CONTROLLING FUNCTIONAL DRIFT IN REPLAY-FREE CONTINUAL OPEN-VOCABULARY SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary segmentation models are strongly shaped by their training data, category taxonomy, and annotation granularity. Continual adaptation to new domains can therefore distort the inherited open-vocabulary prediction function and accumulate functional interference across stages. To address this problem, we propose **Open-Prior Functional Residual Surgery** (OP-FRS), which freezes the source model and historical modules while introducing lightweight structural, readout, and text-conditioned semantic sidecars. OP-FRS constructs protected outputs under joint competition among historical, requested, and no-object classes, and uses an input-conditioned composed Jacobian to estimate how candidate residuals affect the full prediction pathway. A proximal filter suppresses functionally sensitive directions, while magnitude recovery restores the cumulative residual scale after directional correction. At their acquisition checkpoints, OP-FRS improves newly encountered target domains by up to 6.96 mIoU over the frozen source model in a three-target continual stream. At the final checkpoint, OP-FRS retains substantially stronger source-domain, previous-target, and evaluation-vocabulary performance than existing replay-free adaptation methods. These results show that direction-selective residual control in function space provides a favorable replay-free balance between retention and adaptation.

1 INTRODUCTION

Open-vocabulary segmentation (OVS) predicts dense masks for vocabularies that may differ from those used during segmentation training. Although modern OVS models inherit broad visual–language knowledge, their predictions remain shaped by the source data, category taxonomy, annotation granularity, and domain distribution. Their open-vocabulary capability is therefore embodied in an end-to-end prediction function shaped by the source data and training protocol.

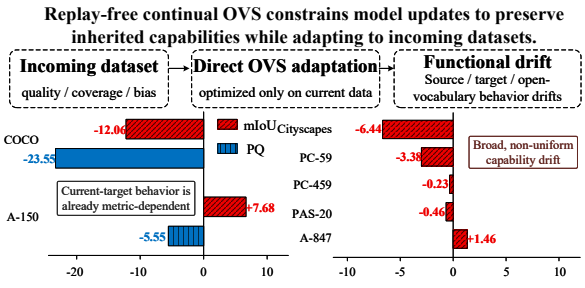

Figure 1: Dataset-induced functional drift in OVS. Direct target-specific adaptation alters current-dataset behavior in a metric-dependent way, potentially pulling the model away from its source and non-optimized open-vocabulary behavior.

We study *replay-free continual OVS*, where an binherited model is sequentially adapted to datasets with different domains, taxonomies, and annotation conventions, without retaining source or previous-target images, features, logits, or exemplars. This setting differs from closed-set continual segmentation because the capability to be retained extends beyond predictions over a fixed historical label set. It also includes proposal generation, mask-conditioned visual aggregation, region–text matching, no-object competition, and dense mask–score composition. We refer to this inherited end-to-end behavior as the *source-induced open prior*.

Figure 1 illustrates this difficulty. In a controlled source-to-target adaptation experiment, optimization on the current dataset produces metric-dependent target changes: mIoU improves while PQ decreases. The same update also causes broad and non-uniform variation on the source domain and on vocabu-

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.

laries excluded from training. This observation suggests that adaptation does not simply append a new capability. Instead, it reshapes the inherited prediction function beyond the current target. Under sequential updates, these changes may accumulate and interfere with source, previous-target, and vocabulary-flexible behavior.

Freezing the source model and historical modules prevents direct parameter overwriting, but not functional interference. Parameter-efficient residuals still modify intermediate masks, region representations, and semantic scores, whose effects propagate through the downstream prediction path. Furthermore, residual norm alone is an incomplete measure of interference: a small residual may align with a sensitive functional direction, whereas a larger residual may have limited effect on the protected behavior. Continual OVS therefore requires controlling the composed effect of each candidate residual, rather than only its parameters, or local activations.

To this end, we propose *Open-Prior Functional Residual Surgery* (OP-FRS). OP-FRS introduces lightweight sidecars at the structural, readout, and text-conditioned semantic stages while freezing the source path and historical sidecars. It constructs a protected output under joint competition among historical, requested, and no-object classes. A semantic relation graph represents the relative structure of historical predictions, while an absolute confidence anchor captures their overall probability level.

For each stage, an input-conditioned composed Jacobian estimates how its accumulated residual affects the protected output through the complete downstream pathway. A proximal filter suppresses functionally sensitive components instead of uniformly shrinking the residual, and magnitude recovery restores the cumulative residual scale after directional correction. This separates interference control from residual attenuation and supports replay-free adaptation without stored historical predictions.

We evaluate OP-FRS on a three-target continual stream and on external-domain and alternate-taxonomy vocabularies. Experiments with FC-CLIP and MAFT, together with controlled ablations and functional diagnostics, evaluate its retention, adaptation, and functional-interference control.

Our main contributions are:

- We characterize accumulated functional interference in replay-free continual OVS, showing that target-specific adaptation can alter the current target, the source domain, and evaluation vocabularies not used for adaptation.
- We formulate continual OVS as protecting a source-induced open prior and introduce a joint-vocabulary operator capturing historical relations and absolute historical confidence.
- We propose composed functional residual surgery, combining input-conditioned downstream sensitivity, direction-selective filtering, and magnitude recovery.

2 Related Work

Open-vocabulary segmentation. Open-vocabulary segmentation (OVS) extends dense prediction beyond a fixed label space by aligning visual regions with text. Representative methods include MaskCLIP Zhou et al. (2022), ZegFormer Ding et al. (2022), OVSeg Liang et al. (2023), SAN Xu et al. (2023b), ODISE Xu et al. (2023a), FC-CLIP Yu et al. (2023), and recent OVS variants Li et al. (2025); Chen et al. (2026); Niu et al. (2025). These works mainly improve static OVS models; we instead study how an inherited OVS model can undergo sequential dataset adaptation under replay-free constraints while limiting interference with its inherited open-vocabulary behavior.

Continual and parameter-efficient segmentation. Continual segmentation mitigates forgetting via background modeling, distillation, replay, or representation regularization Cermelli et al. (2020); Douillard et al. (2021); Michieli & Zanuttigh (2021). Classical controls such as EWC and LwF Kirkpatrick et al. (2017); Li & Hoiem (2016) have recently been extended to universal segmentation, instance replay, adversarial retention, and decoupled designs Lin et al. (2025); Yin et al. (2025); Yu et al. (2025); Guo et al. (2026). ECLIPSE adapts continual panoptic segmentation with visual prompts and logit manipulation Kim et al. (2024a), whereas ConOVS combines expert decoders via an input-distribution estimator Hwang et al. (2025). Adapters, visual prompt tuning, and LoRA enable parameter-efficient adaptation Houlsby et al. (2019); Jia et al. (2022); Hu et al. (2022), including OVS-specific variants Peng et al. (2025); Xu et al. (2025); Kim et al. (2024b); Qorbani et al. (2025);

Mazzucco et al. (2025). However, parameter-efficient residual adaptation can still substantially alter the composed OVS prediction function.

Function-space preservation and implicit optimization. Gradient projection and functional regularization established the idea of controlling interference in parameter or output space Saha et al. (2021); Titsias et al. (2020). More recent work has developed scalable function-space representations for sequential learning, allowing prior predictive information to be retained and new observations to be incorporated with limited access to past data Scannell et al. (2024). In parallel, differentiable optimization layers increasingly embed constrained or linear solvers into end-to-end learning systems, including augmented-Lagrangian differentiation through quadratic programs Bambade et al. (2024) and efficient differentiable sparse linear solvers Wang et al. (2025). OP-FRS differs from these approaches by filtering input-conditioned residuals according to their composed downstream effect on a joint-vocabulary protected output. Its proximal system is solved matrix-free without replay or stored historical predictions.

3 METHODOLOGY

We propose **Open-Prior Functional Residual Surgery** (OP-FRS) for replay-free continual open-vocabulary segmentation. A source-trained OVS model is retained as a frozen source-induced open prior. Figure 2 summarizes OP-FRS at each continual step.

3.1 PROBLEM AND FROZEN SOURCE PATH

Let $\mathcal{D}_0$ be the source dataset and $\mathcal{D}_{1:T}$ the target stream. At step t, only $\mathcal{D}_t$, inherited parameters, and category names are available; no source or previous-target images, visual features, logits, or exemplars are retained. The source path and sidecars $1{:}t-1$ are frozen, and only

$$\phi_t = \{\phi_t^{\mathrm{str}}, \phi_t^{\mathrm{read}}, \phi_t^{\mathrm{sem}}\} \tag{1}$$

is optimized. Given image x, the frozen source model computes

$$F_s = \mathrm{Enc}_s(x), \quad (Q_s, M_s) = \mathrm{Struct}_s(F_s), \quad z_s = \mathrm{Readout}_s(F_s, M_s). \tag{2}$$

For vocabulary $\mathcal{C}$, source region scores are obtained from normalized region–text similarity together with the source no-object logit. Stable normalization and the exact augmented score construction are provided in the supplement. Freezing this path prevents parameter overwriting, but newly added residuals can still change the composed prediction.

3.2 VOCABULARY-AGNOSTIC FUNCTIONAL SIDECARS

Structural sidecar i predicts a mask-logit residual $r_{i,\mathrm{str}}^{M}$ from (F_s, Q_s, M_s). After structural surgery, readout sidecar i predicts a region-embedding residual $r_{i,\mathrm{read}}^{z}$ from $(F_s, M_t^*, z_{t,0}^*)$. Their cumulative raw residuals are

$$r_{t,\mathrm{str}}^{\Sigma} = \sum_{i=1}^{t} r_{i,\mathrm{str}}^{M}, \qquad r_{t,\mathrm{read}}^{\Sigma} = \sum_{i=1}^{t} r_{i,\mathrm{read}}^{z}. \tag{3}$$

A fixed-class score head does not naturally support evaluation vocabularies that change across stages. We therefore use a text-conditioned, category-permutation-equivariant semantic sidecar with row-wise linear projections:

$$r_{t,\mathrm{sem}}^{\Sigma}(z, E) = \frac{1}{\sqrt{d_a}} \sum_{i=1}^{t} h_i^U(z)\, h_i^V(E)^\top, \tag{4}$$

Here, $h_i^U : \mathbb{R}^{d_z} \to \mathbb{R}^{d_a}$ and $h_i^V : \mathbb{R}^{d_e} \to \mathbb{R}^{d_a}$ are row-wise linear projections associated with semantic sidecar i. Sharing h_i^V across category rows makes the semantic residual equivariant to category permutation and defined for vocabularies supplied at evaluation time. Although previous sidecar parameters remain frozen, their outputs are included in the cumulative residual and re-filtered for each input and requested vocabulary. The sidecars are implemented with lightweight low-rank and bottleneck modules; detailed placements and parameter counts are provided in the supplement.

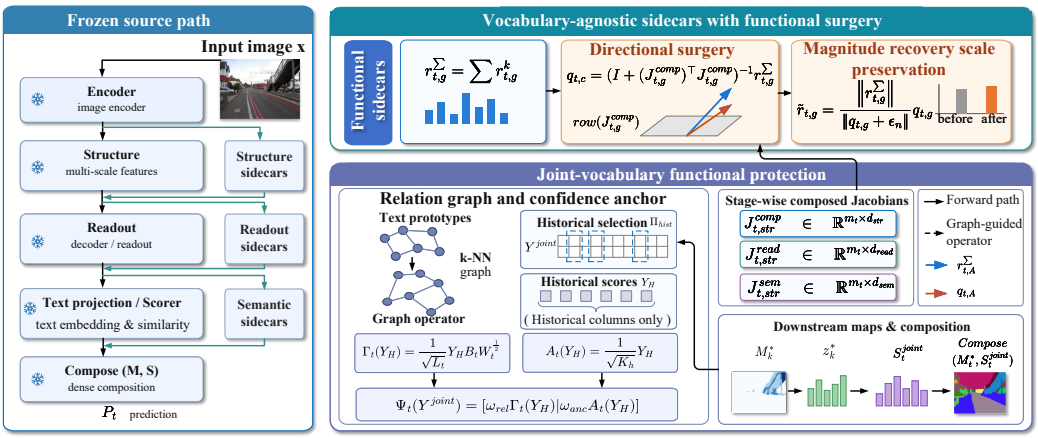

Figure 2: Overview of OP-FRS at continual step t. Filtered residuals are injected into the frozen source path and previous sidecars, followed by joint softmax, historical-column selection, and composed-Jacobian-guided surgery.

3.3 Joint-vocabulary Protocol and Graph Timing

Let

$$\mathcal{C}_{<t} = \cup_{j=0}^{t-1} \mathcal{C}_j \tag{5}$$

be the strictly historical vocabulary and let $\mathcal{C}^{\text{eval}}$ be the vocabulary requested for prediction. We form the canonicalized, de-duplicated joint vocabulary

$$\mathcal{C}_t^{\text{joint}} = \text{unique}(\mathcal{C}_{<t} \cup \mathcal{C}^{\text{eval}}), \qquad E_t^{\text{joint}} = E(\mathcal{C}_t^{\text{joint}}). \tag{6}$$

Selection matrices Π_{hist} and Π_{eval} extract the historical and requested evaluation columns, respectively. Given the dense joint prediction Y^{joint}, the output for the requested vocabulary is

$$Y^{\text{eval}} = Y^{\text{joint}} \Pi_{\text{eval}}^{\top}, \tag{7}$$

whereas Π_{hist} is used only by the protected functional operator. During training at step t, we set $\mathcal{C}^{\text{eval}} = \mathcal{C}_t$; during evaluation, it is replaced by the requested benchmark vocabulary. PAS-20 data and the alternate A-847 and PC-59 vocabularies and annotations are used only for evaluation and are excluded from training, hyperparameter selection, and model selection. Both selectors act after a single joint softmax, so current or evaluation-only classes and no-object compete with historical classes under the same normalization used by the final predictor.

3.4 Historical Relation Graph and Confidence Anchor

Let $K_h = |\mathcal{C}_{<t}|$ and let $E_{<t} = \{e_1, \ldots, e_{K_h}\}$ be the normalized historical text prototypes. Each category is connected to its k nearest semantic neighbors and the graph is symmetrized. For the L_t undirected edges $\mathcal{E}_t = \{(i_\ell, j_\ell)\}_{\ell=1}^{L_t}$, define

$$\widehat{w}_\ell = \exp(e_{i_\ell}^{\top} e_{j_\ell} / \tau_v), \qquad w_\ell = \widehat{w}_\ell \Big/ \frac{1}{L_t} \sum_{m=1}^{L_t} \widehat{w}_m. \tag{8}$$

Let $B_t \in \mathbb{R}^{K_h \times L_t}$ be the oriented incidence matrix and $W_t = \text{diag}(w_1, \ldots, w_{L_t})$. For historical dense scores Y_H,

$$\Gamma_t(Y_H) = \frac{1}{\sqrt{L_t}} Y_H B_t W_t^{1/2} \tag{9}$$

protects local old–old relations. A graph difference alone has a common-shift null space, so we add an absolute historical-confidence anchor

$$\mathcal{A}_t(Y_H) = \frac{1}{\sqrt{K_h}} Y_H. \tag{10}$$

For joint dense scores Y^{joint} and $Y_H = Y^{\text{joint}}\Pi_{\text{hist}}^{\top}$, the protected functional output is

$$\Psi_t(Y^{\text{joint}}) = \Big[\omega_{\text{rel}}\Gamma_t(Y_H) \;\Big|\; \omega_{\text{anc}}\mathcal{A}_t(Y_H)\Big], \tag{11}$$

where $[\cdot \mid \cdot]$ concatenates feature columns. The graph term preserves historical inter-class geometry; the anchor detects a common loss of historical probability caused by newly introduced classes or no-object. Here, confidence denotes the complete historical dense-score vector rather than a scalar maximum probability. Both blocks are normalized by their sizes, and their fixed weights are selected once on the validation protocol described in the supplement.

3.5 Joint Composed Functional Interference

Let $K_u = |\mathcal{C}_t^{\text{joint}}|$. For mask logits $M \in \mathbb{R}^{N\times HW}$ and augmented joint scores $S^+ = [S, s^{\varnothing}] \in \mathbb{R}^{N\times(K_u+1)}$, define

$$\pi_{\text{sem}}(S^+_{n,:}) = \big[\text{Softmax}(S^+_{n,:})\big]_{1:K_u}. \tag{12}$$

The no-object logit therefore participates in normalization and is removed only after the joint competition. Dense joint scores are

$$[\text{Compose}(M, S^+)]_{p,c} = \sum_{n=1}^{N} \sigma(M_{n,p})\, \pi_{\text{sem}}(S^+_{n,:})_c. \tag{13}$$

Define the anchored historical readout

$$\mathcal{H}_t(M, S^+) = \Psi_t\big(\text{Compose}(M, S^+)\big). \tag{14}$$

This composite map first converts region masks and augmented joint logits into dense joint scores, then selects the historical columns inside Ψ_t and evaluates their relation-graph and confidence-anchor outputs. Let $\overline{\mathcal{R}}_t$ and $\overline{\mathcal{S}}_t^+$ denote detached downstream readout and augmented semantic maps containing all active sidecars. Their parameters are constants while input derivatives are retained. The composed continuations are

$$\begin{gathered}\Phi^{\text{comp}}_{t,\text{str}}(M) = \mathcal{H}_t\Big(M, \overline{\mathcal{S}}_t^+(\overline{\mathcal{R}}_t(M), E_t^{\text{joint}})\Big), \Phi^{\text{comp}}_{t,\text{read}}(z) = \mathcal{H}_t\Big(M_t^*, \overline{\mathcal{S}}_t^+(z, E_t^{\text{joint}})\Big), \\ \Phi^{\text{comp}}_{t,\text{sem}}(S) = \mathcal{H}_t\big(M_t^*, [S, s_s^{\varnothing}(z_t^*)]\big).\end{gathered} \tag{15}$$

The stages are applied sequentially as structural, readout, and semantic, with base states

$$h^0_{t,\text{str}} = M_s, h^0_{t,\text{read}} = z^*_{t,0}, h^0_{t,\text{sem}} = S^{\text{joint}}_{t,0}. \tag{16}$$

The input-conditioned composed Jacobian is

$$J^{\text{comp}}_{t,g}(x) = \left.\frac{\partial \,\text{vec}(\Phi^{\text{comp}}_{t,g}(h; x))}{\partial \,\text{vec}(h)}\right|_{h=h^0_{t,g}}. \tag{17}$$

Here, $g \in \{\text{str}, \text{read}, \text{sem}\}$. The Jacobian is recomputed for every training and evaluation image and measures the composed first-order effect of a stage residual on historical relations and confidence under the actual joint-vocabulary competition. Unlike output distillation, OP-FRS does not match stored historical logits or teacher targets.

For a nonzero residual v, define its composed interference energy as

$$\mathcal{E}_{t,g}(v) = \big\|J^{\text{comp}}_{t,g} v\big\|_2^2. \tag{18}$$

3.6 Magnitude-recovered Directional Surgery

For cumulative raw residual $r^{\Sigma}_{t,g}$, OP-FRS solves

$$q_{t,g} = \arg\min_q \left[\frac{1}{2}\|q - r^{\Sigma}_{t,g}\|_2^2 + \frac{1}{2}\|J^{\text{comp}}_{t,g} q\|_2^2\right], \tag{19}$$

whose formal closed-form solution is

$$q_{t,g} = \big(I + (J^{\text{comp}}_{t,g})^{\top} J^{\text{comp}}_{t,g}\big)^{-1} r^{\Sigma}_{t,g}. \tag{20}$$

We recover the cumulative residual scale:

$$\widetilde{r}_{t,g} = \frac{\|r^{\Sigma}_{t,g}\|_2}{\|q_{t,g}\|_2 + \epsilon_n} q_{t,g}. \tag{21}$$

The linear system is approximated with matrix-free conjugate gradient (CG) using Jacobian–vector products (JVPs) and vector–Jacobian products (VJPs). For diagnostics, we compare the recovered residual with the raw cumulative residual through

$$\rho^{\mathrm{int}}_{t,g} = \frac{\mathcal{E}_{t,g}(\widetilde{r}_{t,g})}{\mathcal{E}_{t,g}(r^{\Sigma}_{t,g}) + \epsilon_d}, \qquad \rho^{\mathrm{norm}}_{t,g} = \frac{\|q_{t,g}\|_2}{\|r^{\Sigma}_{t,g}\|_2 + \epsilon_n}. \tag{22}$$

Exact norm computation, implicit differentiation, finite-CG approximation, and recovery-factor diagnostics are detailed in the supplement.

Training objective. For each current batch $(x, y) \sim \mathcal{D}_t$, only the new sidecar group ϕ_t is optimized using the unchanged Hungarian-matched segmentation objective

$$\mathcal{L}_t = \lambda_{\mathrm{cls}}\mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{bce}}\mathcal{L}_{\mathrm{bce}} + \lambda_{\mathrm{dice}}\mathcal{L}_{\mathrm{dice}}, \tag{23}$$

computed from the surgically corrected region scores and masks. The linear-system operator is detached when constructed, while gradients pass through the finite CG solution and magnitude recovery to ϕ_t. Full OP-FRS uses neither replay nor an auxiliary distillation loss.

Table 1: Three-target continual evaluation on COCO→A-150→Cityscapes→PC-459.

Method	COCO		A-150		Cityscapes		PC-459	PAS-20	A-847	PC-59
	mIoU ↑	PQ ↑	mIoU ↑	PQ ↑	mIoU ↑	PQ ↑	mIoU ↑	mIoU ↑	mIoU ↑	mIoU ↑
Mask-adapter	48.27	29.45	21.63	15.74	45.83	33.29	14.53	91.85	10.95	45.12
LoRA-only	38.75	34.28	11.42	12.37	36.89	27.13	9.27	62.37	8.46	25.64
Adapter-only	42.13	38.45	13.76	14.92	39.58	29.86	10.04	66.84	9.81	28.93
EWC	30.14	24.56	14.28	12.39	37.95	30.71	6.45	64.72	6.31	27.85
LwF	35.89	30.27	19.62	16.48	41.73	32.95	7.93	71.38	7.52	34.56
ECLIPSE	44.62	32.13	8.17	14.81	40.52	29.27	10.61	69.51	10.28	30.84
ConOVS	52.38	41.02	19.21	17.18	37.81	32.53	12.58	81.21	13.23	40.12
OP-FRS	**62.94**	**53.63**	**33.52**	**25.21**	**51.82**	**37.39**	**19.61**	**96.08**	**14.73**	**60.08**

4 Experiments

4.1 Experimental Setup

The source model is trained on COCO Caesar et al. (2018). The main continual protocol follows COCO→A-150 Zhou et al. (2017)→Cityscapes Cordts et al. (2016)→PC-459 Mottaghi et al. (2014), with three optimized targets from distinct image corpora. PAS-20 Everingham et al. (2010) serves only as an external evaluation probe, excluded from training, hyperparameter selection, and model selection. A-847 and PC-59 provide alternate-taxonomy evaluation on the A-150 and PC-459 corpora, complementing the external-domain PAS-20 probe. Table 1 uses FC-CLIP/ConvNeXt-L, while Table 2 repeats the same stream under MAFT Jiao et al. (2023). Within each framework, all methods share the same source initialization, data pipeline, prompt templates, input resolution, augmentations, optimizer family, training schedule, and evaluation protocol; method-specific losses and regularizers follow their implementations. Baselines include LoRA-only Hu et al. (2022), adapter-only Houlsby et al. (2019), EWC Kirkpatrick et al. (2017), LwF Li & Hoiem (2016), ECLIPSE Kim et al. (2024a), and ConOVS Hwang et al. (2025).

Diagnostic metrics. The new-class stealing rate is the percentage of labeled historical pixels whose prediction under the evaluated checkpoint falls in a newly introduced class:

$$\mathrm{NCS} = \frac{|\{p : y_p \in \mathcal{C}_{<t},\ \widehat{y}^{\mathrm{after}}_p \in \mathcal{C}_t \setminus \mathcal{C}_{<t}\}|}{|\{p : y_p \in \mathcal{C}_{<t}\}|}. \tag{24}$$

It is measured only on labeled source or historical validation images and is never used for training.

For functional diagnostics, let $Y_{H,g}^{\text{base}}$ and $Y_{H,g}^{\text{surg}}$ denote the historical dense scores before residual injection and after injecting the surgically corrected residual at stage g, respectively. We define the relative historical-confidence drop as

$$\delta_{t,g}^{\text{conf}} = \frac{\left[\|\mathcal{A}_t(Y_{H,g}^{\text{base}})\|_F - \|\mathcal{A}_t(Y_{H,g}^{\text{surg}})\|_F\right]_+}{\|\mathcal{A}_t(Y_{H,g}^{\text{base}})\|_F + \epsilon_a}, \tag{25}$$

where $[z]_+ = \max(z, 0)$.

4.2 Main Three-Target Continual Results

Table 1 reports performance at the final checkpoint of the complete three-target stream. OP-FRS achieves a favorable stability–plasticity balance, retaining 62.94 mIoU and 53.63 PQ on COCO while obtaining the strongest results among continual methods on A-150 and PC-459. It also achieves the strongest result on the external PAS-20 probe and on both alternate-taxonomy probes, A-847 and PC-59. OP-FRS provides the strongest overall final-checkpoint profile among the evaluated continual methods.

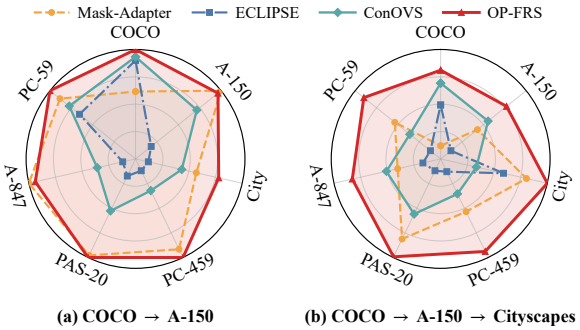

Figure 3: Intermediate capability profiles after adaptation to A-150 and Cityscapes. Axis is normalized jointly over the displayed methods and both checkpoints.

Evaluation-vocabulary transfer. The probe results provide complementary evidence beyond the optimized targets. OP-FRS matches or exceeds the frozen source on all three evaluation-vocabulary probes, with the clearest gain observed under the A-847 alternate taxonomy. Since PAS-20 uses an external image corpus whereas A-847 and PC-59 alter the taxonomy on corpora shared with A-150 and PC-459, this pattern is consistent with preserved vocabulary-flexible behavior under external-domain and alternate-taxonomy evaluation across distinct evaluation conditions.

Intermediate capability profiles. Figure 3 visualizes the capability profiles at the first two continual checkpoints. After the A-150 update, OP-FRS combines effective target adaptation with broad preservation across the source domain and evaluation vocabularies. This advantage remains after the subsequent Cityscapes update, where OP-FRS expands its capability on the incoming domain without concentrating performance on the current dataset alone. The two profiles show that the preservation–adaptation balance is maintained throughout the stream under changing dataset conditions rather than emerging only at the final checkpoint.

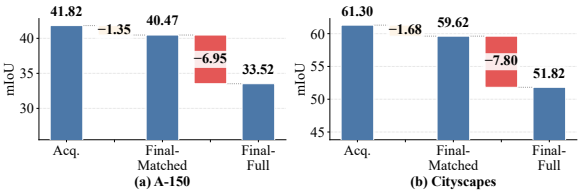

Figure 4: Previous-target performance decomposed into temporal change and final-vocabulary competition.

Vocabulary-expansion decomposition. Figure 4 separates the effects of subsequent continual updates from those of final-vocabulary expansion. For both previous targets, the final checkpoint remains close to the corresponding acquisition performance when evaluated with the matched vocabulary. The larger change appears only after introducing the complete final vocabulary, indicating that expanded class competition accounts for most of the apparent previous-target difference. Because the two final evaluations use the same model and accumulated sidecars, this decomposition distinguishes vocabulary-induced competition from direct capability overwriting.

4.3 Cross-Framework Validation on MAFT

Table 2 shows that OP-FRS also extends effectively to the MAFT framework. Compared with the strongest baseline, MAFT+ConOVS, OP-FRS improves COCO retention, previous-task average, PC-459 adaptation, and PAS-20 transfer by 5.09, 2.24, 1.09, and 4.92 mIoU, respectively. These results support the cross-framework applicability of functional residual surgery across FC-CLIP and MAFT. Architecture-specific insertion points, parameter budgets, and baseline adaptations under MAFT are detailed in the supplement.

Table 2: Cross-framework validation with MAFT on the COCO→A-150→Cityscapes→PC-459 stream. Results are measured at the final checkpoint; Prev. Avg. averages A-150 and Cityscapes.

Method	COCO	Prev. Avg.	PC-459	PAS-20
Frozen Source	40.12	25.23	19.35	60.17
MAFT + ECLIPSE	28.50	22.67	14.80	75.32
MAFT + ConOVS	31.48	24.89	18.57	80.14
MAFT + OP-FRS	**36.57**	**27.13**	**19.66**	**85.06**

4.4 Ablation Studies

Accumulated residual control. As shown in Table 3, raw sidecar accumulation causes substantial functional interference. The comparison supports the benefit of controlling residual direction rather than residual magnitude alone. Magnitude recovery further improves the overall preservation–adaptation balance and reduces new-class stealing.

Table 3: Final-checkpoint ablation of directional filtering and magnitude recovery. "Prev. Avg." averages A-150 and Cityscapes mIoU;

Variant	Inject	COCO	Prev. Avg.	PC-459	NCS ↓
Frozen		65.82	44.96	18.63	–
+ Raw	r^{Σ}	47.23	29.36	17.67	12.84
+ Filter	q	59.34	40.05	19.84	5.93
+ Recov.	$\widetilde{r}$	**62.94**	**42.67**	**19.61**	**3.12**
Norm-matched	αr^{Σ}	52.41	33.78	18.94	9.56

Table 4: Core operator ablation on COCO→A-150: progressive variants add joint competition and the confidence anchor; controls remove the graph or use local sensitivity.

Variant	A-150 ↑	COCO ↑	PC-59 ↑	NCS↓
Frozen	35.57	65.82	59.89	–
+ Hist. graph	38.76	62.98	58.62	7.83
+ Joint comp.	41.13	64.31	60.05	5.46
+ Conf. anchor	**41.82**	**65.05**	**60.38**	**3.79**
Anchor only	41.45	64.72	60.23	4.67
Local sensitivity	41.64	64.88	60.31	4.12

Core operator ablation. Table 4 examines three key design choices in the protected operator. A historical-only softmax ignores competition from newly introduced classes, leading to excessive reassignment of historical pixels. Joint competition corrects this mismatch between the protected operator and the final predictor, while the confidence anchor captures global probability loss that cannot be observed from pairwise graph relations alone. The weaker controls support both the complementarity of the relation and anchor terms and the benefit of composed rather than local sensitivity.

4.5 Functional Diagnostics and Efficiency

Stage-wise interference selectivity. Figure 5 shows distinct stage-dependent interference profiles. Structural residuals produce the largest composed interference energy because their effects propagate through all downstream modules, whereas semantic residuals are more localized. Surgery reduces interference energy at every stage, with a widening separation between interference suppression and norm preservation toward the semantic stage.

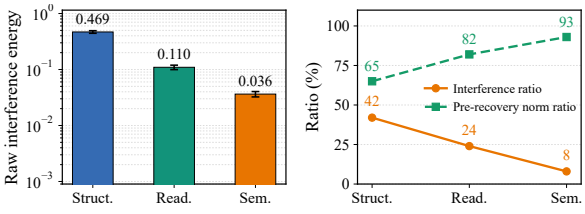

Figure 5: Stage-wise selectivity of OP-FRS: composed interference energy (left), and interference and pre-recovery norm ratios (right).

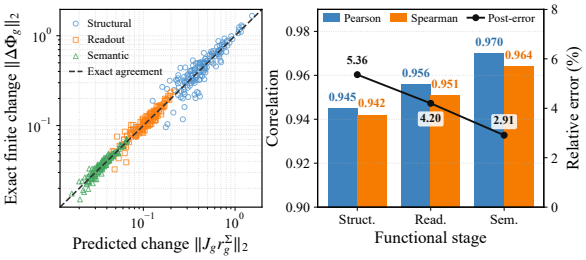

Figure 6: Composed-operator faithfulness: predicted vs. exact finite protected-output changes (left); stage-wise correlations and post-surgery relative errors (right).

Composed-operator faithfulness. Figure 6 evaluates the faithfulness of the composed operator at both sample and stage levels. The predicted first-order changes closely follow the exact finite protected-output changes across structural, readout, and semantic residuals. The consistently high correlations and low post-surgery errors indicate that the composed Jacobian provides a reliable relative ordering of residual directions by functional interference. The closer agreement at later stages is consistent with their shorter downstream prediction paths.

CG budget analysis. Figure 7 exhibits a typical diminishing-return pattern. Additional iterations improve the linear solve and produce small downstream gains across source, target, and evaluation-vocabulary metrics throughout the continual stream, but performance becomes nearly stable once the main interference-sensitive directions have been resolved. Beyond this point, numerical accuracy continues to improve without a comparable segmentation benefit. We therefore use five iterations as a practical balance between solver fidelity and inference cost.

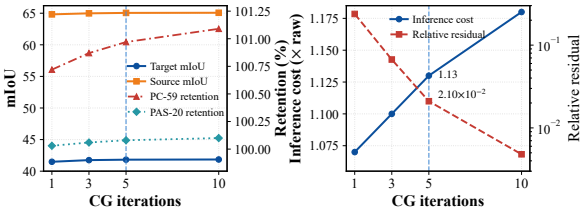

Figure 7: Effect of the CG solver budget. The left plot reports segmentation performance and retention, while the right plot shows the linear-system residual and inference cost.

Computational efficiency. Table 5 shows that OP-FRS remains lightweight despite its input-conditioned surgery. It uses only 1.2M trainable parameters, fewer than Mask-Adapter and ConOVS, while increasing the frozen-source FLOPs from 12.8G to 13.5G. The additional JVP/VJP and CG operations raise inference latency to 54.0 ms per image, but peak memory remains close to the frozen model at 8.9 GB. Overall, OP-FRS achieves a favorable balance between continual performance and computational overhead.

Method	Params	FLOPs	Infer ms/image	Peak GB
Frozen source	0	12.8G	45.2	8.6
Mask-Adapter	16.3M	18.5G	55.8	11.2
ConOVS	2.8M	15.2G	49.5	9.8
OP-FRS	1.2M	13.5G	54.0	8.9

Table 5: Absolute efficiency comparison . We report trainable parameters, FLOPs, inference latency per image, and peak GPU memory.

5 Conclusion

We presented OP-FRS, a replay-free continual OVS approach that preserves an inherited open prior by controlling sidecar residuals in function space. Rather than relying on parameter freezing or uniform residual attenuation, OP-FRS suppresses directions that strongly interfere with the composed joint-vocabulary prediction while retaining adaptation strength. Across the FC-CLIP stream and the MAFT validation, OP-FRS consistently strengthens source retention, previous-target performance, and evaluation-vocabulary transfer while remaining competitive on the final target. Controlled ablations further show that these gains arise from joint-vocabulary monitoring, composed sensitivity, and direction-selective filtering. These results support functional residual surgery as a practical design principle that extends across the two evaluated OVS frameworks.

REFERENCES

Antoine Bambade, Fabian Schramm, Adrien Taylor, and Justin Carpentier. Leveraging augmented-lagrangian techniques for differentiating over infeasible quadratic programs in machine learning. In *International Conference on Learning Representations*, volume 2024, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/3d7dc6f86ad82d94a540a521de17d44b-Paper-Conference.pdf.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1209–1218, 2018.

Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulo, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9233–9242, 2020.

Lin Chen, Qi Yang, Kun Ding, Zhihao Li, Gang Shen, Fei Li, Qiyuan Cao, and Shiming Xiang. Efficient redundancy reduction for open-vocabulary semantic segmentation. *Neurocomputing*, 665: 132229, 2026. ISSN 0925-2312.

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

Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11583–11592, 2022.

Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. Plop: Learning without forgetting for continual semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4040–4050, 2021.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. doi: 10.1007/s11263-009-0275-4.

Yifu Guo, Yuquan Lu, Wentao Zhang, Zishan Xu, Dexia Chen, Siyu Zhang, Yizhe Zhang, and Ruixuan Wang. Decoupling continual semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4476–4484, 2026. doi: 10.1609/aaai.v40i6.42446.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

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

Dongjun Hwang, Yejin Kim, Minyoung Lee, Seong Joon Oh, and Junsuk Choe. Ovs meets continual learning: Towards sustainable open-vocabulary segmentation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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*, pp. 709–727. Springer, 2022.

Siyu Jiao, Yunchao Wei, Yaowei Wang, Yao Zhao, and Humphrey Shi. Learning mask-aware clip representations for zero-shot segmentation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 35631–35653, 2023.

Beomyoung Kim, Joonsang Yu, and Sung Ju Hwang. ECLIPSE: Efficient continual learning in panoptic segmentation with visual prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3346–3356, 2024a.

Sungyeon Kim, Boseung Jeong, Donghyun Kim, and Suha Kwak. Efficient and versatile robust fine-tuning of zero-shot models. In *European Conference on Computer Vision*, pp. 440–458. Springer, 2024b.

James Kirkpatrick et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Yongkang Li, Tianheng Cheng, Bin Feng, Wenyu Liu, and Xinggang Wang. Mask-adapter: The devil is in the masks for open-vocabulary segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14998–15008, 2025.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision*, pp. 614–629, 2016.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7061–7070, 2023.

Zihan Lin, Zilei Wang, and Xu Wang. Towards continual universal segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29417–29427, 2025. doi: 10.1109/CVPR52734.2025.02739.

Silvio Mazzucco, Carl Persson, Mattia Segu, Pier Luigi Dovesi, Federico Tombari, Luc Van Gool, and Matteo Poggi. Lost in translation? vocabulary alignment for source-free adaptation in open-vocabulary semantic segmentation. In *British Machine Vision Conference*, 2025. BMVC.

Umberto Michieli and Pietro Zanuttigh. Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1114–1124, 2021.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 891–898, 2014.

Hongwei Niu, Jie Hu, Jianghang Lin, Guannan Jiang, and Shengchuan Zhang. Eov-seg: Efficient open-vocabulary panoptic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6254–6262, 2025.

Zelin Peng, Zhengqin Xu, Zhilin Zeng, Yu Huang, Yaoming Wang, and Wei Shen. Parameter-efficient fine-tuning in hyperspherical space for open-vocabulary semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15009–15020, 2025.

Reza Qorbani, Gianluca Villani, Theodoros Panagiotakopoulos, Marc Botet Colomer, Linus Härenstam-Nielsen, Mattia Segu, Pier Luigi Dovesi, Jussi Karlgren, Daniel Cremers, Federico Tombari, et al. Semantic library adaptation: Lora retrieval and fusion for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9804–9815, 2025.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021.

Aidan Scannell, Riccardo Mereu, Paul Edmund Chang, Ella Tamir, Joni Pajarinen, and Arno Solin. Function-space parameterization of neural networks for sequential learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Michalis K. Titsias, Jonathan Schwarz, Alexander G. de G. Matthews, Razvan Pascanu, and Yee Whye Teh. Functional regularisation for continual learning with gaussian processes. In *International Conference on Learning Representations*, 2020.

Yu Wang, Mazdak Abulnaga, Yaël Balbastre, and Bruce Fischl. Schwarz–schur involution: Lightspeed differentiable sparse linear solvers. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 62172–62221, 2025.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2955–2966, 2023a.

Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2945–2954, 2023b.

Wenhao Xu, Changwei Wang, Xuxiang Feng, Rongtao Xu, Longzhao Huang, Zherui Zhang, Li Guo, and Shibiao Xu. Generalization boosted adapter for open-vocabulary segmentation. *IEEE Trans. Cir. and Sys. for Video Technol.*, 35(1):520–533, January 2025. ISSN 1051-8215. doi: 10.1109/TCSVT.2024.3454227. URL https://doi.org/10.1109/TCSVT.2024.3454227.

Hongmei Yin, Tingliang Feng, Fan Lyu, Fanhua Shang, Hongying Liu, Wei Feng, and Liang Wan. Beyond background shift: Rethinking instance replay in continual semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9839–9848, 2025. doi: 10.1109/CVPR52734.2025.00919.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36:32215–32234, 2023.

Zhidong Yu, Xiaoman Liu, Jiajun Hu, Zhenbo Shi, and Wei Yang. AAKR: Adversarial attack-based knowledge retention for continual semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22245–22253, 2025. doi: 10.1609/AAAI.V39I21.34379.

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

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European conference on computer vision*, pp. 696–712. Springer, 2022.

Ethics Statement

This work studies model adaptation and evaluation on established segmentation benchmarks and introduces no new personally identifiable data. As with other open-vocabulary perception systems, deployment may inherit biases from vision–language pretraining and may fail under ambiguous, culturally specific, or safety-critical vocabularies. In our evaluations, the frozen-base fallback reduces negative specialization changes and retains access to the base prediction path on retention domains. High-stakes uses warrant application-specific validation and human oversight.

Statement on the Use of Generative AI

Generative AI tools were used to assist with manuscript drafting, editing, and improving clarity and organization, and to provide feedback on the presentation and analysis of experimental controls and results. All experiments, numerical results, and final methodological decisions were produced or verified by the authors. Generative AI tools were not used to generate experimental data or reported results. All AI-assisted content was reviewed and revised by the authors, who take full responsibility for the final content of this work.

A SUPPLEMENTARY OVERVIEW

This supplement provides the complete experimental record and implementation details for Open-Prior Functional Residual Surgery (OP-FRS). In addition to the results shown in the main paper, it reports dataset-role and label-space diagnostics, independent single-step adaptations, full intermediate checkpoint matrices, a controlled short trajectory, protected-set timing, vocabulary-expansion decomposition, extended full-stream ablations, sidecar-capability diagnostics, operator-faithfulness measurements, solver analysis, and absolute efficiency. It also specifies the joint-vocabulary operator, matrix-free implementation, baseline protocols, formal properties, and reproducibility settings.

B PROTOCOL, DATASET ROLES, AND METRICS

Let $\mathcal{D}_0$ be the source dataset and $\mathcal{D}_{1:T}$ the target stream. At step t, the learner accesses only $\mathcal{D}_t$, frozen inherited parameters, and category names. Source or previous-target images, features, logits, and exemplars are not retained. A single task-ID-free model is evaluated after every update.

The main continual stream is COCO→A-150→Cityscapes→PC-459, whose three target datasets use distinct image corpora. PAS-20 is an external evaluation-only probe and is never used for training, hyperparameter selection, or model selection. A-847 and PC-59 are alternate-taxonomy probes sharing image corpora with A-150 and PC-459, respectively; they are reported separately and are not combined into a held-out aggregate. A separately trained COCO→A-150→Cityscapes trajectory is retained for controlled retention and protected-set timing diagnostics.

B.1 LABEL-SPACE SIMILARITY

To characterize the average semantic proximity between dataset vocabularies without using images, we use the pretrained CLIP ViT-L/14 text encoder corresponding to `openai/clip-vit-large-patch14`. Category definitions are taken from `openseg_classes.py`: A-847 (847 classes), A-150 (150), Cityscapes (19), COCO-Stuff (171), PAS-20 (21), PC-59 (59), and PC-459 (459). Each category name is tokenized directly without prompt engineering, and the 768-dimensional `pooler_output` is used as its text representation.

For category sets $\mathcal{C}_i$ and $\mathcal{C}_j$, the simple-mean similarity is the average cosine similarity over all cross-dataset category pairs:

$$\mathrm{Sim}(\mathcal{C}_i, \mathcal{C}_j) = \frac{1}{|\mathcal{C}_i||\mathcal{C}_j|} \sum_{c \in \mathcal{C}_i} \sum_{\hat{c} \in \mathcal{C}_j} \frac{u_c^\top u_{\hat{c}}}{\|u_c\|_2 \, \|u_{\hat{c}}\|_2}, \tag{26}$$

where u_c denotes the pooled CLIP text embedding of category name c. Tiny floating-point asymmetries between the two stored matrix directions are averaged before visualization, and diagonal entries are fixed to 1.0.

Figure 8 shows moderate but non-uniform average similarity across the seven vocabularies. The off-diagonal mean is 0.3997 ± 0.0284, with values ranging from 0.3516 to 0.4493. PAS-20 and PC-59 have the highest pairwise mean, whereas A-847 and A-150 have the lowest. Because this statistic averages every cross-vocabulary category pair, it measures global semantic proximity rather than taxonomy inclusion, best-match coverage, visual domain similarity, or segmentation difficulty. In particular, a large and diverse vocabulary can obtain a modest mean similarity even when it contains many categories related to a smaller vocabulary. The diagnostic therefore complements, rather than replaces, the separate external-domain and alternate-taxonomy evaluations used in the main paper.

B.2 PROTECTED-SET TIMING

The protected category set at step t is

$$\mathcal{C}_{<t} = \bigcup_{j=0}^{t-1} \mathcal{C}_j. \tag{27}$$

The same $\mathcal{C}_{<t}$ is used during training, validation, and immediate post-update evaluation. Current labels $\mathcal{C}_t$ enter the protected set when step $t+1$ begins, ensuring that surgery protects inherited behavior while leaving the incoming task free to establish its own residual capability.

Benchmark	Underlying image corpus	Role in study	Optimized?	Relationship to trained targets	Aggregate?
COCO-Stuff	COCO	Source	Source training	Source initialization	Source only
A-150	ADE20K	Target 1	Yes	Distinct target corpus	Target only
Cityscapes	Cityscapes	Target 2	Yes	Distinct target corpus	Target only
PC-459	PASCAL Context	Target 3	Yes	Distinct target corpus	Target only
PAS-20	PASCAL VOC	External evaluation probe	No	Not used for training or selection	No
A-847	ADE20K	Alternate-taxonomy probe	No	Same image corpus as A-150	No
PC-59	PASCAL Context	Alternate-taxonomy probe	No	Same image corpus as PC-459	No

Table 6: Dataset-role audit for the evaluation protocol. PAS-20 is evaluation-only. A-847 and PC-59 test alternate taxonomies on corpora already represented by A-150 and PC-459, respectively, and are therefore reported separately rather than described as independent held-out datasets.

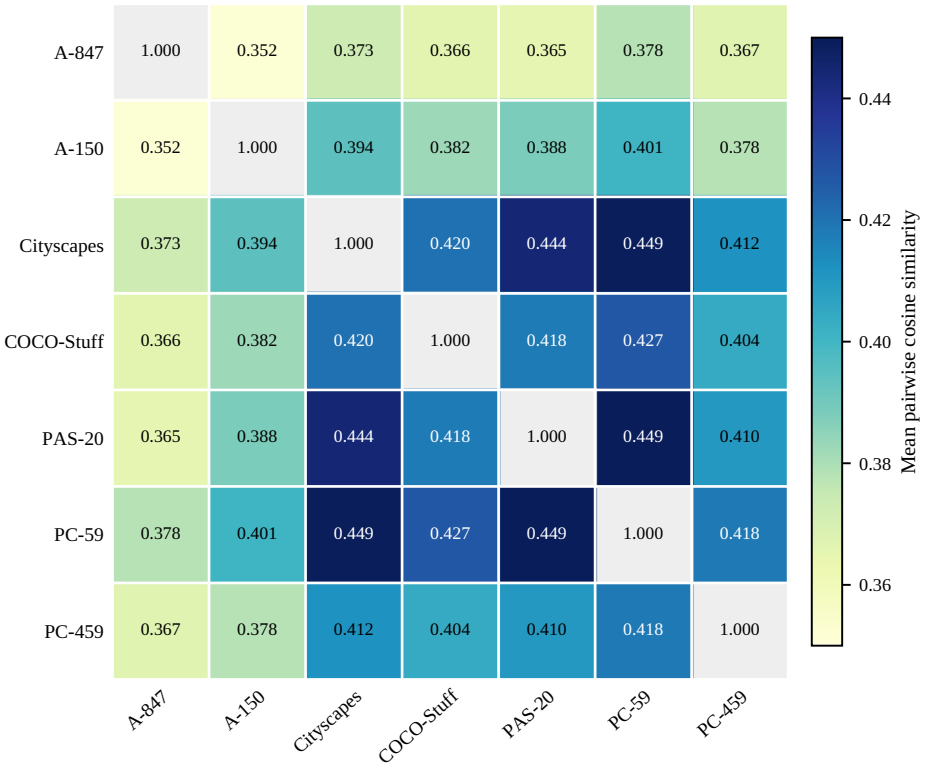

Figure 8: Label-space similarity from CLIP ViT-L/14 text embeddings. Each off-diagonal cell is the mean cosine similarity over all category pairs between two datasets. Diagonal entries are fixed to 1.0 and excluded from the color scale.

B.3 METRICS

Let $\mathcal{M}_D(f)$ denote the selected segmentation metric on dataset D. For single-step adaptation,

$$\begin{aligned} G_{\text{tgt}} &= \mathcal{M}_{\text{tgt}}(f) - \mathcal{M}_{\text{tgt}}(f_s), \\ \Delta_{\text{src}} &= \mathcal{M}_{\text{src}}(f) - \mathcal{M}_{\text{src}}(f_s). \end{aligned} \tag{28}$$

The main table reports absolute results for every probe and does not mix alternate vocabularies into a held-out aggregate.

Let $a_{k,j}$ be performance on target j after update k. Signed forgetting remains

$$F_T^{\text{signed}} = \frac{1}{T-1} \sum_{j=1}^{T-1} \left(\max_{k \in \{j,\ldots,T-1\}} a_{k,j} - a_{T,j} \right), \tag{29}$$

and may be negative, indicating positive backward transfer. To prevent cancellation, the evaluation additionally reports

$$F_T^{+} = \frac{1}{T-1} \sum_{j=1}^{T-1} \max\left(0, \max_{k \in \{j,\dots,T-1\}} a_{k,j} - a_{T,j}\right), \tag{30}$$

$$F_T^{\max} = \max_{j<T} \left(\max_{k \in \{j,\dots,T-1\}} a_{k,j} - a_{T,j} \right), \tag{31}$$

together with the complete performance matrix.

For decision-level diagnostics, the new-class stealing rate is

$$\mathrm{NCS} = \frac{\#\{p : y_p \in \mathcal{C}_{<t},\ \widehat{y}_p^{\mathrm{after}} \in \mathcal{C}_t \setminus \mathcal{C}_{<t}\}}{\#\{p : y_p \in \mathcal{C}_{<t}\}}, \tag{32}$$

measured only on labeled source or historical validation images used for evaluation, never for training. The relative historical-confidence drop follows the main paper:

$$\delta_{t,g}^{\mathrm{conf}} = \frac{\left[\|\mathcal{A}_t(Y_{H,g}^{\mathrm{base}})\|_F - \|\mathcal{A}_t(Y_{H,g}^{\mathrm{surg}})\|_F\right]_+}{\|\mathcal{A}_t(Y_{H,g}^{\mathrm{base}})\|_F + \epsilon_a}. \tag{33}$$

B.4 Single-Update Functional Drift

The frozen FC-CLIP/ConvNeXt-L model is a strong no-update reference: it preserves the complete source-trained mask generation, region–text alignment, no-object calibration, and dense composition pathway. All rows use the same requested vocabulary, prompt templates, no-object normalization, mask–score composition, and post-processing, so differences reflect the effect of adaptation rather than a change in evaluation rules.

C Implementation Details

C.1 Optimization, Training Objective, and Sidecar Capacity

The implementation uses FC-CLIP with a ConvNeXt-L encoder and a Mask2Former-style decoder. Each step trains only the new sidecar group with AdamW, learning rate 10^{-5}, batch size 8, and 150k iterations. Previous sidecars and the source path remain frozen. Structural and readout sidecars use rank-16 LoRA residuals and bottleneck-64 adapters, respectively. Forward and backward implicit solves use five CG iterations in float32. The graph degree, graph temperature, relation/anchor weights, and semantic projection rank are fixed across the complete stream and are not retuned per dataset.

All five-run results use the fixed random seeds $\{42, 32, 1, 14, 27\}$. The same seed set is used for every compared method so that run-level comparisons are aligned across methods. Experiments are conducted on a server with eight NVIDIA GeForce RTX 3090 GPUs (24 GB each), an Intel(R) Xeon(R) Gold 6248 CPU at 2.50 GHz, and 256 GB system memory. The software environment is Ubuntu 22.04.1, PyTorch 2.0.1, and CUDA 11.8.

The training loss is the unchanged source implementation's Hungarian-matched objective,

$$\mathcal{L}_t = \lambda_{\mathrm{cls}}\mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{bce}}\mathcal{L}_{\mathrm{bce}} + \lambda_{\mathrm{dice}}\mathcal{L}_{\mathrm{dice}}, \tag{34}$$

applied to the surgically corrected region scores and masks. OP-FRS introduces no replay, stored teacher outputs, or auxiliary distillation loss. The inherited classification and mask-matching coefficients are kept unchanged across all continual steps.

C.2 Cross-framework Realization under MAFT

The MAFT validation preserves the same three functional intervention points used under FC-CLIP. The structural sidecar is attached to the proposal-mask logits, the readout sidecar modifies the mask-conditioned CLIP region representation before text scoring, and the semantic sidecar modifies the

Component	Injection point	Parameterization	Update rule
Structural sidecar	Proposal mask logits	Rank-16 LoRA residual	Current group only
Readout sidecar	Mask-conditioned region embedding	Bottleneck-64 adapter	Current group only
Semantic sidecar	Region–text scores before joint softmax	Text-conditioned bilinear residual	Current group only
Graph, anchor, and surgery	Protected dense output / stage residual	Parameter-free	Recomputed per image
Source path and earlier groups	Inherited forward path	Frozen	No update

Table 7: Sidecar placement and update audit. Each target adds one structurally identical sidecar group, and only the newest group is optimized at each step. Under the counting convention used in the efficiency comparison, OP-FRS has a 1.20M trainable-parameter budget; earlier groups remain frozen but persistent.

region–text score matrix before appending no-object and applying the joint softmax. The historical graph, confidence anchor, composed Jacobians, CG solve, and magnitude recovery are unchanged. MAFT's source modules and all earlier sidecar groups remain frozen; only the newest task-specific sidecar group is optimized.

The OP-FRS capacity is matched across frameworks by retaining rank-16 structural LoRA modules, bottleneck-64 readout adapters, and the same text-conditioned semantic projection dimensions. The OP-FRS trainable-parameter budget remains 1.20M under the same counting convention as the main efficiency table; at step t, only the newest task-specific group receives gradients. Earlier groups remain frozen and persistent, while the graph, anchor, and surgery operator introduce no trainable parameters.

For the MAFT baselines, the underlying source checkpoint, target order, prompts, resolution, augmentation, optimizer family, schedule, and evaluation pipeline are identical. Sequential fine-tuning updates the framework-specific trainable path directly. ECLIPSE retains its visual-prompt and logit-control mechanism on the MAFT proposal/score path, while ConOVS retains its expert-decoder and routing mechanism on the same MAFT features. We do not add OP-FRS sidecars to these baselines or enlarge their native mechanisms. All methods are evaluated without replay or oracle task identity.

C.3 Stable Scores, Joint Vocabulary, and Selection

For a row vector a,

$$\mathrm{Norm}_\epsilon(a) = \frac{a}{\sqrt{\|a\|_2^2 + \epsilon_{\mathrm{norm}}}}. \tag{35}$$

For the joint vocabulary $\mathcal{C}_t^{\mathrm{joint}}$, the source semantic and no-object scores are

$$\begin{aligned} S_s(\mathcal{C}_t^{\mathrm{joint}}) &= \tau\, \mathrm{Norm}_\epsilon(z_s)\, \mathrm{Norm}_\epsilon(E_t^{\mathrm{joint}})^\top, \\ S_s^+(\mathcal{C}_t^{\mathrm{joint}}) &= [S_s(\mathcal{C}_t^{\mathrm{joint}}), s_s^\varnothing(z_s)] \in \mathbb{R}^{N\times(K_u+1)}. \end{aligned} \tag{36}$$

The semantic selectors satisfy

$$\Pi_{\mathrm{hist}} \in \{0,1\}^{K_h\times K_u}, \qquad \Pi_{\mathrm{eval}} \in \{0,1\}^{K_e\times K_u}. \tag{37}$$

They act on dense semantic outputs *after* the shared joint softmax. No-object remains in the softmax denominator but is removed from the semantic output columns.

C.4 Vocabulary Canonicalization

Every label is canonicalized by Unicode normalization, lowercasing, whitespace and punctuation normalization, and the benchmark-provided alias dictionary. Exact duplicates are removed while preserving a reversible index map. Semantic synonyms are not merged unless an official benchmark mapping declares them equivalent. Prompt templates and alias dictionaries remain fixed across all methods.

C.5 Dense Joint Composition and Anchored Functional Output

For $S^+ = [S, s^\varnothing] \in \mathbb{R}^{N\times(K_u+1)}$,

$$\pi_{\mathrm{sem}}(S_{n,:}^+) = [\mathrm{Softmax}(S_{n,:}^+)]_{1:K_u}, \tag{38}$$

$$[\mathrm{Compose}(M, S^+)]_{p,c} = \sum_{n=1}^{N} \sigma(M_{n,p}) \pi_{\mathrm{sem}}(S^+_{n,:})_c. \tag{39}$$

Let $Y^{\mathrm{joint}} = \mathrm{Compose}(M, S^+)$ and $Y_H = Y^{\mathrm{joint}} \Pi_{\mathrm{hist}}^\top$. The protected output is

$$\Psi_t(Y^{\mathrm{joint}}) = \left[\omega_{\mathrm{rel}} \frac{Y_H B_t W_t^{1/2}}{\sqrt{L_t}} \;\middle|\; \omega_{\mathrm{anc}} \frac{Y_H}{\sqrt{K_h}} \right]. \tag{40}$$

The first block protects old–old relations. The second removes the common-shift null space and detects probability transferred from historical classes to current/evaluation classes or no-object through the joint denominator.

C.6 Text-conditioned Semantic Sidecar

For semantic sidecar i, the two row-wise projection maps are linear:

$$\begin{aligned} U_i(z) &= h_i^U(z) = z W_i^U + \mathbf{1}_N (b_i^U)^\top, \\ V_i(E) &= h_i^V(E) = E W_i^V + \mathbf{1}_{K_u} (b_i^V)^\top, \\ r^S_{i,\mathrm{sem}}(z, E) &= U_i(z) V_i(E)^\top / \sqrt{d_a}. \end{aligned} \tag{41}$$

Because the same h_i^V is applied independently to every text prototype, category permutation only permutes semantic columns. The frozen no-object score is appended after semantic residual injection.

C.7 Composed and Local Functional Operators

Let $\mathrm{sg}_\theta[f]$ detach parameters while retaining derivatives with respect to inputs, and let $\mathcal{R}_s(M) = \mathrm{Readout}_s(F_s, M)$. The detached maps are

$$\begin{aligned} \overline{\mathcal{B}}_t(M) &= \mathcal{R}_s(M) + \sum_{i=1}^{t} \mathrm{sg}_\theta[r^z_{i,\mathrm{read}}(F_s, M, \mathcal{R}_s(M))], \\ \overline{\mathcal{A}}_t(z, E) &= \tau \, \mathrm{Norm}_\epsilon(z) \, \mathrm{Norm}_\epsilon(E)^\top \\ &\qquad + \sum_{i=1}^{t} \mathrm{sg}_\theta[r^S_{i,\mathrm{sem}}(z, E)], \\ \overline{\mathcal{A}}^+_t(z, E) &= [\overline{\mathcal{A}}_t(z, E), s^\varnothing_s(z)]. \end{aligned} \tag{42}$$

Detaching parameters does not detach input dependence. For a continuation $\Phi(h)$ with Jacobian J, the matrix-free products are

$$\begin{aligned} \mathrm{JVP}_\Phi(v) &= \frac{\mathrm{d}}{\mathrm{d}\epsilon} \Phi(h + \epsilon v)\Big|_{\epsilon=0} = Jv, \\ \mathrm{VJP}_\Phi(u) &= \nabla_h \langle u, \Phi(h) \rangle = J^\top u. \end{aligned} \tag{43}$$

The composed JVP/VJP therefore traverses downstream sidecars, text scoring, the single joint softmax, dense composition, historical selection, graph relations, and the confidence anchor. Applying a JVP followed by a VJP computes $J^\top J v$ without explicitly materializing J. Later surgery layers are not recursively nested inside earlier Jacobians.

The local-J control terminates at the current stage output before downstream maps and composition, but uses the same residual architecture, optimizer, and CG budget. This isolates the contribution of the composed continuation.

C.8 Input-conditioned Jacobians

$J^{\mathrm{comp}}_{t,g}(x)$ is constructed independently for each image. During current-task training it is evaluated on current images. During source, previous-target, or probe evaluation it is recomputed on those

evaluation images themselves. No Jacobian estimated on one domain is cached and reused on another. Domain-wise diagnostics report

$$\rho_D = \frac{\|J_D \widetilde{r}_D\|_2^2}{\|J_D r_D^{\Sigma}\|_2^2} \tag{44}$$

for current, source, and previous-target images.

C.9 Magnitude Recovery and Implicit Solves

Magnitude recovery is image-wise and stage-wise. Its norm spans the complete residual tensor excluding the batch dimension. For stage g,

$$A_g = I + (J_g^{\text{comp}})^\top J_g^{\text{comp}}. \tag{45}$$

Forward CG solves $A_g q_g = r_g^{\Sigma}$ and backward CG solves $A_g u_g = \bar{q}_g$. No gradient is propagated through the construction of A_g, and no second-order Jacobian derivative is evaluated. Logs record the linear-system residual, $\|r\|$, $\|q\|$, recovery factor $\|r\|/(\|q\| + \epsilon_n)$, median, 90th/99th percentiles, maximum, and non-finite events.

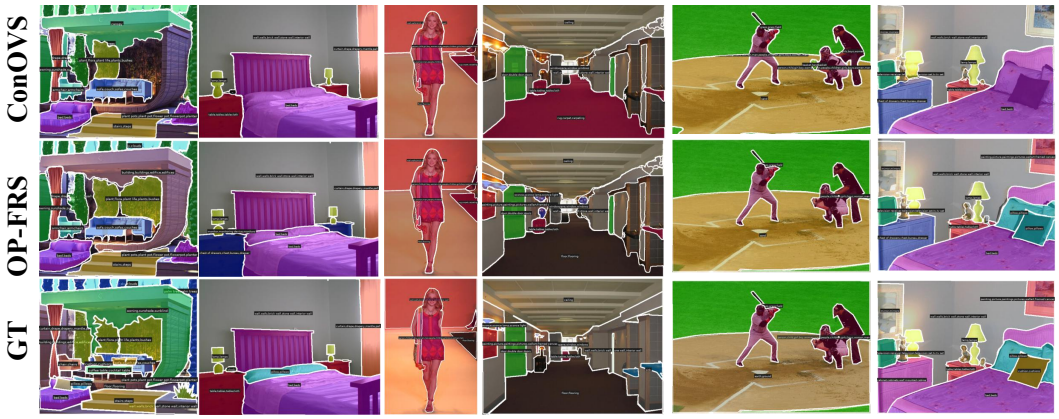

Figure 9: Qualitative comparison. OP-FRS yields more coherent semantic regions and cleaner boundaries than ConOVS, particularly in cluttered indoor scenes and fine-grained object regions.

Qualitative results. Figure 9 presents representative ADE20K predictions from ConOVS and OP-FRS. Compared with ConOVS, OP-FRS shows more coherent region assignment in several complex scenes, including the floor–wall partition in the corridor and the separation of bedding, pillows, and furniture in the bedroom examples. It also maintains clear foreground boundaries in the person and baseball scenes. The qualitative results are consistent with the quantitative advantage of OP-FRS on A-150 and suggest that its functional residual control improves semantic organization without sacrificing prominent object delineation.

D Replay-free Baseline Protocol

All adapted baselines receive the same current images, prompt templates, requested and joint vocabularies, input resolution, augmentations, optimizer family, training schedule, evaluation pipeline, and tuning budget. Architecture-specific trainable modules and method-specific losses follow their corresponding implementations. The frozen source uses the same scoring and evaluation pipeline but undergoes no update.

Operator controls. The reported variants isolate the two design choices that define OP-FRS: joint graph-plus-anchor protection and composed downstream sensitivity. They compare historical-only versus joint competition, relation-only versus anchor-only protection, and local versus composed Jacobians under the same residual architecture and optimization budget.

[t] One continual step of OP-FRS [1] Current batch (x, y), frozen source path, sidecars $1{:}t-1$, historical names $\mathcal{C}_{<t}$, requested vocabulary $\mathcal{C}^{\text{eval}}$ Add ϕ_t and freeze inherited parameters. Build $\mathcal{C}_t^{\text{joint}} = \text{unique}(\mathcal{C}_{<t} \cup \mathcal{C}^{\text{eval}})$. Build the historical text graph and confidence anchor. $g \in \{\text{str}, \text{read}, \text{sem}\}$ in structural-to-semantic order Sum all active sidecar residuals at stage g. Compose to joint dense scores, select historical columns, and form Ψ_t. Construct input-conditioned J_g^{comp} through downstream maps. Solve $(I + J_g^\top J_g)q_g = r_g^\Sigma$ by CG. Recover scale, inject $\widetilde{r}_g$, and continue. Use the same joint softmax, select evaluation columns, compute $\mathcal{L}_t$, and update only ϕ_t. Insert $\mathcal{C}_t$ into the protected set when step $t+1$ begins.

E Formal Analysis of Directional Selectivity

Let J denote a fixed composed Jacobian and $A = J^\top J \succeq 0$. For nonzero v, define

$$\mathcal{R}_J(v) = \frac{v^\top A v}{v^\top v}. \tag{46}$$

Proposition 1 (proximal solution). The objective

$$\mathcal{L}(q) = \frac{1}{2}\|q - r\|_2^2 + \frac{1}{2}\|Jq\|_2^2 \tag{47}$$

is strictly convex and has the unique minimizer

$$q = (I + A)^{-1} r. \tag{48}$$

Proof. The Hessian is $I + A \succ 0$ because $A \succeq 0$. Setting the gradient $(q - r) + Aq$ to zero gives $(I + A)q = r$, proving existence and uniqueness. □

Proposition 2 (Rayleigh quotient cannot increase). Let $r \neq 0$ and $q = (I + A)^{-1} r$. Then

$$\mathcal{R}_J(q) \leq \mathcal{R}_J(r). \tag{49}$$

Equality holds when r lies in a single eigenspace of A, or more generally when all nonzero spectral components of r correspond to the same eigenvalue.

Proof. Let $A = U \,\text{diag}(\mu_1, \ldots, \mu_n) U^\top$, where $\mu_i \geq 0$, and write $r = Ua$ with weights $w_i = a_i^2 \geq 0$. Then

$$q = U \,\text{diag}\left(\frac{1}{1 + \mu_i}\right) a. \tag{50}$$

Set $f_i = (1 + \mu_i)^{-2}$, which is non-increasing in μ_i. The two quotients are

$$\mathcal{R}_J(r) = \frac{\sum_i \mu_i w_i}{\sum_i w_i}, \qquad \mathcal{R}_J(q) = \frac{\sum_i \mu_i w_i f_i}{\sum_i w_i f_i}. \tag{51}$$

The sign of their difference is the sign of

$$\begin{aligned} &\left(\sum_i \mu_i w_i f_i\right)\left(\sum_j w_j\right) - \left(\sum_i \mu_i w_i\right)\left(\sum_j w_j f_j\right) \\ &= \frac{1}{2}\sum_{i,j} w_i w_j (\mu_i - \mu_j)(f_i - f_j). \end{aligned} \tag{52}$$

Because f is non-increasing, $(\mu_i - \mu_j)(f_i - f_j) \leq 0$ for every pair (i, j). Hence the sum is non-positive and $\mathcal{R}_J(q) \leq \mathcal{R}_J(r)$. □

Corollary 1 (magnitude recovery preserves direction selectivity). For any scalar $\alpha \neq 0$,

$$\mathcal{R}_J(\alpha q) = \mathcal{R}_J(q). \tag{53}$$

Therefore the positive scaling in magnitude recovery cannot undo the Rayleigh-quotient reduction of an exact proximal solve.

Proof. Both numerator and denominator are multiplied by α^2. □

Corollary 2 (absolute composed interference does not increase). Let $r \neq 0$, $q = (I + J^\top J)^{-1} r$, and

$$\widetilde{r} = \frac{\|r\|_2}{\|q\|_2 + \epsilon_n} q, \qquad \epsilon_n \geq 0. \tag{54}$$

Then

$$\|J\widetilde{r}\|_2^2 \leq \|Jr\|_2^2. \tag{55}$$

Proof. The recovery rule gives $\|\widetilde{r}\|_2 \leq \|r\|_2$. By Proposition 2 and Corollary 1,

$$\frac{\|J\widetilde{r}\|_2^2}{\|\widetilde{r}\|_2^2} = \mathcal{R}_J(q) \leq \mathcal{R}_J(r) = \frac{\|Jr\|_2^2}{\|r\|_2^2}. \tag{56}$$

Multiplying by $\|\widetilde{r}\|_2^2 \leq \|r\|_2^2$ proves the claim. □

Finite-CG accuracy. The propositions characterize the exact proximal solution. For an approximate CG solution $\widehat{q}$ with $e = \widehat{q} - q$, $I + A \succeq I$ yields

$$\|e\|_2 \leq \|(I + A)^{-1}\|_2 \|r - (I + A)\widehat{q}\|_2 \leq \|r - (I + A)\widehat{q}\|_2. \tag{57}$$

Therefore the logged linear-system residual directly controls the approximation error. The solver study reports both this residual and the measured pre/post interference, and shows that five iterations already reach the performance-saturation regime.

Numerical stabilizers. The implementation uses ϵ_d and ϵ_n only to stabilize ratios and norm recovery. Formal claims are stated with the scale-invariant Rayleigh quotient $\mathcal{R}_J$, while the recovery denominator changes only the positive scalar applied to q and therefore preserves directional selectivity. Solver logs record residuals, norm ratios, recovery-factor percentiles, and non-finite events to verify that the numerical realization remains within the regime analyzed above.

F Additional Experimental Results

F.1 Controlled Short-sequence Trajectory

We use a separately trained COCO→A-150→Cityscapes stream to inspect retention after each update without introducing the additional PC-459 stage. This diagnostic is reported separately from the main three-target protocol and is not used for hyperparameter or model selection.

The trajectory shows that the source, the first target, and the evaluation probes remain close to their corresponding reference checkpoints while the incoming targets improve. This controlled trajectory is used only for mechanism and timing diagnostics and is not merged with the main three-target results. Its source reference and all derived retention/gain values are aligned with the main paper.

Protected-set consolidation. At the last update of the controlled trajectory, the current Cityscapes classes have not yet entered the strictly historical set $\mathcal{C}_{<T}$. Table 10 evaluates the same learned model after expanding the protected set to $\mathcal{C}_{\leq T}$ and recomputing the input-conditioned operator. The small changes show that the timing convention can be consolidated at deployment without replay or additional optimization.

After learning	COCO	A-150	City.	PC-59	PAS-20
Source init.	65.82	35.57	54.34	59.89	96.07
A-150	65.05	41.82	54.42	60.38	96.08
Cityscapes	64.28	39.76	61.66	60.71	95.94

Table 8: Complete performance matrix for the controlled COCO→A-150→Cityscapes trajectory. The source row uses the same frozen-source reference as the main paper. All entries are mIoU.

Checkpoint	COCO ret. (%)	A-150 ret. (%)	PC-59 ret. (%)	PAS-20 ret. (%)	Current gain
Source	100.0	–	100.0	100.0	–
A-150	98.8	100.0	100.8	100.0	+6.25
Cityscapes	97.7	95.1	101.4	99.9	+7.32

Table 9: Retention and current-target gains for the controlled short trajectory, recomputed using the frozen-source reference reported in the main paper.

Deployment mode	COCO	A-150	City.	PC-59	PAS-20
Immediate ($\mathcal{C}_{<T}$)	64.32	39.83	61.45	60.21	95.81
Consolidated ($\mathcal{C}_{\leq T}$)	64.28	39.76	61.66	60.71	95.94

Table 10: Protected-set timing diagnostic on the final controlled COCO→A-150→Cityscapes model. Both rows use identical learned parameters. Consolidated inference expands the protected set to $\mathcal{C}_{\leq T}$ and recomputes the input-conditioned operator without replay or optimization.