# COOPER: CLASS-OUTPUT OPTIMIZATION VIA PER-EXAMPLE EMBEDDING REFINEMENT

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

## ABSTRACT

Methods that learn class embeddings for frozen vision–language models typically optimize them jointly across labeled examples. We compare this approach with local optimization followed by aggregation: fitting one embedding per example, then combining the fits within each class. COOPER builds a reusable class bank from these local refinements; its classification counterpart, COOPER-CLS, fits the bank jointly. Both freeze model weights, select banks using training data, and deploy fixed banks without test-time optimization. Our theory explains why the two approaches can produce different embeddings and when weighting can reduce bias from ineffective refinements. We certify that no convex combination of the sampled text-prompt embeddings reproduces the tested COOPER corrections. We evaluate COOPER on eleven segmentation datasets with SAM3 and COOPER-CLS across nine classification backbones. COOPER yields positive gains over SAM3's original embeddings on all eleven datasets, averaging $+16.30$ mIoU points, and its banks also transfer to other datasets and video. Although controlled comparisons give dataset-dependent local–joint rankings, COOPER outperforms the tested fixed-recipe joint-fitting baselines on average, including joint fitting initialized with COOPER embeddings. For cosine classification with CLIP, PE, and SigLIP, however, our controls favor joint fitting over local aggregation. Our work shows that the choice between local aggregation and joint fitting shapes how well learned class embeddings generalize.

## 1 INTRODUCTION

Vision–language models (VLMs) can segment objects specified by prompts. SAM3 (Carion et al., 2025) supports *promptable concept segmentation* (PCS): finding all instances of a concept specified by text or visual examples. With text prompts, a class embedding guides mask predictions. However, the original class embedding can limit segmentation accuracy even when the frozen model can produce better masks.

Figure 1 shows the effect of replacing the "polyp" embedding. On Kvasir-SEG, SAM3 scores 0.00 held-out mIoU with its original class embedding (*Original*). Our approach constructs a replacement from labeled training images, raising the dataset mean to 78.46 without changing model weights. The original class representation can therefore be a bottleneck. We compare joint fitting across training images with aggregation of per-example fits.

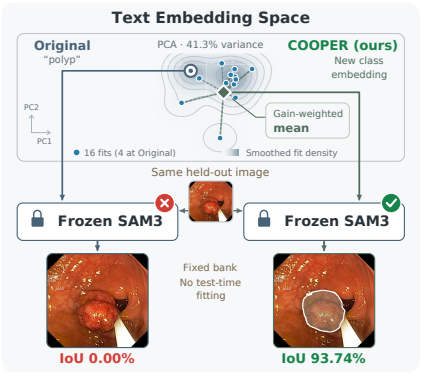



Figure 1: **Recovering the missed polyp.** COOPER merges per-example fits; SAM3 stays frozen.

Several methods use *joint optimization* to learn shared class embeddings for frozen VLMs using combined losses from multiple labeled examples (Yu et al., 2023; Silva-Rodríguez et al., 2024; Huang et al., 2024). Methods for SAM3 likewise fit shared concept embeddings (Zhou et al., 2026; Venkataramani & Sathish, 2026). Yet our SAM3 experiments show that joint fitting can lower training loss while worsening segmentation on held-out images (Figure 7). These results motivate testing local optimization followed by aggregation.



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.

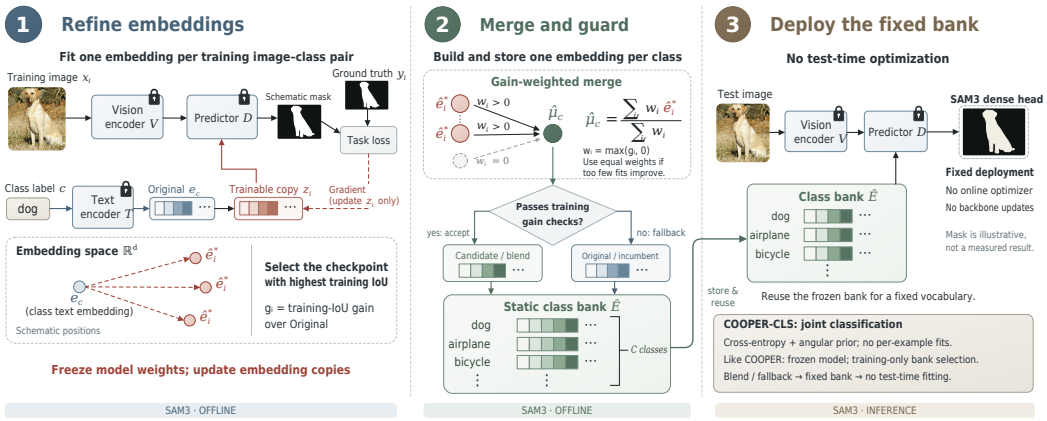



Figure 2: **Two class-bank constructions.** Both freeze model weights, use training-only selection, and deploy fixed banks. **(1–3)** COOPER refines and aggregates SAM3 embeddings; COOPER-CLS instead fits its bank jointly.

**Optimize jointly, or locally and then aggregate?**
Joint optimization fits class embeddings using multiple training examples. Local optimization instead fits a separate embedding to each example. Aggregation then combines these fits into one embedding per class.

To answer this question, we propose COOPER. Across eleven datasets, it outperforms both tested fixed-recipe joint-fitting baselines on average, including joint fitting initialized with COOPER embeddings (Table 2). It constructs a separate embedding for each training image–class pair using only that pair's supervision, with SAM3 frozen. This is *local optimization*. We then estimate one embedding per class by aggregating these local fits. Each fit's *endpoint* is its checkpoint with the highest training IoU. Gain weighting and a training-only guard produce a fixed class bank (Figure 2). Matched ablations test these policies (Table 3; Appendix E.4). The bank is fixed before deployment; the guard does not guarantee generalization.

No convex combination of the sampled text-prompt embeddings reproduces the tested corrections. This motivates adapting embeddings directly without ruling out useful continuous prompts (Section 2). Our theory explains why averaging local fits can differ from joint optimization and when weighting can reduce persistent bias from ineffective fits (Section 3).

Across eleven benchmarks, COOPER improves on SAM3's original embeddings in every dataset, averaging $+16.30$ mIoU points across natural-scene, aerial, and medical segmentation.

We test the same local-versus-joint question in classification, where class scores compete. Controls on CLIP, PE, and SigLIP favor joint fitting (Appendix J.3; Figure 10), motivating COOPER-CLS. It retains COOPER's frozen weights, training-only bank selection and fixed deployment, but fits its bank jointly with an anchor to Original. These comparisons do not establish geometry as the cause of the difference. Together, the findings show that effective adaptation depends on how embeddings are fitted, not only on which representation is updated.

**Contributions.**

- **Learning reusable class embeddings.** COOPER combines per-example refinements into a reusable SAM3 class bank. COOPER-CLS uses anchored joint fitting for classification, where matched controls favor joint fitting over local aggregation (Sections 4 and 5).
- **Understanding the limits of prompt adaptation.** We show that convex combinations of sampled prompt embeddings cannot reproduce the tested corrections, and distinguish this limitation from prompt-tuning failures that depend on the training recipe (Section 2).
- **Why local and joint fitting differ.** Our theory explains when local and joint fitting produce different embeddings and when weighting can reduce bias. Experiments test how well each approach generalizes in SAM3 and cosine classification (Sections 3 and 5).
- **Evaluation across datasets and models.** We test eleven segmentation datasets with SAM3 and nine classification backbones, measure construction cost, and evaluate transfer to new datasets and video. Ablations compare components with simpler alternatives (Section 5).

## 2 WHY ADAPT THE PREDICTOR-FACING REPRESENTATION?

SAM3's training labels may not fully describe a downstream task; its original study reports difficulty with specialized medical terms (Carion et al., 2025, Sec. 4; App. B). COOPER replaces only the *polyp* class embedding, raising held-out Kvasir-SEG mIoU from 0.00 to 78.5 with all model weights frozen. Figure 3 motivates adapting this embedding directly rather than relying only on changes to the text prompt.

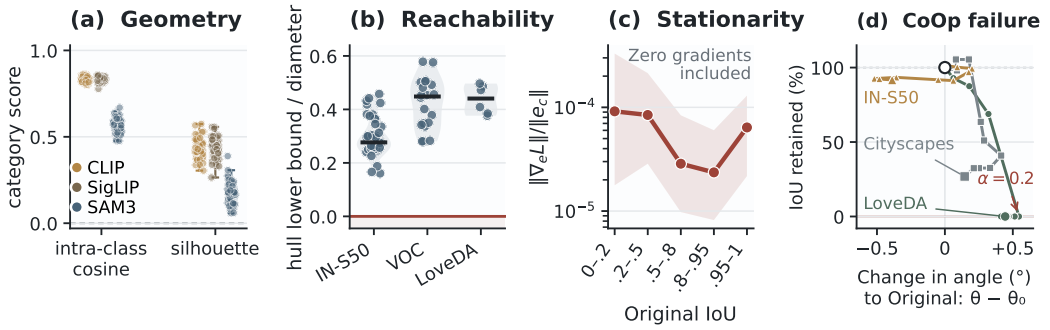



Figure 3: **Predictor-input diagnostics. (a)** Same class prompts, different representation geometries. **(b)** Certified lower distance bounds to sampled prompt mixtures. **(c)** Relative gradients by Original IoU; median and interquartile range. **(d)** CoOp probe IoU retained versus change in angle to Original, both relative to initialization.

**Prompt similarity depends on how the model uses text.** CLIP and SigLIP use class directions to score images; SAM3 uses text features to guide masks. Using the same class prompts across models, pooled CLIP/SigLIP features have higher within-category alignment (median cosine $0.83/0.84$ versus $0.56$) and category separation (median silhouette $0.42/0.44$ versus $0.18$) than SAM3's predictor-facing features (Figure 3a). Appendix B tests how embedding direction and length affect the gain. These differences motivate adapting the representation that the predictor actually uses. For SAM3, we therefore refine the text features that directly guide mask prediction, rather than treating pooled text similarity as a proxy.

**Averaging sampled prompts cannot reproduce the corrections.**
Could rewording a class prompt recover a COOPER embedding? We encode class names, templates and descriptions with frozen $\mathcal{T}$, forming $\mathcal{P}_c$. Its convex hull, $\operatorname{conv}\mathcal{P}_c$, contains every weighted average with nonnegative weights summing to one. For COOPER embeddings $b_c$,

$$\operatorname{dist}(b_c, \operatorname{conv}\mathcal{P}_c) \geq \ell_c > 0 \qquad \text{for all 76/76 tested classes.}$$

The positive bounds account for solver error (duality gap; Appendix B.3). No convex combination reaches these corrections in the three tested domains (Figure 3b). This excludes the sampled combinations, not all strings, continuous prompts, or other embeddings with equally useful predictions.

**A learning signal does not ensure a useful update.** Low-IoU pairs have larger median relative gradients, although some failures have zero gradient (Figure 3c). CoOp can improve IN-S50 yet degrade road and aerial segmentation (Table 1). In LoveDA's first-update scan, probe IoU falls to zero after only a $0.55°$ change in angle to Original and stays zero when queries return; IN-S50 retains at least 91% of its initial IoU (Figure 3d). The trace does not isolate an angle-only cause. Smaller updates or query fallback recover LoveDA probe IoU with the same frozen model (Appendices M.6 and E). The fitting procedure therefore matters for turning the model's existing capability into useful mask predictions.

Table 1: **Recipe-dependent tuning outcomes.** SAM3 gains (pp), same runs as Table 2. Small-step/fallback controls recover LoveDA probe IoU (Appendix E).

| Dataset | COOPER | LoRA (r32) | CoOp-C | CoOp-B |
|---|---|---|---|---|
| VOC | **+6.55** | +4.36 | +2.62 | -2.28 |
| LoveDA | **+18.41** | +13.10 | -27.14 | -10.45 |
| CamVid | **+12.47** | +0.00 | -33.80 | -26.60 |
| IN-S50 | +8.64 | -1.42 | **+11.01** | +9.06 |
| Mean | **+11.52** | +4.01 | -11.83 | -7.57 |

**From a correction to a class embedding.** Should we fit one embedding jointly across examples, or fit locally and then aggregate? Section 3 explains why these can differ; Section 4 develops COOPER and its classification counterpart.

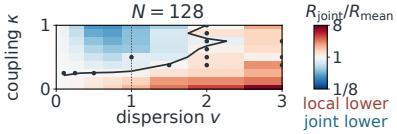
## 3 Local and Joint Estimators: Targets and Risk

Does averaging local fits give the same embedding as joint optimization, and should every fit count equally? Our theory separates these two questions. For a measurable per-example minimizer $e^\star$ and class distribution $\mathcal{D}_c$, define (assuming expectations and minima exist)

$$\mu_c^{\mathrm{pop}} := \arg\min_e \mathbb{E}_{\mathcal{D}_c}[L(e; x, y)], \qquad \mu_c^{\mathrm{avg}} := \mathbb{E}_{\mathcal{D}_c}[e^\star(x, y)]. \tag{1}$$

**Proposition 1** (Curvature–endpoint coupling). *Let $L_i(e) = a_i + \frac{1}{2}(e - e_i^\star)^\top H_i (e - e_i^\star)$ on $\mathbb{R}^d$, with $H_i \succeq 0$ and $\bar{H}_N = N^{-1}\sum_i H_i \succ 0$. Define $\mu_c^{\mathrm{avg}}[N] = N^{-1}\sum_i e_i^\star$, $\xi_i = e_i^\star - \mu_c^{\mathrm{avg}}[N]$, and $e_c^{\mathrm{joint}} = \arg\min_e N^{-1}\sum_i L_i(e)$. Then*

$$\underbrace{e_c^{\mathrm{joint}} - \mu_c^{\mathrm{avg}}[N]}_{\text{estimator gap}} = \bar{H}_N^{-1} \underbrace{\frac{1}{N}\sum_i (H_i - \bar{H}_N)\xi_i}_{\text{curvature–endpoint coupling } C_N} \tag{2}$$




The heatmap beside Equation 2 shows Corollary 1's model. Here $v$ measures curvature spread; $\kappa \in [0, 1]$ controls how strongly curvature and endpoint position are associated. It is not the sample coupling $C_N$. Risk is expected loss on new examples above the population optimum, averaged over training sets: $R_{\mathrm{joint}}$ for joint fitting and $R_{\mathrm{mean}}$ for uniform local averaging (lower is better).

**Corollary 1** (Variance and target bias). *For $N$ endpoints in Appendix M.3's $d$-dimensional model, $R_{\mathrm{mean}} = d/(2N) + (\kappa v)^2/2$. At $\kappa = 0$, $R_{\mathrm{joint}} \geq R_{\mathrm{mean}}$; for fixed $\kappa v \neq 0$, $R_{\mathrm{joint}} = O(N^{-1}) \to 0$.*

**Reading the two plots.** Figure 4 follows the heatmap's dashed $v = 1$ line at $N = 16, 32, 128$, using risk differences rather than ratios. Positive values favor local; negative values favor joint. Even with fixed curvature spread, local and joint fitting can differ in population risk; coupling can reverse which performs better. At $\kappa = 0$, curvature and endpoints are independent, so joint fitting adds variance. For $\kappa > 0$, local bias persists; joint fitting wins for large $N$. Lower training loss need not imply lower risk (Appendix M.3).

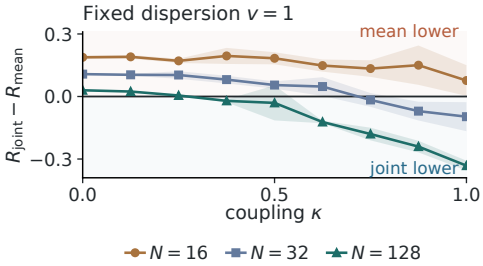



Figure 4: **Synthetic crossover.** Simulation of Corollary 1 with uncertainty bands.

**Not every endpoint supplies a transferable repair.** Should every returned fit contribute equally? Ineffective endpoints can bias the class average even with more examples. We study this using an assumed repair target and evaluation loss $Q$, not COOPER's training objective.

**Proposition 2** (Heterogeneous repair). *Let $(Z_i, B_i, W_i)$ be i.i.d., with $\mathbb{E}\|Z\|_2^2 < \infty$, repair indicator $B \in \{0, 1\}$, $\Pr(B = 1) = \pi \in (0, 1)$, and $\mu_b = \mathbb{E}[Z \mid B = b]$. Interpret $\mu_1$ as the repair target and $\mu_0$ as the ineffective-endpoint mean. For $H \succ 0$ and $Q(e) = \frac{1}{2}\|e - \mu_1\|_H^2$, the uniform mean satisfies*

$$\mathbb{E}Q(\bar{Z}_N) = \frac{\mathrm{tr}(H\,\mathrm{Cov}\,Z)}{2N} + \underbrace{\frac{(1-\pi)^2}{2}\|\mu_0 - \mu_1\|_H^2}_{\text{bias not removed by more endpoints}}.$$

*For a fixed rule assigning weight $W \geq 0$ with $0 < \mathbb{E}W < \infty$ and $\mathbb{E}[W\|Z\|_2] < \infty$, the normalized merge converges almost surely to $m_W = \mathbb{E}[WZ]/\mathbb{E}W$. Writing $a_b = \mathbb{E}[W \mid B = b]$ and $b_W = \mathbb{E}[W(Z - \mu_B)]$ gives*

$$m_W - \mu_1 = \frac{(1-\pi)a_0(\mu_0 - \mu_1) + b_W}{\pi a_1 + (1-\pi)a_0}.$$

*If $b_W = 0$ (no net within-component shift), $a_0 = 0 < a_1$ excludes ineffective fits and recovers $\mu_1$; positive $a_0$ leaves bias when the component means differ.*

**Fitting and aggregation address different sources of error.**
Local and joint fitting can differ in target and generalization risk; uniform local averaging can also retain bias from ineffective fits. These distinctions motivate two comparisons in Section 5: local versus joint fitting, and uniform versus weighted aggregation. The results show domain-dependent local–joint rankings and a large weighting benefit on Kvasir, with smaller, mixed effects elsewhere. Proofs appear in Appendix M.

# 4 METHOD

We first describe COOPER for SAM3: refine one embedding per training example, aggregate the refinements, and deploy a fixed class bank (Figure 5; Appendix E.2). We then introduce COOPER-CLS, our joint-fitting classification counterpart. It retains frozen weights, training-only selection and fixed deployment, but fits class embeddings jointly for cosine classification (Section 5).

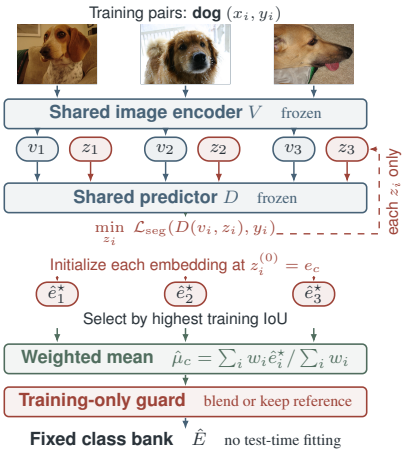


Figure 5: COOPER: local fits to a class bank.

**Step 1: local refinement.** Each image–class pair $(x_i, c_i)$ starts from its own copy $z_i^{(0)} = e_{c_i} = \mathcal{T}(c_i)$. Adam $\mathcal{A}_t$ fits it to mask $y_i$ through the frozen predictor $F_{x_i}$. The star denotes selection, not an exact minimizer. Since lower loss need not mean better masks, select the state with the highest training IoU $S_i \in [0, 1]$ among scored states $\mathcal{I}_i$, including initialization.

$$z_i^{(t+1)} = \mathcal{A}_t\Big(z_i^{(t)}, \nabla_z \mathcal{L}_i(F_{x_i}(z_i^{(t)}), y_i)\Big), \qquad \hat{e}_i^\star = z_i^{(t^\star)}, \quad t^\star = \arg\max_{t \in \mathcal{I}_i} S_i(z_i^{(t)}). \tag{3}$$

**Mask and query-activity loss.** Each query proposes a mask. From mask logits $m_{iqp}$ and query logits $a_{iq}$, set $Q_i = \{q : a_{iq} > 0\}$ and $\hat{m}_{ip} = \max_{q \in Q_i} \sigma(m_{iqp})$. Here $p$ indexes pixels. If $Q_i = \emptyset$, use the top-scoring query during training to keep mask gradients available. The full objective (dataset recipes: Appendix J.7) is

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{BCE}}(\hat{m}_i, y_i) + \mathcal{L}_{\text{Dice}}(\hat{m}_i, y_i) + 0.05 \operatorname{softplus}\Big(-\max_q a_{iq}\Big). \tag{4}$$

BCE penalizes pixelwise errors; Dice (Milletari et al., 2016) emphasizes foreground overlap. In SAM3's DETR-style decoder (Carion et al., 2020; 2025), a useful mask can still be excluded by an inactive query. The last term therefore encourages query activity for each positive training pair.

**Step 2: merge and guard.** To limit dilution by weak repairs, merge selected fits (endpoints) using their gains in training IoU $S$ over Original:

$$\begin{aligned} g_i &:= S(\hat{e}_i^\star; x_i, y_i) - S(e_c; x_i, y_i), \quad A_c := \sum_{i:c_i=c} \mathbf{1}[g_i > \tau_{\text{act}}], \\ w_i &:= \begin{cases} \max(g_i, 0), & A_c \geq 3, \\ 1, & A_c < 3, \end{cases} \qquad \hat{\mu}_c := \frac{\sum_{i:c_i=c} w_i \hat{e}_i^\star}{\sum_{i:c_i=c} w_i}. \end{aligned} \tag{5}$$

The gain-weighting gate requires at least three gains above $\tau_{\text{act}} = 0.005$, then weights all positive gains. Otherwise, uniform averaging avoids relying on too few clear improvements (Table 3).

**Training-only guard.** Per-example improvements need not survive aggregation. We select the deployed embedding $b_c$ on training data: a blend $\tilde{\mu}_c = (1 - \alpha_c)e_c + \alpha_c \hat{\mu}_c$, Original, or a previously accepted embedding (the *incumbent*). This is an empirical check, not a held-out guarantee (Appendix E.4). Weighting combines local fits; blending and fallback control the correction's strength.

**Step 3: deploy the fixed bank.** Store one embedding $b_c$ per class—a token sequence for SAM3—and reuse it with the frozen predictor; no test-time fitting.

**COOPER-CLS: joint classification fitting.** For $N$ images and $K$ classes, normalize class embeddings $E$, Original $E^{(0)}$, and frozen features $v_i$:

$$\mathcal{L}_{\text{cls}}(E) = \frac{1}{N} \sum_{i=1}^{N} \text{CE}\big(\tau v_i E^\top, y_i\big) + \frac{\beta}{K} \sum_{c=1}^{K} \Big[1 - E_c^\top E_c^{(0)}\Big]. \tag{6}$$

$\tau$ scales logits; $\beta$ weights the angular prior. We select one recipe per bank using training data; one shot retains Original because no example remains for a separate check (Appendix J.3). Cross-entropy fits competing class scores; anchoring limits drift from Original when support is scarce.

Table 2: **SAM3 across eleven domains.** Gains use matched Original references. Joint-R/O: fixed-checkpoint replay from Original/COOPER. Protocols: Appendix F.

| Dataset | Benchmark protocol | | | | **Joint optimization** | | Benchmark |
|---|---|---|---|---|---|---|---|
| | Original mIoU (%) | CoOp-C | CoOp-B | LoRA | Joint-R | Joint-O | **COOPER** |
| | | Gain (pp); [95% interval] and time (min) where available | | | | | |
| ADE-847<br>Scene parsing | 35.19 | -5.74<br>$[-7.58, -3.94]$<br>219.3 | -28.23<br>$[-30.31, -26.10]$<br>185.8 | +0.00<br>$[0.00, 0.00]$<br>202.8 | +2.32<br>$[1.40, 3.28]$<br>168.4 | +2.40<br>$[1.50, 3.34]$<br>166.8 | **+2.75**<br>$[1.86, 3.68]$<br>135.4 |
| PC-59<br>Scene parsing | 60.75 | +7.29<br>$[3.17, 11.49]$<br>564.0 | -10.20<br>$[-17.15, -3.71]$<br>543.5 | +2.97<br>$[0.72, 5.25]$<br>241.8 | +3.84<br>$[-0.10, 7.98]$<br>147.7 | +3.11<br>$[-0.39, 6.79]$<br>147.2 | **+9.88**<br>$[7.23, 12.89]$<br>139.3 |
| VOC<br>Object masks | 75.98 | +2.62<br>$[-3.14, 7.39]$<br>175.7 | -2.28<br>$[-12.52, 6.52]$<br>243.4 | +4.36<br>$[1.41, 7.40]$<br>195.7 | -6.98<br>$[-15.67, 0.16]$<br>78.2 | -1.53<br>$[-7.95, 5.61]$<br>78.1 | **+6.55**<br>$[2.51, 12.06]$<br>41.4 |
| IN-S50<br>Object masks | 78.39 | **+11.01**<br>$[5.28, 17.75]$<br>24.3 | +9.06<br>$[3.12, 15.92]$<br>30.4 | -1.42<br>$[-3.27, 0.22]$<br>7.7 | +1.94<br>$[-6.13, 9.91]$<br>28.6 | +0.99<br>$[-6.17, 8.67]$<br>28.2 | +8.64<br>$[4.02, 14.38]$<br>$\geq$7.5 |
| COCO<br>Object masks | 50.81 | -7.81<br>$[-12.00, -3.41]$<br>232.1 | -20.94<br>$[-26.04, -15.66]$<br>216.1 | **+10.90**<br>$[8.57, 13.45]$<br>171.8 | +5.31<br>$[1.97, 8.87]$<br>256.5 | +3.77<br>$[0.26, 7.46]$<br>256.0 | **+10.90**<br>$[8.11, 14.02]$<br>228.1 |
| Cityscapes<br>Road scenes | 55.46 | -55.52<br>$[-65.24, -45.25]$<br>343.2 | -54.21<br>$[-64.07, -43.92]$<br>350.5 | **+7.78**<br>$[5.52, 10.47]$<br>487.5 | +2.41<br>$[-6.55, 9.78]$<br>87.5 | +6.11<br>$[2.45, 10.96]$<br>85.7 | +5.65<br>$[2.92, 8.86]$<br>58.8 |
| CamVid<br>Road scenes | 35.23 | -33.80<br>$[-45.70, -22.35]$<br>87.2 | -26.60<br>$[-39.16, -14.48]$<br>94.7 | +0.00<br>$[0.00, 0.00]$<br>66.6 | +18.89<br>$[11.69, 26.41]$<br>61.6 | +21.92<br>$[14.69, 29.45]$<br>61.1 | **+12.47**<br>$[7.47, 18.01]$<br>58.2 |
| LoveDA<br>Aerial | 27.14 | -27.14<br>$[-40.99, -13.51]$<br>281.3 | -10.45<br>$[-25.23, 2.05]$<br>327.3 | +13.10<br>$[8.02, 19.32]$<br>288.5 | +10.11<br>$[-2.07, 21.73]$<br>41.9 | +12.56<br>$[4.42, 23.06]$<br>39.6 | **+18.41**<br>$[8.46, 30.06]$<br>$\geq$43.5 |
| Kvasir-SEG<br>Medical | 0.00 | +79.73<br>$[75.86, 83.23]$<br>53.3 | **+85.78**<br>$[83.53, 87.78]$<br>53.3 | +0.00<br>$[0.00, 0.00]$<br>11.4 | +0.00<br>$[0.00, 0.00]$<br>0.7 | +80.30<br>$[77.09, 83.32]$<br>0.6 | +78.46<br>$[74.86, 81.78]$<br>2.3 |
| PascalPart<br>Object parts | 30.15 | -2.67<br>$[-6.44, 0.98]$<br>121.5 | -14.30<br>$[-17.90, -10.69]$<br>114.9 | +14.80<br>$[12.62, 17.04]$<br>90.3 | +18.96<br>$[16.50, 21.49]$<br>128.0 | +20.58<br>$[17.97, 23.18]$<br>127.5 | **+18.41**<br>$[16.09, 20.74]$<br>257.5 |
| LVIS<br>Long tail | 53.24 | -18.37<br>$[-20.40, -16.34]$<br>189.1 | -32.37<br>$[-34.78, -29.91]$<br>186.2 | **+14.85**<br>$[13.32, 16.40]$<br>133.5 | +4.05<br>$[2.28, 5.87]$<br>134.5 | +3.14<br>$[1.36, 4.92]$<br>133.4 | +7.16<br>$[5.93, 8.45]$<br>267.8 |
| **Mean** | 45.67 | -4.58 | -9.52 | +6.12 | +5.53 | +13.94 | **+16.30** |

## 5 Experimental Evaluation

**Evaluation protocol.** We construct and select banks on training data; guard development uses LoveDA/VOC (Appendix J). We report class-mean binary IoU (SAM3), top-1 accuracy (classification), and gains in percentage points (pp). Banks and backbones stay fixed during held-out evaluation.

**Consistent gains across all eleven datasets.** COOPER improves frozen SAM3 across all eleven evaluated datasets, averaging +16.30 mIoU points (Table 2). Gains span Original's failure on Kvasir and its strong VOC/IN-S50 performance. Refinement thus helps recover missed foreground and improve already useful predictions.

Table 3: **COOPER weighting ablation.** $\Delta = W - U$: weighted minus uniform. Same fits, support, guard rules and held-out pairs; selection is rerun. Low-gain mass: weight on fits with $g \leq 0.005$, before selection. Appendix E.3.

| **Dataset** | $\Delta$ (pp) | 95% CI | Low-gain mass (%) $U \to W$ |
|---|---|---|---|
| ADE-847 | +0.67 | $[-0.02, 1.38]$ | $4.3 \to < 0.001$ |
| PC-59 | +0.04 | $[-0.62, 0.80]$ | $13.0 \to 0$ |
| VOC | +0.12 | $[-0.30, 0.53]$ | $18.1 \to 0$ |
| IN-S50 | +1.93 | $[-0.19, 4.79]$ | $34.2 \to 6.6$ |
| COCO | -1.05 | $[-1.84, -0.38]$ | $6.0 \to < 0.001$ |
| Cityscapes | -0.21 | $[-0.91, 0.35]$ | $7.9 \to 0$ |
| CamVid | +0.89 | $[0.07, 1.89]$ | $3.8 \to < 0.001$ |
| LoveDA | -0.38 | $[-0.68, -0.05]$ | $0.5 \to 0.003$ |
| **Kvasir** | +37.53 | $[31.85, 43.28]$ | $25.0 \to 0.006$ |
| PascalPart | +0.58 | $[0.10, 1.22]$ | $4.4 \to 0.008$ |
| LVIS | -1.53 | $[-2.38, -0.72]$ | $18.2 \to 2.862$ |
| **Mean** | +3.51 | — | $12.3 \to 0.862$ |

**What the baselines test.** CoOp-B underperforms CoOp-C on average despite extra class-specific parameters (Zhou et al., 2022b): capacity alone does not ensure adaptation. LoveDA's CoOp collapse (Figure 3d) is recoverable: smaller updates or query fallback restore probe IoU (Appendix E). CoOp's Kvasir gains confirm useful prompt capacity. For LoRA (Hu et al., 2022), the three zero gains are

*selected Original fallbacks*; Kvasir enters a zero-gradient loss branch, whereas Cityscapes/LVIS benefit (Appendix K). Together, these controls show why usable gradients and controlled updates matter alongside trainable capacity.

**When weighting helps.** Proposition 2 links bias to the fraction of ineffective fits and the gap between their mean and the repair target. Low gain can reflect an effective Original embedding (VOC/IN-S50). Weighting avoids dilution on Kvasir (+37.53 pp), but can discard useful support or shift the target elsewhere (−1.53 to +1.93 pp; Table 3); low gain alone does not make a fit useless. Uniform local averaging also improves all eleven datasets in the separate fixed-checkpoint replay (+12.96 pp mean; Appendix E), so the benefit is not confined to gain weighting.

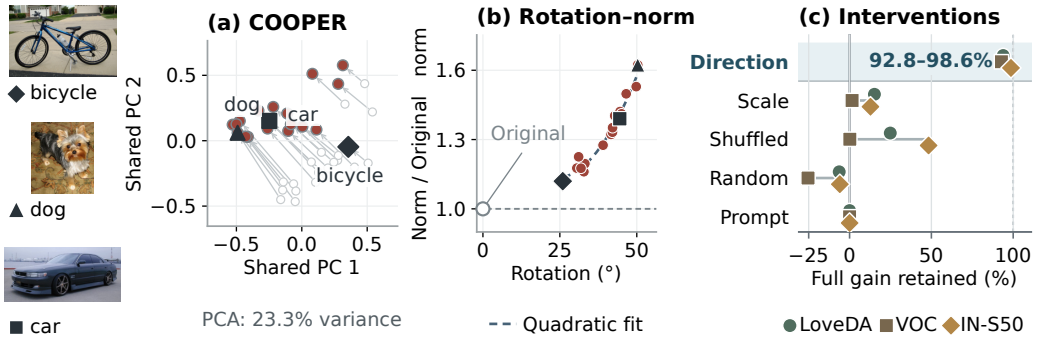



Figure 6: **Geometry and predictor interventions. (a)** VOC PCA (23.3% variance); arrows join Original and COOPER endpoints. **(b)** Native rotation and norm; curve: descriptive quadratic fit across classes. **(c)** Gain retained under five interventions. Photos identify classes.

**What carries the correction?** Figure 6 describes the displacement (a,b) and tests which changes carry the gain (c). Restoring Original's norm preserves 92.8%–98.6% of the gain, favoring changed direction over increased embedding length. Random moves with matched norms fail, and shuffling class assignments weakens gains. ImageNet-S50 retains a positive effect after shuffling, suggesting a domain-wide component alongside class-specific repairs. Appendix C extends the rotation–norm analysis to three datasets. Together, the controls support structured, mainly directional corrections that partly depend on class identity.

**Better fitting need not generalize.** Figure 7 compares direct deployment of the local bank with joint fitting from Original or local initialization. On VOC and LoveDA, both joint initializations reach lower training loss but higher held-out loss and lower mIoU than the local bank. VOC's Original-initialized run cuts training loss 39.2% over steps 50–200 while held-out loss rises 32.0% and mIoU drops by 7.91 points. This late divergence is consistent with overfitting support-specific patterns.

CamVid instead gains 9.34 mIoU points with lower held-out loss. A separate three-seed, tuning-matched study retains modest local advantages on VOC/LoveDA and favors joint fitting on CamVid (Appendix E.1). A local bank can therefore be deployed directly or serve as an initialization; the benefit of further joint fitting depends on the domain.

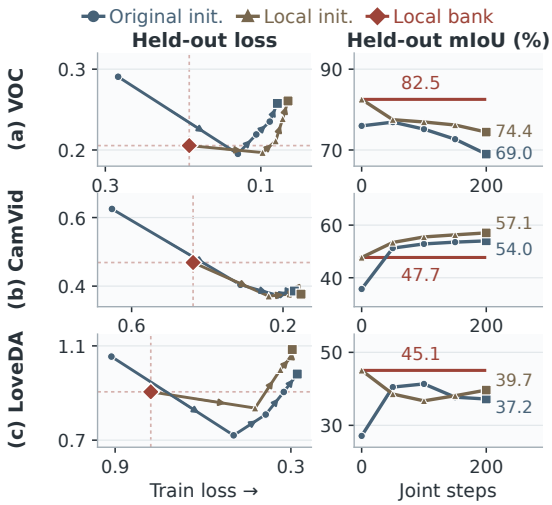



Figure 7: **Joint fitting across domains.** Loss paths (left), held-out mIoU (right). Red diamonds/lines: fixed shared local banks, not per-example minima; squares: step 200. LoveDA: gain-mean replay. Protocols: Appendix E.

**Held-out repairs.** Figure 8 shows recovered foreground and reduced spill with frozen weights. Embedding adaptation can therefore adjust the target boundary in both directions, recovering missed regions and removing excess foreground.

**Offline construction, fixed deployment.** Local fitting scales with support: $M$ pairs and $T$ steps require at most $MT$ single-pair updates, plus extraction, scoring and calibration. Cached features avoid encoder backpropagation; parallelism saves time, not work. Construction is faster than CoOp-C on $7/9$ fully timed domains and LoRA on $6/9$, without GPU-matched budgets. Joint-O's time excludes bank construction (Appendix F). Reuse spreads fitting cost over predictions, for COOPER and cached-prompt baselines.

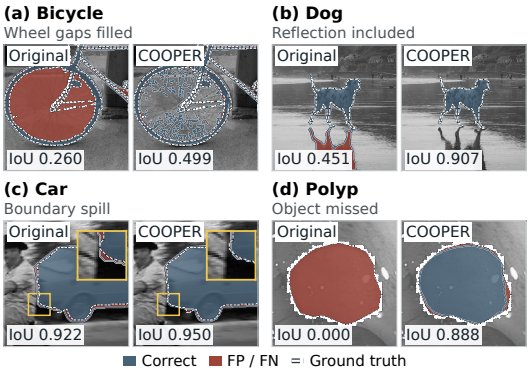



Figure 8: **Four illustrative held-out repairs.** Blue: correct foreground; red: false positives/negatives; white dashes: ground truth. Matched crops and car insets; full-image IoUs. Selected examples illustrate capability, not frequency.

**Cross-domain transfer without target fitting.** We keep SAM3 fixed and reuse source-trained banks for shared class names (Figure 9a). All nine domain means improve over Original; the breadth of class-level gains varies (Appendix H.1), and target-trained banks usually do better. These gains support the purpose of local aggregation: turning image-specific refinements into conditioning that remains useful beyond the construction domain.

**Image-to-video transfer without video training.** We deploy the Pascal Context-59 image bank in SAM3's frozen video predictor on 343 VSPW validation clips (Miao et al., 2021); no VSPW frames are used to fit or select the bank. Figure 9b averages class–frame IoU within successive frame ranges, including absent targets.

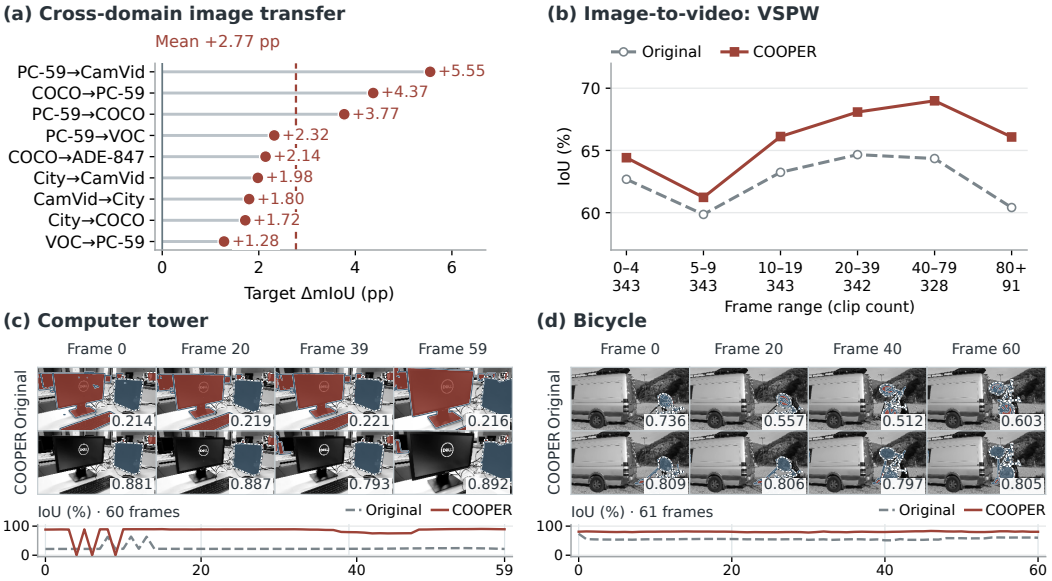



Figure 9: **Cross-domain image and video transfer.** Four evenly spaced frames and full traces illustrate VSPW transfer (c,d). Full-image IoUs; mask colors follow Figure 8.

COOPER's advantage generally widens at later frames: $+1.73$ pp at 0–4 versus $+5.67$ pp at 80+ (Figure 9b). Later bins contain fewer clips; pairing the first and last ten frames within 329 clips gives a $+2.02$-pp mean gain increase (95% CI $[+0.82, +3.25]$), though the median interval crosses zero (Appendix H.2). Image-trained corrections can persist: the computer correction recovers after initial failures; bicycle gains persist through motion (c,d).

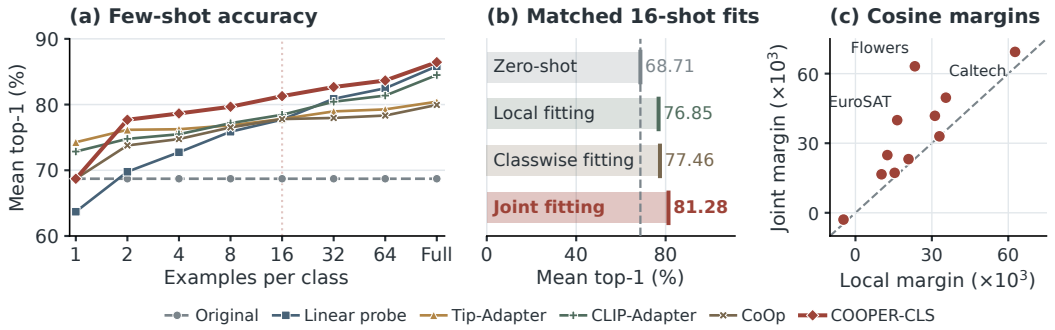



Figure 10: **CLIP: accuracy and margins.** Eleven datasets. **(b,c)** Matched support, selection, guard and refit; ranking uses extra training labels. Appendices J.3, G.

**Cosine classifiers favor joint fitting.** We test the same local-versus-joint choice with competing cosine class scores. COOPER-CLS leads the plotted baselines from two shots (Figure 10a), approaching linear probing with more support, consistent with anchoring noisy estimates; one shot retains Original. Classwise fitting avoids endpoint averaging yet recovers only 0.61 pp of joint fitting's 4.43 pp lead over local fitting (b). Local and classwise fitting use fixed Original competitors that change when the bank is assembled; joint fitting updates them together. Wider margins despite lower mean true-class scores (c) reflect larger reductions in rival scores.

**The benefit of local versus joint fitting depends on the setting.**
Local aggregation improves SAM3 across domains; matched controls reveal domain-dependent rankings, while cosine-head controls favor joint fitting.

## 6 Related Work

**Adapting embeddings.** TaskRes, CLAP, and LP++ fit class embeddings (Yu et al., 2023; Silva-Rodríguez et al., 2024; Huang et al., 2024); TPS and DPE refine them per test image (Sui et al., 2025; Zhang et al., 2024). CLIP-Adapter and Tip-Adapter use an adapter or visual cache (Gao et al., 2021a; Zhang et al., 2022). FS-CPL learns a shared decoder-facing prompt (Venkataramani & Sathish, 2026); M2C jointly fits an embedding correction within an annotation loop (Zhou et al., 2026). COOPER instead fits per-example embeddings and then aggregates them. Joint-R/O test joint fitting, not these methods' full recipes.

**Adapting prompts.** Prompt learning changes tokens at different layers or in different branches (Zhou et al., 2022b;a; Khattak et al., 2023a; Jia et al., 2022; Zang et al., 2022), with regularization and distillation (Yao et al., 2023; Zhu et al., 2023; Khattak et al., 2023b; Li et al., 2024). Textual Inversion, PALAVRA, and LiteEmbed learn composable tokens (Gal et al., 2023; Cohen et al., 2022; Agarwal et al., 2026).

**Local estimators and geometry.** Local optimization and averaging are established in distributed optimization (McMahan et al., 2017; Karimireddy et al., 2020; Stich, 2019) and model soups (Wortsman et al., 2022). COOPER separates per-example fitting from class-level aggregation with frozen weights. Proposition 1 isolates curvature–endpoint coupling as the difference from joint fitting in an idealized model, not a held-out risk ranking. Anisotropy and modality gaps motivate diagnostics rather than impossibility claims (Liang et al., 2022; Schrodi et al., 2025; Ethayarajh, 2019; Bogolin et al., 2022).

## 7 Conclusion

In this paper, we introduced COOPER, turning per-example embedding repairs into reusable class banks without relearning vision. Geometry and theory motivate adapting predictor-facing embeddings through local optimization and aggregation. Table 2's fixed-recipe comparisons favor COOPER on average; matched cosine-head controls favor joint fitting, motivating COOPER-CLS. COOPER improves SAM3 on all eleven segmentation datasets and reuses banks across domains and video. This tailors adaptation to the predictor: banks are constructed offline, model weights stay frozen, and inference requires no test-time optimization.

## AI USE STATEMENT

In this work, we used generative AI tools to draft portions of the manuscript and assist in producing result visualisations. We have not used generative AI tools to develop research ideas, design experiments, implement the proposed method, analyse experimental results, or formulate conclusions. Additionally, we used generative AI tools to improve readability through grammar, spelling, and word-choice edits. We have reviewed all AI-assisted work, and manuscript drafts produced with AI assistance were subsequently revised by the authors. All research ideas, experimental design, implementations, experiments, analyses, and conclusions are the authors' own. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

## REFERENCES

Aishwarya Agarwal, Srikrishna Karanam, and Vineet Gandhi. LiteEmbed: Adapting CLIP to rare classes. *arXiv preprint arXiv:2601.09661*, 2026.

Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025. doi: 10.48550/arXiv.2504.13181.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *ECCV*, 2014. URL https://data.vision.ee.ethz.ch/cvl/datasets_extra/food-101/.

Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009. doi: 10.1016/j.patrec.2008.04.005.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *CVPR*, pp. 1209–1218, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Caesar_COCO-Stuff_Thing_and_CVPR_2018_paper.html.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. doi: 10.48550/arXiv.2511.16719.

Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, pp. 1971–1978, 2014. URL https://openaccess.thecvf.com/content_cvpr_2014/html/Chen_Detect_What_You_2014_CVPR_paper.html.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pp. 3606–3613, 2014. URL https://openaccess.thecvf.com/content_cvpr_2014/html/Cimpoi_Describing_Textures_in_2014_CVPR_paper.html.

Niv Cohen, Rinon Gal, Eli A. Meirom, Gal Chechik, and Yuval Atzmon. "This Is My Unicorn, Fluffy": Personalizing frozen vision-language representations. In *European Conference on Computer Vision (ECCV)*, 2022.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, pp. 3213–3223, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Cordts_The_Cityscapes_Dataset_CVPR_2016_paper.html.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Mark Everingham, Luc Van Gool, Chris K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes challenge 2012 (VOC2012) results, 2012. URL https://www.robots.ox.ac.uk/~vgg/projects/pascal/VOC/voc2012/.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In *CVPR Workshop on Generative-Model Based Vision*, 2004. URL https://authors.library.caltech.edu/records/wj654-7v623.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations (ICLR)*, 2023.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021a.

Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. *arXiv preprint arXiv:2106.03149*, 2021b. doi: 10.48550/arXiv.2106.03149.

Agrim Gupta, Piotr Dollár, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, pp. 5356–5364, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Gupta_LVIS_A_Dataset_for_Large_Vocabulary_Instance_Segmentation_CVPR_2019_paper.html.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. URL https://github.com/phelber/EuroSAT.

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 (ICLR)*, 2022.

Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. LP++: A surprisingly strong linear probe for few-shot CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

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 (MMM)*, pp. 451–462, 2020. doi: 10.1007/978-3-030-37734-2_37.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, 2022.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, 2023a.

Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023b.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *ICCV Workshops*, 2013. URL https://openaccess.thecvf.com/content_iccv_workshops_2013/W19/papers/Krause_3D_Object_Representations_2013_ICCV_paper.pdf.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. PromptKD: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *NeurIPS*, 2022.

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 *ECCV*, 2014. URL https://www.microsoft.com/en-us/research/wp-content/uploads/2014/09/LinECCV14coco.pdf.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. doi: 10.48550/arXiv.1306.5151.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. VSPW: A large-scale dataset for video scene parsing in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4133–4143, 2021.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision (3DV)*, pp. 565–571, 2016. doi: 10.1109/3DV.2016.79.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, pp. 891–898, 2014. URL https://openaccess.thecvf.com/content_cvpr_2014/html/Mottaghi_The_Role_of_2014_CVPR_paper.html.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, 2008. URL https://www.robots.ox.ac.uk/~vgg/data/flowers/102/.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012. URL https://www.robots.ox.ac.uk/~vedaldi/assets/pubs/parkhi12cat.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Julio Silva-Rodríguez, Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. A closer look at the few-shot adaptation of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations (ICLR)*, 2019.

Elaine Sui, Xiaohan Wang, and Serena Yeung-Levy. Just shift it: Test-time prototype shifting for zero-shot generalization with vision-language models. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

Rahul Venkataramani and Rachana Sathish. Few-shot concept prompt learning for segmentation foundation models via visual grounding. *arXiv preprint arXiv:2608.01663*, 2026.

Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In *NeurIPS Datasets and Benchmarks*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/4e732ced3463d06de0ca9a15b6153677-Abstract-round2.html.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. URL https://vision.princeton.edu/projects/2010/SUN/.

Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *CVPR*, 2023.

Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.

Ce Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Dual prototype evolving for test-time generalization of vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *ECCV*, 2022.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Zhou_Scene_Parsing_Through_CVPR_2017_paper.html.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130:2337–2348, 2022b. doi: 10.1007/s11263-022-01653-1.

Quan Zhou, Shaoqing Zhai, Qiang Hu, Jia Chen, Qiang Li, and Zhiwei Wang. Mask to concept: Auto-promptable SAM3 via efficient test-time concept embedding search for few-shot annotation. *arXiv preprint arXiv:2606.26711*, 2026.

Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

# A A GUIDE TO THE SUPPLEMENTARY MATERIAL

Replacing a class embedding can change a frozen model's predictions without relearning its visual features. The central challenge is to turn that change into a correction that remains useful beyond the examples used to fit it. A lower training loss alone does not answer this question, nor should the same optimization strategy be assumed to suit independently applied SAM3 prompts and competing cosine-classification scores.

Geometry is the starting point: we examine how embeddings are arranged, then change their direction and scale to test what carries the repair. Certified distance bounds test whether mixtures of sampled prompt embeddings can reproduce the corrections. Matched weighting and local-versus-joint controls compare fitting and aggregation, retaining both improvements and counterexamples. Cross-domain and video tests examine reuse; protocols and proofs define the scope of the evidence. Geometry motivates the adaptation target, while controlled fitting comparisons establish local–joint rankings in the tested settings. The central distinction remains: choosing what to adapt and choosing between local and joint optimization are separate decisions.

**Geometry: what the predictor consumes**

**B** Prompt topology, coordinate controls, and reachability p. 16
**C** Original and adapted class-bank geometry p. 18

**Evidence: fitting, aggregation, and reuse**

**D** Aggregation rules and reference retention p. 19
**E** Joint fitting, matched weighting, and component controls p. 20
**F** SAM3 comparison and construction-cost audit p. 25
**G** Classification across CLIP, PE, and SigLIP; CLAP controls p. 26
**H** Cross-domain and image-to-video transfer p. 28

**Reproducibility: data, protocols, and baselines**

**I** Datasets, splits, metrics, and support p. 30
**J** Losses, models, support, and selection protocols p. 33
**K** Prompt-tuning and adapter controls p. 38
**L** Wall time and aggregate GPU-hours p. 39

**Formal support: estimators, refinement, and scope**

**M** Estimator proofs, risk crossover, and joint refinement p. 40
**N** Limitations and open questions p. 49
**O** Terminology and notation p. 50

# B TOPOLOGY OF THE PROMPT SPACE

The relevant geometry is that of the representation used by the predictor. We distinguish SAM3's input from pooled text features, then ask which part of the correction carries its gain and whether sampled prompts can reproduce it.

## B.1 PROMPT TOPOLOGY AND THE CONSUMED REPRESENTATION

Figure 3a compares the same 4,708 ImageNet-S50 prompt entries across CLIP, SigLIP and SAM3, grouping them by class. Its median within-category cosines are $0.83/0.84/0.56$, and median silhouettes are $0.42/0.44/0.18$. The superclass analysis below uses a different label grouping and prompt set.

Figure B.1 shows that the same ImageNet superclass labels have different arrangements across pooled text outputs. Silhouettes are measured in the original normalized spaces; UMAP is used only for visualization. SAM3's decoder consumes a different representation, audited below.

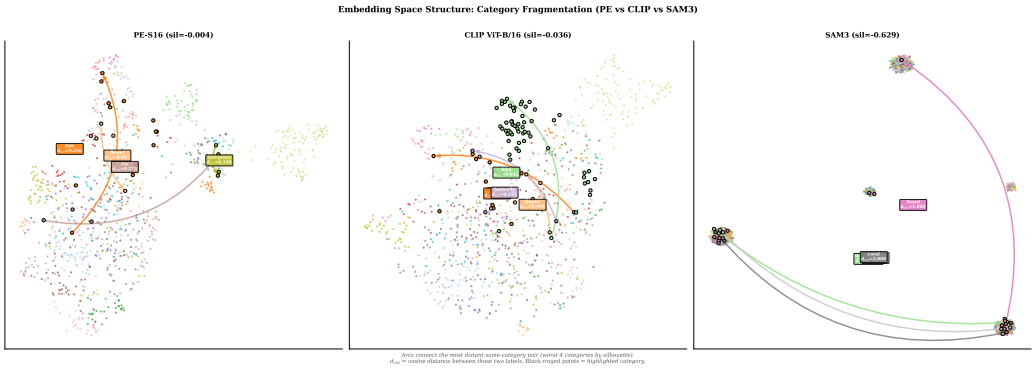



Figure B.1: **Pooled-output category geometry.** Shared underscore-separated superclass labels; native-space cosine silhouettes: PE-Core-S16-384 $-0.004$, CLIP ViT-B/16 $-0.036$, SAM3 $-0.629$. These are pooled encoder outputs, not SAM3's detector input.

**Detector-interface audit.** Let

$$p_c := E_{\text{text}}^{\text{pool}}(c) \in \mathbb{R}^{512}$$

denote the pooled output above, and let

$$e_c := \text{vec}\big(P_{\text{text}} E_{\text{text}}^{\text{tokens}}(c)\big) \in \mathbb{R}^{32\cdot 256}$$

denote the flattened projected token tensor consumed by DETR and optimized by our method. These representations summarize text differently and have different dimensions. The pooled silhouette $-0.629$ in Figure B.1 uses underscore-separated labels, whereas the $-0.963$ row below uses naturalized labels. These are distinct prompt protocols rather than duplicate estimates of one quantity.

In the Hugging Face SAM3 implementation, the DETR encoder updates only the vision stream: vision features query the prompt tokens, and the returned `text_features` field equals the input prompt tensor. The measured object is therefore the image-independent *detector-input projection*, not a post-fusion text representation.

For a finite label set $\mathcal{C}$ and either representation, write $\bar{s} = s/\|s\|$ and define the normalized embedding set

$$S = \{\bar{s}_c : c \in \mathcal{C}\}, \qquad \mu(S) := \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \bar{s}_c.$$

We summarize the angular concentration of $S$ by the *mean resultant length*

$$R(S) := \|\mu(S)\|,$$

and by the empirical $95\%$ cone half-angle

$$\theta_{95}(S) := Q_{0.95}\left(\angle\left(\bar{s}_c, \frac{\mu(S)}{\|\mu(S)\|}\right)\right),$$

where $Q_{0.95}$ is the 95th percentile over $c \in \mathcal{C}$. Small $R$ and large $\theta_{95}$ indicate a wide spread of embedding directions.

Table B.1: **SAM3 text-interface geometry.** Shared naturalized superclass labels; token projections are frozen and image-independent.

| Representation | Cos. sil. | Med. intra-cos | $R$ | $\theta_{95}$ |
|---|---|---|---|---|
| Pooled global output (512d) | $-0.963$ | 0.008 | 0.055 | 180.0° |
| Projected-token mean (256d) | $-0.015$ | 0.459 | 0.460 | 73.7° |
| Projected tokens (8192d) | $-0.013$ | 0.433 | 0.501 | 69.4° |

The detector input is more angularly concentrated, yet its near-zero silhouette still indicates weak superclass separation (Table B.1). Thus concentration is not semantic organization, and pooled similarity is not a substitute for measuring SAM3's input. Whether a change helps prediction requires the interventions below, not a ranking of encoders from this geometry.

### B.2 Correction geometry and coordinate controls

Table B.2: **SAM3 correction geometry.** $\theta$ is angular displacement, $\rho$ the norm ratio, and Rec. the recovered oracle gap.

| Dataset | $\theta$ | $\rho$ | Rec. |
|---|---|---|---|
| Pascal VOC | 51.4° | 1.87 | 27% |
| Cityscapes | 40.1° | 1.53 | 28% |
| CamVid | 32.3° | 1.41 | 49% |
| LoveDA | 24.4° | 1.34 | 86% |
| ADE-847 | 31.5° | 1.35 | 11% |

Table B.3: **Geometry across prediction heads.** Normalized heads discard changes in embedding length; SAM3 uses the full token sequence.

| Model | mean $\theta$ | range | med. $\rho$ |
|---|---|---|---|
| SAM3 | 36.0° | 24.4–51.4° | 1.32 |
| CLIP | 55.6° | 54.2–56.9° | 1.00 |
| PE | 62.1° | 60.6–63.4° | 1.00 |
| SigLIP | 56.3° | 53.7–57.9° | 1.00 |

Corrections rotate by tens of degrees; SAM3 also preserves changes in norm that normalized cosine heads discard. Class centroids recover part of the image-specific gain, suggesting shared structure rather than only individual repairs. Here "oracle" means finite-budget image-specific fitting, not a certified optimum. These historical cosine-head fits describe geometry; COOPER-CLS instead fits its reported bank jointly.

**Direction carries the repair.** Could norm growth alone explain the gain? We retain only direction or scale, then compare matched-norm random moves, shuffled class assignments, and nearest prompts. All arms share one validation vision sweep (Table B.4).

Table B.4: **Coordinate interventions on deployed SAM3 banks.** Parentheses in the Direction column give the fraction of full-bank gain retained; all other entries are mIoU changes in percentage points.

| Dataset | Full | Direction | Scale | Shuffled | Random | Nearest prompt |
|---|---|---|---|---|---|---|
| LoveDA | +18.41 | +17.28 (94%) | +2.79 | +4.58 | -1.17 | +0.00 |
| Pascal VOC | +6.55 | +6.08 (93%) | +0.10 | +0.01 | -1.66 | +0.00 |
| ImageNet-S50 | +8.64 | +8.52 (99%) | +1.10 | +4.17 | -0.52 | +0.00 |

Direction retains 92.8–98.6% of the full gain. The small deficit relative to the full correction is resolved on LoveDA ($-1.13$ pp, 95% CI $[-2.09, -0.20]$) and VOC ($-0.47$, $[-0.83, -0.16]$), but not ImageNet-S50 ($-0.12$, $[-0.58, +0.29]$). Scale alone is positive on LoveDA and ImageNet-S50 but is far weaker, and a matched isotropic displacement never helps. Correct class pairing matters: full correction beats the shuffled bank on every dataset. Yet shuffling retains a resolved 48% of the ImageNet-S50 gain and essentially none on VOC: useful conditioning can contain both domain-shared and class-specific information. The polar interaction varies ($-1.67/+0.37/-0.99$ pp on LoveDA/VOC/ImageNet-S50), so direction and scale do not contribute additively to task performance.

### B.3 FINITE-POPULATION PROMPT REACHABILITY

This test supports Section 2: the learned corrections are not simply mixtures of the sampled text prompts. For each class we encode the canonical 80 CLIP templates, sense-controlled WordNet synonyms and hypernyms, and deterministic definition-based descriptions in fp32. Distances use the flattened detector-facing $32 \times 256$ tensor and are normalized by each prompt cloud's diameter. We retain the stored Original embedding as the class-name vertex, so the tested hull includes the actual initialization, not only its re-encoding. Nearest-prompt selection uses only distance to the bank built from training data.

**How the distance is certified.** Let $P$ contain the sampled embeddings as rows and let $b$ be the learned class embedding. We minimize

$$f(\lambda) = \tfrac{1}{2}\|P^\top \lambda - b\|_2^2, \qquad \lambda \geq 0, \quad \mathbf{1}^\top \lambda = 1.$$

At a feasible mixture $q = P^\top \lambda$, define the Frank–Wolfe gap $G = (q-b)^\top q - \min_j (q-b)^\top p_j$. Convexity gives $f(\lambda) - G \leq f^\star \leq f(\lambda)$, hence

$$\sqrt{\max(2f(\lambda) - 2G, 0)} \;\leq\; \operatorname{dist}(b, \operatorname{conv} \mathcal{P}_c) \;\leq\; \|q - b\|_2.$$

Thus an unfinished minimization still supplies a lower bound, not just the distance to the best mixture found. We use float64 arithmetic, initialize at the nearest vertex, and take exact line-search steps. We stop at $G \leq 10^{-7} \max(f(\lambda_0), 10^{-12})$ or 5,000 iterations. The implementation also computes distance to the affine hull by an eigendecomposition and takes the larger lower bound; the dual bound is at least as large for all 76 reported classes.

**Numerical verification.** We check feasible weights and lower–upper ordering. The outside-hull threshold is $\max(10^{-8}, 10^{-5} D_c)$, where $D_c$ is the cloud diameter. An independent check renormalizes the saved mixture weights and evaluates the separating-direction bound $\min_j u^\top (b - p_j)$, with $u = (b-q)/\|b-q\|_2$, in extended precision. All 76 bounds remain positive and support the reported lower bounds within $10^{-10} D_c$ numerical tolerance. The smallest normalized separation is $0.161$, well above the $10^{-5}$ threshold. This verifies separation of the stored prompt population without requiring the optimizer to reach its minimum.

Table B.5: **Distance from the learned bank to a broad finite prompt population.** Hull lower is a certified lower bound on distance to the sampled convex hull; Outside counts classes with a positive certified bound.

| Dataset | Prompts/class | Nearest/diam. | Affine/diam. | Hull lower/diam. | Outside |
|---|---|---|---|---|---|
| LoveDA | 98 | 0.4409 | 0.4208 | 0.4406 | 100% |
| Pascal VOC | 97 | 0.4486 | 0.4195 | 0.4483 | 100% |
| ImageNet-S50 | 93 | 0.2773 | 0.2597 | 0.2767 | 100% |

All 76/76 learned class vectors have a positive sampled-hull separation certificate, and convex mixing barely improves over the nearest prompt. The nearest member is the original class label for every class, so its validation result is exactly the baseline on all three datasets. ImageNet-S50 has the smallest normalized separation, consistent with its stronger shared component, while LoveDA and VOC are farther away. This is evidence about the declared finite population, not proof that no possible string or learned continuous prompt can realize the bank.

These tests motivate adapting the consumed representation directly: direction accounts for most of the gain, and sampled prompt mixtures cannot reproduce the corrections. Geometry identifies properties of useful corrections; the aggregation controls test how to combine per-example repairs. Local–joint rankings come from fitting comparisons, not from a projection or a demonstrated geometric cause.

## C ORIGINAL AND ADAPTED CLASS-BANK GEOMETRY

This section extends Figure 6's direction-versus-scale diagnostic to three datasets; no held-out predictions or labels determine the projection coordinates.

With original and deployed flattened token sequences $v_c^{(0)}, v_c^{(*)} \in \mathbb{R}^{8192}$, we report

$$\theta_c = \arccos\left(\frac{\langle v_c^{(0)}, v_c^{(*)}\rangle}{\|v_c^{(0)}\|_2 \|v_c^{(*)}\|_2}\right), \qquad \rho_c = \frac{\|v_c^{(*)}\|_2}{\|v_c^{(0)}\|_2}.$$

For visualization, each dataset uses one PCA fitted to both banks after unit normalization of each class embedding. Original and deployed coordinates share its basis and axis limits; native-space norm ratios are computed before normalization. Arrows join endpoints, not optimization trajectories.

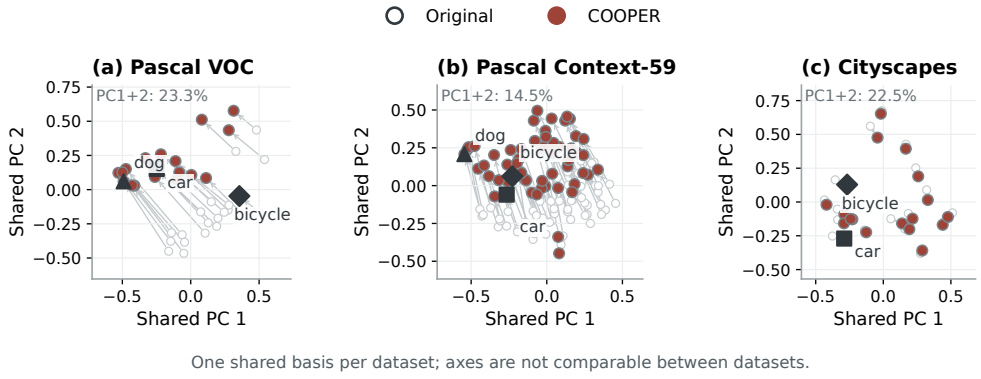



Figure C.1: **Class-bank geometry across three datasets.** Open markers: original embeddings; filled markers: deployed embeddings. Each dataset uses a shared PCA, not a common cross-dataset coordinate system.

For Pascal VOC, Pascal Context-59, and Cityscapes, median native-space rotations are $42.1°$, $30.9°$, and $31.9°$; median norm ratios are 1.335, 1.184, and 1.185. The rotations rule out a pure rescaling. PCA makes the displacement visible but retains only 23.3% of VOC's directional variance. Prediction quality is assessed separately: Table B.4 tests which changes retain the gain. The projection describes the correction; interventions test its functional contribution rather than inferring it from visual separation.

**Illustration provenance.** The masks and embedding positions in Figure 2 are schematic, not measured predictions. Photograph: Skansken, CC BY 3.0 (cropped; silhouette traced). Figure 5 uses the first three Pascal VOC training images in split order with at least 3% dog foreground. These photographs illustrate construction inputs, not prediction results. Photographs beside class-bank projections identify class markers, not projected image features.

## D Aggregation Operators

These are the aggregation rules used in Section 4: repairing one example does not ensure a useful class embedding. Figure D.1 measures improvement on each fit's training example; the controls in Appendix E.3 test what survives aggregation. SAM3 uses a gain-weighted candidate and training-only adoption, with additional global and hierarchical candidates for ImageNet-S50 (Appendix J.7). COOPER-CLS instead fits its bank jointly.

**Notation.** For a class $c$, let $I_c$ index its training image–class pairs, $\hat{e}_i^\star$ be their selected endpoints, and $e_c$ the original class embedding. As in Eq. (5),

$$g_i = S_i(\hat{e}_i^\star) - S_i(e_c), \qquad S_i \in [0, 1].$$

Thus $100g_i$ is the gain in percentage points. Aggregation acts on the complete predictor-facing token sequence, not a token-pooled vector.

**Base aggregation rule.** The gain-weighted estimate is

$$A_c = \sum_{i \in I_c} \mathbf{1}[g_i > \tau_{\text{act}}], \qquad \tau_{\text{act}} = 0.005,$$

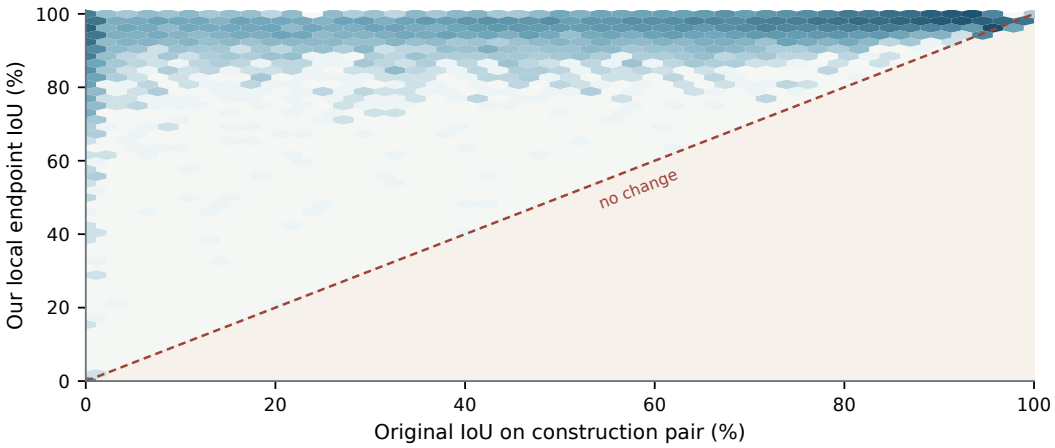



Figure D.1: **Local repair before aggregation.** Finite-budget endpoint density over 8,640 training image–class pairs across six domains. Above-diagonal mass indicates improvement on the endpoint's own construction pair, not held-out performance; deployment is evaluated only after merging and training-only selection.

$$\hat{\mu}_c = A_{\mathrm{gw}}(c) = \frac{\sum_{i \in I_c} w_i \hat{e}_i^\star}{\sum_{i \in I_c} w_i}, \qquad w_i = \begin{cases} \max(g_i, 0), & A_c \geq 3, \\ 1, & A_c < 3. \end{cases}$$

The threshold is $0.5$ pp of IoU gain. It counts improved examples, not active decoder queries, and does not discard smaller positive gains: once $A_c \geq 3$, every positive-gain endpoint receives weight. If fewer than three examples pass the gain threshold, we use the unweighted class mean for the candidate centroid. The deployed class embedding is then selected on the training split only:

$$\tilde{\mu}_c(\alpha) = (1 - \alpha) e_c + \alpha \hat{\mu}_c, \qquad \alpha \in \{0.3, 0.5, 0.7, 0.85, 1.0\}.$$

The basic blend rule accepts $\tilde{\mu}_c(\alpha)$ only if its training mean gain is positive, its win rate is at least $0.5$, and at least three training examples are available; otherwise it retains $e_c$. The selected embedding is $b_c$. The nested and ImageNet-S50 selectors extend this basic rule, including incumbent and hierarchical candidates where recorded in Table J.5; they are not one universal guard recipe.

The uniform control replaces the weights by one, $A_{\mathrm{mean}}(c) = |I_c|^{-1} \sum_{i \in I_c} \hat{e}_i^\star$. Appendix E.3 reruns the same selection policy for both merge rules; Appendix E.4 separately isolates selection, weighting, blending, and rejection on shared trajectories.

**A training gain need not survive deployment.** A post-lock LoveDA audit tests uniform averaging after the reported estimator was frozen. Uniform averaging improves a disjoint training confirmation set by $+0.560$ pp (95% CI $[+0.365, +0.757]$), but its held-out change is only $+0.024$ pp ($[-1.408, +1.090]$); frozen VOC transfer changes by $-0.053$ pp ($[-0.381, +0.233]$). The training preference therefore does not establish a better deployed estimator. The benchmark policy remains frozen; this distinct construction is not a repeat of the matched replay below.

## E Joint Fitting and Component Controls

These controls support Figure 7: a local bank can be a useful endpoint or an initialization for further fitting. We then isolate weighting, endpoint selection and guarding under their declared protocols.

**Local refinement helps without gain weighting.** Uniform local averaging improves on Original on all eleven datasets, with an equal-dataset mean gain of $+12.96$ pp (Table E.1). It exceeds Joint-R on nine datasets and Joint-O on seven, although its mean mIoU is lower than Joint-O's (58.62 versus 59.60). Thus gain weighting is not required for substantial local-refinement gains, but uniform averaging is not uniformly preferable to joint fitting. This comparison retains per-example checkpoint selection; it does not isolate locality from selection or regularization. Unlike Table 3's within-pipeline weighting intervention, it compares a uniform-local construction with saved banks and fixed joint recipes.

Table E.1: **Fixed-checkpoint SAM3 replay.** Held-out class-mean mIoU (%). Joint-R/O start from Original/saved COOPER; $T$ is the final step, with Adam learning rate $0.05$ and no evaluation-based arm selection. All arms share the displayed coverage. Uniform local retains per-example checkpoint selection; saved COOPER uses dataset-specific policies. LoveDA uses the reconstructed gain mean (45.06), not the guarded bank (45.55). Protocols: Appendix E.

| Dataset | $T$ | Classes / prompts | Original | Joint-R | Joint-O | Uniform local | Saved COOPER |
|---|---|---|---|---|---|---|---|
| ADE-847 | 200 | 847 / 22313 | 35.19 | 37.52 | 37.59 | 38.05 | 37.93 |
| PC-59 | 200 | 59 / 25975 | 60.76 | 64.59 | 63.87 | 70.72 | 70.61 |
| VOC | 200 | 20 / 2148 | 75.97 | 68.98 | 74.43 | 82.01 | 82.53 |
| IN-S50 | 50 | 50 / 759 | 78.39 | 80.33 | 79.39 | 86.09 | 87.03 |
| COCO | 200 | 133 / 34730 | 50.80 | 56.12 | 54.57 | 63.32 | 61.71 |
| Cityscapes | 200 | 19 / 6005 | 55.45 | 57.87 | 61.57 | 60.67 | 61.11 |
| CamVid | 200 | 29 / 1562 | 35.13 | 54.01 | 57.05 | 46.97 | 47.71 |
| LoveDA | 200 | 6 / 6959 | 27.10 | 37.21 | 39.66 | 45.58 | 45.06 |
| Kvasir-SEG | 50 | 1 / 200 | 0.00 | 0.00 | 80.30 | 40.13 | 78.46 |
| PascalPart | 50 | 212 / 22075 | 30.16 | 49.12 | 50.74 | 48.12 | 48.55 |
| LVIS | 50 | 804 / 15491 | 53.31 | 57.36 | 56.46 | 63.20 | 60.37 |

**Initialization does not settle the choice of estimator.** The replay still favors joint fitting on CamVid and PascalPart (Table E.1), but a COOPER initialization does not make joint fitting preferable everywhere. This is why local construction and subsequent joint fitting are separate choices. Main-table Joint-R/O gains subtract replay Original, not the local initialization. LoveDA uses the reconstructed gain mean (45.06%), not its guarded headline bank (45.55%). Uniform local averaging retains per-example checkpoint selection; saved COOPER can include incumbent or hierarchical candidates. These are fixed recipes, not a fully tuned, regularization- and selection-matched comparison; Table E.3 reports a separate three-domain control.

Joint fitting uses BCE for ImageNet-S50 and BCE, Dice, and query-activity loss elsewhere. The sampler uses the full usable support pool when it contains at most 32 images; requested minibatches apply to larger pools. Warm fitting inherits its initial bank's selection and construction cost. Equal update counts therefore do not establish equal construction budgets.

**Better optimization and better transfer are different tests.** VOC exposes two failures that should not be conflated. At step 50, the Original-initialized fit has lower held-out loss than the local bank ($0.1952$ versus $0.2054$), yet lower mIoU ($76.89$ versus $82.53$): loss and IoU disagree. After 200 warm steps, training loss falls from $0.1921$ to $0.0651$ while held-out loss rises to $0.2607$, consistent with support overfitting. Figure 7 scores the shared local bank, not per-example minima. CamVid improves on both held-out measures; the three-seed study in Appendix M.7 asks whether that opportunity can be identified before deployment.

These trajectories each come from one run with fixed checkpoints, not held-out-selected models. Training IoU uses 288-pixel objective masks; held-out IoU uses native-resolution postprocessing, so their numerical gap is not a generalization estimate. VOC/LoveDA use 20/6 classes; CamVid uses 23 supported training classes and 29 evaluation classes, including fallbacks.

**Recoverable CoOp collapse.** The four arms below share seed 42, four fit and two disjoint probe pairs per class, 16 context tokens, and 1,000 AdamW updates (learning rate $10^{-3}$, weight decay $10^{-4}$, cosine schedule). The cap limits each update's L2 norm to $0.01$, not total drift. Top-query fallback changes the no-active-query branch during fitting and probing. All final fallback-arm probes have active queries and use the common active-query rule; earlier steps need not.

Restoring a usable gradient or limiting the update rescues LoveDA (Table E.2). The production endpoint has zero gradient support; all interventions retain it on every probed class. The failure is therefore recipe-dependent, not proof that the representation cannot support the task.

**Other controls and remaining scope.** An unregularized, training-selected 50-step screen completed eight of nine dataset–seed points (ImageNet-S50, CamVid, Kvasir; seeds 7/17/42; ImageNet seed 7 incomplete). This historical screen is excluded from the completed matched study in Table E.3. A separate matched-minibatch intervention starts every arm from the same local mean. Uniform,

Table E.2: **Fixed-step anti-collapse controls.** Probe IoU (%); LoveDA uses 24 fit/12 probe pairs, ImageNet-S50 200/100. These are training/probe controls, not a tuned baseline evaluated on the full held-out split.

| CoOp variant | LoveDA probe IoU | ImageNet probe IoU | LoveDA active (%) |
|---|---|---|---|
| Production objective | 0.00 | 91.36 | 0.00 |
| Update cap | 42.61 | 90.74 | 91.67 |
| Top-query fallback | 40.66 | 90.77 | 100.00 |
| Cap + fallback | 41.84 | 92.44 | 100.00 |

inverse-winsorized-radial-proxy, and within-class-permuted weights obtain $38.54/37.65/37.82$ mIoU on LoveDA and $79.64/75.96/75.77$ on ImageNet-S50. The scalar proxy (loss improvement divided by squared displacement) does not improve joint fitting; it is not a Hessian measurement or evidence identifying curvature as the cause.

### E.1 Completed matched local–joint controls

Table E.3 compares local and joint optimization under training-only tuning on VOC, LoveDA and CamVid. This separate construction does not replace Table 2's eleven-domain benchmark.

Table E.3: **Matched local and joint fitting.** Held-out mIoU (%), mean $\pm$ sample SD over seeds 7/17/42. $\Delta$ = joint minus local (pp); negative values favor local fitting.

| Dataset | Joint start | Local | Joint | $\Delta$ (pp) |
|---|---|---|---|---|
| *Selected checkpoint; before the common guard* | | | | |
| VOC | Cold | $81.89 \pm 0.40$ | $80.54 \pm 0.69$ | $-1.35 \pm 0.58$ |
| VOC | Warm | $81.72 \pm 0.74$ | $80.31 \pm 1.93$ | $-1.41 \pm 1.24$ |
| LoveDA | Cold | $41.93 \pm 1.57$ | $40.37 \pm 0.62$ | $-1.56 \pm 1.18$ |
| LoveDA | Warm | $41.62 \pm 1.49$ | $40.62 \pm 0.50$ | $-1.00 \pm 1.50$ |
| CamVid | Cold | $45.80 \pm 0.99$ | $50.88 \pm 3.75$ | $+5.07 \pm 3.52$ |
| CamVid | Warm | $46.72 \pm 0.73$ | $54.17 \pm 1.99$ | $+7.45 \pm 1.34$ |
| *Deployment with training-selected blending and the common guard* | | | | |
| VOC | Cold | $81.55 \pm 0.41$ | $80.58 \pm 0.65$ | $-0.97 \pm 0.81$ |
| VOC | Warm | $81.47 \pm 0.70$ | $80.92 \pm 1.12$ | $-0.55 \pm 0.97$ |
| LoveDA | Cold | $39.26 \pm 1.66$ | $38.28 \pm 2.19$ | $-0.98 \pm 2.74$ |
| LoveDA | Warm | $40.41 \pm 0.45$ | $39.28 \pm 2.63$ | $-1.13 \pm 3.06$ |
| CamVid | Cold | $45.04 \pm 1.18$ | $49.90 \pm 2.36$ | $+4.86 \pm 1.89$ |
| CamVid | Warm | $45.03 \pm 0.96$ | $51.72 \pm 1.49$ | $+6.69 \pm 2.22$ |

**Matched selection, distinct local references.** Training images are globally image-disjoint across fit, selection and guard roles, capped at $8/4/4$ per class (minima $3/2/2$); unsupported classes retain Original. Uniform-local and cold-joint fits, and warm-joint continuation, share Adam, loss (4), learning rates $\{0.005, 0.05\}$, drift penalties $\{0, 0.1\}$ and checkpoints $\{0, 25, 50, 100, 200\}$. The penalty is squared embedding displacement from Original divided by its squared norm; selection maximizes native training IoU on the selection split. The cold-start comparison uses uniformly averaged local snapshots at a common checkpoint and Original-initialized joint fitting. The warm start uses local fits selected by IoU on their own training images, gain-weighted and blended using selection data at learning rate $0.05$ with no penalty; it compares that initialization with joint continuation. Thus the two local references are not interchangeable.

Blends use $\alpha \in \{0.3, 0.5, 0.7, 0.85, 1\}$, selected before guarding. Both arms retain a candidate only when its mean guard-IoU gain over Original is positive and at least half the pairs improve strictly; otherwise they retain Original. The warm initialization is already blended in the selected panel. Held-out masks use unchanged native inference, without query fallback; evaluation never selects a candidate. Per-configuration image-gradient counts match, but warm fitting additionally pays for local construction: total pipeline compute is not equal.

**The ranking depends on the domain.** Local fitting has modest mean advantages on VOC/LoveDA; CamVid favors both joint initializations in every seed, including after guarding. CamVid's larger

benefit makes the equal-domain mean favor joint fitting in this three-domain study, unlike Table 2's eleven-domain fixed-recipe comparison. Support recipes also differ, so the smaller VOC gap cannot be attributed to tuning alone. These previously inspected domains support domain-dependent local–joint rankings, not independent confirmation or a geometry-based selection rule.

### E.2 Component-level rationale and evidence

A useful bank does not make every construction choice necessary. Table E.4 separates the evidence for each choice; the matched studies below test simpler alternatives without changing the other components.

Table E.4: **Evidence for each component.** Distinct recipes are not interchangeable ablations.

| Choice | Evidence and scope |
|---|---|
| **Per-example fitting** | Fixed joint replays retain both local- and joint-favorable regimes. Completed three-seed controls match tuning options, selection and guards on three domains (Table E.3). Warm fitting adds initialization cost, so total pipeline compute is not equal. |
| **IoU selection** | On identical trajectories it improves CamVid in all three seeds; VOC and LoveDA effects are small and seed-dependent (Table E.6). |
| **Gain weighting** | The benchmark replay gives $+37.53$ pp on Kvasir and mixed $-1.53$ to $+1.93$ pp elsewhere. The separate forced-deployment factorial favors uniform means in all three domain averages. |
| **Guard and blend** | Rejection reduces harmed-pair frequency but also mean IoU in all nine dataset–seed cells: reference retention is a trade-off, not a safety guarantee (Table E.7). |
| **Angular prior** | Three-backbone controls compare a prior grid including zero with zero prior alone. This tests access to regularized candidates, not a fixed penalty or the nine-backbone refit protocol (Table J.4). |

### E.3 Matched benchmark weighting contrast

Weighting can prevent ineffective fits from diluting a correction, but it can also discard useful support. The eleven-domain replay tests both possibilities (Table 3); Table E.5 gives the absolute scores.

Both arms reuse the same endpoint pools, training calibration records, held-out image–class pairs, and inference settings. All affected means, including incumbents, are rebuilt; the same training-only selection procedure is rerun separately per arm. Thus weights change throughout construction and selection, not only in the final mean with selection held fixed. The nested guard uses $z = 1$ in both arms. ImageNet-S50 instead reruns its hierarchical-family selector over the same candidates and folds: it selects $z = 1$ for weighted and $z = 1.5$ for uniform. Neither choice uses held-out labels.

Table E.5: **Matched weighting across domains.** Replay mIoU (%); $\Delta$ = weighted minus uniform.

| Dataset | Weighted | Uniform | $\Delta$ (pp) | 95% CI for $\Delta$ |
|---|---|---|---|---|
| ADE-847 | 37.89 | 37.23 | +0.67 | $[-0.02, +1.38]$ |
| PC-59 | 70.57 | 70.53 | +0.04 | $[-0.62, +0.80]$ |
| VOC | 82.48 | 82.36 | +0.12 | $[-0.30, +0.53]$ |
| IN-S50 | 86.95 | 85.02 | +1.93 | $[-0.19, +4.79]$ |
| COCO | 61.72 | 62.77 | -1.05 | $[-1.84, -0.38]$ |
| Cityscapes | 61.12 | 61.33 | -0.21 | $[-0.91, +0.35]$ |
| CamVid | 47.79 | 46.90 | +0.89 | $[+0.07, +1.89]$ |
| LoveDA | 45.55 | 45.92 | -0.38 | $[-0.68, -0.05]$ |
| Kvasir-SEG$^{\dagger}$ | 79.72 | 42.19 | +37.53 | $[+31.85, +43.28]$ |
| PascalPart | 48.56 | 47.98 | +0.58 | $[+0.10, +1.22]$ |
| LVIS | 60.38 | 61.91 | -1.53 | $[-2.38, -0.72]$ |
| Mean (11 complete) | 62.07 | 58.56 | +3.51 | — |

Intervals use 20,000 paired class-bootstrap draws (seed 20260916), resampling the same class indices in both arms after averaging image IoUs within each class. †Kvasir has one foreground class, so its interval resamples the 200 paired held-out images instead. These pointwise intervals condition on the fitted banks; they do not quantify construction-seed or acquisition-cluster uncertainty, correct for

multiple comparisons, or establish equivalence. The mean weights all completed datasets equally; no aggregate interval is inferred.

**Weighting helps when the discarded mass matters.** Weighting improves Kvasir by $37.53$ pp and sharply reduces low-gain mass. Elsewhere the effects are mixed and much smaller ($-1.53$ to $+1.93$ pp). CamVid and PascalPart favor weighting; COCO, LoveDA and LVIS favor uniform, with pointwise intervals excluding zero. The remaining intervals cross zero. Proposition 2 explains why mass alone is insufficient: bias also depends on the separation between useful and ineffective fits. Low-gain endpoints near a working Original can still be useful. The Kvasir audit (Appendix M.2.1) supports the dilution interpretation, but neither identifies the latent mixture nor establishes equivalence on the other domains. The separate uniform-local replay improves all eleven domains (Appendix E): refinement supplies broad gains, while weighting addresses a regime-specific dilution problem.

**What the mass column measures.** Low gain means $g_i \leq 0.005$. We compare its uniform and gain-weighted shares on the same endpoint pool, averaging equally over supported classes and then datasets. Each domain uses its largest candidate pool; ImageNet-S50 uses full-refit class-mean primitives. Nested pools are not combined.

Mass is measured before selection, blending, and hierarchy, not on the final bank. Arms can select different supports or fallbacks; Kvasir is the exception, with both choosing the same $n = 32$ mean. ImageNet-S50 retains 6.6% low-gain mass because some gates remain inactive. Thus the column describes what weighting changes, not the latent fraction of failed repairs.

The separate post-lock LoveDA audit in Appendix D uses a different construction: uniform averaging improves held-out mIoU by $0.024$ pp, with an interval crossing zero. It is not an independent repetition of this replay's $0.38$-pp uniform advantage ($W - U = -0.38$).

These controls do not replace the historical scores in Table 2. Both ADE arms exclude eleven previously identified duplicate training images and retain all evaluation images; its training population therefore differs from the archived benchmark. Inference drift and arm-specific training-only decisions are recorded in the audit. These previously inspected domains provide development evidence, not independent confirmation.

### E.4 Controlled component factorial

This study supports Section 4's distinction between a useful bank and necessary policies. CamVid/VOC/LoveDA use seeds 7/17/42 and one trajectory per FIT image: three endpoint rules (final, minimum loss, maximum IoU), two merges (uniform, gain), and three deployment rules (forced, blend, guarded), giving 18 banks plus Original per cell. All nine cells are complete; this separate fitting/selection recipe on development domains is not the matched benchmark ablation or independent confirmation.

**Protocol.** Whole decoded images enter disjoint FIT/SELECT/GUARD roles before class sampling (caps 8/4/4; minima 3/2/2). Fits use 200 Adam steps, lr 0.05, float32 embeddings, bf16 forwards, BCE + Dice + $0.05$ query-activity loss, and no regularizer. A top-query fallback preserves the mask-loss gradient when no query is active during fitting. Endpoint rules use the same checkpoints $\{0, 25, 50, 100, 200\}$ and metrics on each fit's own training image; ties favor earlier checkpoints. SELECT chooses $\alpha \in \{0.3, 0.5, 0.7, 0.85, 1\}$; GUARD requires positive mean native-IoU gain and at least half strict wins. Unsupported classes and rejected candidates retain Original. Evaluation uses the full 29/20/6-class ontologies, including CamVid's 25/26/27 supported classes across seeds, not its declared 31-class vocabulary.

Endpoint contrasts fix uniform/forced deployment; weighting fixes IoU-selected/forced deployment; blend and guard fix the IoU-selected, gain-weighted candidate. These contrasts summarize the full factorial without selecting a winner on evaluation data; uncertainty is three-seed sample SD, not a confidence interval.

Table E.6: **Isolated component effects.** Paired mIoU differences over all evaluated classes (pp), mean ± sample SD over three seeds.

| Change (others fixed) | CamVid | VOC | LoveDA |
|---|---|---|---|
| IoU vs final | $+0.75 \pm 0.46$ | $+0.05 \pm 0.08$ | $+0.07 \pm 0.50$ |
| IoU vs minimum loss | $+1.04 \pm 0.88$ | $+0.05 \pm 0.08$ | $+0.09 \pm 0.52$ |
| Gain vs uniform | $-0.11 \pm 0.83$ | $-0.44 \pm 0.69$ | $-0.54 \pm 1.12$ |
| Blend vs forced | $-0.07 \pm 0.54$ | $+0.27 \pm 0.46$ | $-0.93 \pm 1.27$ |
| Guard vs blend | $-0.69 \pm 0.07$ | $-0.34 \pm 0.22$ | $-1.89 \pm 1.32$ |

**Selection gains vary; weighting is not always needed.** IoU selection helps CamVid in every seed, but not consistently on VOC or LoveDA; gain weighting improves none of the three domain means. This contrasts with Kvasir's matched replay and supports a regime-specific weighting policy.

Table E.7: **Guarding trades repairs for fewer harms.** IoU and harmed-pair fractions (%) are seed means; final columns count harmful accepted and useful rejected class–seed cases.

| Domain | Blend IoU | Guard IoU | Blend harm | Guard harm | Harm/accept | Useful/reject |
|---|---|---|---|---|---|---|
| CamVid | 46.51 | 45.82 | 18.16 | 11.63 | 6/53 | 12/25 |
| VOC | 81.79 | 81.45 | 25.22 | 21.34 | 1/52 | 7/8 |
| LoveDA | 42.11 | 40.22 | 20.89 | 14.92 | 0/14 | 4/4 |

**Fewer harmful changes can mean fewer useful repairs.** Guarding lowers mean IoU in all nine cells while reducing harmed-pair frequency; the table retains both useful rejections and harmful acceptances. With only 2–4 potentially correlated GUARD images, reference retention is a trade-off, not a calibrated safety guarantee.

# F SAM3 Comparison and Construction-Cost Audit

This audit defines the references and uncertainty behind Table 2; construction-cost accounting is consolidated in Appendix L. Benchmark gains use arm-matched Original references. Bold marks the largest rounded benchmark-protocol gain, not significance or a cross-protocol ranking.

The main table's separate replay columns report signed gains in percentage points at fixed final checkpoints: both subtract the replay Original score before rounding. Joint-R starts from replay Original; Joint-O starts from saved COOPER. Joint-O does not subtract its COOPER initialization. The matched replay Original can differ slightly from the displayed benchmark Original. The replay references and saved policies are distinct from their historical benchmark counterparts: LoveDA uses the reconstructed saved bank (45.06%) rather than the guarded headline bank (45.55%). Full matched references, local means, saved COOPER, and evaluation coverage appear in Table E.1.

The joint columns' bracketed intervals are recomputed from the same saved predictions, paired against replay Original, not COOPER. We resample class-mean gains (20,000 draws; seed 20260916), including zero-gain fallbacks; Kvasir-SEG instead resamples 200 paired images. These pointwise percentile 95% intervals condition on fixed banks, without training-seed uncertainty, shared-image cluster adjustment, or simultaneous coverage.

The main table's means weight datasets equally across all eleven domains within each protocol. All adaptation columns average point gains; only Original averages absolute mIoU. We do not average interval endpoints or report an aggregate confidence interval.

# G FULL CLASSIFICATION RESULTS

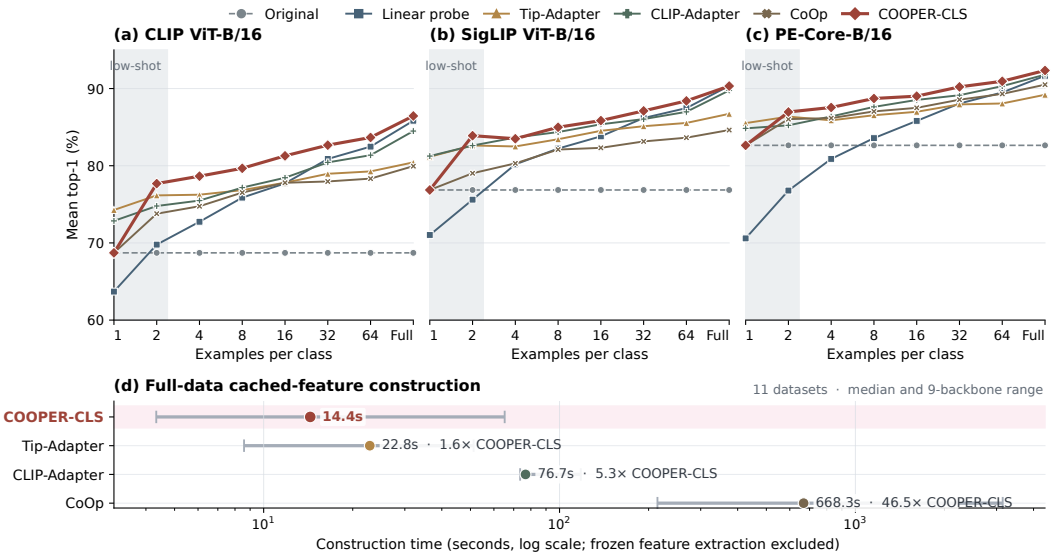



Figure G.1: **Classification accuracy and recorded construction cost. (a–c)** Mean top-1 (%) over 11 datasets on identical nested supports. **(d)** Median full-data construction time (seconds) and range over nine backbones, excluding the shared frozen-feature pass.

Classification tests class-embedding adaptation with a different predictor: cosine scores compete across classes. COOPER-CLS therefore fits an anchored bank jointly. Its low-support advantage over a linear probe narrows as labels increase (Figure G.1), consistent with an anchor that stabilizes noisy estimates rather than a large full-data gain.

All methods share eleven datasets and nested support IDs; CoOp uses each backbone's native frozen text tower. The nine-backbone table reports the guarded COOPER-CLS bank. Training-only selection and refitting follow Appendix J.3; one shot retains the zero-shot anchor, and adaptation starts at two shots.

Accuracy curves alone do not justify joint fitting. The curve-matched 16-shot CLIP comparison holds support, selection, guard, and refit fixed (Table J.2); the separate three-seed 8/4/4 study compares local and joint fitting without refitting (Table J.3). Both favor joint fitting, consistent with coordinated class separation, without isolating geometry as its cause. Original-code CLAP supplies the stronger low-shot baseline below. Construction timings are offline costs; fixed-bank inference is audited separately in Appendix G.2.

**Breadth, rather than a large full-data margin.** COOPER-CLS leads on eight of nine backbone averages and is within 0.03 pp of the linear probe on SigLIP ViT-B/16. Dataset-paired gains against the strongest compatible control are resolved on CLIP B/16 ($+0.62$ pp, 95% CI $[+0.25, +1.05]$) and PE-Core B/16 ($+0.54$, $[+0.11, +1.09]$), not SigLIP B/16 ($-0.02$, $[-0.34, +0.25]$). The consistent low-support advantage over a linear probe therefore supports anchoring across backbones, not superiority to all anchored classifiers. At one shot it is the unchanged zero-shot bank.

## G.1 ORIGINAL-CODE CLAP: MATCHED DATA, NATIVE RECIPES

We run the original CLAP implementation at pinned commit `dbdf4095b3d8` with OpenAI CLIP ViT-B/16. Nine shared native-loader datasets (Pets, Food-101, DTD, Flowers, Cars, Aircraft, EuroSAT, Caltech101, SUN397), eight support settings and seeds 1/2/3 yield 216 matched CLAP–COOPER-CLS pairs, all complete. Both methods use the same pretrained weights, TRAIN images and TEST split. CLAP retains its native prompts, 20 augmentations, SGD schedule and fixed epoch-300 checkpoint; the original evaluator supplies the score, not the highest monitored test accuracy. COOPER-CLS retains its template ensemble, single-view features, anchored fitting and training-only selection/refit. This is a matched-data comparison of native recipes, not a component ablation.

Table G.1: **Full-support classification comparison.** Mean top-1 (%) over the same 11 datasets. Each panel contains one backbone family.

*(a) CLIP*

| Method | ViT-B/16 | ViT-B/32 | RN50x4 |
|---|---|---|---|
| Zero-shot | 68.71 | 64.85 | 62.85 |
| Linear probe | 85.83 | 83.49 | 84.37 |
| Tip-Adapter | 80.43 | 77.18 | 76.72 |
| CLIP-Adapter | 84.49 | 81.44 | 82.80 |
| CoOp | 79.94 | 76.64 | 76.53 |
| **COOPER-CLS** | **86.45** | **83.80** | **84.40** |

*(b) SigLIP*

| Method | ViT-B/16 | ViT-B/16-384 | SO400M/14-384 |
|---|---|---|---|
| Zero-shot | 76.86 | 77.89 | 84.55 |
| Linear probe | **90.35** | 91.05 | 93.54 |
| Tip-Adapter | 86.72 | 87.64 | 90.81 |
| CLIP-Adapter | 89.75 | 91.02 | 93.59 |
| CoOp | 84.62 | 85.21 | 88.09 |
| **COOPER-CLS** | 90.32 | **91.25** | **93.72** |

*(c) PE-Core*

| Method | S/16-384 | B/16 | L/14-336 |
|---|---|---|---|
| Zero-shot | 74.76 | 82.64 | 85.91 |
| Linear probe | 89.15 | 91.65 | 94.14 |
| Tip-Adapter | 84.84 | 89.19 | 91.78 |
| CLIP-Adapter | 87.61 | 91.80 | 94.22 |
| CoOp | 86.09 | 90.50 | 92.99 |
| **COOPER-CLS** | **89.40** | **92.34** | **94.33** |

Table G.2: **Original-code CLAP comparison.** Mean top-1 (%) over nine datasets and three seeds; wins count dataset means. 32/64/full are recipe extensions.

| Support | CLAP | COOPER-CLS | Δ (pp) | Dataset wins |
|---|---|---|---|---|
| 1 | 70.30 | 65.32 | -4.98 | 1/9 |
| 2 | 73.59 | 71.54 | -2.04 | 4/9 |
| 4 | 76.61 | 76.85 | +0.24 | 6/9 |
| 8 | 78.57 | 79.88 | +1.31 | 7/9 |
| 16 | 81.20 | 81.66 | +0.46 | 4/9 |
| 32 | 82.48 | 83.54 | +1.06 | 9/9 |
| 64 | 82.84 | 84.69 | +1.85 | 9/9 |
| Full | 82.89 | 85.18 | +2.29 | 9/9 |

**The low-shot comparison depends on the baseline.** CLAP wins at one and two shots; COOPER-CLS uses its unchanged anchor at one. The $+0.46$-pp mean at 16 shots includes only four dataset wins and a $+4.29$-pp EuroSAT gain, so it is not broad dominance. Wins on every dataset appear at 32, 64 and full support, beyond CLAP's published 1–16-shot range. Those extensions use available TRAIN images only; some 64-shot and full memberships coincide. These descriptive native-recipe results concern joint COOPER-CLS, not the per-example SAM3 estimator.

### G.2 Matched-feature diagnostics and deployment cost

The separate 16-shot, 11-dataset, three-seed study on CLIP, PE and SigLIP uses a class-weighted text-regularized probe, denoted *CLAP-style diagnostic*. It is not original-code CLAP: both controls use single-view cached features, normalized anchors and a common temperature. Table G.3 separates its accuracy and CPU fitting times from the native-recipe comparison. COOPER-CLS includes candidate search, selection, guarding and refit; feature extraction and evaluation are excluded for both. Timings are descriptive means over the same 33 dataset–seed cells; peak CPU RAM was not recorded.

Table H.1: **Same-decoder cross-domain transfer.** Shared-class mIoU (%). Target zero-shot uses Original; Target-opt. is the target-trained reference; Transfer deploys the source bank without target fitting. $\Delta$ = Transfer minus target zero-shot.

| Source | Target | Shared | Target zero-shot | Target-opt. | Transfer | $\Delta$ pp |
|---|---|---|---|---|---|---|
| Pascal Context | COCO | 27 | 64.33 | 69.49 | 68.10 | **+3.77** |
| Pascal Context | CamVid | 8 | 63.33 | 74.52 | 68.88 | **+5.55** |
| COCO | Pascal Context | 27 | 69.35 | 76.10 | 73.73 | **+4.37** |
| COCO | ADE-847 | 30 | 53.60 | 52.72 | 55.74 | **+2.14** |
| Cityscapes | COCO | 9 | 68.51 | 72.33 | 70.22 | +1.72 |
| CamVid | Cityscapes | 7 | 66.78 | 68.70 | 68.58 | +1.80 |
| Cityscapes | CamVid | 7 | 65.99 | 75.35 | 67.98 | +1.98 |
| Pascal Context | Pascal VOC | 19 | 77.02 | 82.43 | 79.34 | +2.32 |
| Pascal VOC | Pascal Context | 19 | 74.84 | 83.56 | 76.12 | +1.28 |
| | | | | | **Mean** | **+2.77** |

Table G.3: **Cached-feature diagnostic, not original CLAP.** Mean 16-shot top-1 (%) and CPU fit time (s) over 11 datasets and three seeds.

| Backbone | Diagnostic acc. | COOPER-CLS acc. | Diagnostic fit (s) | COOPER-CLS fit (s) |
|---|---|---|---|---|
| CLIP B/16 | 82.12 | 82.14 | 11.15 | 7.25 |
| PE-Core B/16 | 89.68 | 89.50 | 22.42 | 11.75 |
| SigLIP B/16 | 87.18 | 87.12 | 16.96 | 7.68 |

COOPER-CLS's accuracy differences are small and mixed ($+0.02/-0.18/-0.06$ pp on CLIP/PE/SigLIP), while recorded CPU fitting time is lower for all three. This diagnostic supports economical bank construction with similar observed accuracy, not an accuracy advantage over original CLAP.

**Inference is a shared fixed-bank computation.** A separate inference-only microbenchmark uses these frozen SUN397 seed-1 banks (397 classes), an RTX A6000, FP32 with TF32 disabled, and synthetic $224\times224$ RGB inputs already on GPU. Both methods retain only the same vision encoder and one normalized bank; feature normalization, bank multiplication and argmax are timed. Three rounds each use five warmups and 25 event-timed calls, reversing method order in the middle round. Batch-one latencies, computed as the median of the round medians, are 4.08/4.08 ms (CLIP), 6.39/6.42 ms (PE), and 4.27/4.26 ms (SigLIP), diagnostic/COOPER-CLS. Peak allocated GPU memory is identical within each pair: 350/377/374 MiB. Batch-32 peaks are 654/610/606 MiB, also identical. Allocation includes visual weights and live forward tensors but excludes CUDA context/driver memory; the unused text tower is removed for both. Loading, preprocessing, transfers and serving overhead are not timed; sampled worker-memory measurements have a different scope. This is not a benchmark of the untouched original CLAP evaluator. The fitting-cost difference ends at deployment: both diagnostic banks use the same inference computation, with matching allocated memory and near-identical measured latency.

# H Transfer Audits and Empirical Boundaries

A reusable bank must work beyond the images used to fit it. Domain and video transfer test that requirement with the predictor fixed. The final diagnostics ask where reuse fails and whether simple geometry summaries explain the difference.

## H.1 Same-Predictor Cross-Domain Transfer

We hold SAM3's encoders and decoder fixed, deploy a source bank without target retraining, and evaluate class names shared by both datasets. This tests reuse at the same interface; changing encoders would also change the space in which the bank is defined.

**Transfer can be broad or concentrated in one concept.** All nine domain means improve, but class-level checks distinguish two stories. Context→COCO improves 16/27 classes, harms none by

more than 1 pp, and retains $+3.30$ pp after removing its best class. VOC→Context instead falls to $-1.12$ pp without *pottedplant*; the two VOC/Context gains are concentrated in that class. A positive mean alone is therefore insufficient evidence of broadly reusable corrections.

Target-trained banks usually remain stronger. COCO→ADE-847 is the exception, exceeding the target-trained bank (55.74 versus 52.72), consistent with a benefit from source support when target per-class data are sparse. Together, these results support reusable conditioning at a fixed predictor, with both class-specific and domain-specific limits.

### H.2 Image-to-Video Transfer

The same correction can be carried through SAM3's frozen video interface without estimating a video-specific bank. Table H.2 reports positive-frame IoU and paired intervals that resample complete clips, sequences, or videos rather than treating correlated frames as independent.

Table H.2: **Image-to-video transfer with no video training.** The first effect is class-macro positive-frame IoU (0–1); the second resamples complete temporal clusters.

| Image bank → video | Macro IoU (Δ pp) | Cluster Δ [95% CI] |
|---|---|---|
| CamVid→CamVid | 0.353→0.403 (+5.00) | +6.77 [+5.72, +7.80] |
| Cityscapes→CamVid | 0.662→0.672 (+0.99) | +1.01 [-0.33, +2.45] |
| COCO→DAVIS | 0.651→0.670 (+1.87) | +1.87 [-0.19, +5.27] |
| PC–59→VSPW | 0.628→0.656 (+2.71) | +4.19 [+2.79, +5.63] |

Context-59→VSPW (Miao et al., 2021) tests semantic reuse without video-specific fitting: 343 validation videos, 24,502 frames and 33 exact-name classes contribute to a $+2.71$-pp class-macro gain (95% CI $[+0.32, +5.00]$). The complete-video gain is also positive, $+4.19$ pp $[+2.79, +5.63]$, so the result does not depend on treating frames as independent. No VSPW frames are used to train, blend, or select the bank. Same-domain CamVid also improves; Cityscapes→CamVid and COCO→DAVIS remain inconclusive because their clustered intervals include zero.

**The correction need not fade with propagation.** Later-frame bins show larger gains, but contain different videos. The stronger test pairs the first and last ten frames within 329 clips: the mean gain grows by $+2.02$ pp (95% CI $[+0.82, +3.25]$; Figure H.1). The median change is only $+0.59$ pp ($[-0.03, +2.11]$). Thus some clips sustain or increase the benefit without online fitting; improvement over time is not uniform.

These temporal diagnostics average all matched prompt-frame IoUs, including absent targets, across 39 matched classes and 343 clips. They are distinct from the positive-frame, exact-class macro-average in Table H.2. The paired intervals use 20,000 complete-video bootstrap replicates (seed 0).

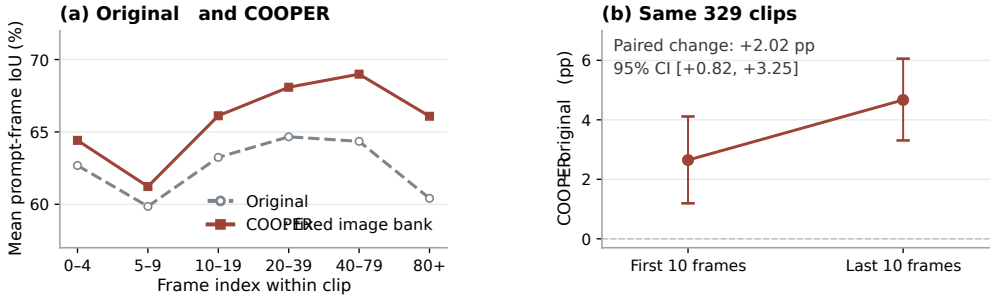



Figure H.1: **Image-to-video scores and paired advantages.** **(a)** Original and deployed IoU by frame-depth bin. **(b)** First-to-last advantage within 329 clips; 95% intervals resample complete videos.

Table H.3: **Seven-dataset cache-only curvature audit.** CV columns are medians over classes; Winsor CV clips each class at its 1st/99th percentiles, and $\Delta$ is the deployed bank's held-out mIoU gain.

| Dataset | Traj. | $\hat{h}>0$ | CV | Winsor CV | $\Delta$ pp |
|---|---|---|---|---|---|
| Pascal VOC | 82.4% | 56.0% | 1.542 | 1.252 | +6.55 |
| LoveDA | 98.3% | 57.6% | 3.287 | 2.319 | +18.41 |
| CamVid | 96.8% | 47.0% | 1.956 | 1.687 | +12.47 |
| Cityscapes | 93.5% | 58.7% | 2.754 | 1.953 | +5.65 |
| Pascal Context-59 | 85.4% | 51.1% | 2.815 | 1.898 | +9.88 |
| ImageNet-S50 | 66.2% | 45.0% | 0.931 | 0.912 | +8.64 |
| COCO | 95.3% | 50.0% | 2.389 | 2.193 | +10.90 |

The selected computer-tower trace in Figure 9 includes early failures before later repair. Both arms use identical inference and the full prompt set; full-image IoUs reproduce all 60 saved frames. It illustrates persistence, not a representative or monotonic gain.

### H.3 What Scalar Curvature Does and Does Not Predict

Proposition 1 identifies curvature–endpoint coupling as the exact difference between local and joint fits in the quadratic model. We test the cheaper scalar proxy $\mathrm{CV}_c(h)$ using cached optimization artifacts and no new model forwards. The proxy uses loss-unit decrease, $\hat{h}_i = 2\Delta_i^{\mathrm{loss}}/\|\hat{e}_i^\star - e_c\|_2^2$, over the flattened $(32, 256)$ SAM3 tensor. This is deliberately separate from the reported merge rule, whose gain and $0.5$-pp gain threshold are measured in IoU percentage points.

Across the 54 classes with room for validation improvement, CV versus recovery has Spearman $\rho = 0.096$ ($p = 0.491$); across seven dataset-level gains, $\rho = 0.393$ ($p = 0.383$). The prespecified negative relationship is therefore not supported. Raw CV is also tail-sensitive, motivating the winsorized companion in Table H.3.

The quadratic control explains why dispersion alone is insufficient: holding $\mathrm{CV}(h)$ fixed while changing curvature–endpoint pairing can reverse the estimator ordering (Figure M.1). The cache audit does not measure that coupling. Its negative result therefore rejects this scalar proxy as a selector, not Proposition 1; neither result identifies the cause of the SAM3–classification difference.

### H.4 Where the Bank Does Not Transfer

Jointly fitted class embeddings depend on the competing source labels. Transplanting them is therefore a different test from reusing independently consumed SAM3 prompts. The tested transplant is non-positive in 25 of 27 pairs, and MCM OOD detection worsens despite better in-distribution top-1 accuracy. These results limit that reuse procedure; they do not isolate geometry as the cause or establish that the representation itself cannot transfer.

Support also bounds the correction. Training-only guards can retain Original or the incumbent when a candidate lacks evidence for adoption (Appendix J.7). This differs from the gain-weighting gate: fewer than three above-threshold gains trigger uniform averaging, not automatic rejection. On the verified ADE-847 rerun, unsupported classes retain Original and the dataset still gains $+2.75$ pp. Fallback limits unsupported corrections but does not guarantee that an accepted embedding cannot degrade on unseen data; gains shrink toward Original as per-class support decreases.

## I Datasets and Protocols

This section defines the data and metrics for Section 5. COOPER and COOPER-CLS keep all model weights frozen. Training data alone are used for per-example SAM3 fits, aggregation, joint classification fitting, and selection. Training holdouts are subsets reserved for selection or guarding; the reported held-out evaluation split is not used for either. For SAM3, the table counts evaluated image–class pairs, each scored by binary IoU. COOPER-CLS jointly fits the class embeddings for CLIP, PE, and SigLIP and reports closed-set top-1 accuracy on the listed evaluation split. The separate local classification controls follow Appendix J.3. No external images beyond these datasets are used.

SAM3 mIoU first averages image–class IoUs within each evaluated class, then weights classes equally; *class-mean* and *class-macro* refer to this same metric. It is not a single IoU over pooled dataset pixels. Scores use percentages unless marked as fractions; gains use percentage points (pp).

Table I.1: **Segmentation benchmarks used with SAM3.**

| Dataset | Classes | Bank data | Reported eval | Prompts | Protocol notes |
|---|---|---|---|---|---|
| Pascal VOC 2012 (Everingham et al., 2012) | 20 | full train | val | 2,148 | Foreground semantic classes only. Evaluation is class-local binary mask prediction with standard VOC void handling. |
| Cityscapes (Cordts et al., 2016) | 19 | 1,300 train images | validation | 6,005 | Semantic label-ID protocol from the HF mirror; class names follow the standard Cityscapes taxonomy. |
| CamVid (Brostow et al., 2009) | 29 | full train | val | 1,562 | Evaluation covers the 29 non-empty/valid semantic classes observed in the split, with ignore-mask handling. |
| LoveDA (Wang et al., 2021) | 6 | full train | val | 6,959 | Land-use segmentation with six foreground classes; the no-data label is excluded from supervision and evaluation. |
| Pascal Context 59 (Mottaghi et al., 2014) | 59 | 3,400 train images | val | 25,975 | Standard 59-class semantic segmentation label set with void handling; evaluation uses the full validation prompt set. |
| COCO 2017 (Lin et al., 2014; Caesar et al., 2018) | 133 | 2,000 train images | val | 34,730 | Object/stuff prompts include merged stuff labels; evaluation uses the 5k-image validation split. |
| ImageNet-S50 (Gao et al., 2021b) | 50 | full train | val | 759 | Fifty ImageNet-style object categories with class-local binary mask prompts; evaluation is over held-out validation images containing those categories. |
| ADE-847 (Zhou et al., 2017) | 847 | 12,000 train prompts | validation | 22,313 | Official OV-Seg 847-class ontology loaded via metadata; class presence is dynamic per image, so only present classes are prompted. |
| Kvasir-SEG (Jha et al., 2020) | 1 | 800 train images | 200-image validation | 200 | Deterministic SHA256 split of all 1,000 images; uncertainty resamples images rather than the single class. |
| PascalPart (Chen et al., 2014) | 213 | 8,120 train images | 1,983-image validation | 22,075 | Object and compound part masks; the compositional test estimates shared part corrections leave-one-object-out. |
| LVIS v1 (Gupta et al., 2019) | 1,203 | 99,388 train images | 4,752-image validation | 15,491 | Official rare/common/frequent ontology; classes below the minimum construction support retain the original embedding. |

For the main SAM3 domain table, the large training splits use the training-only budgets shown above to keep the bank-estimation pass tractable. These budgets are fixed before validation and do not select on validation results. The prompt count column is the final held-out evaluation workload, not the number of optimized training prompts.

The three boundary extensions split whole training images 50/50 into endpoint construction and independent training-only calibration. Kvasir-SEG and PascalPart use nested supports $n \in \{16, 32\}$; LVIS uses $n \in \{4, 8\}$. All use $T = 50$, and the one-standard-error guard retains the original class tensor unless the train-calibration improvement is positive. Validation is scored once after that policy is frozen. PascalPart and LVIS use paired class bootstraps; Kvasir-SEG uses a paired image bootstrap because it has one class.

Table I.2: **LVIS long-tail and support audit.** Absolute class-macro mIoU (0–1) and paired class-bootstrap differences (pp). Frequency rows report accepted/ontology classes; support rows report accepted validation-present classes.

| Slice | Cls. | Accepted | Original | COOPER | Δ pp [95% CI] |
|---|---|---|---|---|---|
| *Official frequency* | | | | | |
| Rare | 70 | 23/337 | 0.4568 | 0.4878 | +3.10 [+0.24,+6.91] |
| Common | 345 | 223/461 | 0.4853 | 0.5687 | +8.34 [+6.12,+10.72] |
| Frequent | 389 | 192/405 | 0.5878 | 0.6563 | +6.84 [+5.39,+8.40] |
| *Effective fit support* | | | | | |
| 0 | 49 | 0 | 0.3891 | 0.3891 | +0.00 [+0.00,+0.00] |
| 4 | 57 | 20 | 0.5492 | 0.6211 | +7.19 [+2.54,+12.80] |
| 8 | 698 | 333 | 0.5411 | 0.6177 | +7.66 [+6.27,+9.09] |

The guarded policy accepts 438/1203 ontology classes; 353 are present in validation. All 765 fallbacks are exact (maximum absolute replay difference 0.0). The zero-support row is therefore unchanged, while both supported rows improve. Rare, common and frequent groups all gain, so the benefit is not confined to frequent concepts. Support count is only weakly associated with class gain ($\rho = 0.121$): available support governs eligibility for adaptation, but does not by itself explain the size of the repair.

Phase A stores 6888 endpoints, including 1 tiny target that is empty at decoder resolution and recorded with zero gain. Early stopping reduces 50 nominal steps to 41.91 realized steps on average. Phase A costs 3.95 GPU-hours; complete training-only construction, including calibration and selection, costs 4.46 GPU-hours.

Table I.3: **PascalPart leave-one-object-out composition.** Absolute class-macro mIoU (0–1) on the 136 validation-present part classes with at least one donor from another object.

| Embedding bank | mIoU |
|---|---|
| Original | 0.2582 |
| Object only | 0.2932 |
| Shared part only | 0.3423 |
| Additive, LOO | 0.3768 |
| Direct compound | 0.4502 |

The additive bank improves the original embedding by $+11.86$ pp [$+9.31$,$+14.43$] and improves the object-only transfer by $+8.36$ pp [$+6.26$,$+10.46$]. It remains $-7.35$ pp [$-9.69$,$-5.21$] below direct compound correction. On the stricter 99-class subset where the target object, target compound, and cross-object donor corrections are all learned, the median normalized additive defect is 1.107; its Spearman association with the direct-minus-additive mIoU deficit is 0.279. This supports reusable structure but rejects exact embedding-space additivity in the interaction-heavy cases. Direct and compositional banks were evaluated in separate sweeps, so their interval compares paired gains relative to each sweep's own original baseline; the original-baseline replay drift is only $-0.009$ pp class-macro.

Two protocol differences matter when comparing with published prompt-learning results. Our Caltech101 protocol uses the cleaned 100-class variant described above, which differs from some published CoOp setups. Food-101 follows the HF naming convention, where `validation` is the standard evaluation split. Comparisons with external results therefore do not always use identical protocols.

Table I.4: **Eleven classification benchmarks used across backbones.**

| Dataset | Classes | Bank data | Reported eval | Protocol notes |
|---|---|---|---|---|
| CIFAR-10 (Krizhevsky, 2009) | 10 | train | test | Official torchvision split with the standard class names. |
| CIFAR-100 (Krizhevsky, 2009) | 100 | train | test | Official torchvision split with the standard fine-grained class names. |
| Oxford-IIIT Pets (Parkhi et al., 2012) | 37 | train | test | HF mirror with the standard species labels and the dataset-specific pet template. |
| Stanford Cars (Krause et al., 2013) | 196 | train | test | HF parquet mirror of the standard car-model benchmark. |
| Flowers102 (Nilsback & Zisserman, 2008) | 102 | train+validation | test | Canonical CLIP/CoOp protocol that merges train and validation for training and uses the standard test split for evaluation. |
| FGVC-Aircraft (Maji et al., 2013) | 100 | train | test | Variant-level 100-class aircraft recognition. |
| DTD (Cimpoi et al., 2014) | 47 | train | test | Describable textures benchmark with the standard train/test split. |
| EuroSAT (Helber et al., 2019) | 10 | train | test | HF parquet mirror; class strings are normalized to the verbose OpenAI CLIP names used in the paper protocol. |
| Food-101 (Bossard et al., 2014) | 101 | train | validation | HF uses the validation split as the canonical test set; we follow that convention. |
| Caltech101 (Fei-Fei et al., 2004) | 100 | train | test | HF mirror with CoOp-style cleanup: `background_google` and `faces_easy` are excluded, leaving 100 semantic classes with CLIP-style name normalization. |
| SUN397 (Xiao et al., 2010) | 397 | train | test | Standard scene-recognition protocol over the HF shard mirror. |

# J Implementation Details

This section specifies the estimators and selection protocols behind Section 4 and Figure 10. COOPER uses per-example SAM3 fitting; the classification extension, COOPER-CLS, uses anchored joint fitting.

## J.1 Protocol and selection rules

All COOPER fits freeze the vision encoder, text encoder and predictor. We use training data alone to construct and select the bank: SAM3 selection compares class-level candidates; COOPER-CLS selection chooses one bank-wide recipe. Evaluation labels enter neither procedure. Guards are empirical fallback rules, not guarantees on unseen data. Deployment uses the stored bank without an online optimizer. CoOp and decoder LoRA change context tokens and cross-attention weights, respectively; their recipes and Original fallbacks are in Appendix K.

## J.2 Models and predictor inputs

- **SAM3** (`facebook/sam3`). The decoder consumes the 32-token, 256-dimensional text sequence after projection, not a pooled vector. Vision features are shared across classes; forwards use bf16.
- **CLIP**. The representative head is ViT-B/16 (`openai/clip-vit-base-patch16`), dimension 512 at resolution 224. The complete table also includes ViT-B/32 (512, 224) and RN50x4 (640, 288).
- **PE-B/16** (Bolya et al., 2025) (`PE-Core-B16-224`): dimension 1,024, context length 32. Additional models are S/16-384 (512 dimensions) and L/14-336 (1,024 dimensions).
- **SigLIP-B/16** (`google/siglip-base-patch16-224`): text is padded to the maximum context length. Additional models are B/16-384 and SO400M/14-384 (768 and 1,152 dimensions).

### J.3 COOPER-CLS CLASSIFICATION PROTOCOL

COOPER-CLS jointly updates all $K$ class embeddings in a free $K \times D$ bank by full-batch AdamW on Eq. (6); it does not merge per-example fits. Table J.1 separates the published curves, their exact 16-shot replay, and the three-seed local–joint control. Each candidate is formed by normalizing the class embeddings in the blend $\mathrm{normalize}((1-\alpha)E^{(0)} + \alpha E_T)$, with one $\alpha$ shared across the bank. All features and class embeddings are normalized; rejected candidates retain the complete Original bank.

Table J.1: **Classification protocols.** Support counts are per class; selection and guarding use training data only.

| Study | Support / seed | Optimization | Selection / deployment |
|---|---|---|---|
| Nine-backbone curves | Nested supports; split seed 2026. Roles depend on minimum class support. | AdamW; grid A. Runtime scale and decay were not recorded. | Rule A: 1 shot retains Original; 2 shots use 2-fold fits, $z = 1.64$; 4/8 shots use 4 folds, $z = 1$; larger supports use 50/25/rest% FIT/SELECT/GUARD, $z = 1$. After acceptance, refit on all support examples. |
| Curve-matched CLIP | Same 16-shot supports; 8/4/4 roles; split seed 2026. | AdamW; grid A; logit scale 100; decay $10^{-4}$. | Rule A, $z = 1$; after acceptance, refit on all 16 examples. |
| 99-setting control | 11 datasets; CLIP/PE/SigLIP B/16; support seeds 1/2/3; 8/4/4 roles. | Adam; grid B; logit scale 100; zero decay. | Rule B, $z = 1$; no all-support refit. |

**Grid A** uses learning rate $0.003$, priors $\beta \in \{0.3, 3, 30\}$, checkpoints $T \in \{10, 25, 50, 100\}$ and blends $\alpha \in \{0.25, 0.5, 0.75, 1\}$: 48 candidates sharing trajectories. **Grid B** uses learning rates $\{0.003, 0.01\}$, priors $\{0, 0.3, 3, 30\}$, checkpoints $\{0, 10, 25, 50, 100, 200\}$ and blends $\{0, 0.25, 0.5, 0.75, 1\}$. Rule A ranks accuracy, margin, smaller movement, shorter horizon, stronger prior, then learning rate; rule B ranks accuracy, lower cross-entropy, smaller movement, earlier checkpoint, then fixed recipe order. Both require $\bar{d} - z s_d/\sqrt{n} > 0$, with image-level correctness difference $d$ against Original. Rule A additionally requires improved accuracy or tied accuracy with improved margin. Low-shot cross-fitting reuses outcomes for selection and guarding; subsequent all-support refitting is not independently certified.

**Support and metadata limitation.** The curve-matched CLIP supports are endpoint-ranked: ranking used training endpoint information beyond the selected subset, so its 16-shot point is not a strict 16-label-total acquisition study. Original nine-backbone logs also omit runtime logit scale and AdamW decay; 100 and $10^{-4}$ are verified defaults, not recovered run values. The CLIP 16-shot replay fixes both and reproduces every selected recipe and held-out prediction, resolving that point but not the missing metadata for other backbones or support sizes.

**Curve-matched 16-shot comparison.** Figure 10b uses the exact CLIP ViT-B/16 supports underlying panel a's 16-shot point, with matched selection, guard and refit (Table J.1).

The joint arm updates all class embeddings. The local arm fits one true-class embedding per example against fixed Original competitors, selects its lowest cross-entropy on that example among the initial embedding and checkpoints up to $T$, then averages normalized endpoints within each class and renormalizes. Classwise fitting optimizes one embedding on a class's positive examples against the same fixed competitors. Accepted recipes are refitted on all 16 examples; the local refit retains the same endpoint checkpoints up to the selected $T$. This matches support and deployment selection, not parameter counts or negative-class gradient paths. Each arm selects its own recipe using training data; no held-out result selects a hyperparameter or overrides the guard.

All eleven joint recipes and 98,814 held-out predictions match the original curve point (81.2781%). Joint fitting beats local on ten datasets and ties on CIFAR-10, where both retain Original. Its $+4.43$ pp mean has a paired-dataset bootstrap 95% interval $[2.29, 6.76]$ (20,000 resamples, seed 2026); this captures between-dataset variation for fixed banks, not construction-seed uncertainty. The separate three-seed study below is retained rather than pooled into it.

Table J.2: **Local and joint controls matched to the CLIP curve.** Held-out top-1 (%) at 16 shots. $\Delta m$: joint minus local mean cosine margin; each image's margin is its true-class score minus its strongest rival.

| Dataset | Original | Local mean | Classwise | Joint | $\Delta m \times 10^3$ |
|---|---|---|---|---|---|
| CIFAR-10 | 90.78 | 90.78 | 90.78 | 90.78 | +0.00 |
| CIFAR-100 | 68.22 | 69.96 | 70.72 | 72.90 | +1.84 |
| Pets | 89.07 | 89.07 | 89.07 | 93.49 | +10.61 |
| Flowers | 71.31 | 86.34 | 89.58 | 96.44 | +39.84 |
| Aircraft | 24.60 | 33.18 | 33.30 | 44.91 | +1.76 |
| DTD | 45.05 | 63.94 | 64.73 | 67.77 | +6.36 |
| EuroSAT | 53.61 | 78.94 | 79.31 | 83.24 | +23.55 |
| Food101 | 88.70 | 88.70 | 88.70 | 89.18 | +14.13 |
| SUN397 | 65.53 | 74.24 | 74.08 | 75.69 | +2.30 |
| Caltech101 | 94.28 | 96.85 | 97.00 | 97.58 | +6.71 |
| Cars | 64.67 | 73.36 | 74.82 | 82.08 | +12.36 |
| Mean | 68.71 | 76.85 | 77.46 | 81.28 | +10.86 |

**Separation improves without stronger positive alignment.** Joint fitting lowers both true-class cosine ($-0.05393$) and strongest-rival cosine ($-0.06479$) relative to local averaging. The rival falls further, so the margin grows by $0.01086$ (dataset-bootstrap 95% CI $[0.00498, 0.01815]$). This explains why independent positive-class improvement is an incomplete classification objective.

The diagnostic uses the same fixed banks and held-out images, with no refit or projection. Means average images within datasets, then datasets equally; the strongest rival can change. Since margins determine decisions, this is descriptive support for coordinated class separation, not an independent causal test of geometry or a quantity directly comparable to SAM3's decoder.

**Matched local, classwise, and joint classification fitting.** Table J.3 retains all 99 settings of the separate three-seed study. The 8/4/4 roles, grid B and absence of refitting (Table J.1) distinguish it from the exact curve replay and from original-code CLAP.

Table J.3: **Local, classwise, and joint classification fitting.** Mean top-1 (%) over eleven datasets and three seeds. Shared support roles, search grid, and training-only guard; no refit.

| Backbone | Local mean | Classwise fitting | Joint bank |
|---|---|---|---|
| CLIP B/16 | 77.88 | 77.48 | 80.21 |
| PE B/16 | 87.39 | 87.36 | 88.65 |
| SigLIP B/16 | 84.74 | 84.46 | 86.05 |

All fits start from Original. Table J.4 also tests gain-weighted local merging and joint fitting restricted to $\beta = 0$. Without a prior, joint fitting still leads uniform local merging by $1.92/0.83/1.02$ pp on guarded CLIP/PE/SigLIP means. Allowing prior selection adds $0.41/0.43/0.29$ pp, whereas gain weighting lowers all three local means. Joint fitting also leads before guarding: neither the prior nor rejection creates the local–joint ordering in these controls.

For class scores $s_k$ and $p_k = \exp(s_k)/\sum_j \exp(s_j)$, cross-entropy gives $\partial \mathcal{L}_{\mathrm{CE}}/\partial s_k = p_k - \mathbf{1}[k = y]$. Joint fitting updates the incorrect classes as well as the true class. Local and classwise fits instead improve against Original competitors, which change when the fitted embeddings are assembled. The derivative identifies what the controls omit; their measured ranking establishes which recipe wins.

The joint advantage is consistent with a role for cross-class competition; it does not isolate geometry, architecture, or task as its cause. The benchmarks are previously inspected development datasets, and legacy cached features provide feature-record identities rather than verifiable image-byte identities.

Table J.4: **All local, classwise, and joint classification controls.** Each cell gives training-selected / guarded mean top-1 (%). Original is unchanged by either operation.

| Estimator | CLIP B/16 | PE B/16 | SigLIP B/16 |
|---|---|---|---|
| Original | 68.71 | 82.64 | 76.86 |
| Local mean | 78.25 / 77.88 | 87.52 / 87.39 | 84.77 / 84.74 |
| Local gain-weighted | 77.91 / 77.59 | 87.25 / 87.13 | 84.41 / 84.49 |
| Classwise fitting | 77.70 / 77.48 | 87.60 / 87.36 | 84.43 / 84.46 |
| Joint bank | 80.38 / 80.21 | 88.69 / 88.65 | 86.17 / 86.05 |
| Joint, no prior | 79.99 / 79.80 | 88.21 / 88.22 | 85.85 / 85.76 |

### J.4 SAM3 PER-EXAMPLE OPTIMIZATION

- **SAM3**: Adam, lr $= 5 \times 10^{-2}$, with 8,192 free parameters per image–class pair. The refined objective is

$$\mathcal{L}_i^{\text{ref}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}} + 0.05 \ \text{softplus}(-\max_q a_q), \tag{7}$$

  where $a_q$ is the frozen head's objectness logit. Refined trajectories clip the embedding gradient to norm one and select the highest training-IoU iterate among scored states $0, \ldots, T-1$; the post-update state $T$ is unscored. Optimization exits early when Original IoU exceeds $0.95$. Table J.5 gives each domain's objective, support and horizon; the refined objective was fixed before transfer to CamVid, Cityscapes, PC-59 and COCO.

For completeness, the expanded cross-entropy terms suppressed in the main method section are

$$\text{CE}\big(\tau v_i E^\top, y_i\big) = -\log \frac{\exp(\tau v_i^\top E_{y_i})}{\sum_{c=1}^{K} \exp(\tau v_i^\top E_c)}, \tag{8}$$

$$\mathcal{L}_{\text{BCE}}(\hat{m}_i, y_i) = -\frac{1}{P} \sum_p \left[ y_{ip} \log \hat{m}_{ip} + (1 - y_{ip}) \log(1 - \hat{m}_{ip}) \right]. \tag{9}$$

Here $P$ is the number of pixels; the remaining notation follows Eqs. (6) and (4). The soft Dice term is

$$\mathcal{L}_{\text{Dice}}(\hat{m}_i, y_i) = 1 - \frac{2 \sum_p \hat{m}_{ip} y_{ip} + 1}{\sum_p \hat{m}_{ip} + \sum_p y_{ip} + 1}.$$

Both smoothing constants are 1; BCE uses a pixel mean and the query-activity term is $0.05 \log(1 + \exp[-\max_q a_{iq}])$. With mask logits $m_{iqp}$, set $Q_i^+ = \{q : a_{iq} > 0\}$ and $\hat{m}_{ip} = \max_{q \in Q_i} \sigma(m_{iqp})$. For the full training loss, $Q_i = Q_i^+$ unless it is empty, in which case $Q_i = \{\arg\max_q a_{iq}\}$. This fallback supplies a mask gradient; it does not change scoring. Training IoU uses $\mathbf{1}[\exists q \in Q_i^+ : m_{iqp} > 0]$, with an empty mask if no query is active. Held-out evaluation retains the native postprocessor: query probabilities are multiplied by concept-presence probability when available, filtered at $0.5$, and their sigmoid masks are bilinearly resized to native image resolution, thresholded at $0.5$, and unioned. Neither scoring path uses the top-query fallback. The earlier BCE-only recipe has no such fallback during training either.

SAM3's DETR-style decoder (Carion et al., 2020) separates per-query scores from a global concept-presence score (Carion et al., 2025, Sec. 3). Since $\text{softplus}(-\max_q a_{iq}) = -\log \sigma(\max_q a_{iq})$, our activity term is a positive-label logistic penalty on the highest-scoring query. Each fitted pair contains foreground, motivating at least one such proposal. The term does not directly supervise the global presence logit or guarantee that a query passes the final presence-weighted threshold.

### J.5 AGGREGATION AND SUPPORT SELECTION

Appendix D defines gain weighting, its $\tau_{\text{act}} = 0.005$ gate, and the blend/guard rule. Uniform controls still average metric-selected endpoints. Nested support selection fits the largest pool once, compares its predefined subsets on training pairs unused for fitting, and changes the incumbent only when the paired gain exceeds the declared standard-error penalty. The domain-specific candidate families are listed below; COOPER-CLS uses neither this mean nor a per-class blend.

### J.6 Construction and reporting conventions for Table 2

Table J.5 specifies COOPER's construction families; Appendix K specifies CoOp and fixed LoRA. Joint-R and Joint-O are fixed-checkpoint replays initialized from Original and COOPER, respectively. Gain references and uncertainty follow Appendix F; wall-time and worker-time definitions follow Appendix L.

### J.7 Protocols underlying the reported SAM3 results

Table J.5 separates the common estimator from its declared support regimes. "Stable" means exact fallback to the previously fixed embedding; it is not a validation-time choice. ImageNet-S50 is the only benchmark whose candidate family includes a shared global correction and low-rank hierarchical shrinkage.

Table J.5: **Exact protocol family behind each SAM3 result.** All selection columns use training data only; activity denotes the query-activity loss.

| Reported rows | Local objective | Support / horizon | Frozen candidate and guard |
|---|---|---|---|
| ImageNet-S50 | BCE | 10/class, $T = 50$ | Five-fold selector over gain-weighted class, global, and low-rank hierarchical candidates; penalty $z = 1$. |
| LoveDA | BCE, Dice, activity | $n \in \{16, 32\}$, $T = 200$ | Per-class support choice; development row for this refinement. |
| VOC, CamVid, Cityscapes, PC-59, COCO | BCE, Dice, activity | $n \in \{16, 32\}$, $T = 200$ | One-standard-error nested candidate versus the fixed incumbent. The four non-VOC rows are frozen transfers. |
| ADE-847 | BCE, Dice, activity | $n \in \{16, 32\}$, $T = 200$ | Same nested rule; unsupported or unconfirmed classes retain the fixed embedding. |
| Kvasir-SEG, PascalPart | BCE, Dice, activity | $n \in \{16, 32\}$, $T = 50$ | Predeclared scale-boundary protocol; image-disjoint calibration and one-standard-error fallback. |
| LVIS | BCE, Dice, activity | $n \in \{4, 8\}$, $T = 50$ | Low-support scale boundary; 765/1,203 ontology classes retain the original embedding. |

**ImageNet-S50: from local corrections to candidate banks.** This construction makes the ImageNet-S50 entry in Table J.5 explicit. It retains local optimization: the candidates transform already fitted endpoints, rather than jointly optimizing a shared embedding. Flatten each correction as $d_i = \hat{e}_i^\star - e_{c_i}$, let $\delta_c$ be its class gain-weighted mean with the gate in Eq. (5), and define the equal-class mean $\bar{\delta} = C^{-1} \sum_c \delta_c$. Stack $d_i - \bar{\delta}$ across fitting pairs; its leading right singular vectors define the rank-$r$ orthogonal projector $\Pi_r$. The seven candidates, in selector order, are

$$h_c^{96,.4}, \quad e_c + \delta_c, \quad e_c + \bar{\delta}, \quad h_c^{\text{full},.5}, \quad h_c^{160,.4}, \quad h_c^{160,.5}, \quad h_c^{96,.5},$$
$$h_c^{r,s} = e_c + \bar{\delta} + s\Pi_r(\delta_c - \bar{\delta}).$$

For the full-rank candidate, $\Pi_{\text{full}} = I$. These combine a domain-wide correction with a class-specific residual; low-rank projection and shrinkage limit the residual estimated from few fits. All use the recorded BCE, $T = 50$ endpoints; this is the candidate family behind the reported row, not an additional joint-training stage.

**Training-only selection and refitting.** Five images per class fit the candidate banks; five other training images per class score them. The default is $h_c^{96,.4}$. For each alternative, compute its paired IoU gain over that default and select the largest positive $\bar{d} - z s_d/\sqrt{n}$; otherwise retain the default. Ties retain the earlier candidate in the order above. Five-fold selection leaves out one calibration image per class, ordered by image ID, to choose $z \in \{0, 0.5, 1, 1.5, 2\}$ by mean left-out IoU (ties favor smaller $z$). The selected $z = 1$ is then applied to all five calibration images per class. Finally, recompute the candidates, including their means and basis, from endpoints fitted on all ten training images per class and apply the fixed class choices, with no extra blend. Held-out evaluation selects nothing. The 50 choices comprise 13 class means, 9 global corrections and 28 hierarchical candidates: local fits support both shared and class-specific corrections within one fixed deployment bank.

### J.8 PROMPTS

SAM3 uses plain noun phrases (its native training convention); CLIP and PE use "a photo of a {}." Radford et al. (2021); SigLIP uses "This is a photo of a {}." Zhai et al. (2023).

### J.9 COMPUTE RESOURCES

All experiments were run on a local Linux workstation exposing 16 CPU threads, 57 GiB system RAM, and two NVIDIA RTX A6000 GPUs (49,140 MiB each, marketed as 48 GB). Unless otherwise stated, each per-example optimization worker used one GPU slot, with SAM3 jobs parallelized across the two GPUs when memory allowed. Cached image features and saved Phase-A results were reused across sweeps. Appendix L reports recorded construction wall time and aggregate GPU-hours; it does not use deployment latency as evidence.

## K PROMPT-TUNING AND ADAPTER CONTROLS

The baseline question is not whether prompts or adapters can ever repair SAM3, but what the reported recipes actually test. We separate fixed benchmark settings from tuned diagnostics, then trace the zero-gain cases to their selection and optimization behavior.

**CoOp variants.** We follow the terminology of Context Optimization (CoOp) Zhou et al. (2022b), which learns continuous prompt-context vectors while freezing the pretrained model. Let $\tau_c = (w_{c,1}, \ldots, w_{c,L_c})$ denote the tokenized class name for class $c$, and let $M = 16$ be the number of learned context vectors. Our SAM3 CoOp baseline feeds the frozen text encoder

$$\big[\texttt{SOS},\, v_{c,1}, \ldots, v_{c,M},\, \tau_c,\, \texttt{EOS}\big].$$

The original CoOp paper calls the shared-context constraint $v_{c,j} = v_j$ for all classes *unified context*; we label this variant CoOp-C. It calls the class-indexed context $v_{c,j}$ *class-specific context* (CSC); we label this variant CoOp-B. Thus CoOp-C shares context vectors across classes, whereas CoOp-B fits separate vectors for each class. In both cases, the class-name tokens $\tau_c$ remain fixed and only the continuous context vectors are optimized.

**More context parameters do not supply more evidence.** CoOp-B adds class-specific capacity to CoOp-C, not new labels or descriptions; both optimize through the frozen text encoder. B's lower mean in Table 2 therefore shows that class specificity alone does not solve adaptation. The original CoOp comparison is also task-dependent (Zhou et al., 2022b); CoCoOp's base-to-new-class overfitting result (Zhou et al., 2022a) concerns a different shift, not a diagnosis of these same-class failures.

Both variants repair Kvasir, and C leads on ImageNet-S50, ruling out blanket prompt inadequacy. Kvasir has one class, so its C/B difference cannot establish a benefit from class specificity. The LoveDA interventions provide the sharper diagnosis: small updates and a differentiable fallback rescue a collapsed recipe (Table E.2). The trajectories in Appendix M.6 test this optimization failure, separately from finite sampled-prompt reachability. Cosine-head CoOp results remain in Appendix G.

**Dataset-tuned decoder-LoRA diagnostic.** LoRA is included as a compact model-adaptation baseline rather than as part of the embedding-estimator theory. The main domain matrix uses one globally fixed recipe (rank 32, alpha 64, learning rate $3 \times 10^{-4}$), trains fresh weights per dataset, and selects only the checkpoint on an image-disjoint training holdout. The recipe is chosen once from the ImageNet-S50, LoveDA, and VOC training screens (Kvasir-SEG is included but all candidates tie there) and is frozen before the remaining domains. For diagnosis, Table K.1 retains the more expensive per-dataset rank–learning-rate searches. These are labeled separately and are not mixed into the main LoRA row; the reported validation split is opened once after training-only selection in either protocol. The minimax recipe has 0.72 pp mean and 1.62 pp worst training-holdout mIoU regret relative to choosing a separate screened configuration on each informative development domain.

**Why the fixed LoRA row can equal Original.** The zero-gain ADE-847, CamVid and Kvasir entries in Table 2 all select epoch zero, the exact frozen no-op, from the image-disjoint training

holdout. The 15-epoch trajectories were run; zero is the selected deployment outcome, not evidence that no adapter was trained. This also explains their degenerate paired gain intervals. In the fixed Kvasir trajectory, mean pre-clipping gradient norm is nonzero in epoch one ($0.0524$), then exactly zero in epochs 2–15; training loss is $1$ throughout those later epochs and every evaluated training-holdout checkpoint has zero IoU. The baseline's hard active-query mask loss returns a parameter-independent constant when no query exceeds its threshold. That branch explains the lack of an optimization signal, not the absence of a representable polyp.

LoRA changes the decoder's text-conditioning cross-attention (Hu et al., 2022), so it can in principle compensate for a poor input embedding. SAM3 also scores concept presence separately from localization (Carion et al., 2025, Sec. 3; App. A.1); mask improvement alone need not survive scoring. We have not isolated those scores for LoRA. The full COOPER loss adds a differentiable no-query fallback and query-activity penalty that this LoRA recipe lacks, so Kvasir is not a loss-matched capacity test. LoRA's gains on Cityscapes/LVIS and CoOp's Kvasir gains establish the useful distinction: embedding repair can be sufficient without being the only successful route.

Table K.1: **Dataset-tuned SAM3 LoRA diagnostic (appendix only).** The interval is a paired class bootstrap for bank minus LoRA, in percentage points; absolute IoUs are on a 0–1 scale. These cells are not the fixed-recipe main-table row.

| Dataset | Original | LoRA | COOPER | COOPER−LoRA [95% CI] |
|---|---|---|---|---|
| ImageNet-S50 | 0.7836 | 0.7856 | **0.8696** | +8.40 [+3.31,+14.69] |
| LoveDA | 0.2714 | 0.4216 | **0.4555** | +3.38 [-2.64,+10.06] |

**Tuning changes the comparison, not the interface question.** Training-only selection chooses rank 4, learning rate $3 \times 10^{-4}$, epoch 3 on ImageNet-S50, and rank 16, epoch 5 on LoveDA. The bank's advantage over tuned LoRA is resolved on ImageNet-S50, but not on six-class LoveDA (Table K.1); LoRA's own gain on LoveDA is resolved (95% CI $[9.06, 22.09]$). Combining the bank and adapter loses $2.32$ pp on ImageNet-S50 and $0.16$ on LoveDA relative to the bank alone. Separate successful adaptations therefore need not compose: the adapter changes the predictor receiving the bank.

The LoveDA screen plus three complete 15-epoch refinement trajectories costs 14.13 aggregate GPU-hours; the selected full trajectory accounts for 4.82 GPU-hours, and its selected epoch is reached after 42,285 steps. Training BCE continues to fall after epoch 5 while training-holdout mIoU deteriorates, and rank 32 attains lower BCE than rank 16 but worse mIoU. This loss–IoU disagreement motivates selecting checkpoint and rank by image-disjoint training mIoU rather than BCE. Both fixed controls use seed 42, so they are finite-grid tuned diagnostics, not an exhaustive LoRA upper bound or a seed-variance study. The shared evaluators reproduce the reported domain-table banks within 0.075 pp.

## L Offline Construction Efficiency

This section tests whether the construction savings in Section 5 reflect less GPU work as well as parallelism. Table L.1 compares four completed jobs on the same workstation: COOPER includes endpoint fitting and training-only calibration on two GPUs; CoOp-C uses one GPU for 15 epochs. Held-out scoring is excluded.

Table L.1: **Offline bank construction versus CoOp-C.** A ratio above $1\times$ favors COOPER.

| | Wall time | | | Aggregate GPU-hours | | |
|---|---|---|---|---|---|---|
| Dataset | COOPER min | CoOp-C min | Ratio | COOPER | CoOp-C | Ratio |
| Pascal VOC | 41.4 | 175.7 | 4.24× | 1.38 | 2.93 | 2.12× |
| Cityscapes | 58.8 | 343.2 | 5.84× | 1.96 | 5.72 | 2.92× |
| CamVid | 58.2 | 87.2 | 1.50× | 1.94 | 1.45 | 0.75× |
| PC-59 | 139.3 | 564.0 | 4.05× | 4.64 | 9.40 | 2.02× |

Across the four jobs, COOPER reduces total wall time ($3.93\times$) and aggregate GPU-hours ($1.97\times$), but uses more GPU-hours on CamVid. Across Table 2's nine fully timed domains, COOPER is faster on $7/9$ against CoOp-C and $6/9$ against CoOp-B or LoRA, but slower on PascalPart/LVIS against all three and on COCO against CoOp-B/LoRA. ImageNet-S50 and LoveDA have incomplete timers. GPU counts, searches and stopping rules differ: these are recorded costs, not a budget-matched Pareto frontier or evidence of an inference advantage.

**Reading the main-table timers.** Benchmark entries are construction wall-minutes; fixed LoRA covers one 15-epoch trajectory, not recipe development. Joint-R/O instead sum per-class optimizer-loop seconds and divide by 60: *fitting-only worker-minutes*, including in-loop checkpoints but excluding model loading, feature extraction, preloading and training/held-out scoring. Unsupported fallbacks contribute zero, and Joint-O excludes construction of its COOPER initialization. Thus the joint timers cannot be compared as total construction costs. $\geq$ denotes a lower bound, *n.r.* an unrecorded time, and a blank an unexecuted configuration.

The practical benefit is reuse: per-example fitting occurs during bank construction and is not repeated for each test image. This amortization also applies to cached-prompt baselines; the measured difference is construction cost, not a distinct inference-time operation.

# M Estimator Proofs and Conditional Joint Refinement

Local averaging and joint fitting can target different embeddings; weighting can change the target again. The proofs establish these distinctions under the assumptions of Propositions 1, 2 and Corollary 1. The synthetic crossover illustrates their consequences. Kvasir and the refinement studies then test what the measured evidence supports in SAM3.

## M.1 Proof of Proposition 1

*Proof.* Under the exact quadratic model, $\nabla_e L(e; x_i, y_i) = H_i(e - e_i^\star)$. The empirical objective $\hat{F}(e) = N^{-1}\sum_i L(e; x_i, y_i)$ has positive definite Hessian $\bar{H}_N$, so its unique minimizer satisfies the normal equations $\bar{H}_N e_c^{\text{joint}} = N^{-1}\sum_i H_i e_i^\star$. Thus $e_c^{\text{joint}} = \bar{H}_N^{-1}\,\frac{1}{N}\sum_i H_i e_i^\star$, and substituting $e_i^\star = \mu_c^{\text{avg}}[N] + \xi_i$, with $\sum_i \xi_i = 0$, gives

$$e_c^{\text{joint}} - \mu_c^{\text{avg}}[N] = \bar{H}_N^{-1}\frac{1}{N}\sum_i H_i\xi_i = \bar{H}_N^{-1}\frac{1}{N}\sum_i (H_i - \bar{H}_N)\xi_i.$$

□

## M.2 Heterogeneous Repair: Proof, Weight Conditions, and a Matched Case

Proposition 2 concerns the distribution of returned endpoints, including nonstationary iterates or unchanged original embeddings. It does not require them to minimize the per-example loss. The latent indicator $B = 1$ denotes membership in a population of transferable repairs with mean $\mu_1$; that population and the quadratic target are assumptions, not labels recovered from the measured training gains. Endpoints with $B = 0$ may lie near the original embedding or be diffuse. Diffuseness alone does not create the stated bias: their mean must differ from $\mu_1$.

**Proof of Proposition 2.** Write $\Sigma = \mathrm{Cov}(Z)$ and $m = \mathbb{E}Z = \pi\mu_1 + (1-\pi)\mu_0$. Independence gives $\mathrm{Cov}(\bar{Z}_N) = \Sigma/N$, hence the bias–variance identity yields

$$\mathbb{E}Q(\bar{Z}_N) = \frac{\mathrm{tr}(H\Sigma)}{2N} + \frac{1}{2}\|m - \mu_1\|_H^2 = \frac{\mathrm{tr}(H\Sigma)}{2N} + \frac{(1-\pi)^2}{2}\|\mu_0 - \mu_1\|_H^2.$$

The moment assumptions and strong law give $N^{-1}\sum_i W_i \to \mathbb{E}W > 0$ and $N^{-1}\sum_i W_i Z_i \to \mathbb{E}[WZ]$ almost surely. The denominator is therefore eventually positive; any fixed fallback on samples with a zero denominator leaves the limit unchanged. Decompose $Z = \mu_B + (Z - \mu_B)$ to obtain

$$\mathbb{E}[WZ] = \pi a_1\mu_1 + (1-\pi)a_0\mu_0 + b_W, \qquad \mathbb{E}W = \pi a_1 + (1-\pi)a_0.$$

Subtracting $\mu_1$ proves the identity. In particular, $b_W = 0$ when $\mathbb{E}[W(Z - \mu_b) \mid B = b] = 0$ for both components, a sufficient condition weaker than within-component independence. Under this centering condition,

$$Q(m_W) = \frac{1}{2}\left(\frac{(1-\pi)a_0}{\pi a_1 + (1-\pi)a_0}\right)^2 \|\mu_0 - \mu_1\|_H^2 .$$

For $a_1 > 0$ the coefficient increases with $a_0/a_1$. Fixed small positive contamination therefore reduces, but does not eliminate, asymptotic bias. The case $a_0 = 0$ gives the claimed consistency. This proves almost-sure convergence of the estimator and its quadratic risk, not convergence of expected weighted finite-sample risk without further uniform-integrability conditions. □

**An oracle weight and the proxy gap.** Let $U$ collect per-example information and let $p(Z, U) = \Pr(B = 1 \mid Z, U)$. Taking the endpoint itself among the conditioning variables is important. By the tower property,

$$\mathbb{E}[pZ] = \mathbb{E}[BZ] = \pi\mu_1, \qquad \mathbb{E}p = \pi.$$

Thus posterior-weighted averaging is consistent for $\mu_1$, even for overlapping components and without imposing $b_W = 0$. More generally, for any fixed nonnegative proxy $W = w(Z, U)$ with positive mean and the required moments,

$$m_W - \mu_1 = \frac{\mathbb{E}[(W-p)(Z-\mu_1)]}{\mathbb{E}W}, \qquad \|m_W - \mu_1\|_H \le \frac{\sqrt{\mathbb{E}(W-p)^2\,\mathbb{E}\,\|Z-\mu_1\|_H^2}}{\mathbb{E}W}.$$

The inequality assumes the displayed second moments are finite. It identifies what a probability-proxy argument would need: calibration to transferable repair, including possible correlation with endpoint direction. Training IoU gain is not a measured repair probability; this identity does not make it risk-optimal.

**Why homogeneous repair does not force equality.** For every fixed weight rule with the stated moments,

$$m_W - \mathbb{E}Z = \frac{\mathrm{Cov}(W, Z)}{\mathbb{E}W}.$$

Constant weights give identical sample means; zero covariance gives identical population targets, but not necessarily identical finite-sample means. Even a single repair population can be biased by weighting: if $Z \in \{-1, 1\}$ is equiprobable and $W = 2 + Z$, then $\mathbb{E}Z = 0$ but $m_W = 1/2$. Weight concentration can also increase sampling variance. There is no general finite-sample theorem that risk-optimal weights are monotone in repair probability when endpoint variances differ.

**Relation to the gain-weighting gate.** Within a class, COOPER uses $W_i = \max(g_i, 0)$ once at least three gains exceed $\tau_{\mathrm{act}}$; otherwise it uses the uniform mean. If the per-example pairs are i.i.d. and $\Pr(g > \tau_{\mathrm{act}}) > 0$, the gate becomes active almost surely as $N$ grows, so the fixed positive-gain rule determines the asymptotic target. If that probability is zero, the gate stays inactive and the estimator is uniform. The proposition does not remove finite-support gate effects, training-only candidate selection, or the empirical guard, and does not assert that a finite benchmark support satisfies its latent-mixture assumptions.

### M.2.1 Matched Kvasir Weighting Replay

**Protocol and uncertainty.** This fresh paired replay changes the weights in every affected mean while reusing the same per-example endpoint caches. Both arms use the same candidate supports, training-only nested/guard rule ($z = 1$), inference settings, and held-out split. Each arm independently selects $n = 32$: 32 fitting images, 64 calibration images, and 200 evaluation images are disjoint. Reconstructed means and fixed banks agree bitwise with independent reductions. The gain-weighting gate activates (24 of 32 gains exceed $0.005$). It does not discard every below-threshold endpoint: all positive gains receive weight once activated.

Table E.5 reports the absolute scores. The weighted-minus-uniform difference is $+37.53$ pp, with paired image-bootstrap 95% interval $[+31.85, +43.28]$ (20,000 draws, seed 20260916). Kvasir has one foreground class, so class resampling would be degenerate. This interval conditions on the fitted banks and treats held-out images as resampling units; it does not measure construction-seed or acquisition-cluster uncertainty. These are previously inspected data, not independent confirmation.

**A few strong repairs dominate the merge.** Eight of 32 endpoints have gain at most $0.005$; seven are exactly unchanged original embeddings with zero gain. Their combined mass falls from 25% under uniform averaging to $0.006\%$ under gain weighting. Seven endpoints with training gain above 50 IoU points receive 79.38% of the weighted mass, versus 21.88% under uniform averaging. Effective sample size is 10.42 out of 32. This is direct evidence of repair-gain concentration, not an estimate of the latent $\pi$ or component means in Proposition 2. In particular, low gain need not mean a poor endpoint: it can also mean that the original embedding already performs well on that training example.

**Weighting changes both direction and magnitude.** The weighted displacement is 1.445 times as long as the uniform displacement, with cosine 0.934 between them; both magnitude and direction change. Thus this contrast does not isolate a unique geometric cause. Training gains use decoder-resolution, query-gated IoU; held-out inference uses full-resolution postprocessing with query scores multiplied by presence scores. Sixteen of the 32 training original scores are nonzero (training mean 41.19%), despite zero native held-out reference mIoU. It would be incorrect to call every per-example reference broken. The matched contrast establishes a benefit of weighting for this construction, not calibration of gains to transfer or universal necessity.

**Relation to the historical benchmark.** Table 2 retains its archived Kvasir score of 78.46%. Fresh training-only selection chooses $n = 32$ in both replay arms, whereas the archived bank used $n = 16$; that archived bank scores 78.47% in this inference pass. The weighting contrast is therefore 79.72 versus 42.19, not a subtraction from the historical benchmark. No per-example trajectories were retrained for this comparison.

### M.3 Controlled Curvature–Direction Crossover

**Exact supplementary examples.** Figure M.2 contains analytic examples, with no model training or Monte Carlo error. Panel (a) uses $h = (1, 1)$ and $z = (-1, 1)$: both the endpoint mean and the minimizer of the averaged quadratic are zero. Panel (b) fixes $h = (1, 1, 4, 4)$ and enumerates all six distinct permutations of $z = (-1, -1, 1, 1)$. Every pairing has $\bar{h} = 2.5$, $\mathrm{CV}(h) = 0.6$, $\bar{z} = 0$, and endpoint variance one. The empirical coupling $C$ takes values $-1.5, 0, 1.5$ (in one, four, and one pairings), giving discrepancies $C/\bar{h} = -0.6, 0, 0.6$. Thus even fixing both marginal distributions does not fix the estimator gap. Panel (c) keeps the positive-coupling fit unchanged: the endpoint mean is zero and the joint fit is $0.6$. Under the separately declared synthetic evaluation loss $R_\theta(e) = \frac{1}{2}(e - \theta)^2$, their loss difference is $0.18 - 0.6\theta$, which changes sign at $\theta = 0.3$. This is a counterexample to inferring risk order from training geometry alone, not an empirical oracle or evidence that a particular domain has that target.

**Larger Monte Carlo crossover.** Consider $n$ training examples from one class, with per-example preferred embeddings $z_i$. Endpoint averaging estimates the class embedding as $n^{-1} \sum_i z_i$, assigning every endpoint the same weight. Under the scalar quadratic loss $L_i(e) = \frac{1}{2} h_i \|e - z_i\|^2$, joint fitting instead returns $\sum_i h_i z_i / \sum_i h_i$, so endpoints with larger local curvature $h_i > 0$ exert more influence. Whether this weighting helps depends on the sampling law and evaluation target; alignment among high-curvature examples alone does not establish that their common direction transfers. The relevant quantity is therefore the coupling between curvature and endpoint direction, not curvature dispersion alone.

The historical simulation below uses eight-dimensional exact quadratics. For $v = \mathrm{CV}(h) > 0$, draw independent standard Gaussians $u, \epsilon_1, \ldots, \epsilon_8$ and set

$$h = \exp(su - s^2/2), \quad s^2 = \log(1 + v^2), \qquad z_1 = \kappa(h - 1)/v + \sqrt{1 - \kappa^2}\, \epsilon_1, \quad z_j = \epsilon_j \ (j > 1).$$

Then $\mathbb{E}[h] = 1$, $\mathbb{E}[z] = 0$, and the population quadratic minimizer is $\mathbb{E}[hz]/\mathbb{E}[h] = \kappa v (1, 0, \ldots, 0)$. At $v = 0$ the simulation instead uses $h = 1$ and independent standard-Gaussian endpoints. The plotted risks are excess population loss $\frac{1}{2} \|\hat{e} - \mu_c^{\mathrm{pop}}\|_2^2$, averaged over 3,000 independent replications per cell. This same population model generates Figures 4 and M.1. The separate illustration in Figure M.2(c) instead varies its evaluation target while keeping the fitted pair fixed. In Figure M.1, dots identify paired risk-difference estimates within $1.96$ Monte Carlo standard errors of zero, and the slice bands in Figure 4 show the paired mean difference plus or minus $1.96$ standard errors over

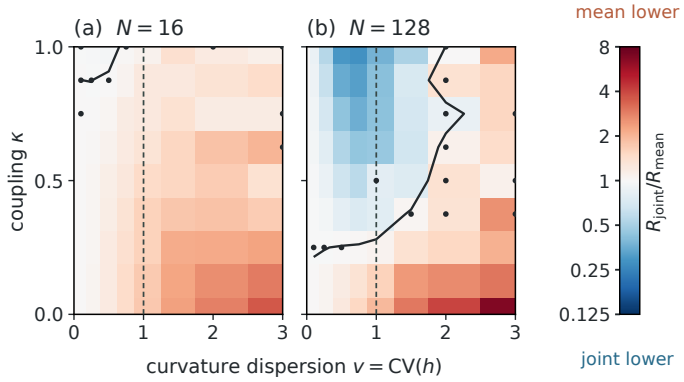



Figure M.1: **Synthetic risk across support and coupling.** Joint/endpoint-mean risk ratios at $N = 16$ and 128; black contours mark estimated ties, dots unresolved paired differences. Dashed lines select the $v = 1$ slice retained in Figure 4.

those 3,000 replications. These are pointwise normal approximations to simulation uncertainty, not confidence intervals for SAM3. The black curves interpolate equal estimated risk, excluding the structural $v = 0$ tie column; the dashed lines select the $v = 1$ slice.

**Proof of Corollary 1.** Fix the dimension $d \geq 1$, dispersion $v \geq 0$, and coupling $\kappa \in [0, 1]$, and take $N \geq 1$ independent copies of the population above; the saved simulation uses $d = 8$. Write $\bar{z} = N^{-1} \sum_i z_i$, $\hat{e}_{\mathrm{joint}} = \sum_i h_i z_i / \sum_i h_i$, and $\mu = \mu_c^{\mathrm{pop}} = \kappa v \mathbf{e}_1$, where $\mathbf{e}_1$ is the first coordinate vector. Since $\mathbb{E}[h] = 1$, the population loss $F(e) = \frac{1}{2}\mathbb{E}[h \, \|e - z\|_2^2]$ satisfies $F(e) - F(\mu) = \frac{1}{2} \|e - \mu\|_2^2$. Thus $R_{\mathrm{mean}}$ and $R_{\mathrm{joint}}$ average this excess loss over independent training samples from the stipulated population, not over SAM3 test predictions.

The construction gives $\mathbb{E}[z] = 0$ and $\mathrm{Cov}(z) = I_d$: for $v > 0$, the variance of its first coordinate is $\kappa^2 \, \mathrm{Var}(h)/v^2 + (1 - \kappa^2) = 1$, and the independent Gaussian coordinates have zero cross-covariance. The $v = 0$ branch has the same moments. Consequently,

$$R_{\mathrm{mean}} = \tfrac{1}{2}\mathbb{E} \, \|\bar{z} - \mu\|_2^2 = \frac{d}{2N} + \frac{\kappa^2 v^2}{2}.$$

When $\kappa = 0$, the endpoints are independent standard Gaussian vectors, independent of the curvatures. Conditional on $h_1, \ldots, h_N$, the joint fit is therefore centered Gaussian with covariance $[\sum_i h_i^2 / (\sum_i h_i)^2] I_d$, giving

$$R_{\mathrm{joint}} = \frac{d}{2}\mathbb{E}\left[\frac{\sum_i h_i^2}{(\sum_i h_i)^2}\right] \geq \frac{d}{2N} = R_{\mathrm{mean}}.$$

The inequality follows from Cauchy–Schwarz and is strict for $v > 0$ and $N \geq 2$, since independent nondegenerate lognormal curvatures are not all equal with probability one.

To justify convergence of the expected joint risk, rather than only of the estimator, set $X_i = h_i(z_i - \mu)$, $\bar{X} = N^{-1} \sum_i X_i$, and $\bar{h} = N^{-1} \sum_i h_i$. Then $\mathbb{E}[X_i] = 0$ and $\hat{e}_{\mathrm{joint}} - \mu = \bar{X}/\bar{h}$. All positive and negative lognormal moments are finite:

$$\mathbb{E}[h^r] = \exp\!\big(\tfrac{1}{2}(r^2 - r)s^2\big) \quad (r \in \mathbb{R}), \qquad \mathbb{E}[h^{-4}] = (1 + v^2)^{10}.$$

The affine dependence of $z$ on $h$ and Gaussian noise therefore ensures $M_4 := \mathbb{E} \, \|X_i\|_2^4 < \infty$. With $\Sigma_X = \mathbb{E}[X_i X_i^\top]$, independence and centering give the exact fourth-moment expansion

$$\mathbb{E} \, \big\|\bar{X}\big\|_2^4 = \frac{M_4}{N^3} + \frac{N - 1}{N^3} \left[(\mathrm{tr}\, \Sigma_X)^2 + 2 \, \mathrm{tr}(\Sigma_X^2)\right] \leq \frac{3M_4}{N^2}.$$

Here $\mathrm{tr}(\Sigma_X^2) \leq (\mathrm{tr}\, \Sigma_X)^2 \leq M_4$. Convexity of $t \mapsto t^{-4}$ gives $\mathbb{E}[\bar{h}^{-4}] \leq \mathbb{E}[h^{-4}]$. Applying Cauchy–Schwarz to the ratio, without assuming independence of its numerator and denominator, yields

$$R_{\mathrm{joint}} = \tfrac{1}{2}\mathbb{E}\Big[\big\|\bar{X}\big\|_2^2 \, \bar{h}^{-2}\Big] \leq \tfrac{1}{2}\sqrt{\mathbb{E} \, \big\|\bar{X}\big\|_2^4 \, \mathbb{E}[\bar{h}^{-4}]} \leq \frac{\sqrt{3M_4 \, \mathbb{E}[h^{-4}]}}{2N} = O(N^{-1}).$$

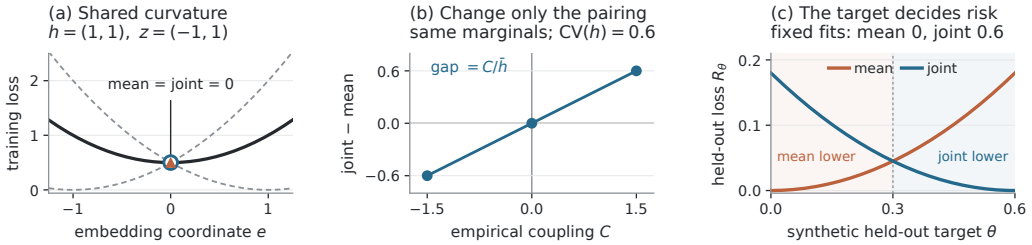



Figure M.2: **Exact supplementary examples. (a)** Shared curvature gives equal estimators (gray: instance losses; black: joint loss). **(b)** All six pairings of $h = (1, 1, 4, 4)$ and $z = (-1, -1, 1, 1)$ share their marginals, yet their empirical coupling changes the gap. **(c)** With the fitted pair fixed, changing the synthetic evaluation target in $R_\theta(e) = \frac{1}{2}(e - \theta)^2$ reverses risk order.

For fixed $\kappa v \neq 0$, $R_{\text{mean}}$ tends to the positive limit $\kappa^2 v^2/2$, whereas $R_{\text{joint}}$ tends to zero, so joint fitting has lower risk for all sufficiently large $N$. At $v = 0$, the two estimators coincide exactly for every $N$; at $N = 1$, they coincide for every admissible $v, \kappa$. The finite constant above depends on the stipulated population; no uniform crossover threshold is asserted when its parameters change with $N$, and these results do not certify a deployment selector. □

At fixed curvature dispersion, changing the coupling crosses the risk boundary (Figure M.1). More support expands the joint-favorable region in this stipulated model. The lesson is that dispersion alone cannot choose the estimator, not that COOPER should threshold a synthetic contour. The following audits ask what gradients and actual fitting paths reveal.

### M.4 Empirical Non-Stationarity at the Original Embedding

We measure the Phase-A training-loss gradient at the original embedding on 400 class-stratified training pairs per dataset. This gives 2,800 measurements across the seven allowed segmentation datasets using the same SAM3 commit as the deployed banks. VOC, LoveDA, and ImageNet-S50 replay complete cached vision features. Table M.1 reports the median of per-class median gradients, its scale-normalized companion, and held-out bank gain.

Table M.1: **Seven-dataset stationarity audit.** Gradients use stratified training pairs; Δ is held-out class-macro mIoU gain in percentage points.

| Dataset | Classes | Median $\|\nabla L(e_c)\|_2$ | Relative grad. | Δ pp |
|---|---|---|---|---|
| Pascal VOC | 20 | 1.94e-03 | 3.05e-05 | +6.55 |
| LoveDA | 6 | 8.53e-03 | 1.26e-04 | +18.41 |
| CamVid | 29 | 1.42e-03 | 2.04e-05 | +12.47 |
| Cityscapes | 19 | 7.67e-04 | 1.13e-05 | +5.65 |
| PC-59 | 59 | 3.21e-03 | 4.52e-05 | +9.88 |
| ImageNet-S50 | 50 | 3.92e-03 | 5.11e-05 | +8.64 |
| COCO | 133 | 2.28e-03 | 2.91e-05 | +10.90 |

Across datasets the association is positive but unresolved ($\rho = 0.500$, $p = 0.253$, $n = 7$). Ranking classes within each dataset gives $\rho = 0.400$ ($p = 1.46 \times 10^{-13}$) over 316 classes; the relative gradient gives $\rho = 0.379$ ($p = 3.17 \times 10^{-12}$). The relationship is strong in LoveDA, CamVid, Pascal Context-59, and COCO, but unresolved in VOC, Cityscapes, and ImageNet-S50. Radial gradient-energy share is not associated with gain in any dataset. Non-stationarity is therefore a useful within-regime diagnostic, not a universal dataset-level predictor or evidence that scale alone drives the correction.

### M.5 Non-Smoothness of the SAM3 Objective

SAM3's admissible mask queries and pixelwise winning queries can change as the embedding moves. The pixel maximum is piecewise smooth, and thresholded query admission can be discontinuous. Fixed indices at recorded checkpoints do not establish smoothness along the intervening path. This

limits the interpretation of both the exact-quadratic model and finite-difference curvature probes; the CoOp controls below test the active-query failure directly.

### M.6 Convergence Geometry of Unified CoOp

We trace unified-context CoOp on ImageNet-S50 (where it gains $+11.01$ mIoU), LoveDA ($-27.14$), and Cityscapes ($-55.52$). This is a controlled mechanism diagnostic rather than a replacement benchmark: every dataset uses the same random initialization, AdamW schedule, 1,000-update budget, four fixed fit pairs and two disjoint probe pairs per class. Cached frozen FPN features restrict every optimization step to the text encoder and decoder; no validation labels enter the trace.

For class gradients $g_{c,t} = \nabla_u F_c(u_t)$ of the shared context $u$, define

$$\kappa_t = \frac{\|C^{-1}\sum_c g_{c,t}\|}{C^{-1}\sum_c \|g_{c,t}\|}, \qquad q_t = C^{-1}\sum_c \mathbf{1}[\|g_{c,t}\| > 0]. \tag{10}$$

Set $\kappa_t = 0$ when its denominator is zero. Otherwise it measures alignment among class gradients, but can equal one when only one class remains active; $q_t\kappa_t$ combines measured support and alignment. It is a diagnostic score, not a bound on improvement or generalization.

The failure is not initial gradient disagreement. At $t = 0$, raw coherence is $0.56$ on ImageNet-S50, $0.96$ on LoveDA, and $0.90$ on Cityscapes. Instead, random CoOp contexts begin at very different distances from a useful active component: relative to the original embedding on identical probes, the initial context retains $86.7\%$ of IoU on ImageNet-S50, but only $33.5\%$ on LoveDA and $19.5\%$ on Cityscapes.

Adam amplifies this difference geometrically. On its first connected update, bias correction gives

$$u_1 - u_0 \simeq -\eta\,\mathrm{sign}(g_0), \qquad \|u_1 - u_0\| \simeq \eta\sqrt{16{,}384} = 0.128,$$

which is the measured norm on all three datasets despite very different raw gradient norms. Along this direction LoveDA falls to at most $50\%$ active support by $\alpha = 0.10$ and to zero by $\alpha = 0.20$, whereas ImageNet-S50 remains 96–98% active through $\alpha = 1.5$. Thus the realized LoveDA step is at least five times its observed zero-support radius. Cityscapes' first direction is less abrupt, but subsequent sparse directions reduce it to $2.6\%$ support by step 50–100.

Figure 3d retains all eleven first-direction scan points per dataset. Let $\theta_\alpha$ be the median class angle between the flattened $32\times256$ consumed tensor and its own Original tensor. The horizontal coordinate is $\theta_\alpha - \theta_0$, not angular distance from initialization, token-mean rotation, or CoOp–COOPER alignment. The vertical coordinate is $100\,\mathrm{IoU}_\alpha/\mathrm{IoU}_0$ on the same fixed training probes, using a ratio of class means. Initial probe IoUs are $13.01\%/11.38\%/78.10\%$ for LoveDA/Cityscapes/IN-S50: equal retained percentages need not mean equal absolute quality. The hollow marker denotes each random initialization at $(0, 100)$; lines follow $\alpha$, including angle reversals. The callout marks LoveDA's first observed zero IoU at $\alpha = 0.2$, when its native angle changes from $86.88°$ to $87.43°$. Subsequent query support partly recovers, but IoU stays zero; IN-S50 retains at least $91\%$ of its initial IoU throughout. Since direction, norm, and token coordinates co-vary, these paths establish neither an angle-only cause nor a transferable angular threshold. They are training-probe diagnostics, not held-out test results.

The early regime split repeats across context seeds 7, 17, and 42 with the image pairs held fixed: by step 100, LoveDA and Cityscapes collapse in $3/3$ runs, ImageNet-S50 in $0/3$; connected-update fractions over those steps are $0.080 \pm 0.040$, $0.107 \pm 0.059$, and $0.983 \pm 0.006$, respectively. Finally, log distance to the terminal bank is nearly linear even for the failed runs ($R^2 = 0.969, 0.960, 0.915$). Apparent geometric contraction therefore does not certify convergence to a useful optimum. Active support must be checked when interpreting these traces, and does not by itself ensure strong convexity or improvement. The no-query branch returns a disconnected constant one; with the traced implementation's disconnected-step handling, the failed runs stop making loss-driven updates instead of reaching a verified useful minimum.

### M.7 Case study: when does joint refinement help?

CamVid poses two distinct questions: can further fitting improve a local bank, and can we identify those opportunities before deployment? The following three-seed study supports the first claim but

not a reliable cross-domain switching rule. It tests both questions with disjoint training-selection images, actual refinement trajectories, and simpler controls.

**Design and scope.** We freeze SAM3 and reuse local banks from a separate component study on CamVid, VOC, and LoveDA with support seeds 7, 17, and 42. Training images are assigned globally by decoded image identity to disjoint FIT (3–8 images/class), SELECT (2–4), and GUARD (2–4) roles. The starting bank uses per-example maximum-IoU endpoints, gain weights, and a SELECT-chosen blend; it is frozen *before* GUARD acceptance. SELECT has therefore already influenced initialization; GUARD has not. All nine constructions and the analysis rules are fixed before new held-out scoring. All 156 supported class–seed outcomes are retained (78/60/18 by domain). These banks differ from the historical main-table banks; their local mIoUs over all evaluated classes are 46.51/81.79/42.11%.

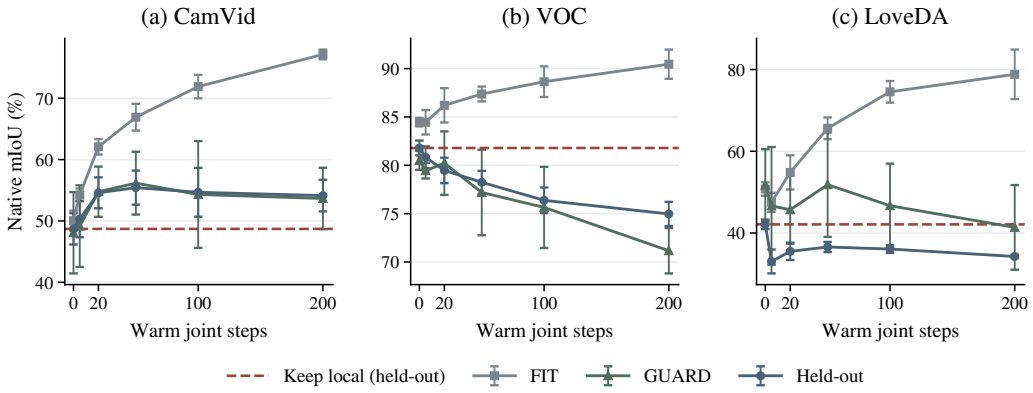



Figure M.3: **Fitting progress versus transfer.** Warm-joint native mIoU on supported classes, mean ± SD over three support seeds. Red: fixed local held-out score. GUARD images enter selection, not gradient updates.

**Matched continuation and selection controls.** Warm fitting starts from the same local embedding $e_c^0$ and minimizes FIT BCE–Dice loss for 200 Adam steps (learning rate 0.05, betas $(0.9, 0.999)$, $\epsilon = 10^{-8}$, no weight decay or clipping). Embedding parameters and losses are float32; forward computation uses bf16 autocast with TF32 disabled. A fixed regularized arm adds $\|e_c - e_c^0\|^2/(\|e_c^{\text{orig}}\|^2 + 10^{-12})$ with coefficient one, where $e_c^{\text{orig}}$ is the Original embedding; the coefficient is not tuned. The uniform reference averages the same per-example endpoints without blending or rejection, so it is a simpler bank reference, not an isolated weighting ablation.

Each early gate either retains $e_c^0$ or deploys the identical unregularized step-200 endpoint. For per-GUARD-image signals $q_i$, it switches only when $\bar{q} > \text{SE}(q)$ and at least half of the signals are positive. The IoU and loss gates use observed improvements at step 20; the initial gradient gate uses $q_i = -\nabla \ell_i(e_c^0)^\top v_c$, where $v_c$ is the first Adam step defined by FIT. We also test their conjunction. Ordinary checkpoint selection maximizes GUARD IoU over steps $0, 5, 20, 50, 100, 200$, breaking ties by lower GUARD loss, then earlier step. It sees the full trajectory, unlike the early gates. All trajectories are actually computed; hypothetical early stopping is not measured runtime savings.

**A joint-favorable dataset still contains harmful refinements.** Warm fitting helps CamVid in every seed, yet only 10 of the 25 classes supported in every seed improve consistently; six deteriorate consistently. The dataset mean is therefore not a rule for refining every class. On VOC and LoveDA, better FIT performance accompanies worse held-out transfer. Regularization reduces the damage, while ordinary checkpoint selection gives the largest CamVid gain among the fixed controls (Table M.2). Neither outperforms the local bank in every domain. A geometry-based selector must improve on these simpler alternatives.

**Does geometry add transferable decision value?** We compare four prespecified ridge predictors of warm-minus-local held-out IoU, leaving out the entire target dataset, including all its seeds. Headroom features describe the room for improvement: initial GUARD IoU/loss, FIT loss/count,

Table M.2: **Fixed refinement controls.** Full-ontology mIoU gains over the same local bank (pp; mean ± SD over three support seeds). Keeping local gives zero.

| Control | CamVid | VOC | LoveDA |
|---|---|---|---|
| Uniform local reference | $+0.18 \pm 1.37$ | $+0.17 \pm 0.25$ | $+1.47 \pm 0.17$ |
| Warm joint, step 200 | $+4.88 \pm 0.64$ | $-6.82 \pm 0.82$ | $-7.85 \pm 1.52$ |
| Warm joint + fixed $L_2$ | $+4.63 \pm 1.21$ | $-1.49 \pm 1.20$ | $-3.08 \pm 1.26$ |
| Step-20 GUARD IoU gate | $+3.62 \pm 2.38$ | $-2.92 \pm 1.75$ | $-0.65 \pm 0.43$ |
| Step-20 GUARD loss gate | $+3.56 \pm 1.44$ | $-4.11 \pm 1.44$ | $-4.12 \pm 3.34$ |
| Initial gradient gate | $+2.30 \pm 1.56$ | $-3.94 \pm 1.91$ | $-5.70 \pm 2.98$ |
| IoU AND gradient gate | $+2.47 \pm 1.10$ | $-1.98 \pm 2.04$ | $-0.56 \pm 0.57$ |
| GUARD checkpoint selection | $+6.86 \pm 0.47$ | $-1.56 \pm 0.79$ | $-2.23 \pm 3.22$ |

foreground area, and training prevalence. Early features measure step-20 GUARD IoU/loss changes, IoU-gain uncertainty, and win fraction. Geometry variants add FIT–GUARD gradient alignment, per-example endpoint dispersion/rank, and finite-secant coupling/branch stability; the initialization-only variant also includes first-step directional gain and endpoint-span energy. Ridge strength is fixed at 10, preprocessing uses training folds only, and weights balance datasets, then seeds, then classes. Refinement is chosen if and only if the predicted gain is positive; other domains' held-out outcomes train the predictor, never the target domain's outcomes.

Table M.3: **Cross-dataset switching.** Supported-class gains over local (pp; mean ± SD). Each target dataset, including all its seeds, is excluded from detector training.

| Detector features | CamVid | VOC | LoveDA |
|---|---|---|---|
| Headroom | $0.00 \pm 0.00$ | $-3.95 \pm 0.93$ | $-7.47 \pm 1.14$ |
| Headroom + geometry | $0.00 \pm 0.00$ | $-1.40 \pm 0.70$ | $-6.53 \pm 2.77$ |
| Headroom + early validation | $0.00 \pm 0.00$ | $-1.63 \pm 1.33$ | $-5.41 \pm 4.13$ |
| Headroom + early validation + geometry | $+0.57 \pm 0.98$ | $-0.42 \pm 0.47$ | $-3.67 \pm 2.62$ |

Table M.3 uses supported classes (CamVid has 25/26/27 across seeds), unlike the full evaluation label set of Table M.2 (29 CamVid evaluation classes). Geometry reduces prediction error in these folds, but improved regression is not enough: the combined model selects only 1/78 CamVid cases, missing 47/48 improvements larger than 0.1 IoU points. It still makes 4/6 harmful switches on VOC and 8/10 on LoveDA, using the same 0.1-point threshold for harm. The fixed IoU gate is less conservative on CamVid (20 switches, one harmful) but makes 22/24 harmful switches on VOC. The failure is therefore not confined to one regression fit.

**Headroom is not a mechanism; secants are not necessarily Hessians.** CamVid's initial FIT loss correlates positively with warm gain (Spearman $0.35, 0.45, 0.41$), but negatively with absolute warm performance ($-0.81, -0.79, -0.78$). Partial-rank adjustment for the held-out local score changes the loss–gain correlations to $-0.31, 0.06, -0.26$, consistent with a baseline/headroom contribution rather than causal identification. Held-out local scores enter this explanatory audit only, never detector inputs.

We also measure central full-gradient secants at the local embedding along each per-example endpoint minus their unweighted mean, at coefficients $0.001, 0.003, 0.01$, using float32 decoder probes of cached bf16 features. The registered detector uses coefficient $0.003$. All 156 probes are numerically valid, but only 3/78 CamVid, 5/60 VOC, and 0/18 LoveDA cases preserve all monitored branches across every probe image. Query selection, fallback, pixel-winner, and presence-argmax changes prevent interpreting branch-crossing secants as clean Hessian measurements. The observed domain split is compatible with the theory's conditional risk ordering, but does not identify curvature as its cause.

**Useful initialization, unresolved selection.** Local banks can enable better subsequent fits without revealing when to continue. The tested detectors do not add reliable cross-domain decision value, so conditional refinement remains an extension rather than another validated method. These are three inspected development domains, not 156 independent tasks; 2–4 potentially correlated GUARD images cannot certify safe adoption. Better detectors or tuned regularization remain possible. Appendix M.8 checks whether class influence, alignment, or numerical instability changes this conclusion.

### M.8 SENSITIVITY OF THE REFINEMENT FINDINGS

Could the refinement split be driven by one class, noisy gradients, or the harm threshold? These post-hoc checks reuse all 156 class–seed cases without retraining or detector tuning; Table M.4 retains every seed.

Table M.4: **Every support seed.** Harm exceeds 0.1 IoU points. Stable: aligned under every paired FIT/GUARD gradient deletion. Turn and paired displacement ratios are class medians.

| Dataset / seed | Cases | Harm / aligned | Harm / stable | Turn (°) | $L_2$/warm drift |
|---|---|---|---|---|---|
| CamVid / 7 | 25 | 10/21 | 6/10 | 69.7 | 0.152 |
| CamVid / 17 | 26 | 9/21 | 3/7 | 71.1 | 0.139 |
| CamVid / 42 | 27 | 7/21 | 3/10 | 70.9 | 0.140 |
| VOC / 7 | 20 | 15/17 | 7/8 | 76.2 | 0.059 |
| VOC / 17 | 20 | 18/19 | 8/9 | 74.2 | 0.063 |
| VOC / 42 | 20 | 14/17 | 11/14 | 73.7 | 0.081 |
| LoveDA / 7 | 6 | 6/6 | 5/5 | 75.1 | 0.120 |
| LoveDA / 17 | 6 | 3/3 | 2/2 | 73.7 | 0.171 |
| LoveDA / 42 | 6 | 5/6 | 2/2 | 74.7 | 0.153 |

**What changes under regularization.** For the float32 probes, the first-step direction is $v_0 = -0.05\, g_F/(|g_F| + 10^{-8})$, with elementwise division and FIT gradient $g_F$. Alignment is measured against the negative mean GUARD gradient; preconditioning changes its sign in 9/156 cases. This probe need not equal the first bf16-autocast fitting step. Despite initial alignment, 26/63 CamVid, 47/53 VOC, and 14/15 LoveDA cases lose more than 0.1 held-out IoU points after warm fitting. The fixed $L_2$ penalty reduces displacement in all 156 cases: median drift relative to the Original norm falls from 1.19 to 0.17, 1.70 to 0.12, and 1.62 to 0.24, respectively. This shows an effect of regularization, not that drift alone determines performance.

**Alignment does not suffice.** Across harm thresholds $0, 0.1, 0.5, 1, 2, 5$ IoU points, the conclusion persists. Even at five points, initially aligned refinement harms 4/21 CamVid cases in each seed, 10/17, 11/19, and 5/17 on VOC, and 5/6, 2/3, and 3/6 on LoveDA. Recomputing the first-step direction and GUARD gradient under all 4,906 paired single-image deletions leaves 27/31/9 cases aligned under every deletion; 12/26/9 still lose more than 0.1 pp. These are derivative-influence checks, not refitted leave-one-image-out banks.

From step-5 to step-200, displacement directions turn by a median of $69.7°$–$76.2°$ across dataset–seed cells (all 780 checkpoint measurements defined). Initial alignment therefore does not describe the direction of the final correction, or certify its held-out benefit.

**The domain ordering is not driven by one class.** Deleting each class jointly across seeds leaves mean warm gains in $[+4.72, +6.30]$ pp on CamVid, $[-7.28, -6.39]$ on VOC, and $[-9.31, -6.58]$ on LoveDA. Restricting CamVid to 25 classes present in all seeds gives $+5.69$ pp, with deletion range $[+4.97, +6.58]$. These supported-class influence ranges are not confidence intervals and differ from Table M.2's full evaluation label set.

**Stable branches do not certify a Hessian.** All 3,699 per-example secant probes and 312 adjacent-scale comparisons are defined at coefficients $0.001, 0.003, 0.01$. Fully branch-stable class counts remain 3/78, 5/60, and 0/18. Even within those CamVid/VOC cases, median relative secant changes $2\|C_h - C_{h'}\|/(\|C_h\| + \|C_{h'}\|)$ are $0.48/0.40$ and $0.78/0.67$ for adjacent scales. Dividing changes by the initial FIT-gradient norm instead gives $1.59/1.01$ and $2.70/2.05$. At coefficient 0.003, pixel-winner-only changes affect 367/609 CamVid, 164/480 VOC, and 108/144 LoveDA per-example probes; other monitored changes (query, fallback, or presence) affect 42/609, 2/480, and 20/144. These per-example counts are not class-level trials, and a changed pixel winner does not by itself mean active-query collapse. Neither scale check identifies the numerical cause or establishes a smoothness constant. The audits support the direction–transfer mismatch and reduced drift under regularization, not a transferable geometry-based switching rule.

## N LIMITATIONS AND FUTURE WORK

The results establish a useful fixed-bank intervention, subject to the following limitations.

**Predictor and vocabulary scope.** We do not compare text-encoder or full-model fine-tuning. A class-level bank cannot capture every image-specific correction, and transfer is tested at a fixed predictor and interface. Its embeddings cover a declared vocabulary, not arbitrary composable language. Other fixed-vocabulary prompt methods can cache outputs too; bypassing the encoder does not establish an inference-speed advantage.

**Reachability and baseline coverage.** Certificates cover the declared finite prompt population, not all strings or learned continuous prompts (Appendix B). The main LoRA matrix uses one fixed recipe, training-only checkpoint selection, and one training seed. Per-dataset finite-grid searches are diagnostics, not an exhaustive upper bound or a seed-variance study. External prompt-learning protocols also differ in support and selection (Appendices J and K).

**Relation to prior adaptation interfaces.** Class-embedding, test-time prototype, and shared SAM3-prompt adaptation have precedents (Yu et al., 2023; Silva-Rodríguez et al., 2024; Sui et al., 2025; Zhang et al., 2024; Zhou et al., 2026; Venkataramani & Sathish, 2026). COOPER's distinction is independent per-example fitting followed by class-level aggregation, supported by theory and local–joint comparisons, rather than a new embedding interface. COOPER-CLS fits anchored class embeddings jointly; it tests the interface, not the local estimator.

**Selection and uncertainty.** The training-only guard is an empirical acceptance and fallback rule, not a generalization certificate. The main SAM3 construction can reuse training examples across endpoint construction, gain estimation, and class-level acceptance; the nested refinement and low-shot classification protocols use separate training examples for fitting and selection, or cross-fitting, where stated. We do not describe every table entry as disjoint. Headline class-bootstrap intervals resample classes with banks fixed; they are not training-seed intervals. Repeated deployment timing and limited support-seed studies do not establish construction uncertainty for every benchmark bank.

**Scope of the estimator and theory.** Proposition 1 identifies curvature–endpoint coupling, not scalar curvature spread alone, as the quadratic discrepancy. The seven-dataset audit finds no significant CV–recovery relationship, so we do not claim a measured rule that selects a universal estimator. On cosine heads a linear probe can outperform COOPER-CLS (Figure G.1). The SAM3 objective is only piecewise smooth, while Proposition 1 concerns exact quadratic optima and Proposition 2 assumes a latent repair population. Neither establishes convergence of Adam, calibrates training gain to transferable repair, or certifies the guard. The refinement study measures training-selected continuation directly and does not establish a transferable geometry-based switching rule. Neither a low-dimensional projection nor a finite perturbation probe establishes a globally smooth semantic space.

**Future work.** Extending the completed matched local–joint controls beyond three domains (Table E.3) and testing a second dense predictor remain open. Mixture-style summaries, OOD detection and class-incremental deployment require separate objectives and evaluation.

# O TERMINOLOGY AND NOTATION

**Reading conventions.** *Original* denotes the unmodified text-derived class embedding or bank; it is the zero-shot anchor in classification. A *class bank* stores one class embedding per class: a token sequence for SAM3 and a vector for cosine classification. This term refers to the predictor's input, not a matrix orientation. An *incumbent* is a previously accepted bank entry, which need not equal Original.

*Local optimization* or *per-example fitting* optimizes an embedding for one training image–class pair. Its selected iterate is an *endpoint*, not necessarily its last iterate or an exact optimum. *Joint optimization* or *joint fitting* optimizes class embeddings using combined losses from multiple training examples; COOPER-CLS additionally updates competing class embeddings together. *Classwise fitting* in the classification controls fits each class only on its positive examples, with competitors fixed. *Token pooling* instead combines positions within an encoder output; it is not a fitting rule.

*Aggregation* or *merging* combines endpoints into a candidate embedding; its weighted mean is also called a *centroid*. A *uniform* mean is the unweighted mean. *Selection* chooses a candidate or blend; a *guard* may reject it and keep the reference. Both use training data, including training holdouts where specified, not the held-out evaluation split. A *fixed bank* is unchanged during deployment, not during its offline construction.

| Symbol | Meaning |
|---|---|
| $\mathcal{T}, \mathcal{V}, \mathcal{D}$ | Frozen text mapping, vision encoder, and decoder; $\mathcal{T}$ includes the projection into the predictor-facing space |
| $F_x$ | Frozen predictor conditioned on image $x$ |
| $e_c = \mathcal{T}(c)$ | Original embedding for class $c$; SAM3 tensors are flattened when vector algebra is used |
| $z_i^{(t)}, \hat{e}_i^\star$ | Per-example iterate and training-IoU-selected endpoint |
| $S_i, g_i$ | Training IoU in $[0, 1]$ and its improvement over Original |
| $\tau_{\text{act}}, A_c$ | Gain threshold $0.005$ ($0.5$ pp) and number of gains above it; not decoder-query activity |
| $w_i, \hat{\mu}_c$ | Merge weight and weighted endpoint mean before selection |
| $\alpha_c, \tilde{\mu}_c, b_c$ | Blend coefficient, blended candidate, and selected deployed embedding |
| $\hat{E}$ | Deployed SAM3 class bank containing the embeddings $b_c$ |
| $E, E^{(0)}$ | Classification bank and its original anchor, with each class embedding unit-normalized |
| $\tau, \beta$ | Classification logit scale (inverse temperature) and angular-prior coefficient |
| $\mathcal{P}_c, \ell_c$ | Sampled prompt embeddings and certified lower distance bound to their convex hull |
| $L, \mathcal{L}$ | Task loss; objectives are defined in Section 4 |
| $e^\star$ | Measurable choice of exact per-example optimum; distinct from $\hat{e}_i^\star$ |
| $\mathcal{D}_c, \mu_c^{\text{pop}}, \mu_c^{\text{avg}}$ | Class distribution, population joint optimum, and expected per-example optimum |
| $\bar{H}_N$ | Empirical mean Hessian in the exact-quadratic model |
| $Z, B, \pi, \mu_0, \mu_1$ | Endpoint, latent repair indicator, repair fraction, and component means in Proposition 2 |
| $\theta, \rho$ | Angular displacement and norm ratio relative to Original |
| $\Delta$ | Locally defined difference (pp in gain tables) |