kirillov et al., 2023) under three adaptations of its mask decoder, all trained with Dice plus binary cross-entropy (BCE) loss at the canonical prompt $p_A$ only, for 4,000 steps on about 200 training images: rank-4 LoRA (Hu et al., 2022) on the attention layers in the decoder's two-way transformer (47,104 parameters); partial tuning of the mask tokens, hypernetwork multilayer perceptrons (MLPs), and output upscaling, with the transformer and IoU head frozen (634,208 parameters); and full tuning of the mask decoder (4,058,340 parameters). We also partially tune SAM2 Hiera-L (Ravi et al., 2025) and audit two released micro-sam models (Archit et al., 2025), micro-sam LM and the micro-sam LIVECell specialist. As a non-SAM, single-output reference we specialize SimpleClick ViT-B (Liu et al., 2023) by tuning its click embedding, neck, and head (11.0M parameters; backbone frozen) for 2,000 steps with the same canonical-only objective. The SAM adaptations train the single-mask output, whereas the audit scores SAM's three multimask candidates; scoring the single-mask output also shows the reversal (Appendix J.6).

Every controlled adaptation on PanNuke and every SAM2 and SimpleClick adaptation uses seeds 42, 123, and 456. On LIVECell, the SAM LoRA, partial, and full-decoder endpoints train on one pool of official test-split images outside the audit set (zero overlap), with the same images for all three adaptations of a seed; this is an audit design, not official-test benchmark evaluation. Each adaptation has three seeds (full-decoder seeds 123 and 456 were added after the analysis was frozen;

Table 1: **Negative-point response at the B core.** $o_B$: fraction of distractor $B$ covered before and after a negative point at the center of $B$; $s_{\mathrm{Bcore}}$ is their difference, so negative values mean the negative point adds $B$. Single endpoints show 95% CIs; multi-seed rows show the seed mean and the range of seed means. "CI $< 0$" counts seeds whose CI lies below zero; with about 100 images even increases below 0.01 are detectable, so the magnitude matters. SRAUC averages five locations. Dice uses all pairs (400 on LIVECell); the other columns use the pairs with all six negative-point locations. [§]SimpleClick uses the corrected pair bank; its spatial scores are omitted (Appendix K). PanNuke SAM2 partial: canonical Dice not recorded (target Dice in Table S18). [‡]Trained after the analysis plan was frozen (Appendix J.4).

| Endpoint | Seeds | Dice | $o_B(p_A)$ | $o_B(p_A{+}n_B)$ | $s_{\mathrm{Bcore}}$ [CI or seed range] | CI $< 0$ | SRAUC |
|---|---|---|---|---|---|---|---|
| *LIVECell (100 images, 387 pairs)* | | | | | | | |
| SAM (reference) | 1 | 0.308 | 0.601 | 0.480 | +0.121 [+0.093, +0.152] | 0/1 | +0.090 |
| SAM + LoRA | 3 | 0.599 | 0.091 | 0.078 | +0.012 (−0.028 to +0.045) | 1/3 | +0.064 |
| SAM + partial | 3 | 0.617 | 0.067 | 0.122 | −0.055 (−0.068 to −0.044) | 3/3 | 0.000 |
| **SAM + full decoder** | **1** | **0.693** | **0.026** | **0.526** | **−0.500 [−0.535, −0.464]** | **1/1** | **−0.173** |
| seeds 123, 456[‡] | 2 | 0.719 | 0.003 | 0.370 | −0.367 (−0.411 to −0.324) | 2/2 | −0.103 |
| SAM2 (reference) | 1 | 0.302 | 0.601 | 0.422 | +0.179 [+0.152, +0.207] | 0/1 | +0.133 |
| SAM2 + partial | 3 | 0.693 | 0.007 | 0.000 | +0.007 (+0.006 to +0.007) | 0/3 | +0.007 |
| micro-sam LM (released) | 1 | 0.703 | 0.043 | 0.007 | +0.036 [+0.020, +0.055] | 0/1 | +0.030 |
| SimpleClick (reference)[§] | 1 | 0.040 | 0.948 | 0.129 | +0.818 [+0.791, +0.844] | 0/1 | – |
| SimpleClick + tuning[§] | 3 | 0.695 | 0.001 | 0.003 | −0.002 (−0.005 to −0.001) | 2/3 | – |
| *PanNuke fold 3 (93 images, 349 pairs)* | | | | | | | |
| SAM (reference) | 1 | 0.723 | 0.140 | 0.066 | +0.074 [+0.047, +0.105] | 0/1 | +0.058 |
| SAM + LoRA | 3 | 0.831 | 0.000 | 0.005 | −0.005 (−0.006 to −0.003) | 3/3 | −0.001 |
| SAM + partial | 3 | 0.833 | 0.004 | 0.047 | −0.043 (−0.061 to −0.034) | 3/3 | −0.016 |
| **SAM + full decoder** | **3** | **0.849** | **0.002** | **0.527** | **−0.525 (−0.721 to −0.204)** | **3/3** | **−0.168** |
| SAM2 (reference) | 1 | 0.753 | 0.114 | 0.040 | +0.074 [+0.043, +0.109] | 0/1 | +0.070 |
| SAM2 + partial | 3 | – | 0.015 | 0.002 | +0.013 (+0.011 to +0.017) | 0/3 | +0.012 |
| SimpleClick (reference)[§] | 1 | 0.074 | 0.929 | 0.050 | +0.879 [+0.838, +0.914] | 0/1 | – |
| SimpleClick + tuning[§] | 3 | 0.834 | 0.001 | 0.014 | −0.013 (−0.024 to −0.006) | 3/3 | – |

Appendix J.4). An endpoint enters the analysis only if it improves held-out canonical Dice (the specialization gate), and no prompt-relation outcome is used to select checkpoints; all endpoints are final checkpoints of fixed budgets (Appendices B and J).

# 5 Results

## 5.1 Similar canonical quality, different negative-point semantics

Table 1 and Figure 1 summarize the B-core response. On LIVECell, the seed-42 full-decoder endpoint reaches canonical Dice 0.693, yet a negative point inside $B$ raises its overlap with $B$ from 0.026 to 0.526 ($s_{\mathrm{Bcore}} = -0.500$, 95% CI $[-0.535, -0.464]$, Holm $p = 0.0005$). Target retention falls from 0.694 to 0.436 Dice, so the negative point both adds $B$ and degrades the mask of $A$. Across the three full-decoder seeds, the mask even comes to cover the clicked pixel itself in 31% to 57% of pairs, against at most 5.4% for the other SAM-family adaptations (a descriptive check; Appendix J.5). The response goes in the direction of a positive click, only weaker: in an earlier label-swap audit (20 images), a negative point added 0.542 of $B$, a positive point at the same coordinate 0.858 (Appendix J.7).

Similar canonical Dice does not determine these responses: we compared the seed-42 full decoder with every endpoint of another SAM-family adaptation whose canonical Dice lies within 0.03 of it. Six endpoints qualify: micro-sam LM, three SAM2 partial seeds, and one partial and one LoRA seed of SAM. All but the LoRA seed have a Dice difference whose CI contains zero, and all six exceed the full decoder's B-core response by 0.46 to 0.54 (all Holm $p \leq 0.0003$; Table S19). Within the same training pool, all three partial seeds exceed it by 0.43 to 0.46.

The reversal is typical, not driven by a few pairs (Figure 3A). For the seed-42 full decoder, the negative point adds more than a quarter of $B$ in 73% of LIVECell pairs and more than half of $B$ in 50%; for every endpoint and seed of the other SAM-family adaptations, these fractions are at

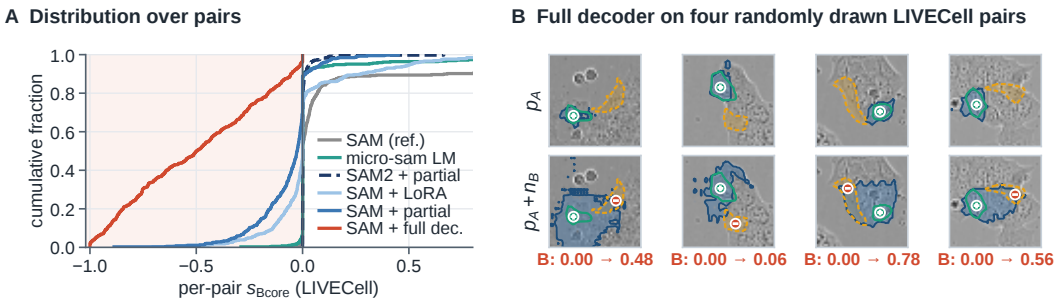



Figure 3: **The reversal is typical across pairs.** (A) Cumulative distribution of the per-pair B-core response on LIVECell (seed 42 for multi-seed adaptations); shaded: the negative point adds $B$. (B) The seed-42 full-decoder endpoint on four pairs drawn at random from the 280 LIVECell pairs with canonical Dice $\geq 0.6$ (Appendix J); blue/green/amber: mask/$A$/$B$; red dot: $n_B$; labels give the fraction of $B$ covered.

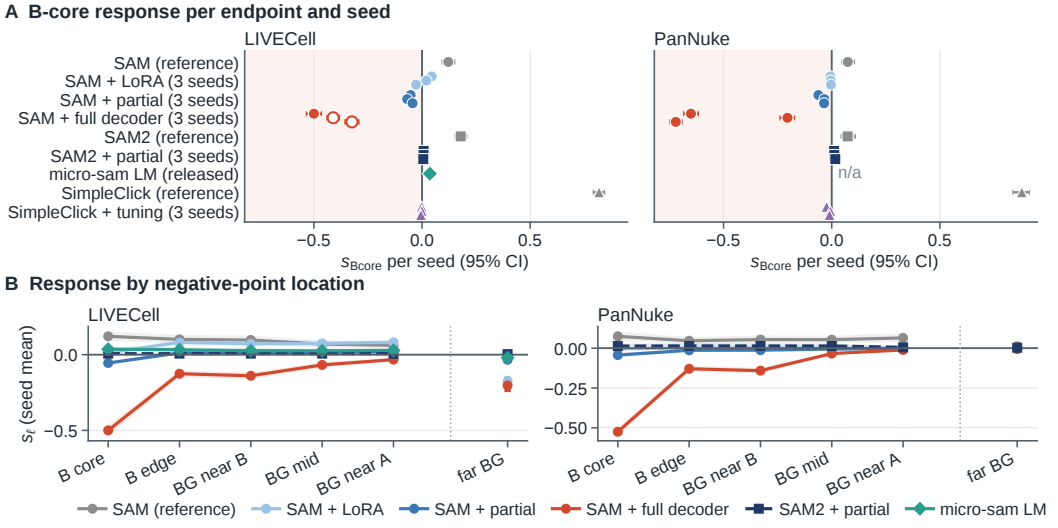



Figure 4: **The reversal recurs across seeds and domains for full-decoder tuning.** (A) B-core response per endpoint and seed (95% CIs); shaded: the negative point adds $B$; open: LIVECell full-decoder seeds added later; SimpleClick uses the corrected pair bank. (B) Seed-averaged $s_\ell$ at the five local locations and the far-background control (95% bands; LIVECell full decoder: seed 42; micro-sam LM: LIVECell only).

most 10.1% and 1.3%. On PanNuke, the fraction above a quarter is 30%, 88%, and 95% across the three full-decoder seeds. In all four randomly drawn pairs (Figure 3B), the mask also grows into the background.

The reversal is strongest where the instruction is least ambiguous, but it is not confined to that probe (Figure 4B). On LIVECell the full decoder's response rises from $-0.50$ at the B core to $-0.03$ near $A$. A background point next to $B$, where a user would click to stop leakage into $B$, still gives $-0.14$ in both domains ($-0.08$ and $-0.05$ for the added LIVECell seeds), while no endpoint of another SAM-family adaptation adds more than 0.02 of $B$ there. The far-background control is negative on LIVECell for both the full decoder ($-0.20$) and LoRA ($-0.18$), whose B-core response is near zero, so this location-independent effect is not specific to the full decoder; on PanNuke it is near zero.

### 5.2 Replication and boundaries

Two further LIVECell seeds, trained with the same recipe after the analysis plan was frozen (Appendix J.4), also reverse the negative point ($s_{\mathrm{Bcore}} = -0.411$ and $-0.324$, each Holm $p < 0.001$, at Dice 0.702 and 0.737) and fall 0.36 and 0.28 below the same-seed partial endpoint. On PanNuke,

Table 2: **Changes in the other relations (specialized minus reference).** Estimates with 95% CIs ($\Delta\text{STS}_{target}$: point estimates); $^{\circ}$ inconclusive. $^{*}\text{STS}_{target}$ change confirmed; $^{\dagger}$reported without a test. The first three rows are seed-42 endpoints of the training pool of Table 1 (where LoRA and partial rows average three seeds); their STS and TSO changes are comparable relative to their Dice gains, yet only the full decoder's reversal is large. $s_{\text{Bcore}}$: absolute B-core response of the same endpoints (Tables 1 and S18); –: not in the negative-point analysis. $^{\P}$micro-sam LIVECell specialist, not LM.

| Setting / endpoint | $\Delta$Dice | $\Delta$STS | $\Delta\text{STS}_{target}$ | $\Delta$TSO | $s_{\text{Bcore}}$ |
|---|---|---|---|---|---|
| LIVECell SAM + LoRA (s42) | +0.260 [+0.231, +0.290] | −0.234 [−0.258, −0.210] | +0.156$^{*}$ | +0.255 [+0.228, +0.282] | +0.020 |
| LIVECell SAM + partial (s42) | +0.236 [+0.216, +0.257] | −0.201 [−0.221, −0.180] | +0.163$^{*}$ | +0.235 [+0.213, +0.256] | −0.068 |
| **LIVECell SAM + full decoder (s42)** | +0.384 [+0.351, +0.418] | −0.351 [−0.377, −0.324] | +0.238$^{*}$ | +0.378 [+0.348, +0.408] | −0.500 |
| LIVECell SAM2 + partial (3 seeds) | +0.391 [+0.364, +0.418] | −0.274 [−0.297, −0.250] | +0.271$^{*}$ | +0.395 [+0.367, +0.422] | +0.007 |
| LIVECell micro-sam specialist$^{\P}$ | +0.477 [+0.438, +0.516] | −0.080 [−0.103, −0.058] | +0.425$^{\dagger}$ | +0.511 [+0.473, +0.549] | – |
| LIVECell SimpleClick + tuning (3 seeds) | +0.656 [+0.641, +0.670] | −0.310 [−0.351, −0.268] | +0.416$^{*}$ | +0.547 [+0.530, +0.565] | −0.002 |
| PanNuke SAM + partial (3 seeds) | +0.110 [+0.077, +0.144] | −0.112 [−0.145, −0.078] | +0.068$^{*}$ | +0.112 [+0.076, +0.150] | −0.043 |
| COCO SAM + partial (3 seeds) | +0.070 [+0.048, +0.094] | +0.018$^{\circ}$ [−0.008, +0.045] | +0.080$^{\dagger}$ | +0.076 [+0.052, +0.104] | – |

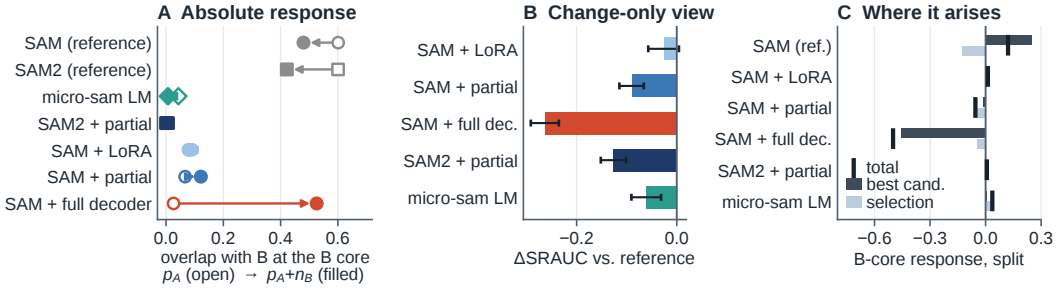



Figure 5: **A change-only view conflates compression and reversal (LIVECell).** Seed means for multi-seed adaptations; full decoder: seed 42 in all panels. Colors and markers follow Figure 4. (A) B-core overlap with $B$ without (open) and with (filled) the negative point; a rightward arrow means the point adds $B$. (B) Change in SRAUC from the reference (95% CIs); all point estimates are negative; LoRA's CI includes zero. (C) B-core response split into best-candidate (dark) and selection (light) terms; ticks: total.

the full decoder reverses it in all three seeds ($s_{\text{Bcore}} = -0.204$, $-0.651$, and $-0.721$; each Holm $p = 0.0005$) at canonical Dice 0.841–0.859, slightly above partial tuning (0.827–0.838). Magnitudes vary threefold across seeds; the same-seed partial endpoint exceeds the full decoder by 0.17, 0.59, and 0.69.

The other adaptations show at most small effects (Figure 4A). SAM2 partial tuning and micro-sam LM keep $B$ excluded. SAM partial tuning adds a small part of $B$ in every seed and both domains ($-0.03$ to $-0.07$), LoRA at most 0.03, and specialized SimpleClick at most 0.024 despite its tuned click modules.

### 5.3 Preservation and switching do not reveal the reversal

Preservation and switching alone would not single out the reversed endpoint (Table 2). For the seed-42 endpoints of the training pool of Table 1, LoRA, partial, and full-decoder tuning raise Dice by 0.260, 0.236, and 0.384, and all three lower STS and raise TSO by amounts comparable to their Dice gains (0.85–0.91 and 0.98–1.00 times), while $\text{STS}_{target}$ rises in all three. Their B-core responses, in contrast, are $+0.020$, $-0.068$, and $-0.500$. All audited specialized endpoints in dense scenes (SAM2, the micro-sam LIVECell specialist, SimpleClick) share this profile; since $\text{STS}_{target}$ rises, lower STS does not mean worse masks. On COCO, the STS change is inconclusive.

### 5.4 A RELATIVE DECLINE CAN MEAN COMPRESSION OR REVERSAL

Every SAM-family specialized endpoint in Figure 5B has a lower SRAUC point estimate than its reference, from $-0.026$ after LoRA, whose CI includes zero, to $-0.263$ after full-decoder tuning; the added full-decoder seeds fall as well (Appendix J.4). The change alone suggests one effect of varying size; the absolute overlaps (Figure 5A) show two. After partial tuning of SAM or SAM2, and in micro-sam LM, the canonical mask already excludes most of $B$, so the negative point has little left to remove; after full-decoder tuning, the negative point adds $B$.

Relative to the reference, the candidate decomposition splits SAM's $-0.090$ SRAUC change after partial tuning into a $-0.238$ best-candidate term and an opposing $+0.149$ score-selection term (SAM2: $-0.344 + 0.218$). Among candidates within 90% of the best target Dice, the best candidate's overlap with $B$ falls from 0.437 (reference) to 0.030 (partial) without the negative point and from 0.194 to 0.019 with it, and likewise for SAM2 (Figure S3). Specialized candidates overlap $B$ less under both prompts; the contrast shrinks because the overlap without the negative point falls further. We call this response-range compression; it lowers SRAUC without making the negative point harmful. The full decoder's reversal, by contrast, lies in its best returned candidate: at the B core, the best-candidate term is $-0.456$ (95% CI $[-0.490, -0.422]$) and the score-selection term only $-0.044$; the other five full-decoder seeds agree ($-0.153$ to $-0.661$ against $-0.015$ to $-0.100$; Appendix J). Partial tuning's small B-core increase is mostly selection ($-0.044$ of $-0.055$).

This matters for remedies: retraining the IoU head can fix ranking errors (Appendix E), but reranking chooses only among returned candidates and cannot remove the $-0.456$ the best one carries.

## 6 IMPLICATIONS FOR RELEASING SPECIALIZED MODELS

We suggest three checks before releasing a specialized promptable segmenter. First, place a negative point inside a neighboring object and report the overlap with it before and after; in our audit this test sharply separated full-decoder tuning from the other audited endpoints, including several with similar Dice. Second, report the absolute levels behind any change, because compression and reversal produce changes of the same sign. Third, audit the extent of adaptation: in our SAM configurations, a large reversal appeared only when the whole mask decoder was tuned on canonical prompts.

## 7 LIMITATIONS

**Scope.** The reversal is established for SAM ViT-B with full mask-decoder tuning on two dense domains, with three seeds each and magnitudes from $-0.20$ to $-0.72$, each with one training recipe (no learning-rate, length, or data-size sweep). SAM3 (Carion et al., 2026) did not pass the specialization gate, and the specialized SimpleClick endpoints showed no reversal of comparable size, so we make no claim about specialization in general. That a large reversal appeared only under full-decoder tuning comes from three configurations with different tuned modules, not a controlled parameter-count comparison.

**Analysis.** The B-core analyses are post-hoc analyses of frozen endpoints, with test families fixed before the final recomputation and an addendum for the two added seeds. Saved SimpleClick predictions had encoded the negative click as positive; we replaced them with a verified pair bank (Appendix K). Two independent pipelines agree on the B-core overlaps of eight LIVECell endpoints.

**What is measured.** Prompts are synthetic; the B-core point is a deliberate probe, not a typical correction click, though background points next to $B$ fail more mildly. We do not measure human effort, usability, or clinical outcomes; released micro-sam models lack per-image training manifests.

## 8 CONCLUSION

SAMLens reports prompt relations as absolute responses and as changes. In the audited SAM family, endpoints with similar canonical Dice can respond to the same negative point in opposite ways: full mask-decoder tuning reverses the instruction, a failure in the returned masks that neither canonical Dice nor preservation and switching reveal. Future work should test negative-point-aware training and whether reversals cost annotators extra corrections.

## ETHICS STATEMENT

This study analyzes released segmentation models, existing image datasets, and publicly released click records; it recruits no participants and collects no new clinical data. Microscopy and medical-image results are research measurements, not evidence of clinical safety or improved human interaction. The supplement does not redistribute image datasets or model checkpoints; users must obtain them under the respective providers' terms.

## REPRODUCIBILITY STATEMENT

Each main-text result is traceable to saved outputs indexed by model, prompt, data split, and instance pair, with explicit validity metadata. The anonymous supplement provides prompt generators, model and dataset wrappers, evaluation and cluster-inference code, figure/table generators, source hashes, and evidence manifests. Setup instructions document the separately obtained checkpoints, predictions, and datasets. We verify no train/evaluation image overlap in controlled comparisons; independently released endpoints retain the provenance limits stated in Section 7. Appendix H details the artifact contents and validity checks.

## AI USE STATEMENT

Generative AI tools assisted research-question and hypothesis refinement; experimental and statistical-analysis design; code drafting and debugging; literature search, organization, and summarization; result interpretation and adversarial review; figure implementation; and manuscript drafting, editing, and Chinese–English translation. The authors independently inspected and tested AI-assisted code, verified statistical outputs against saved artifacts, checked cited claims against primary sources, reviewed all AI-assisted text and figures, and take responsibility for the manuscript and its scientific claims. No model-generated output was treated as experimental evidence without verification.

## REFERENCES

Anton Antonov, Andrey Moskalenko, Denis Shepelev, Alexander Krapukhin, Konstantin Soshin, Anton Konushin, and Vlad Shakhuro. RClicks: Realistic click simulation for benchmarking interactive segmentation. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, pp. 127673–127710, 2024.

Anwai Archit, Luca Freckmann, Sushmita Nair, Nabeel Khalid, Paul Hilt, Vikas Rajashekar, Marei Freitag, Carolin Teuber, Melanie Spitzner, Constanza Tapia Contreras, Genevieve Buckley, Sebastian von Haaren, Sagnik Gupta, Marian Grade, Matthias Wirth, Günter Schneider, Andreas Dengel, Sheraz Ahmed, and Constantin Pape. Segment anything for microscopy. *Nature Methods*, 22(3): 579–591, 2025. doi: 10.1038/s41592-024-02580-4.

Anwai Archit, Luca Freckmann, and Constantin Pape. MedicoSAM: Robust improvement of SAM for medical imaging. *IEEE Transactions on Medical Imaging*, 45(5):1974–1985, 2026. doi: 10.1109/TMI.2025.3644811.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, 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. In *International Conference on Learning Representations*, pp. 138846–138923, 2026.

Wei-Ting Chen, Yu-Jiet Vong, Sy-Yen Kuo, Sizhou Ma, and Jian Wang. RobustSAM: Segment anything robustly on degraded images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4081–4091, 2024.

Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. FocalClick: Towards practical interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1300–1309, 2022.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Christoffer Edlund, Timothy R. Jackson, Nabeel Khalid, Nicola Bevan, Timothy Dale, Andreas Dengel, Sheraz Ahmed, Johan Trygg, and Rickard Sjögren. LIVECell: A large-scale dataset for label-free live cell segmentation. *Nature Methods*, 18:1038–1045, 2021. doi: 10.1038/s41592-021-01249-6.

Qi Fan, Xin Tao, Lei Ke, Mingqiao Ye, Di Zhang, Pengfei Wan, Yu-Wing Tai, and Chi-Keung Tang. Stable segment anything model. In *International Conference on Learning Representations*, pp. 48170–48183, 2025.

Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. PanNuke: An open pan-cancer histology dataset for nuclei instance segmentation and classification. In *Digital Pathology*, pp. 11–19, 2019. doi: 10.1007/978-3-030-23937-4_2.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D. Johansen. Kvasir-SEG: A segmented polyp dataset. In *MultiMedia Modeling*, pp. 451–462, 2020. doi: 10.1007/978-3-030-37734-2_37.

Timo Kaiser, Thomas Norrenbrock, and Bodo Rosenhahn. UncertainSAM: Fast and efficient uncertainty quantification of the segment anything model. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 28670–28688, 2025.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *Advances in Neural Information Processing Systems*, volume 36, pp. 29914–29934, 2023.

Ismail Hakki Kinalioglu. Structural prompt-response consistency for label-free failure screening in promptable medical segmentation. *Pattern Recognition Letters*, 209:60–64, 2026. doi: 10.1016/j.patrec.2026.08.026.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023. doi: 10.1109/ICCV51070.2023.00371.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

Shuang Liang, Zeqing Wang, Yuxian Li, Xihui Liu, and Han Wang. From pixels to concepts: Do segmentation models understand what they segment? *arXiv preprint arXiv:2605.09591*, 2026. URL https://arxiv.org/abs/2605.09591.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48.

Yuqi Lin, Hengjia Li, Wenqi Shao, Zheng Yang, Jun Zhao, Xiaofei He, Ping Luo, and Kaipeng Zhang. SAMRefiner: Taming segment anything model for universal mask refinement. In *International Conference on Learning Representations*, pp. 57368–57392, 2025.

Kangning Liu, Brian Price, Jason Kuen, Yifei Fan, Zijun Wei, Luis Figueroa, Krzysztof J. Geras, and Carlos Fernandez-Granda. Uncertainty-aware fine-tuning of segmentation foundation models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 53317–53389, 2024. doi: 10.52202/079017-1690.

Qin Liu, Zhenlin Xu, Gedas Bertasius, and Marc Niethammer. SimpleClick: Interactive image segmentation with simple vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22290–22300, 2023.

Xueyu Liu, Xiaoyi Zhang, Meilin Liu, Guangze Shi, Jia Shen, Yujie Wang, Cai Zhao, Ziyuan He, Yongfei Wu, Mingqiang Wei, and Yongle Chen. Attack for defense: Adversarial agents for point prompt optimization empowering segment anything model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6591–6600, 2026.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:654, 2024. doi: 10.1038/s41467-024-44824-z.

Andrey Moskalenko, Vlad Shakhuro, Anna Vorontsova, Anton Konushin, Anton Antonov, Alexander Krapukhin, Denis Shepelev, and Konstantin Soshin. TETRIS: Towards exploring the robustness of interactive segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5): 4287–4295, 2024. doi: 10.1609/aaai.v38i5.28225.

Andrey Moskalenko, Danil Kuznetsov, Irina Dudko, Anastasiia Iasakova, Nikita Boldyrev, Denis Shepelev, Andrei Spiridonov, Andrey Kuznetsov, and Vlad Shakhuro. BREPS: Bounding-box robustness evaluation of promptable segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(10):8107–8115, 2026. doi: 10.1609/aaai.v40i10.37757.

Josh Myers-Dean, Kangning Liu, Brian Price, Yifei Fan, Jason Kuen, and Danna Gurari. conSAMmé: Achieving consistent segmentations with SAM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 759–768, 2025.

Shengyang Ping, Zhijie Lin, Liliang Lin, Lei Zhao, Lisha Ye, Bangguo Wang, and Tao Wang. A prompt-preserving SAM ViT-B adaptation framework with historical branch fusion and soft convolutional expert weighting for medical image segmentation. *Bioengineering*, 13(8):914, 2026. doi: 10.3390/bioengineering13080914.

Md Mostafijur Rahman, Mustafa Munir, Debesh Jha, Ulas Bagci, and Radu Marculescu. PP-SAM: Perturbed prompts for robust adaption of segment anything model for polyp segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4989–4995, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, pp. 28085–28128, 2025.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442.

Leo Richter, Xuanli He, Pasquale Minervini, and Matt J. Kusner. An auditing test to detect behavioral shift in language models. In *International Conference on Learning Representations*, pp. 49313–49339, 2025.

Javier Gamazo Tejero, Moritz J. Schmid, Pablo Márquez Neila, Martin Zinkernagel, Sebastian Wolf, and Raphael Sznitman. SAM-DA: Decoder adapter for efficient medical domain adaptation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6775–6784, 2025. doi: 10.1109/WACV61041.2025.00659.

Jinfeng Wang, Sifan Song, Xinkun Wang, Yiyi Wang, Yiyi Miao, Jionglong Su, and S. Kevin Zhou. ProMISe: Promptable medical image segmentation using SAM. *arXiv preprint arXiv:2403.04164*, 2024. URL https://arxiv.org/abs/2403.04164.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7959–7971, 2022.

Cilin Yan, Haochen Wang, Jie Liu, Xiaolong Jiang, Yao Hu, Xu Tang, Guoliang Kang, and Efstratios Gavves. PiClick: Picking the desired mask from multiple candidates in click-based interactive segmentation. *Neurocomputing*, 599:128083, 2024. doi: 10.1016/j.neucom.2024.128083.

# A AUDIT PROTOCOL AND EXPERIMENTAL INVENTORY

Figure S1 summarizes the audit protocol: canonical-only specialization, the paired prompt changes applied to reference and specialized endpoints, and the relations and statistics that SAMLens reports.

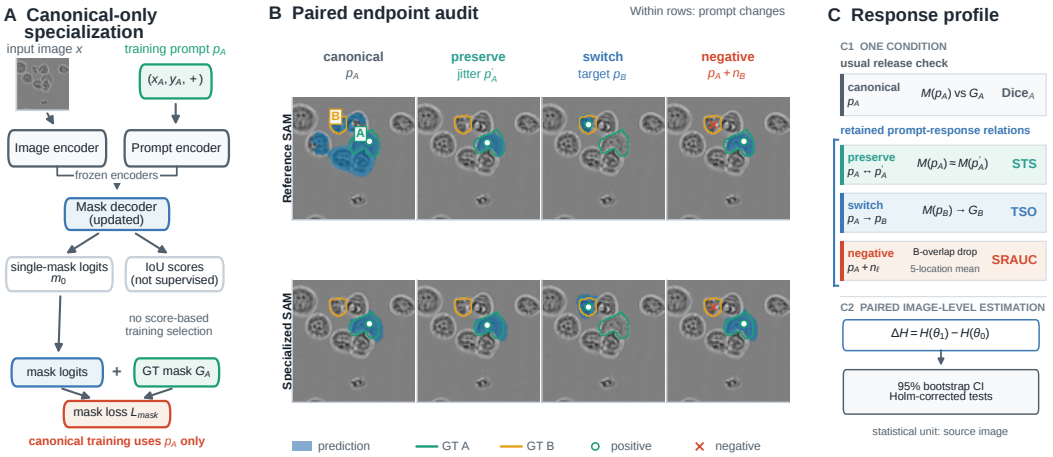



Figure S1: **Audit protocol.** (A) Canonical-only specialization applies the mask loss to the single-mask logits at $p_A$ against $G_A$; IoU scores do not select the training output. (B) Reference and specialized SAM (partial tuning, official split) on one LIVECell scene; within each row only the prompt changes. (C) The relations and the paired image-level estimation used throughout.

## A.1 METRIC ROLES AND PROMPT CONSTRUCTION

Table S1 separates the four headline axes from quality controls, structural diagnostics, and domain-specific companion measures. Canonical quality evaluates the supervised condition; the other three axes compare outputs under the declared prompt transformations.

Table S1: **Metric roles and interpretation boundaries.** The four headline axes answer distinct relation questions; companion and structural metrics constrain interpretation rather than adding independent headline claims.

| Role | Relation | Transformation | Estimand | Interpretation | Boundary |
|---|---|---|---|---|---|
| headline | canonical quality | positive point on target A | Dice_A | quality at the supervised condition | not the retained interface |
| headline | same-target preserve | move the positive point inside A | STS | selected-mask preservation | pair with STS_target; stable-wrong masks exist |
| headline | target switch | move the positive point from A to B | TSO | continuous switch toward B | SwitchSuccess is a companion threshold |
| headline | spatial negative response | add a negative point at six locations | SRAUC | marginal spatial response | background locations diagnose influence, not unconditional suppression |
| quality control | target-conditioned preserve | same-target point movement | STS_target | quality retained on target A | does not replace output stability |
| structural | candidate diversity | same prompt, K returned masks | MMD | binary-mask diversity | lower diversity is not necessarily harmful |
| companion | single-point negative response | one object-internal negative point | NPS | ratio-derived suppression direction | unbounded; do not compare raw magnitudes |
| domain-specific | single-target leakage | background point outside target | TRNP / NPL / BHR | location-dependent target recovery | synthetic Kvasir protocol; not clinical risk |

**Companion definitions.** Let $M_A, M'_A, M_B$ be the selected masks under $p_A, p'_A, p_B$. Target-conditioned stability, the quality control used with STS, is

$$\text{STS}_{target} = \min\{\text{IoU}(M_A, G_A), \text{IoU}(M'_A, G_A)\}.$$

SwitchSuccess is $\mathbf{1}\{\mathrm{IoU}(M_B, G_B) > \mathrm{IoU}(M_B, G_A)\}$, with ties counted as failure. For negative-point suppression (NPS), let $a = |M_A \cap G_B|$ and $b = |M_{A+n_B} \cap G_B|$. The implementation uses

$$\mathrm{NPS} = \begin{cases} (a-b)/a, & a > 0, \\ 0, & a = b = 0, \\ -\min\{1, b/\max(|G_B|, 10^{-8})\}, & a = 0,\ b > 0. \end{cases}$$

This baseline-relative quantity differs from the bounded overlap contrast in SRAUC. Multi-mask diversity (MMD), for $K \geq 2$ returned candidates, is $\mathrm{MMD} = 1 - \binom{K}{2}^{-1} \sum_{k<l} \mathrm{IoU}(m_k, m_l)$; it is zero for $K < 2$. For a background-point mask $M_{bg}$, non-target prompt leakage (NPL) is $\mathrm{NPL} = \max_k \mathrm{IoU}(M_{bg}, G_k)$ over annotated targets, or zero if there are none. Background hallucination rate (BHR), $\mathrm{BHR} = \mathbf{1}\{|M_{bg}| > 0.01HW \ \wedge \ \mathrm{NPL} > 0.1\}$, uses strict inequalities and image dimensions $H, W$. For single-target images, target recovery under a non-target prompt (TRNP) is $\mathrm{TRNP} = \mathrm{Dice}(M_{bg}, G_A)$; near-background versions replace only the prompt. The anonymous artifact guide locates the corresponding metric implementations.

Positive points are distance-transform maxima inside the requested instance. The preserve point is a second interior point sampled under a frozen same-instance rule. The switch point lies inside the paired instance. The negative-response field contains the distractor core and edge, three background locations along the distractor-to-target corridor, and a separate far-background control. All masks used to construct prompts are annotations; no endpoint prediction enters prompt placement.

The spatial prompt set requires all six locations before any endpoint is evaluated. The frozen LIVECell manifest retains 387 of 400 pairs on all 100 images: 12 pairs lack three ordered background corridor locations and one has non-strictly ordered locations. COCO retains 343 of 344 pairs on 122 of 123 images; one pair lacks three ordered background locations. These are geometry-construction exclusions, not inference failures. Reference and specialized endpoints use the same retained prompt bank. Zero inference fallback/skips is reported after this construction stage. Both construction manifests are included in the anonymous artifact.

Figure 2 illustrates the spatial construction beside the main-text metric definitions. Each query contains the same positive point and one negative point; the ordered locations are not accumulated clicks.

### A.2 Cohorts, endpoints, and separation

Table S2 lists the source-image counts and split boundaries. Multiple instance pairs from one image are never treated as independent observations. Controlled adaptations use disjoint training and audit images; released specialists retain the provenance boundary stated in Table S13.

Table S2: **Cohort inventory and separation.** Source images, not prompt pairs, are the inferential units throughout. RClicks is observational and has no training cohort. The LIVECell legacy pool trains the SAM endpoints behind the negative-point analyses of Section 5 (SAM LoRA, partial, and full-decoder endpoints), the SRAUC column of Table S5, Figure S3, the scorer analysis (Appendix E), and the rows marked legacy in Table S6.

| Cohort | Train source | Train images | Audit source | Images | Units | Separation |
|---|---|---|---|---|---|---|
| LIVECell | official train split | 200 (199–200 observed) | official test split | 100 | 400 | zero recorded image overlap |
| LIVECell legacy | official test split, audit images excluded | 200 (196–199 observed) | same 100 test images | 100 | 400 | zero recorded image overlap |
| PanNuke | fold 1 | 199 after filtering | fold 3 | 93 | 349 | official disjoint folds |
| COCO | train2017 | 200 | val2017 | 123 | 344 | zero recorded image overlap |
| Kvasir | three disjoint small pools | 3 x 8 | frozen Kvasir100 | 100 | 100 | zero pool/evaluation overlap |
| RClicks COCO-MVal | none | – | released first clicks | 722 | 800 targets / 38,097 clicks | public observational click coordinates |

Table S3: **Original relation-audit endpoints.** Controlled adaptations use canonical mask supervision; the negative-point endpoint inventory is in Table S17. The released micro-sam model is external evidence, not a budget-matched competitor; SAM3 failed induction and was not effect-evaluated.

| Setting | Updated modules | Params | Objective | Budget | Seeds | Gate |
|---|---|---|---|---|---|---|
| LIVECell / SAM partial | mask tokens + hypernetworks + upscaling | 634,208 | Dice + BCE on canonical p_A | 4,000 steps / 200 images | 42 (main official-train row) | pass |
| PanNuke / SAM partial | mask tokens + hypernetworks + upscaling | 634,208 | Dice + BCE on canonical p_A | step-normalized n=200 | 42, 123, 456 | pass |
| LIVECell / SAM2 partial | mask tokens + hypernetworks + upscaling | 634,208 | Dice + BCE on canonical p_A | 4,000 steps / 200 images | 42, 123, 456 | pass |
| COCO / SAM partial | mask tokens + hypernetworks + upscaling | 634,208 | Dice + BCE on canonical p_A | 4,000 steps / 200 images | 42, 123, 456 | pass |
| Kvasir / SAM partial | mask tokens + hypernetworks + upscaling | 634,208 | Dice + BCE on canonical p_A | 200 steps / 8 images per pool | 3 pools x 3 seeds | pass |
| Kvasir / SAM full decoder | entire mask decoder | 4,058,340 | Dice + BCE on canonical p_A | 80 steps / 8 images per pool | 3 pools x 3 seeds | pass |
| LIVECell / micro-sam specialist | released checkpoint | not controlled here | upstream mixed prompt training | upstream release | not published | external endpoint |
| SAM3 stopped adaptation | output upscaling only | 73,952 | canonical mask loss | 4,000 steps / 200 images | frozen induction run | fail: 8.55% $<$ 10% |

For the controlled specializations in Table S5, the endpoint is the final checkpoint from a frozen budget, not the best checkpoint along a training trajectory. Their objective is Dice plus BCE under $p_A$; prompt-response outcomes do not enter training or checkpoint selection. Partial tuning updates mask tokens, output hypernetworks, and output upscaling while leaving the image encoder, prompt encoder, two-way transformer, and IoU prediction head fixed. Full decoder tuning changes the complete mask decoder except that a mask-only loss still provides no gradient to the IoU head, as verified in Appendix E.

Separate repair controls use prompt-contract-preserving adaptation (PCPA), which adds losses on transformed prompts. Compact repair restricts the trainable decoder modules. These controls are not the canonical-only specializations in Table S5.

# B Conditional estimand and statistical inference

## B.1 Successful-specialization conditioning

Controlled checkpoints use frozen training budgets and enter confirmatory interpretation only after improving held-out canonical Dice. This gate conditions the estimand on successful specialization:

$$\mathbb{E}[\Delta h \mid \Delta \,\mathrm{Dice}_A > 0],$$

where $h$ is a prespecified prompt-response relation. It is never applied to a prompt-response outcome and does not estimate the unconditional effect of an arbitrary adaptation run. Failed fit gates remain in the result ledger but their held-out relation outcomes are not evaluated. The gate and effect audit use the same held-out bank; we therefore make the conditional scope explicit rather than interpreting the reported Dice change as an independently selected effect.

## B.2 Image-cluster resampling and multiplicity

Let $d_{ij}$ denote the adapted-minus-reference effect for prompt unit $j$ in source image $i$. We first compute $\bar{d}_i = n_i^{-1} \sum_j d_{ij}$, then estimate the cohort effect from the mean of $\bar{d}_i$. Confidence intervals resample source images. Paired sign-flip tests flip the image-level effects, and Holm correction is applied within each frozen inferential family. Two multi-seed procedures remain distinct: the spatial-response and quality-decomposition families first average the observed seeds and then resample images; the PanNuke/SAM2 canonical, STS, and TSO replications resample seeds and images. The absolute-overlap control also uses seed–image resampling. The first procedure conditions on the observed seeds and does not estimate between-seed uncertainty. Neither treats pair or seed–image rows as independent observations.

Within a Holm-tested family, a row is *confirmed* only when its interval excludes zero and the corrected test passes. PanNuke/SAM2 multi-seed replication reports provide intervals but no additional

Holm tests; their intervals are not relabeled as corrected evidence. Table S16 identifies both cases. Inconclusive rows do not establish equivalence.

Table S4: **Frozen inferential families.** Each listed family uses 10,000 bootstrap draws and 20,000 sign flips. Confirmed counts refer to the listed family, not the four displayed headline axes. COCO's four-test family includes MMD; its spatial analysis is a separate two-test family. CI-only multi-seed reports are identified in Table S16.

| Family | Members | Resampling | Holm $m$ | Confirmed |
|---|---|---|---|---|
| core relation audit | core metrics and companions | image | 102 | 69 |
| spatial response | 5 contrasts x 3 spatial measures | image; seeds averaged | 15 | 15 |
| external micro-sam | 2 references x 5 measures | image | 10 | 9 |
| COCO same-domain | Dice, STS, TSO, MMD | seed x image | 4 | 3 |
| COCO spatial | SRAUC, quality90 candidate term | seed x image | 2 | 1 |
| Kvasir train-pool replication | 2 variants x 5 measures | seed x image | 10 | 10 |
| RClicks first-click audit | 7 first-click contrasts | image | 7 | 5 |

NPL and BHR fail this standard in most adaptation contrasts and are therefore companion diagnostics. Six metrics in the original core audit (Dice$_A$, STS, TSO, SwitchSuccess, NPS, and MMD) survive image-level clustering. The main paper uses the bounded spatial-response field in place of single-point NPS where the required predictions are available.

# C Extended held-out effects and prompt-policy sensitivity

## C.1 Response profile of the five original settings

Table S5 and Figure S2 report the four headline relations for the four controlled partial adaptations and the released micro-sam LIVECell specialist that formed the original audit.

Table S5: **Response profile of the original settings.** Specialized minus reference on the same images and prompts. For LIVECell SAM, Dice/STS/TSO use official-train seed 42, but SRAUC uses the three-seed legacy pool (Table S16). $n/s$: source images / seeds for Dice, STS, and TSO. Dice levels pool prompt units; paired effects weight source images equally.

| | | **Canonical quality** | | **Retained prompt relations after specialization** | | |
|---|---|---|---|---|---|---|
| Setting / endpoint | $n$/s | Dice ref.→spec. | ΔDice [95% CI] | Preserve ΔSTS<br>target $\Delta STS_{target}$ | Switch ΔTSO | Spatial ΔSRAUC |
| **LIVECell / SAM**<br>partial | 100/1 | 0.308→0.693 | +0.385<br>[+0.356, +0.412] | −0.300 ↓<br>tgt +0.263 ↑ | +0.387 ↑ | −0.090 ↓ |
| **PanNuke / SAM**<br>partial | 93/3 | 0.713→0.829 | +0.110<br>[+0.077, +0.144] | −0.112 ↓<br>tgt +0.068 ↑ | +0.112 ↑ | −0.074 ↓ |
| **LIVECell / SAM2**<br>partial | 100/3 | 0.302→0.693 | +0.391<br>[+0.364, +0.418] | −0.274 ↓<br>tgt +0.271 ↑ | +0.395 ↑ | −0.126 ↓ |
| **COCO / SAM**<br>partial | 123/3 | 0.674→0.744 | +0.070<br>[+0.048, +0.094] | +0.018°<br>tgt +0.080 | +0.076 ↑ | -0.004° |
| **LIVECell / micro-sam**†<br>released specialist | 100/– | 0.308→0.786 | +0.477<br>[+0.438, +0.516] | −0.080 ↓<br>tgt +0.425 | +0.511 ↑ | −0.086 ↓ |

↑ / ↓ mark 95% CIs that exclude zero and a stronger / weaker measured relation, not benefit or harm. ° inconclusive. Gray $STS_{target}$ values are reported without a test. †Released LIVECell specialist, not trained by us.

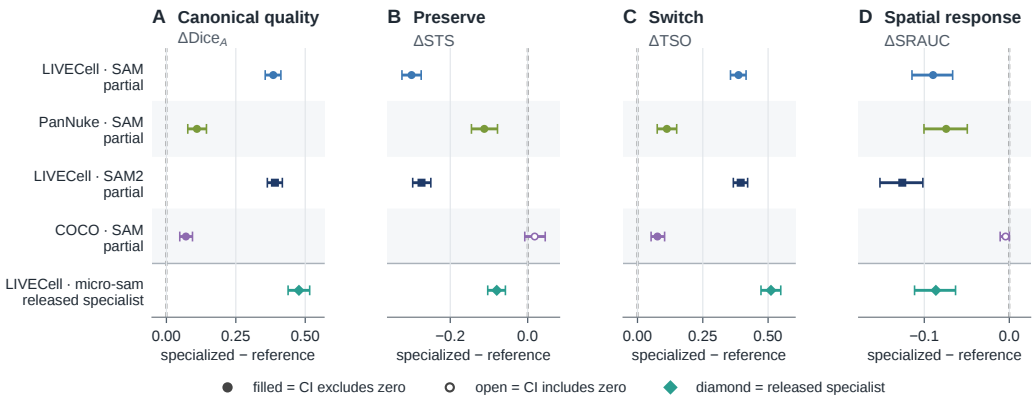



Figure S2: **Effects of Table S5 with 95% CIs.** Filled markers: CI excludes zero; open markers: CI includes zero; diamonds: released micro-sam specialist.

### C.2 Extended inventory

Table S6 retains 47 frozen inventory rows, including single-seed contrasts and the COCO intervals that cross zero. It is not a complete mapping of the multi-seed summaries in Table S5: Table S16 supplies that mapping with original intervals, seed sets, valid sample counts, test families, and source-report keys. Values are specialized-minus-reference effects, not a leaderboard across unrelated training budgets.

Table S6: **Extended frozen effect inventory.** These 47 rows retain the original image-level and COCO reports, including single-seed contrasts. They do not replace the multi-seed estimates of Table S5; Table S16 maps every headline effect of that table to its source and inferential unit. Inconclusive rows do not establish equivalence. LIVECell100 legacy endpoints train on test-split images outside the audit set (Table S2).

| Cohort | Contrast | Metric | $n$/units | Δ [95% CI] | Holm $p$ | Status | Unit |
|---|---|---|---|---|---|---|---|
| COCO val2017 | base → partial (3 seeds) | Dice$_A$ | 123/344 | +0.070 [+0.048, +0.094] | 0.0002 | **confirmed** | seed x image |
| COCO val2017 | base → partial (3 seeds) | STS | 123/344 | +0.018 [-0.008, +0.045] | 0.1231 | inconclusive | seed x image |
| COCO val2017 | base → partial (3 seeds) | TSO | 123/344 | +0.076 [+0.052, +0.104] | 0.0002 | **confirmed** | seed x image |
| COCO val2017 | base → partial (3 seeds) | SRAUC | 122/343 | -0.004 [-0.011, +0.000] | 0.0718 | inconclusive | seed x image |
| Kvasir100 | base → BG-rejection repair | Dice$_A$ | 100/100 | +0.059 [+0.037, +0.085] | 0.0051 | **confirmed** | image cluster |
| Kvasir100 | base → BG-rejection repair | STS | 100/100 | -0.016 [-0.036, +0.006] | 1.0000 | inconclusive | image cluster |
| Kvasir100 | base → Kvasir partial | Dice$_A$ | 100/100 | +0.036 [+0.017, +0.058] | 0.0051 | **confirmed** | image cluster |
| Kvasir100 | base → Kvasir partial | STS | 100/100 | -0.011 [-0.027, +0.007] | 1.0000 | inconclusive | image cluster |
| Kvasir100 | base → Kvasir full decoder | Dice$_A$ | 100/100 | +0.195 [+0.123, +0.271] | 0.0051 | **confirmed** | image cluster |
| Kvasir100 | base → Kvasir full decoder | STS | 100/100 | +0.117 [+0.040, +0.197] | 0.1638 | directional | image cluster |
| LIVECell100 | SAM base → SAM partial | SRAUC | 100/387 | -0.090 [-0.114, -0.067] | 0.0007 | **confirmed** | image cluster |
| LIVECell100 | SAM partial → post-adapt repair | SRAUC | 100/387 | +0.014 [+0.011, +0.018] | 0.0007 | **confirmed** | image cluster |
| LIVECell100 | SAM2 base → SAM2 partial | SRAUC | 100/387 | -0.126 [-0.152, -0.102] | 0.0007 | **confirmed** | image cluster |
| LIVECell100 external | base → micro-sam specialist | Dice$_A$ | 100/400 | +0.477 [+0.438, +0.516] | 0.0005 | **confirmed** | image cluster |
| LIVECell100 external | base → micro-sam specialist | STS | 100/400 | -0.080 [-0.103, -0.058] | 0.0005 | **confirmed** | image cluster |

*Continued on next page*

*Continued from previous page*

| Cohort | Contrast | Metric | $n$/units | $\Delta$ [95% CI] | Holm $p$ | Status | Unit |
|---|---|---|---|---|---|---|---|
| LIVECell100 external | base → micro-sam specialist | TSO | 100/400 | +0.511 [+0.473, +0.549] | 0.0005 | **confirmed** | image cluster |
| LIVECell100 external | base → micro-sam specialist | SRAUC | 100/387 | -0.086 [-0.111, -0.063] | 0.0005 | **confirmed** | image cluster |
| LIVECell100 external | micro-sam LM → micro-sam specialist | $\text{Dice}_A$ | 100/400 | +0.082 [+0.063, +0.103] | 0.0005 | **confirmed** | image cluster |
| LIVECell100 external | micro-sam LM → micro-sam specialist | STS | 100/400 | +0.003 [-0.023, +0.030] | 0.8399 | inconclusive | image cluster |
| LIVECell100 external | micro-sam LM → micro-sam specialist | TSO | 100/400 | +0.102 [+0.079, +0.125] | 0.0005 | **confirmed** | image cluster |
| LIVECell100 external | micro-sam LM → micro-sam specialist | SRAUC | 100/387 | -0.026 [-0.041, -0.013] | 0.0009 | **confirmed** | image cluster |
| LIVECell100 official | base → LoRA r4 (s42) | $\text{Dice}_A$ | 100/400 | +0.399 [+0.366, +0.432] | 0.0051 | **confirmed** | image cluster |
| LIVECell100 official | base → LoRA r4 (s42) | STS | 100/400 | -0.295 [-0.323, -0.267] | 0.0051 | **confirmed** | image cluster |
| LIVECell100 official | base → LoRA r4 (s42) | TSO | 100/400 | +0.399 [+0.369, +0.428] | 0.0051 | **confirmed** | image cluster |
| LIVECell100 official | base → partial (s42) | $\text{Dice}_A$ | 100/400 | +0.385 [+0.356, +0.412] | 0.0051 | **confirmed** | image cluster |
| LIVECell100 official | base → partial (s42) | STS | 100/400 | -0.300 [-0.325, -0.275] | 0.0051 | **confirmed** | image cluster |
| LIVECell100 official | base → partial (s42) | TSO | 100/400 | +0.387 [+0.356, +0.416] | 0.0051 | **confirmed** | image cluster |
| LIVECell100 legacy | base → LoRA r4 (s42) | $\text{Dice}_A$ | 100/400 | +0.260 [+0.231, +0.290] | 0.0051 | **confirmed** | image cluster |
| LIVECell100 legacy | base → LoRA r4 (s42) | STS | 100/400 | -0.234 [-0.258, -0.210] | 0.0051 | **confirmed** | image cluster |
| LIVECell100 legacy | base → LoRA r4 (s42) | TSO | 100/400 | +0.255 [+0.228, +0.282] | 0.0051 | **confirmed** | image cluster |
| LIVECell100 legacy | base → partial (s42) | $\text{Dice}_A$ | 100/400 | +0.236 [+0.216, +0.257] | 0.0051 | **confirmed** | image cluster |
| LIVECell100 legacy | base → partial (s42) | STS | 100/400 | -0.201 [-0.221, -0.180] | 0.0051 | **confirmed** | image cluster |
| LIVECell100 legacy | base → partial (s42) | TSO | 100/400 | +0.235 [+0.213, +0.256] | 0.0051 | **confirmed** | image cluster |
| LIVECell100 legacy | base → full decoder (s42) | $\text{Dice}_A$ | 100/400 | +0.384 [+0.351, +0.418] | 0.0051 | **confirmed** | image cluster |
| LIVECell100 legacy | base → full decoder (s42) | STS | 100/400 | -0.351 [-0.377, -0.324] | 0.0051 | **confirmed** | image cluster |
| LIVECell100 legacy | base → full decoder (s42) | TSO | 100/400 | +0.378 [+0.348, +0.408] | 0.0051 | **confirmed** | image cluster |
| LIVECell100 legacy | partial (s42) → compact repair (s42) | $\text{Dice}_A$ | 100/400 | +0.076 [+0.061, +0.091] | 0.0051 | **confirmed** | image cluster |
| LIVECell100 legacy | partial (s42) → compact repair (s42) | STS | 100/400 | -0.064 [-0.075, -0.052] | 0.0051 | **confirmed** | image cluster |
| LIVECell100 legacy | partial (s42) → compact repair (s42) | TSO | 100/400 | +0.084 [+0.069, +0.099] | 0.0051 | **confirmed** | image cluster |
| PanNuke fold 3 | base → LoRA r4 (s42) | $\text{Dice}_A$ | 93/349 | +0.105 [+0.072, +0.139] | 0.0051 | **confirmed** | image cluster |
| PanNuke fold 3 | base → LoRA r4 (s42) | STS | 93/349 | -0.139 [-0.165, -0.112] | 0.0051 | **confirmed** | image cluster |
| PanNuke fold 3 | base → LoRA r4 (s42) | TSO | 93/349 | +0.115 [+0.075, +0.154] | 0.0051 | **confirmed** | image cluster |
| PanNuke fold 3 | base → partial (s42) | $\text{Dice}_A$ | 93/349 | +0.104 [+0.069, +0.139] | 0.0051 | **confirmed** | image cluster |
| PanNuke fold 3 | base → partial (s42) | STS | 93/349 | -0.137 [-0.161, -0.112] | 0.0051 | **confirmed** | image cluster |
| PanNuke fold 3 | base → partial (s42) | TSO | 93/349 | +0.107 [+0.068, +0.146] | 0.0051 | **confirmed** | image cluster |

*Continued on next page*

*Continued from previous page*

| Cohort | Contrast | Metric | $n$/units | Δ [95% CI] | Holm $p$ | Status | Unit |
|---|---|---|---|---|---|---|---|
| PanNuke fold 3 | SAM base → SAM partial | SRAUC | 93/349 | -0.074 [-0.100, -0.049] | 0.0007 | **confirmed** | image cluster |
| PanNuke fold 3 | SAM2 base → SAM2 partial | SRAUC | 93/349 | -0.058 [-0.083, -0.035] | 0.0007 | **confirmed** | image cluster |

Table S7 reports target-conditioned stability, thresholded switch success, single-point negative response, and candidate diversity. These measures constrain the interpretation of a headline effect. For example, lower STS with higher $\text{STS}_{target}$ means that moving the point changes the selected mask more even though target-conditioned quality improves.

Table S7: **Companion and structural effects.** Asterisks mark rows that pass both the image-cluster interval and the project-level Holm gate; open circles are inconclusive. NPS is ratio-derived, MMD is structural, and generic NPL/BHR rows are not headline outcomes.

| Cohort | Variant | $\Delta\text{STS}_{target}$ | ΔSwitch | ΔNPS | ΔMMD | ΔNPL | ΔBHR |
|---|---|---|---|---|---|---|---|
| Kvasir100 | BG-rejection repair | $+0.045*$ | – | – | $-0.198*$ | $-0.039*$ | $-0.180*$ |
| Kvasir100 | Kvasir partial | $+0.026^{\circ}$ | – | – | $-0.138*$ | $-0.005^{\circ}$ | $-0.030^{\circ}$ |
| Kvasir100 | Kvasir full decoder | $+0.192*$ | – | – | $-0.345*$ | $-0.013^{\circ}$ | $-0.070^{\circ}$ |
| LIVECell100 official | LoRA r4 (s42) | $+0.274*$ | $+0.302*$ | $-0.780*$ | $-0.417*$ | $-0.013^{\circ}$ | $-0.013^{\circ}$ |
| LIVECell100 official | partial (s42) | $+0.263*$ | $+0.305*$ | $-0.564*$ | $-0.308*$ | $-0.012^{\circ}$ | $-0.013^{\circ}$ |
| LIVECell100 legacy | LoRA r4 (s42) | $+0.156*$ | $+0.265*$ | $-1.721*$ | $-0.210*$ | $-0.005^{\circ}$ | $-0.007^{\circ}$ |
| LIVECell100 legacy | partial (s42) | $+0.163*$ | $+0.280*$ | $-2.630*$ | $-0.272*$ | $-0.006^{\circ}$ | $-0.013^{\circ}$ |
| LIVECell100 legacy | full decoder (s42) | $+0.238*$ | $+0.295*$ | $-12.709*$ | $-0.402*$ | $+0.007^{\circ}$ | $-0.007^{\circ}$ |
| LIVECell100 legacy | compact repair (s42) | $+0.049*$ | $+0.007^{\circ}$ | $+2.450*$ | $+0.126*$ | $-0.004^{\circ}$ | $+0.000^{\circ}$ |
| PanNuke fold 3 | LoRA r4 (s42) | $+0.052^{\circ}$ | $+0.073*$ | $-0.193*$ | $-0.499*$ | $-0.017*$ | $-0.027^{\circ}$ |
| PanNuke fold 3 | partial (s42) | $+0.052^{\circ}$ | $+0.073*$ | $-0.218*$ | $-0.595*$ | $-0.016^{\circ}$ | $-0.013^{\circ}$ |

### C.3 POSITIVE-POINT POLICY REPLICATION

The center-point protocol is repeated with random-interior and boundary-interior policies, each using three frozen prompt seeds. Table S8 reports the range across policies. LIVECell directions are stable, whereas PanNuke full tuning contains a boundary-policy Dice reversal. This is a scope boundary: the PanNuke center-point result does not establish policy invariance.

Table S8: **Prompt-policy robustness.** Each row summarizes distance-transform center, random-interior, and boundary-interior policies, each repeated across three prompt seeds. PanNuke Dice/TSO vary with prompt policy; full tuning reverses the Dice gain under the boundary-interior policy.

| Cohort | Comparison | Metric | Pass | Δ range | Boundary Δ | Verdict |
|---|---|---|---|---|---|---|
| LIVECell100 | SAM partial | MMD | 3/3 | [-0.219, -0.166] | -0.166 | policy-robust |
| LIVECell100 | SAM partial | NPS | 3/3 | [-3.057, -2.354] | -2.787 | policy-robust |
| LIVECell100 | SAM partial | STS | 3/3 | [-0.373, -0.253] | -0.373 | policy-robust |
| LIVECell100 | SAM partial | Switch | 3/3 | [+0.287, +0.292] | +0.292 | policy-robust |
| LIVECell100 | SAM partial | TSO | 3/3 | [+0.146, +0.307] | +0.146 | policy-robust |
| LIVECell100 | SAM partial | $\text{Dice}_A$ | 3/3 | [+0.154, +0.303] | +0.154 | policy-robust |
| LIVECell100 | SAM full decoder | MMD | 3/3 | [-0.402, -0.348] | -0.348 | policy-robust |
| LIVECell100 | SAM full decoder | NPS | 3/3 | [-15.086, -11.860] | -11.860 | policy-robust |
| LIVECell100 | SAM full decoder | STS | 3/3 | [-0.422, -0.348] | -0.422 | policy-robust |
| LIVECell100 | SAM full decoder | Switch | 3/3 | [+0.284, +0.296] | +0.296 | policy-robust |
| LIVECell100 | SAM full decoder | TSO | 3/3 | [+0.166, +0.378] | +0.166 | policy-robust |
| LIVECell100 | SAM full decoder | $\text{Dice}_A$ | 3/3 | [+0.176, +0.384] | +0.176 | policy-robust |
| LIVECell100 | SAM2 partial | MMD | 3/3 | [-0.344, -0.243] | -0.243 | policy-robust |
| LIVECell100 | SAM2 partial | STS | 3/3 | [-0.405, -0.283] | -0.405 | policy-robust |
| LIVECell100 | SAM2 partial | Switch | 3/3 | [+0.309, +0.317] | +0.309 | policy-robust |
| LIVECell100 | SAM2 partial | TSO | 3/3 | [+0.181, +0.395] | +0.181 | policy-robust |
| LIVECell100 | SAM2 partial | $\text{Dice}_A$ | 3/3 | [+0.205, +0.391] | +0.205 | policy-robust |
| PanNuke fold 3 | SAM partial | MMD | 3/3 | [-0.588, -0.500] | -0.500 | policy-robust |
| PanNuke fold 3 | SAM partial | NPS | 3/3 | [-1.316, -0.777] | -0.804 | policy-robust |
| PanNuke fold 3 | SAM partial | STS | 3/3 | [-0.150, -0.123] | -0.150 | policy-robust |
| PanNuke fold 3 | SAM partial | Switch | 3/3 | [+0.073, +0.114] | +0.114 | policy-robust |
| PanNuke fold 3 | SAM partial | TSO | 2/3 | [-0.006, +0.112] | -0.006 | policy-sensitive |
| PanNuke fold 3 | SAM partial | $\text{Dice}_A$ | 2/3 | [+0.031, +0.110] | +0.031 | policy-sensitive |
| PanNuke fold 3 | SAM full decoder | MMD | 3/3 | [-0.688, -0.613] | -0.613 | policy-robust |
| PanNuke fold 3 | SAM full decoder | NPS | 3/3 | [-3.738, -1.871] | -3.738 | policy-robust |
| PanNuke fold 3 | SAM full decoder | STS | 3/3 | [-0.279, -0.221] | -0.279 | policy-robust |
| PanNuke fold 3 | SAM full decoder | Switch | 3/3 | [+0.073, +0.115] | +0.115 | policy-robust |
| PanNuke fold 3 | SAM full decoder | TSO | 1/3 | [-0.125, +0.126] | -0.125 | policy-sensitive |
| PanNuke fold 3 | SAM full decoder | $\text{Dice}_A$ | 1/3 | [-0.088, +0.122] | -0.088 | policy-sensitive |

## D SPATIAL RESPONSE AND RETURNED-CANDIDATE DECOMPOSITION

Figure S3 shows the response-range compression behind the SRAUC decline after partial tuning (Section 5.4).

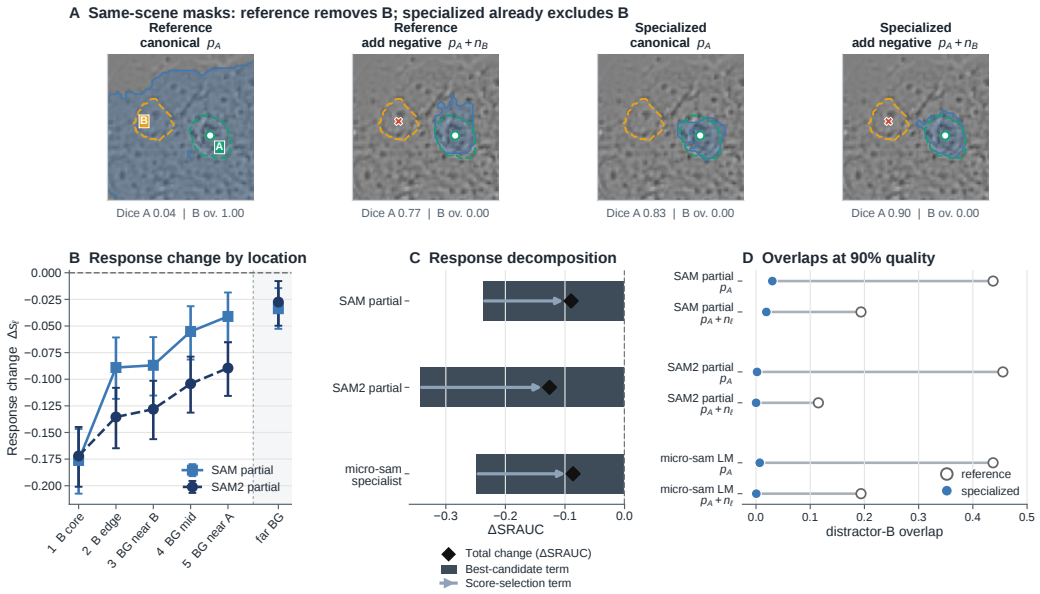



Figure S3: **Response-range compression after partial tuning (LIVECell).** (A) One scene: the reference mask covers $B$ until a negative point on $B$ removes it, whereas the specialized mask excludes $B$ from the start. (B) Specialized-minus-reference change in $s_\ell$ at the five locations and the far-background control. (C) Split of ΔSRAUC (diamond) into the best-candidate term (red) and the score-selection term (green), without a quality constraint; the micro-sam row is its LIVECell specialist. (D) Best-candidate overlap with $B$ without and with the negative point, among candidates within 90% of the best target Dice; the micro-sam rows are micro-sam LM. Both overlaps decrease, with a larger canonical reduction.

Figure S4 expands the main-text spatial summary. Panel A shows the frozen location-wise changes; the disconnected far-background point is separated by a dashed rule. Panel B reports the prespecified SRAUC together with target-retention adjustment. The dense multi-instance effects extend across the response field rather than being driven by one negative-point location.

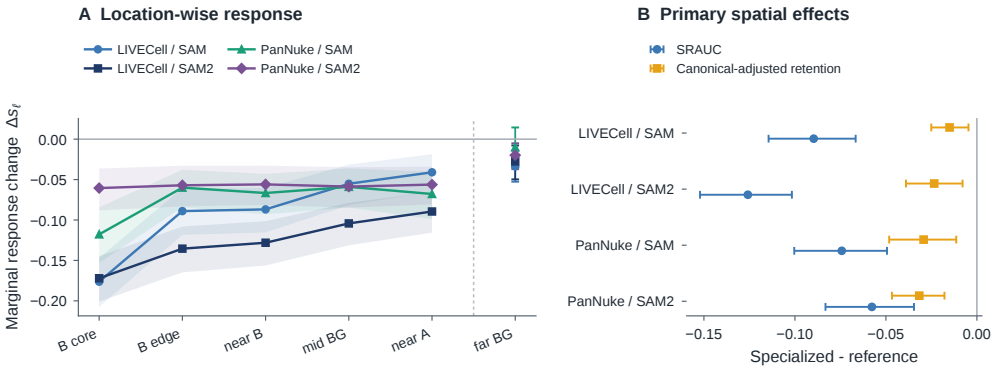



Figure S4: **Spatial-response replication across dense cohorts and SAM generations.** Panel A shows source-image-cluster means and intervals for five ordered local locations plus the detached far-background control. Panel B compares SRAUC and canonical-adjusted retention between endpoints. The latter averages negative-prompt minus canonical target Dice across local locations; it is not absolute target Dice. Negative effects indicate reduced overlap suppression or a less favorable target-Dice response, respectively. Location-wise rows localize the primary AUC result and are not an additional family of headline claims.

For each prompt, the returned candidate set is restricted to masks reaching 80%, 90%, or 95% of that set's oracle target Dice. Within the retained set, the best distractor-suppressing candidate defines the best-candidate response contrast; the difference between score selection and that oracle defines the score-selection contrast. Their sum equals the selected-output response contrast. This is a relative-quality restriction, not a guarantee of high absolute target quality. With $Q^\star = \max_k \mathrm{Dice}(m_k, G_A)$, the implemented eligibility rule is $\mathrm{Dice}(m_k, G_A) \geq \alpha Q^\star$. If $Q^\star = 0$, all candidates remain eligible; no zero-oracle rows are removed retrospectively.

The frozen reliability report counts oracle Dice at most $10^{-12}$ as zero. Across its 29 checkpoint configurations, canonical zero-oracle incidence is zero. Negative-prompt incidence is nonzero only for LIVECell SAM2 partial: 1/2,322, 3/2,322, and 1/2,322 prompt-location records for seeds 42, 123, and 456. These are location records, not independent images, and remain in the reported analysis. The anonymous artifact includes the per-checkpoint oracle-quality counts.

Table S9: **Quality-constrained spatial-response decomposition.** Total equals best-candidate plus score-selection change. Intervals bootstrap source images after averaging seeds. Total-effect intervals, including repair minus partial, come from the spatial-response family in Table S6; component intervals come from the separate 45-test quality-sensitivity family. The relative GT quality constraint is diagnostic, not an absolute-quality guarantee.

| Comparison | Quality | Total | Best-candidate response | Score-selection contrast |
|---|---|---|---|---|
| LIVECell repair − partial | 80% | +0.014 [+0.011, +0.018] | +0.004 [+0.001, +0.007] | +0.011 [+0.007, +0.015] |
| LIVECell repair − partial | 90% | +0.014 [+0.011, +0.018] | +0.003 [+0.001, +0.006] | +0.011 [+0.007, +0.015] |
| LIVECell repair − partial | 95% | +0.014 [+0.011, +0.018] | +0.003 [+0.000, +0.006] | +0.011 [+0.007, +0.015] |
| LIVECell SAM partial − base | 80% | -0.090 [-0.114, -0.067] | -0.235 [-0.272, -0.199] | +0.146 [+0.103, +0.189] |
| LIVECell SAM partial − base | 90% | -0.090 [-0.114, -0.067] | -0.233 [-0.268, -0.197] | +0.143 [+0.100, +0.185] |
| LIVECell SAM partial − base | 95% | -0.090 [-0.114, -0.067] | -0.232 [-0.268, -0.196] | +0.142 [+0.100, +0.183] |
| LIVECell SAM2 partial − base | 80% | -0.126 [-0.152, -0.102] | -0.339 [-0.381, -0.297] | +0.214 [+0.164, +0.262] |
| LIVECell SAM2 partial − base | 90% | -0.126 [-0.152, -0.102] | -0.339 [-0.381, -0.296] | +0.213 [+0.162, +0.262] |
| LIVECell SAM2 partial − base | 95% | -0.126 [-0.152, -0.102] | -0.338 [-0.381, -0.296] | +0.212 [+0.162, +0.261] |
| PanNuke SAM partial − base | 80% | -0.074 [-0.100, -0.049] | -0.024 [-0.040, -0.010] | -0.051 [-0.080, -0.023] |
| PanNuke SAM partial − base | 90% | -0.074 [-0.100, -0.049] | -0.025 [-0.042, -0.011] | -0.049 [-0.079, -0.022] |
| PanNuke SAM partial − base | 95% | -0.074 [-0.100, -0.049] | -0.025 [-0.043, -0.012] | -0.049 [-0.077, -0.020] |

*Continued on next page*

*Continued from previous page*

| Comparison | Quality | Total | Best-candidate response | Score-selection contrast |
|---|---|---|---|---|
| PanNuke SAM2 partial − base | 80% | -0.058 [-0.083, -0.035] | -0.009 [-0.019, -0.001] | -0.049 [-0.076, -0.024] |
| PanNuke SAM2 partial − base | 90% | -0.058 [-0.083, -0.035] | -0.009 [-0.018, -0.001] | -0.049 [-0.076, -0.024] |
| PanNuke SAM2 partial − base | 95% | -0.058 [-0.083, -0.035] | -0.008 [-0.018, -0.000] | -0.049 [-0.076, -0.024] |

The GT oracle in this decomposition is diagnostic and unavailable at deployment. Absolute overlaps at the 90% relative-quality threshold in Table S10 clarify the contrast: the specialized endpoint lowers distractor overlap under both canonical and negative prompts, with a larger reduction under the canonical prompt. The resulting contrast reflects response-range compression despite improved absolute exclusion.

Table S10: **Absolute 90%-quality overlap control.** Both absolute overlaps fall, with a larger canonical gain. SAM/SAM2 intervals resample seeds and images; micro-sam uses image resampling. Table S9 instead averages seeds before image resampling, so identical point estimates need not have identical intervals. These overlaps support response-range compression, not candidate disappearance.

| Endpoint | Images/units | Canonical $B$ overlap | Negative $B$ overlap | Contrast $\Delta$ [95% CI] |
|---|---|---|---|---|
| SAM partial | 100/387 | $0.437 \rightarrow 0.030$ | $0.194 \rightarrow 0.019$ | -0.233 [-0.272, -0.194] |
| SAM2 partial | 100/387 | $0.455 \rightarrow 0.002$ | $0.115 \rightarrow 0.000$ | -0.339 [-0.381, -0.297] |
| micro-sam LM | 100/387 | $0.437 \rightarrow 0.007$ | $0.194 \rightarrow 0.001$ | -0.237 [-0.275, -0.200] |

Tables S9 and S10 use identical mean contrasts but different uncertainty targets: image bootstrap after seed averaging versus seed–image bootstrap. For example, SAM's $-0.233$ contrast has intervals $[-0.268, -0.197]$ and $[-0.272, -0.194]$, respectively. They are separate frozen analyses, not interchangeable bootstrap reruns.

## E GENERATOR–SCORER AUDIT AND BOUNDED INTERVENTION

Section 5.4 separates changes in the returned candidates from changes in their selection. This appendix asks whether SAM's IoU head, which ranks the candidates, adapts at all. Under our mask-only loss it cannot: Dice plus BCE depends on the mask logits, not on the predicted IoU, so in a backward pass all six IoU-head tensors have `grad=None`, while 96 of 114 other mask-decoder tensors receive nonzero gradients. Checkpoint comparisons (Table S11) show that the head stays bitwise identical to SAM's in 17 of 17 mask-only checkpoints, whereas four positive controls trained with a selection loss change all six head tensors (Figure S5A). Mask-only adaptation therefore updates the candidate generator but not the scorer. This is a parameter mismatch, not an internal neural cause: upstream feature changes can still alter score outputs, and unchanged head parameters do not imply unchanged calibration or selection.

Table S11: **Checkpoint-level IoU-head audit.** All 17 mask-only checkpoints leave the six head tensors bitwise unchanged, whereas the same comparison detects all six tensors changing in four selector-trained positive controls. Non-head decoder counts exclude the six IoU-head tensors and refer to the 114 remaining mask-decoder tensors; candidate-generating tensors change in every mask-only endpoint.

| Group | Checkpoint | Head changed | Max $\lvert\Delta\rvert$ | Non-head changed | Status |
|---|---|---|---|---|---|
| mask-only | lora r4 n200 (LIVECell) | 0/6 | 0 | 28/114 | ok |
| mask-only | lora r4 n200 s123 (LIVECell) | 0/6 | 0 | 28/114 | ok |
| mask-only | lora r4 n200 s456 (LIVECell) | 0/6 | 0 | 28/114 | ok |
| mask-only | lora r16 n200 (LIVECell) | 0/6 | 0 | 28/114 | ok |
| mask-only | lora r4 offtrain n200 (LIVECell) | 0/6 | 0 | 28/114 | ok |
| mask-only | partial n200 (LIVECell) | 0/6 | 0 | 31/114 | ok |
| mask-only | partial offtrain n200 (LIVECell) | 0/6 | 0 | 31/114 | ok |
| mask-only | full tune n200 (LIVECell) | 0/6 | 0 | 114/114 | ok |
| mask-only | pcpa v2 repair | 0/6 | 0 | 31/114 | ok |
| mask-only | compact repair | 0/6 | 0 | 30/114 | ok |
| mask-only | compact repair n200 | 0/6 | 0 | 30/114 | ok |
| mask-only | partial n200 (PanNuke) | 0/6 | 0 | 31/114 | ok |
| mask-only | lora r4 n200 (PanNuke) | 0/6 | 0 | 28/114 | ok |

*Continued on next page*

*Continued from previous page*

| Group | Checkpoint | Head changed | Max $\|\Delta\|$ | Non-head changed | Status |
|---|---|---|---|---|---|
| mask-only | full tune n200 (PanNuke) | 0/6 | 0 | 114/114 | ok |
| mask-only | partial (Kvasir) | 0/6 | 0 | 31/114 | ok |
| mask-only | full tune (Kvasir) | 0/6 | 0 | 114/114 | ok |
| mask-only | bg-rejection repair (Kvasir) | 0/6 | 0 | 31/114 | ok |
| selector control | selector on lora n200 | 6/6 | 0.05327 | 28/114 | ok |
| selector control | selector on lora offtrain n200 | 6/6 | 0.04366 | 28/114 | ok |
| selector control | selector on partial n200 | 6/6 | 0.0568 | 31/114 | ok |
| selector control | selector on compact repair n200 | 6/6 | 0.05424 | 30/114 | ok |

### E.1 Frozen-candidate intervention eligibility

We define the oracle gap as the Dice difference between the GT-best and the score-selected candidate, and $P$(top = best) as the probability that the scorer picks the GT-best candidate. Reranking is informative only when a useful candidate remains but the original scores fail to select it, so we intervene only where adaptation both widens the oracle gap and lowers $P$(top = best) (Table S12). Only the seed-42 LIVECell LoRA endpoint meets these criteria. Across three LoRA seeds, best-candidate Dice rises by 0.384 and $P$(top = best) falls by 0.197, and all seeds agree (Figure S5B). Failing the criteria does not establish zero absolute ranking headroom or equivalence to an oracle selector. In variant labels, $r$ denotes LoRA rank and $s$ the training seed.

Table S12: **Intervention eligibility.** The frozen gate requires a confirmed increase in oracle gain and decrease in oracle-selection probability. Only LIVECell LoRA meets the gate; failing it does not imply zero absolute ranking headroom. Released-model and cross-generation rows are descriptive; repair rows answer different questions.

| Cohort | Variant | Role | $\Delta$ oracle gain [CI] | $\Delta P$(top=best) [CI] | Eligible |
|---|---|---|---|---|---|
| LIVECell100 | compact repair | repair | +0.008 [-0.012, +0.028] | +0.085 [+0.003, +0.165] | no |
| LIVECell100 | LoRA r4 (s42) | adaptation | +0.101 [+0.078, +0.124] | -0.190 [-0.258, -0.122] | yes |
| LIVECell100 | partial (s42) | adaptation | -0.005 [-0.019, +0.008] | -0.035 [-0.075, +0.005] | no |
| LIVECell100 | full decoder (s42) | adaptation | -0.033 [-0.052, -0.013] | +0.247 [+0.165, +0.328] | no |
| LIVECell100 | micro-sam LM | released model | +0.025 [+0.002, +0.047] | -0.035 [-0.113, +0.043] | no |
| LIVECell100 | PCPA v2 | repair | +0.005 [-0.013, +0.023] | +0.035 [-0.050, +0.120] | no |
| LIVECell100 | SAM2 Hiera-L | cross-generation | -0.014 [-0.031, +0.001] | +0.125 [+0.070, +0.180] | no |
| PanNuke fold 3 | LoRA r4 (s42) | adaptation | -0.041 [-0.068, -0.016] | -0.099 [-0.185, -0.012] | no |
| PanNuke fold 3 | partial (s42) | adaptation | -0.047 [-0.076, -0.021] | -0.381 [-0.467, -0.292] | no |

### E.2 Retraining the scoring head on fixed candidates

On the qualifying endpoint we retrain only the IoU head. Candidate masks and logits are verified bitwise identical across 2,400 prompt groups (7,200 of each), so only scores and selected indices can change. The original selector obtains 0.569 Dice, a zero-learning rule based on point containment, centroid distance, and compactness obtains 0.707, the retrained head 0.735, and the GT oracle 0.749. The retrained head recovers 92.3% of the oracle gap and the geometric rule 76.75%; learning adds 0.028 Dice [0.018, 0.039], or 15.55 percentage points of recovery, while lowering STS by 0.119. This is bounded evidence for a ranking bottleneck at one endpoint, not a generally applicable adaptation method. Because reranking chooses only among the returned candidates, it cannot remove the full-decoder reversal, which the best returned candidate already carries (Section 5.4).

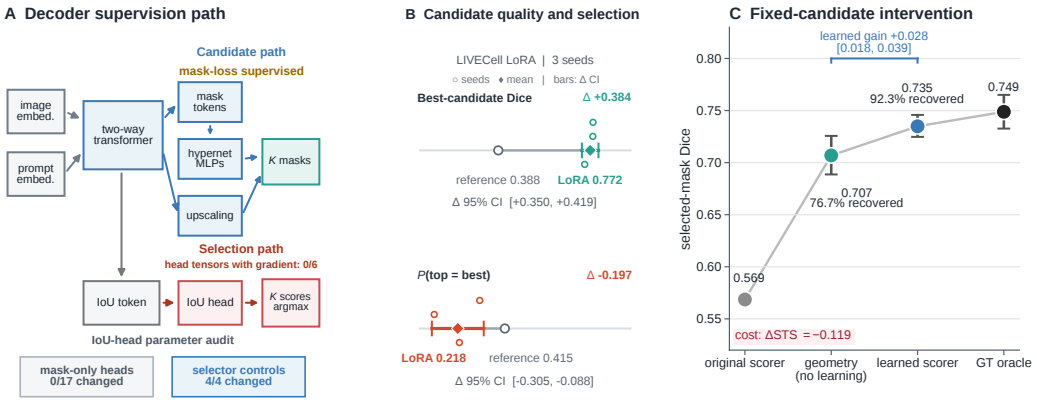



Figure S5: **Mask-only adaptation updates the candidate generator but not the IoU scoring head.** (A) Decoder dataflow; the six IoU-head tensors receive no gradient from the mask loss (unchanged in 17/17 mask-only checkpoints, changed in 4/4 selector controls). (B) LIVECell LoRA, three seeds: best-candidate Dice rises while $P$(top = best), the rate at which the scorer picks the GT-best candidate, falls. (C) Seed-42 LoRA endpoint with fixed candidates: selected Dice under the original scorer, a zero-learning geometric rule, a retrained IoU head, and the GT oracle; percentages give the share of the oracle gap recovered.

# F External validity and model-family boundaries

External evaluations address different questions. The relation estimand is interface-level; the candidate and scorer diagnoses are specific to SAM. micro-sam tests independence from our training code; pretrained SAM2/SAM3 test whether the same protocol runs across generations; RClicks tests sensitivity to released click locations. SimpleClick (Appendix K) provides a non-SAM behavioral boundary for the B-core response and the positive-point relations but does not establish that reversal is architecture-general. None establishes human usability. Table S13 records the corresponding limits.

## F.1 Single-target distance field

Kvasir uses one positive point outside its single annotated target, with no target point. TRNP measures target recovery, not the response to an added negative point. Figure S6 varies GT-derived boundary distance across frozen pretrained endpoints. These synthetic points do not represent user clicks or clinical risk.

Kvasir in-domain adaptation has different effects near and far from the target: far-background BHR/TRNP decrease only in point estimates, while near-boundary TRNP increases. An external box-trained specialist is inconclusive under this point-prompt protocol and is retained as a compatibility boundary rather than scored as a failed method.

Table S13: **External-validity evidence and its scope.** Positive, null, and stopped results answer different questions and are not pooled into one model-family ranking.

| Evidence | Result | Role | Boundary |
|---|---|---|---|
| micro-sam LIVECell specialist | Dice +0.477; TSO +0.511; STS -0.080; SRAUC -0.086 | confirmation outside our training code | one mixed-prompt microscopy specialist; no per-image upstream manifest |
| SAM/SAM2/SAM3 frozen baselines | unchanged harness; 12/12 model-domain evaluations complete | protocol portability and family-specific profiles | not an adaptation replication for SAM3 |
| SAM3 adaptation induction | 8.55% fixed-probe reduction vs frozen 10% gate | stopped induction boundary | effects unopened; no immunity or more-steps claim |
| external Kvasir box-trained specialist | Dice +0.0218, CI [-0.0441, +0.0878]; 0/10 confirmed endpoints | valid but inconclusive point-protocol boundary | not native box performance or equivalence evidence |
| RClicks released first-click locations | human-minus-center Dice +0.026/+0.024/+0.034 for SAM/SAM2/HQ-SAM | prompt-location external validity | no participant IDs, timing, trajectories, or usability outcome |

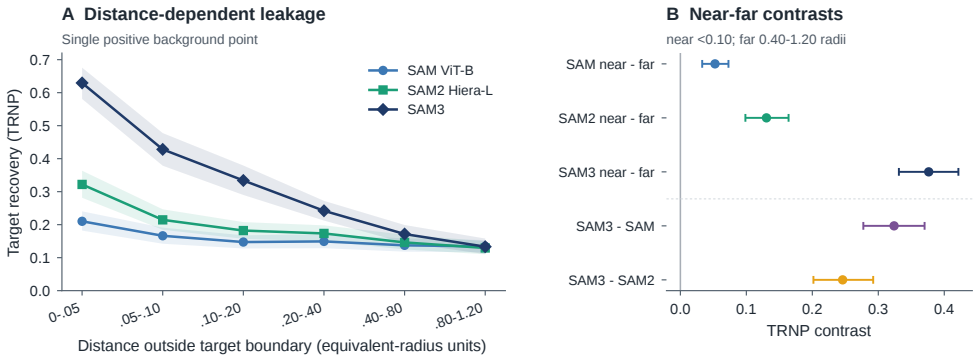



Figure S6: **Kvasir background-point leakage depends on boundary distance.** (A) 100 images, five prompt seeds, six distance bins. (B) Near-minus-far TRNP within each model (first three rows) and differences between those contrasts (last two). Near is below 0.10 equivalent target radii; far is 0.40–1.20. Intervals use source-image bootstrap with a five-test Holm family. SAM3 has the largest near response; far-background levels converge.

### F.2 Closest adjacent work

Table S14 records the closest overlap and the explicit claim boundary used in this paper. No individual operation–prompt perturbation, consistency measurement, multi-mask oracle analysis, or reranking–is presented as novel. The contribution is the paired post-specialization measurement and diagnostic chain.

Table S14: **Closest adjacent work and the remaining measurement gap.** The table distinguishes shared operations from the paired post-specialization audit, with the scope of each comparison stated explicitly.

| Work | Venue | Closest Overlap | Distinguishing Axis | Claim Boundary |
|---|---|---|---|---|
| CAFE (2026) | arXiv | Counterfactual prompts show mask quality is not semantic faithfulness | Text concepts and mask-preserving image edits rather than post-specialization spatial point relations | Accuracy–faithfulness audits precede this work |
| HBF-BCER (2026) | Bioeng. | Prompt-preserving SAM adaptation under point and box perturbations | Same-intent perturbation robustness rather than intent-changing relations and candidate/scorer diagnosis | Prompt robustness during adaptation has prior work |
| Stable-SAM (2025) | ICLR | Prompt-quality stability and a lightweight robustness method | Low-quality same-intent prompts rather than target switching and negative-point response after specialization | Prompt-stability analysis has prior work |
| BREPS (2026) | AAAI | Natural and adversarial bounding-box robustness with human boxes | Box-prompt robustness rather than spatial response reshaping after specialization | Realistic prompt-variation benchmarks have prior work |
| RClicks (2024) | NeurIPS Datasets and Benchmarks | Realistic click distributions and interactive-segmentation robustness | Human click location benchmark rather than fixed counterfactual relation diagnosis | Synthetic proxies are not a human study |
| TETRIS (2024) | AAAI | Robustness benchmark for interactive segmentation | General interaction robustness rather than semantics-separated post-specialization relations | Interactive-robustness benchmarks have prior work |
| Point Prompt Defender (2026) | CVPR | Adversarial point-prompt optimization for robustness | Optimizes prompts while SAM-Lens audits model behavior under frozen prompts | The selector changes mask ranking, not prompts |
| PP-SAM (2024) | CVPR Workshops | Perturbed-box training for robust polyp adaptation | Box perturbation method rather than a relational post-specialization audit | Robust medical SAM adaptation has prior work |
| PiClick (2024) | Neurocomp. | Multiple candidates and learned target reasoning selector | New architecture for ambiguity rather than diagnosis of a stale SAM scorer | Learned candidate selection has prior work |
| SAMRefiner++ (2025) | ICLR | Additional IoU adaptation for mask refinement | Mask-refinement pipeline rather than post-specialization relation auditing | IoU-head adaptation has prior work |
| SAM-DA (2025) | WACV | Decoder adaptation with an explicitly trained IoU head | Shows stale scoring is objective-dependent rather than inevitable | Parameter mismatch is specific to the audited mask-only objectives |
| RobustSAM (2024) | CVPR | Preserves SAM performance on degraded images | Image corruption rather than prompt transformations after specialization | Image-corruption robustness differs from prompt-response relations |

# G NEGATIVE RESULTS AND STOPPED BRANCHES

Table S15 records failed gates and boundary results. We did not change thresholds, substitute a new metric, or continue searching after a failed gate.

PRDD denotes prompt-response delta distillation, an attempt to preserve differences between prompt-conditioned outputs. It remains a failed development control, not a proposed method in this paper.

Table S15: **Frozen negative and stopped branches.** These branches remain visible because they delimit the main claims. None was rescued by threshold changes, post-hoc retuning, or selective omission.

| Branch | Frozen Gate | Outcome | Decision |
|---|---|---|---|
| SAM3 specialization | $\geq$10% fixed-probe loss reduction | 8.55%; technically active but gate failed | stopped; held-out effects not evaluated |
| selector outside LIVE-Cell LoRA | oracle gain increases and oracle-selection probability decreases | five tested adaptation settings do not meet the intervention gate | no selector extension |
| label-free margin trigger | three-seed CI and direction gate | prospective low-margin trigger failed | no replacement threshold search |
| COCO capacity pathway | full spatial-capacity confirmation | fit and TSO/MMD pass; capacity pathway fails | retain structural boundary |
| PCPA / compact repair | Pareto improvement over partial | relation recovery trades Dice; no universal dominance | supplementary intervention evidence only |
| counterfactual selector / PRDD | predeclared development utility | failed quality/specificity gates | stop without confirmation or retuning |
| historical active-compact cache | zero train/evaluation overlap | nine overlapping images | quarantined; excluded from evidence |

SAM3 passed the recorded training checks: gradients reached all six targeted tensors, all six changed, checkpoint reload was equivalent, and the late loss trend remained downward. Fixed-probe loss nevertheless fell by 8.55%, below the prespecified 10% induction gate, so held-out relation outcomes were not evaluated. SAM3's tuned output-upscaling module has 73,952 trainable parameters; the SAM/SAM2 three-module configuration has 8.58 times as many (634,208). SAM3 also starts at a lower loss. The result means insufficient specialization under this frozen, unmatched design. It does not show that SAM3 is immune to interface reshaping.

The historical active-compact cache is excluded because its training and evaluation sets overlap by nine images. Validity metadata marks it invalid; confirmatory analysis requires an explicit valid status.

# H REPRODUCIBILITY AND ARTIFACT PROVENANCE

Predictions are indexed by model, dataset split, image, instance pair, and prompt unit. Checkpoints, datasets, and dense prediction masks are not stored in the paper repository; configuration files, train logs, prompt manifests, checkpoint hashes, generated statistics, and audit summaries are retained.

The anonymous artifact includes frozen statistical inputs, table and figure generators, source-hash manifests, and a synthetic fixture for prompt parsing, pair-keyed loading, metric computation, and report generation. Its guide provides the source-file index, build commands, and public-dataset setup instructions. The submitted figures are included; regenerating image panels requires the separately obtained datasets and prediction artifacts. The fixture tests the implementation, not model accuracy.

## H.1 FAIL-CLOSED VALIDITY CHECKS

We include a report only with explicit valid artifact status and supporting evidence; matching model, split, prompt-library, and checkpoint identities; pair-keyed predictions with zero fallback and explained skips; and no controlled-training/evaluation image overlap. Reports must use the declared image or seed–image statistical unit and retain inconclusive rows in their multiplicity family. The overlap-contaminated cache in Appendix G fails the validity and separation checks and cannot enter confirmatory reports.

### H.2 SOURCE MAPPING OF THE ORIGINAL RESPONSE PROFILE

Table S16 links the four headline effects in every row of Table S5 to its source report. P2/P3 supply CI-only replications without reported $p$ values. P4 pools seeds before image inference, while P2/P3/P5/P6 resample seeds and images. This distinction also explains why the canonical seed count in Table S5 need not describe its spatial-response column.

Table S16: **Statistical provenance of the original response profile.** Each headline effect of Table S5 retains its original inference. Three seeds means 42/123/456; one seed means 42; ext. denotes a released checkpoint. Image* averages the observed seeds before resampling images, unlike seed × image intervals. Not reported means the source provides a CI but no Holm test; it is not a zero $p$ value. Counts are valid images/unique pairs, not seed-replicated rows.

| Setting | Metric | Seeds | $n$/pairs | $\Delta$ [95% CI] | Resampling | Holm $p$ $(m)$ | Src. |
|---|---|---|---|---|---|---|---|
| LIVECell / SAM | $\text{Dice}_A$ | 1 | 100/400 | +0.385 [+0.356, +0.412] | image | 0.0051 (102) | P1 |
| | STS | 1 | 100/400 | -0.300 [-0.325, -0.275] | image | 0.0051 (102) | P1 |
| | TSO | 1 | 100/400 | +0.387 [+0.356, +0.416] | image | 0.0051 (102) | P1 |
| | SRAUC | 3 | 100/387 | -0.090 [-0.114, -0.067] | image* | 0.0007 (15) | P4 |
| PanNuke / SAM | $\text{Dice}_A$ | 3 | 93/349 | +0.110 [+0.077, +0.144] | seed × image | not reported | P2 |
| | STS | 3 | 93/349 | -0.112 [-0.145, -0.078] | seed × image | not reported | P2 |
| | TSO | 3 | 93/349 | +0.112 [+0.076, +0.150] | seed × image | not reported | P2 |
| | SRAUC | 3 | 93/349 | -0.074 [-0.100, -0.049] | image* | 0.0007 (15) | P4 |
| LIVECell / SAM2 | $\text{Dice}_A$ | 3 | 100/400 | +0.391 [+0.364, +0.418] | seed × image | not reported | P3 |
| | STS | 3 | 100/400 | -0.274 [-0.297, -0.250] | seed × image | not reported | P3 |
| | TSO | 3 | 100/400 | +0.395 [+0.367, +0.422] | seed × image | not reported | P3 |
| | SRAUC | 3 | 100/387 | -0.126 [-0.152, -0.102] | image* | 0.0007 (15) | P4 |
| COCO / SAM | $\text{Dice}_A$ | 3 | 123/344 | +0.070 [+0.048, +0.094] | seed × image | 0.0002 (4) | P5 |
| | STS | 3 | 123/344 | +0.018 [-0.008, +0.045] | seed × image | 0.1231 (4) | P5 |
| | TSO | 3 | 123/344 | +0.076 [+0.052, +0.104] | seed × image | 0.0002 (4) | P5 |
| | SRAUC | 3 | 122/343 | -0.004 [-0.011, +0.000] | seed × image | 0.0718 (2) | P6 |
| LIVECell / micro-sam | $\text{Dice}_A$ | ext. | 100/400 | +0.477 [+0.438, +0.516] | image | 0.0005 (10) | P7 |
| | STS | ext. | 100/400 | -0.080 [-0.103, -0.058] | image | 0.0005 (10) | P7 |
| | TSO | ext. | 100/400 | +0.511 [+0.473, +0.549] | image | 0.0005 (10) | P7 |
| | SRAUC | ext. | 100/387 | -0.086 [-0.111, -0.063] | image | 0.0005 (10) | P7 |

**Source key.** P1: core image-paired audit; P2: PanNuke seed replication; P3: SAM2 seed replication; P4: spatial-response audit; P5: COCO canonical/relation audit; P6: COCO spatial audit; P7: released-specialist audit. The anonymous artifact guide maps these keys to the included CSV files.

# I ADDITIONAL QUALITATIVE EVIDENCE

Figure S7 illustrates two diagnostic paths under frozen case-selection rules; case-level gains are not aggregate effects. The single-point NPS contrast does not measure negative-click damage.

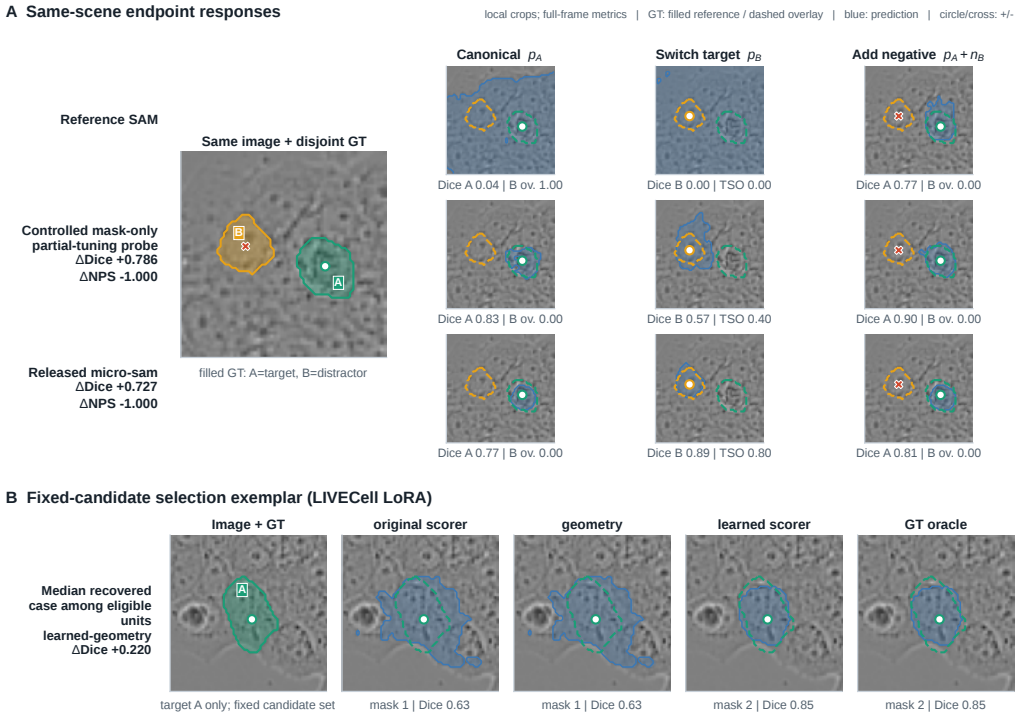



Figure S7: **Actual masks expose two different diagnostic paths.** All images use a shared local crop within their panel group; every printed metric is computed on the full frame. Solid blue contours show predictions, dashed green/amber contours show GT, and compact circles/crosses show positive/negative points. (A) The same LIVECell scene, GT pair, and prompts are used for a controlled mask-only partial-tuning probe and the independently released micro-sam specialist. Green and amber fills mark two disjoint instances ($A$ target, $B$ distractor). Under $p_B$, reference SAM fails to switch ($\text{Dice}_B = 0.00$), whereas the partial probe and micro-sam return $B$ (0.57 and 0.89); switch performance did not enter case selection. Reference SAM covers $B$ under $p_A$ ($B$ overlap 1.00), then removes it under $p_A + n_B$ (0.00); both specialized endpoints already exclude $B$ and barely change after the negative point. A GT-morphology tie-break selects this scene from 19 response-visible cases; model Dice, switch performance, and endpoint ordering are not selection criteria. The displayed partial-probe Dice is 0.83 versus micro-sam's 0.77, but aggregate LIVECell100 Dice is 0.693 versus 0.786 in Table S5. (B) Only target $A$ is shown because the distractor is irrelevant to fixed-candidate selection. The original scorer and zero-learning geometry rule select candidate 1 (Dice 0.63), whereas the learned scorer selects the unchanged GT-oracle candidate 2 (Dice 0.85). This is the median learned-over-geometry gain among 50 eligible units, not frequency evidence.

The controlled partial endpoint in panel A is an identification probe, not a proposed segmentation method or a budget-matched micro-sam competitor. Panel B does not imply that ground truth is available at deployment; ground truth is used only to identify the frozen oracle candidate and to evaluate selectors.

# J NEGATIVE-POINT ANALYSIS OF FROZEN ENDPOINTS

## J.1 ANALYSIS PLAN AND DECISION RULES

The B-core analyses are post-hoc: they were motivated by absolute SRAUC levels that we had already inspected. Before the final recomputation we fixed, in a dated and version-controlled plan, the endpoint list, the primary metric ($s_{\mathrm{Bcore}}$, reported with both overlaps and target retention), the secondary metrics (SRAUC, per-location responses, far-background control), four test families, and the decision rules. F1 tests $s_{\mathrm{Bcore}} = 0$ for the full-decoder endpoints (Holm over ten tests; six SimpleClick tests in this family used the invalid predictions of Appendix K, which only makes the correction for the four SAM tests conservative). F2 compares the full decoder with every LIVECell endpoint of another SAM-family adaptation whose canonical Dice lies within 0.03 of it (Holm over six tests); the full-decoder seeds added later (Appendix J.4) are replications, not comparators. F3 compares full-decoder and partial endpoints within the same training pool (Holm over six tests). F4 decomposes the full-decoder response into best-candidate and score-selection terms (descriptive). All intervals use 10,000 image-cluster bootstrap draws and all tests 20,000 paired sign flips. The plan required a significant LIVECell reversal (F1) and a positive same-pool contrast for all three LIVECell partial seeds against the frozen seed-42 full decoder (F3) before the analysis could enter the paper; both passed. The planned non-SAM replication did not (Appendix K). Two further LIVECell full-decoder seeds were trained after this plan was frozen; a dated addendum, committed before their results were computed, fixed their tests, random streams, and reporting rule (Appendix J.4).

## J.2 ENDPOINTS

Table S17: **Endpoints of the negative-point analysis.** All controlled adaptations train on the canonical prompt only with Dice plus BCE loss and use the final checkpoint of a fixed budget. The LIVECell SAM endpoints train on official test-split images outside the audit set; zero overlap was verified from training logs. micro-sam LM has no per-image training manifest, so its overlap with the audit images cannot be verified.

| Model | Tuned modules | Parameters | Training images | Steps | Seeds |
|---|---|---|---|---|---|
| SAM ViT-B + LoRA | rank-4 LoRA on two-way transformer attention | 47,104 | LIVECell legacy pool (196–199); PanNuke fold 1 (199) | 4,000 | 42, 123, 456 |
| SAM ViT-B + partial | mask tokens, hypernetwork MLPs, upscaling | 634,208 | as above | 4,000 | 42, 123, 456 |
| SAM ViT-B + full decoder | entire mask decoder | 4,058,340 | as above | 4,000 | LIVECell 42 (+123, 456 after the freeze); PanNuke 42, 123, 456 |
| SAM2 Hiera-L + partial | mask tokens, hypernetwork MLPs, upscaling | 634,208 | LIVECell official train (199–200); PanNuke fold 1 (199) | 4,000 | 42, 123, 456 |
| micro-sam LM | released light-microscopy model | – | external | – | – |
| SimpleClick ViT-B | click embedding, neck, head (backbone frozen) | 11,006,273 | LIVECell official train (145–148); PanNuke fold 1 (195–198) | 2,000 | 42, 123, 456 |

## J.3 ALL ENDPOINT RESPONSES AND TESTS

Table S18 lists every endpoint and seed. Table S19 reports the planned tests, and Table S20 the decomposition of the full-decoder response.

Table S18: **Negative-point response of every audited endpoint.** LC: LIVECell; PN: PanNuke. Image-cluster 95% CIs (10,000 draws). [§] SimpleClick values use the corrected pair bank (Appendix K); its spatial scores are omitted. Dice$_A$ columns report target retention at the B-core prompt.

| Endpoint | Dice | $o_B$: $p_A \rightarrow p_A{+}n_B$ | $s_{\mathrm{Bcore}}$ [95% CI] | Dice$_A$: $p_A \rightarrow p_A{+}n_B$ | SRAUC | far BG |
|---|---|---|---|---|---|---|
| LC SAM ref. | 0.308 | 0.601→0.480 | +0.121 [+0.093, +0.152] | 0.311→0.339 | +0.090 | 0.000 |
| LC SAM + LoRA s123 | 0.563 | 0.109→0.065 | +0.045 [+0.024, +0.067] | 0.565→0.569 | +0.090 | −0.125 |
| LC SAM + LoRA s42 | 0.569 | 0.122→0.102 | +0.020 [−0.003, +0.044] | 0.573→0.530 | +0.082 | −0.257 |
| LC SAM + LoRA s456 | 0.666 | 0.041→0.068 | −0.028 [−0.043, −0.012] | 0.670→0.597 | +0.020 | −0.144 |
| LC SAM + partial s123 | 0.629 | 0.053→0.106 | −0.053 [−0.068, −0.040] | 0.634→0.588 | −0.002 | −0.035 |
| LC SAM + partial s42 | 0.545 | 0.124→0.192 | −0.068 [−0.082, −0.055] | 0.550→0.515 | +0.012 | −0.037 |
| LC SAM + partial s456 | 0.677 | 0.023→0.067 | −0.044 [−0.056, −0.033] | 0.661→0.613 | −0.009 | −0.028 |

| Endpoint | Dice | $o_B$: $p_A \rightarrow p_A{+}n_B$ | $s_{\mathrm{Bcore}}$ [95% CI] | Dice$_A$: $p_A \rightarrow p_A{+}n_B$ | SRAUC | far BG |
|---|---|---|---|---|---|---|
| LC SAM + full decoder s42 | 0.693 | 0.026→0.526 | −0.500 [−0.535, −0.464] | 0.694→0.436 | −0.173 | −0.203 |
| LC SAM2 ref. | 0.302 | 0.601→0.422 | +0.179 [+0.152, +0.207] | 0.305→0.339 | +0.133 | +0.031 |
| LC SAM2 + partial s123 | 0.690 | 0.008→0.000 | +0.007 [+0.005, +0.011] | 0.694→0.698 | +0.007 | +0.004 |
| LC SAM2 + partial s42 | 0.696 | 0.007→0.000 | +0.007 [+0.004, +0.012] | 0.700→0.717 | +0.007 | +0.004 |
| LC SAM2 + partial s456 | 0.693 | 0.006→0.000 | +0.006 [+0.004, +0.008] | 0.697→0.713 | +0.006 | +0.003 |
| LC micro-sam LM | 0.703 | 0.043→0.007 | +0.036 [+0.020, +0.055] | 0.706→0.721 | +0.030 | −0.020 |
| LC SimpleClick ref.[§] | 0.040 | 0.948→0.129 | +0.818 [+0.791, +0.844] | 0.039→0.352 | – | – |
| LC SimpleClick + tuning[§] s123 | 0.701 | 0.001→0.002 | −0.001 [−0.003, 0.000] | 0.702→0.696 | – | – |
| LC SimpleClick + tuning[§] s42 | 0.690 | 0.000→0.001 | −0.001 [−0.002, −0.000] | 0.691→0.685 | – | – |
| LC SimpleClick + tuning[§] s456 | 0.694 | 0.002→0.007 | −0.005 [−0.008, −0.002] | 0.695→0.673 | – | – |
| PN SAM ref. | 0.723 | 0.140→0.066 | +0.074 [+0.047, +0.105] | 0.723→0.753 | +0.058 | +0.009 |
| PN SAM + LoRA s123 | 0.833 | 0.001→0.007 | −0.006 [−0.010, −0.003] | 0.833→0.818 | −0.001 | +0.001 |
| PN SAM + LoRA s42 | 0.828 | 0.000→0.004 | −0.004 [−0.008, −0.001] | 0.828→0.819 | −0.001 | 0.000 |
| PN SAM + LoRA s456 | 0.830 | 0.000→0.004 | −0.003 [−0.006, −0.002] | 0.830→0.822 | −0.001 | 0.000 |
| PN SAM + partial s123 | 0.838 | 0.006→0.068 | −0.061 [−0.081, −0.044] | 0.838→0.818 | −0.023 | −0.001 |
| PN SAM + partial s42 | 0.827 | 0.002→0.036 | −0.034 [−0.048, −0.021] | 0.827→0.821 | −0.012 | −0.001 |
| PN SAM + partial s456 | 0.834 | 0.003→0.039 | −0.035 [−0.051, −0.022] | 0.834→0.826 | −0.012 | −0.001 |
| PN SAM + full decoder s123 | 0.859 | 0.000→0.651 | −0.651 [−0.684, −0.616] | 0.859→0.610 | −0.196 | −0.009 |
| PN SAM + full decoder s42 | 0.846 | 0.000→0.205 | −0.204 [−0.240, −0.171] | 0.837→0.746 | −0.069 | −0.002 |
| PN SAM + full decoder s456 | 0.841 | 0.004→0.725 | −0.721 [−0.748, −0.691] | 0.841→0.571 | −0.240 | −0.006 |
| PN SAM2 ref. | 0.753 | 0.114→0.040 | +0.074 [+0.043, +0.109] | 0.753→0.797 | +0.070 | +0.022 |
| PN SAM2 + partial s123 | – | 0.012→0.001 | +0.011 [+0.002, +0.026] | 0.843→0.847 | +0.010 | +0.003 |
| PN SAM2 + partial s42 | – | 0.014→0.002 | +0.012 [+0.002, +0.028] | 0.842→0.849 | +0.011 | +0.003 |
| PN SAM2 + partial s456 | – | 0.018→0.002 | +0.017 [+0.005, +0.034] | 0.838→0.852 | +0.015 | +0.002 |
| PN SimpleClick ref.[§] | 0.074 | 0.929→0.050 | +0.879 [+0.838, +0.914] | 0.074→0.407 | – | – |
| PN SimpleClick + tuning[§] s123 | 0.851 | 0.002→0.026 | −0.024 [−0.031, −0.016] | 0.851→0.836 | – | – |
| PN SimpleClick + tuning[§] s42 | 0.806 | 0.000→0.006 | −0.006 [−0.010, −0.003] | 0.806→0.811 | – | – |
| PN SimpleClick + tuning[§] s456 | 0.846 | 0.001→0.011 | −0.010 [−0.014, −0.006] | 0.846→0.839 | – | – |

Table S19: **Planned tests.** F1: B-core response of the full decoder. F2: endpoints of other SAM-family adaptations with canonical Dice within 0.03 of the full decoder (comparator minus full decoder). F3: same-pool partial minus full decoder. Holm correction within each family. Each test draws its own bootstrap samples, so its CI can differ from Table S18 in the third decimal.

| Family | Contrast | Estimate [95% CI] | ΔDice [95% CI] | Holm $p$ |
|---|---|---|---|---|
| F1 reversal | LIVECell full decoder s42 | −0.500 [−0.534, −0.465] | – | 0.0005 |
| F1 reversal | PanNuke full decoder s42 | −0.204 [−0.239, −0.170] | – | 0.0005 |
| F1 reversal | PanNuke full decoder s123 | −0.651 [−0.685, −0.616] | – | 0.0005 |
| F1 reversal | PanNuke full decoder s456 | −0.721 [−0.748, −0.692] | – | 0.0005 |
| F2 similar Dice | micro-sam LM − full decoder | +0.536 [+0.495, +0.579] | +0.011 [−0.010, +0.032] | 0.0003 |
| F2 similar Dice | SAM LoRA s456 − full decoder | +0.472 [+0.434, +0.510] | −0.026 [−0.046, −0.006] | 0.0003 |
| F2 similar Dice | SAM partial s456 − full decoder | +0.456 [+0.422, +0.491] | −0.015 [−0.042, +0.010] | 0.0003 |
| F2 similar Dice | SAM2 partial s123 − full decoder | +0.507 [+0.471, +0.543] | −0.002 [−0.018, +0.014] | 0.0003 |
| F2 similar Dice | SAM2 partial s42 − full decoder | +0.507 [+0.472, +0.542] | +0.003 [−0.013, +0.020] | 0.0003 |
| F2 similar Dice | SAM2 partial s456 − full decoder | +0.506 [+0.471, +0.540] | +0.001 [−0.015, +0.017] | 0.0003 |
| F3 same pool | LIVECell partial s42 − full s42 | +0.431 [+0.396, +0.466] | −0.148 [−0.171, −0.125] | 0.0003 |
| F3 same pool | LIVECell partial s123 − full s42 | +0.447 [+0.411, +0.483] | −0.064 [−0.090, −0.039] | 0.0003 |
| F3 same pool | LIVECell partial s456 − full s42 | +0.456 [+0.421, +0.491] | −0.015 [−0.042, +0.011] | 0.0003 |
| F3 same pool | PanNuke partial s42 − full s42 | +0.171 [+0.134, +0.207] | −0.020 [−0.030, −0.010] | 0.0003 |
| F3 same pool | PanNuke partial s123 − full s123 | +0.590 [+0.555, +0.623] | −0.020 [−0.029, −0.011] | 0.0003 |
| F3 same pool | PanNuke partial s456 − full s456 | +0.686 [+0.655, +0.716] | −0.007 [−0.015, +0.002] | 0.0003 |

Table S20: **Decomposition of the full-decoder response** into the best-candidate term and the score-selection term, at the B core and averaged over the five local locations.

| Domain | Full decoder | Total $s$ | Best-candidate term [95% CI] | Score-selection term [95% CI] |
|---|---|---|---|---|
| LIVECell | s42, B core | −0.500 | −0.456 [−0.490, −0.422] | −0.044 [−0.061, −0.028] |
| LIVECell | s42, 5-location mean | −0.173 | −0.141 [−0.157, −0.127] | −0.032 [−0.043, −0.021] |
| PanNuke | s42, B core | −0.204 | −0.153 [−0.184, −0.124] | −0.051 [−0.068, −0.037] |
| PanNuke | s42, 5-location mean | −0.069 | −0.052 [−0.065, −0.040] | −0.017 [−0.022, −0.013] |
| PanNuke | s123, B core | −0.651 | −0.636 [−0.669, −0.600] | −0.015 [−0.020, −0.011] |
| PanNuke | s123, 5-location mean | −0.196 | −0.188 [−0.203, −0.173] | −0.008 [−0.011, −0.006] |
| PanNuke | s456, B core | −0.721 | −0.661 [−0.690, −0.632] | −0.059 [−0.072, −0.048] |
| PanNuke | s456, 5-location mean | −0.240 | −0.205 [−0.219, −0.191] | −0.034 [−0.040, −0.029] |

### J.4 TWO FULL-DECODER SEEDS ADDED AFTER THE FREEZE

The addendum defined two tests, each with Holm correction over two: F1b tests $s_{\mathrm{Bcore}} = 0$ for the new endpoints, and F3b compares each with the partial endpoint of the same seed and training pool. We trained the two endpoints with the recipe of seed 42 and the seeds of the existing LoRA and partial endpoints (123, 456). Both pass the specialization gate (canonical Dice 0.702 and 0.737, against 0.308 for reference SAM). Their training logs list the same images as the partial endpoints of the same seeds (199 and 196 images) and none of the 100 audit images. Both endpoints reverse the negative point (Table S21), and the same-seed partial endpoint exceeds them by 0.36 and 0.28 while its canonical Dice is 0.073 and 0.060 lower. The negative point also lowers the Dice of $A$, from 0.703 to 0.473 and from 0.739 to 0.510. The best-candidate term again carries most of the response, and the negative point adds more than a quarter of $B$ in 69% and 53% of pairs (more than half in 34% and 23%). At the other locations the responses are milder: $-0.068$ and $-0.048$ at the B edge, $-0.075$ and $-0.047$ on the background next to $B$, and $-0.134$ and $-0.112$ at the far-background control; SRAUC is $-0.119$ and $-0.088$, against $+0.090$ for reference SAM. As fixed in the addendum, Figures 1, 3, 4B, and 5 keep seed 42.

Table S21: **LIVECell full-decoder seeds trained after the analysis was frozen** (100 images, 387 pairs). Holm $p$: F1b. Best / selection: the two terms of the B-core response (Section 5.4). Partial − full: B-core response of the same-seed, same-pool partial endpoint minus that of the full decoder (F3b, Holm $p < 0.001$ for both seeds).

| Seed | Dice | $o_B$: $p_A \to p_A{+}n_B$ | $s_{\mathrm{Bcore}}$ [95% CI] | Holm $p$ | Best / selection | Partial − full [95% CI] |
|---|---|---|---|---|---|---|
| 123 | 0.702 | 0.004 → 0.415 | −0.411 [−0.437, −0.384] | 0.0001 | −0.363 / −0.048 | +0.358 [+0.330, +0.386] |
| 456 | 0.737 | 0.001 → 0.325 | −0.324 [−0.355, −0.293] | 0.0001 | −0.224 / −0.100 | +0.280 [+0.250, +0.311] |

### J.5 PIXEL-LEVEL CHECK OF THE NEGATIVE POINT

The B-core relation is stated at the object level, whereas SAM's interface labels only the clicked pixel as background. As a descriptive check added after external review, not one of the frozen tests, we read the selected masks of the paired prompt bank for $p_A$ and $p_A{+}n_B$ on the 387 LIVECell pairs and recorded whether the clicked pixel lies inside the mask. For the three full-decoder seeds, the negative point pulls the mask onto its own pixel, which lay outside the mask before, in 56.8%, 44.2%, and 31.0% of pairs, and it never removes that pixel from the mask. The same transition occurs in no pair for reference SAM, reference SAM2, micro-sam LM, or SAM2 partial tuning, in at most 2.8% for LoRA, and in at most 5.4% for SAM partial tuning. Before the negative point, the reference masks already contain the clicked pixel in 65% (SAM) and 67% (SAM2) of pairs because they often cover both objects; the negative point removes it in 11% and 21% of pairs.

### J.6 SINGLE-MASK OUTPUT

The controlled SAM adaptations compute their loss on SAM's single-mask output, whereas the audit scores the three multimask candidates selected by the IoU head, SAM's default for a single click. As a descriptive check added after internal review, not one of the frozen tests, we reran the 387 LIVECell B-core pairs with both outputs (`singlemask_check.py`). The multimask rerun reproduces Table 1 ($+0.121$ for reference SAM, $-0.068$ for seed-42 partial tuning, $-0.500$, $-0.411$, and $-0.324$ for the three full-decoder seeds). With the single-mask output, reference SAM gives $+0.206$, seed-42 partial tuning $-0.004$, and the three full-decoder seeds $-0.408$, $-0.343$, and $-0.319$ (seed 42: overlap $0.002 \to 0.410$). The reversal therefore does not depend on which output is scored.

### J.7 VERIFICATION AND SUPPORTING EVIDENCE

Two independent pipelines produce the SAM overlaps: the candidate decomposition runs SAM during analysis, and the paired prompt bank stores masks from a separate inference run whose combined prompt $p_A{+}n_B$ uses the same B-core coordinate. For eight endpoints (reference, LoRA, partial, the three full-decoder seeds, SAM2 partial, micro-sam LM), the two agree on the B-core overlaps of all 387 LIVECell units (maximum absolute difference 0.0000). Rerunning the seed-42 full-decoder endpoint on 12 randomly drawn units reproduced every saved mask. In an earlier label-swap audit

on 20 LIVECell images, which places a negative or a positive point at the same coordinate inside $B$, the negative point removed 0.139 of $B$ for reference SAM but added 0.542 for the full decoder, compared with 0.858 for a positive point.

### J.8 Qualitative example

Figure 1 shows the unit selected by a rule fixed in the figure script before rendering: among LIVECell units whose canonical Dice with $A$ is at least 0.6 for the SAM2 partial, full-decoder, and SimpleClick endpoints (225 units), we chose the one whose B-core responses under the SAM2 partial and full-decoder endpoints (seed 42) lie closest to those endpoints' means, measured as the sum of the two standardized distances. SAM2 partial and micro-sam LM were chosen as similar-quality comparators; SAM partial was added as a same-base, same-seed training-pool comparator, not a Dice-matched one. A SimpleClick column was removed without changing the selected unit. The rule originally also used SimpleClick's B-core response; after the correction in Appendix K we removed that term and reran the selection once. The four pairs in Figure 3B were drawn uniformly at random (seed 20260924) among the 280 LIVECell pairs whose canonical Dice with $A$ is at least 0.6 for the seed-42 full-decoder endpoint.

## K SimpleClick endpoints and a corrected prediction set

SimpleClick (Liu et al., 2023) is a single-output interactive segmenter that encodes positive and negative clicks as disk maps. We start from the released ViT-B checkpoint trained on COCO and LVIS and tune the click embedding, neck, and head (11,006,273 parameters) with the canonical-only objective for 2,000 steps, on 145–148 LIVECell official-training images or 195–198 PanNuke fold-1 images per seed. Each endpoint passed a canonical-Dice gate on held-out validation images. A COCO specialization failed its gate and was not analyzed.

While rendering Figure 1, we found that the saved SimpleClick predictions for the six-location negative-point prompts had been produced with the negative click entering the model as a positive click. In 60 of 60 checked cases (both domains, two locations), a fresh single-sample inference with the click labeled positive reproduced the saved mask exactly, and inference with the intended label did not. These files came from an earlier state of the batched inference code that the current code does not reproduce. The paired prompt bank, whose combined prompt $p_A{+}n_B$ uses the same B-core coordinate, matched fresh correct inference in 50 of 50 checked units. We therefore report SimpleClick's B-core response from the paired bank (Table S22) and omit its spatial scores. SimpleClick's Dice, STS, and TSO use single positive clicks and are unaffected.

Table S22: **Corrected SimpleClick B-core response** (paired prompt bank, same units and statistics as Table S18). The reference strongly removes $B$; specialized endpoints add at most 0.024 of it.

| Domain | Endpoint | $o_B$: $p_A \rightarrow p_A{+}n_B$ | $s_{\mathrm{Bcore}}$ [95% CI] | Dice$_A$: $p_A \rightarrow p_A{+}n_B$ |
|---|---|---|---|---|
| LIVECell | reference | 0.948 → 0.129 | +0.818 [+0.791, +0.844] | 0.039 → 0.352 |
| LIVECell | s42 | 0.000 → 0.001 | −0.001 [−0.002, −0.000] | 0.691 → 0.685 |
| LIVECell | s123 | 0.001 → 0.002 | −0.001 [−0.003, 0.000] | 0.702 → 0.696 |
| LIVECell | s456 | 0.002 → 0.007 | −0.005 [−0.008, −0.002] | 0.695 → 0.673 |
| PanNuke | reference | 0.929 → 0.050 | +0.879 [+0.838, +0.914] | 0.074 → 0.407 |
| PanNuke | s42 | 0.000 → 0.006 | −0.006 [−0.010, −0.003] | 0.806 → 0.811 |
| PanNuke | s123 | 0.002 → 0.026 | −0.024 [−0.031, −0.016] | 0.851 → 0.836 |
| PanNuke | s456 | 0.001 → 0.011 | −0.010 [−0.014, −0.006] | 0.846 → 0.839 |