CONTROLLING SPURIOUS FOREGROUND EXPANSION IN OPEN-VOCABULARY SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary segmentation identifies image regions corresponding to text queries, even when the queried classes are absent from the image. During single-image test-time adaptation, confidence sharpening can suppress true foreground, while anchor self-training can turn a weak false foreground into its own supervision. The resulting update enlarges the false mask, which then supplies more anchors. This second failure is termed *spurious foreground expansion* (SFE). To control this feedback loop, the proposed Spurious Foreground Expansion Guard (SFE-Guard) provides a two-stage guard. Credit-Denominated Issuance (CDI) freezes zero-shot semantic evidence and bounds the pseudo supervision issued to each foreground class. Hierarchical Risk Correction (HRC) reviews aggregate and background risks and applies a feasible correction before a candidate update proposed by the optimizer is committed and reused as supervision. Across all three datasets, SFE-Guard improves state-of-the-art Present Dice by 3.066 and 3.177 points for one and multiple queries, respectively. It removes 77.9% of the excess false foreground caused by anchor-based self-training in one-query OVS and restores multi-query absent class area to the zero-shot level.

1 INTRODUCTION

Unlike conventional closed-set segmentation with a fixed set of training classes, open-vocabulary segmentation (OVS) transfers arbitrary text-specified concepts to dense prediction without task-specific training (Radford et al., 2021; Li et al., 2022; Rao et al., 2022). Existing OVS methods obtain dense masks through text–pixel or region–text alignment and grouping (Xu et al., 2022; Zhou et al., 2022). More recent training-free methods improve localization by refining frozen VLM attention or feature aggregation (Wang et al., 2024; Lan et al., 2024), with NACLIP supporting arbitrary queries at deployment (Hajimiri et al., 2025). Their zero-shot masks can nevertheless degrade under domain shift. Test-time adaptation (TTA) mitigates this shift by adapting a pretrained model to unlabeled test images. Representative methods minimize prediction entropy, filter unreliable samples, or regularize unstable updates (Wang et al., 2021; Niu et al., 2022; 2023); continual variants further use restoration and augmentation consistency (Wang et al., 2022; Zhang et al., 2022). For vision–language models, prompt adaptation refines textual representations from augmented views or through calibration constraints (Shu et al., 2022; Sharifdeen et al., 2025; Vargas Hakim et al., 2025).

Among these works, *confidence sharpening* is a representative technique. MLMP exemplifies entropy minimization that updates shared visual features to reinforce the current prediction (Noori et al., 2025). Under background dominance, this reinforcement can progressively reduce or eliminate true foreground, a failure termed *foreground suppression*. BAPA instead uses *anchor self-training*, selecting confident pixels from each predicted class as supervision (Wang et al., 2026b). This can preserve supported foreground, but an incorrect anchor can be strengthened, enlarged, and selected again. This recursive failure is termed *spurious foreground expansion* (SFE). Figure 1 contrasts foreground suppression with SFE. To reduce false foreground, existing OVS work augments a query with test-time contrastive concepts or trains a model to abstain under counterfactual images (Wysoczańska et al., 2025; Li et al., 2026). However, they cannot regulate how a prediction becomes subsequent supervision during single-image TTA. This motivates coordinated control of supervision and updates.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

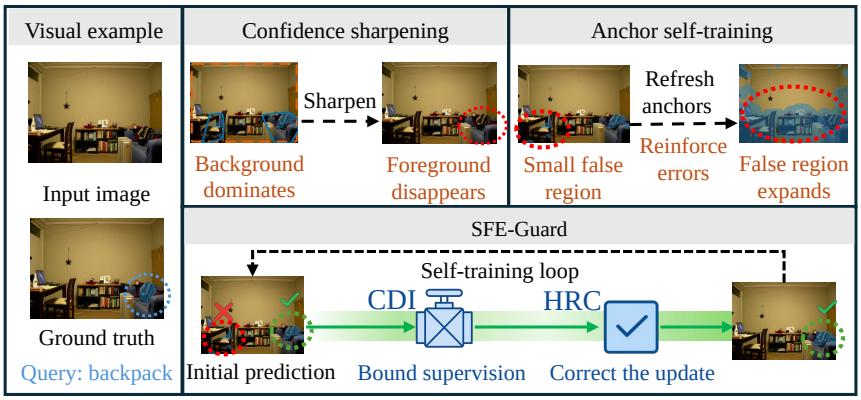


Figure 1: **Motivation for SFE-Guard.** Confidence sharpening suppresses true foreground under background dominance, while anchor self-training expands false foreground through repeated reinforcement. SFE-Guard bounds supervision with CDI and corrects candidate updates with HRC.

SFE-Guard controls the two links in this feedback loop. *Credit-Denominated Issuance* (CDI) uses semantic support fixed at the initial zero-shot prediction to cap the supervision that each class can issue. Anchors may be refreshed, but enlarging a predicted mask cannot increase that class's total supervision weight. *Hierarchical Risk Correction* (HRC) then checks the optimizer's proposed candidate update and corrects changes that would reinforce erroneous predictions. Together, CDI and HRC control supervision and updates in both one- and multi-query OVS. This work makes three contributions.

1. The analysis distinguishes foreground suppression under confidence sharpening from SFE under anchor self-training, showing how dynamic anchor refresh reinforces erroneous supervision and expands false foreground.
2. SFE-Guard combines CDI, which bounds classwise supervision using fixed semantic evidence, with HRC, which corrects candidate updates. Theoretical analysis establishes the supervision bound and correction optimality under linearized risk constraints.
3. SFE-Guard jointly controls foreground suppression and SFE in one- and multi-query OVS, achieving state-of-the-art Present Dice across three benchmarks in both settings while reducing false foreground relative to anchor self-training.

2 Related Work

2.1 Test-Time Adaptation for Open-Vocabulary Segmentation

Early work explored adaptation through self-supervision and source hypothesis transfer (Sun et al., 2020; Liang et al., 2020). TENT minimizes prediction entropy (Wang et al., 2021), while EATA and SAR improve reliability through sample filtering, regularization, or sharpness-aware updates (Niu et al., 2022; 2023). CoTTA addresses continual adaptation through teacher averaging and stochastic restoration (Wang et al., 2022), while MEMO minimizes marginal entropy over augmented views (Zhang et al., 2022). For vision–language models, methods adapt prompts from augmented views (Shu et al., 2022), impose prompt orthogonality (Sharifdeen et al., 2025), or refine text representations iteratively (Vargas Hakim et al., 2025). Calibration-aware prompt tuning further studies prediction calibration during adaptation (Yoon et al., 2024).

Open-vocabulary segmentation extends CLIP (Radford et al., 2021) to dense prediction. Region-based and mask-classification approaches connect text prompts to dense regions (Ghiasi et al., 2022; Ding et al., 2022; Shin et al., 2022). Other methods adapt CLIP through mask-specific, side-network, or diffusion features (Liang et al., 2023; Xu et al., 2023b;a). Fully convolutional and cost-aggregation models further improve dense category matching (Yu et al., 2023; Cho et al., 2024). Subsequent methods improve dense representations through training or feature alignment (Zhou et al., 2023; Shin et al., 2024; Barsellotti et al., 2025). Training-free attention and feature refinements improve

CLIP localization (Wang et al., 2024; Lan et al., 2024; Wysoczańska et al., 2024), with NACLIP producing dense masks through modified attention (Hajimiri et al., 2025). MLMP extends test-time adaptation to OVS by updating visual normalization parameters (Noori et al., 2025), connecting dense prediction with confidence sharpening.

The closest existing studies operate at different stages. Test-time Contrastive Concepts constructs separate alternatives for each query at inference (Wysoczańska et al., 2025), while Counterfactual Segmentation Reasoning learns abstention from paired factual and counterfactual examples (Li et al., 2026). Both address unsupported foreground in the prediction itself. Test-Time Prototype Adaptation transfers evidence across a pool of deployment images by aggregating confident patches with frozen DINO features (Wang et al., 2026a). SFE-Guard instead controls single-image self-training to limit the amplification of false foreground across adaptation steps. Further protocol comparisons are provided in Appendix E.

2.2 Self-Training and Constrained Updates

Pseudo-labeling and confidence selection turn model predictions into supervision (Lee, 2013; Zou et al., 2018). Iterative refresh can improve the targets, but it can also amplify early errors. Class-balanced self-training reduces majority bias (Zou et al., 2018), and BAPA applies this principle to foreground–background segmentation through dynamic classwise anchors (Wang et al., 2026b). Because these anchors are refreshed from the evolving prediction, an erroneous region can repeatedly supply its own supervision. SFE-Guard retains the classwise anchor structure while using fixed semantic evidence to bound each class's supervision and reviewing the resulting update before it changes the next prediction.

Classical gradient projection maps an iterate onto a fixed feasible parameter set (Bertsekas, 1976). PCGrad removes pairwise conflicting components between task gradients (Yu et al., 2020). Constrained policy optimization couples a trust region with surrogate constraints and a policy-specific performance bound (Achiam et al., 2017). HRC constructs constraints from the increases in aggregate and background prediction risk caused by a candidate update. Combined with supervision bounded by semantic evidence, this correction controls both the generation of erroneous supervision and its reinforcement through subsequent updates.

3 Method

3.1 Problem Analysis

Task definition. Let $\mathcal{Q} = \{q_1, \ldots, q_{C-1}\}$ be the foreground query set supplied with an unlabeled image x, and let class 0 denote background, so $C = |\mathcal{Q}| + 1$. A one-query episode contains one foreground query and background. A multi-query run jointly predicts all classes defined by the dataset. In either case, frozen dense visual features and normalized text prototypes produce per-pixel logits $z_{ic}(\delta)$ for pixel i and class $c \in \{0, \ldots, C-1\}$. For SFE-Guard, normalized zero-shot prototypes are constructed for the supplied query set as detailed in Appendix A. The only trainable variable is a text residual $\delta \in \mathbb{R}^{C \times d}$, where d is the text-embedding dimension; both encoders and the cached dense visual representation remain fixed. The common prediction rule is

$$p_{ic}(\delta) = \mathrm{softmax}_c\, z_{ic}(\delta), \qquad \hat{y}_i(\delta) = \arg\max_c p_{ic}(\delta). \tag{1}$$

δ_0 is initialized to zero, with a fresh optimizer for every image. The frozen zero-shot probability $p^0 = p(0)$ is the initial evidence and serves as the *zero-shot reference*. For a single query, the model predicts background versus one text-specified class. For multiple queries, the model jointly predicts several text-specified classes. All classes compete in one softmax and are affected by the same adaptation update. Both settings can produce false foreground. Adaptation uses only the test image and text queries, without source data or ground-truth annotations. Figure 2 quantifies two adaptation failures. Foreground suppression occurs when a present class occupies at least 1% of the initial prediction but less than 1% after adaptation. SFE is measured on absent classes as an increase in predicted area exceeding one percentage point, for which all predicted foreground is erroneous.

Foreground suppression under confidence sharpening. Confidence sharpening minimizes uncertainty in the current prediction. For the pixelwise entropy $H(p_i) = -\sum_c p_{ic} \log p_{ic}$, its logit

derivative is

$$\frac{\partial H(p_i)}{\partial z_{ic}} = -p_{ic}\big(\log p_{ic} + H(p_i)\big). \tag{2}$$

This derivative depends only on the current probabilities; it contains no signal indicating whether the predicted class is present or correct. Gradient descent therefore sharpens the current winner even when that winner is unsupported. With one query, every pixel contributes to the same foreground–background contrast, so dominant background predictions can control the shared update and suppress the queried foreground. Multiple queries distribute this competition across classes. Prediction collapse is therefore more pronounced with one query, as reflected in the higher frequency of foreground suppression in Figure 2.

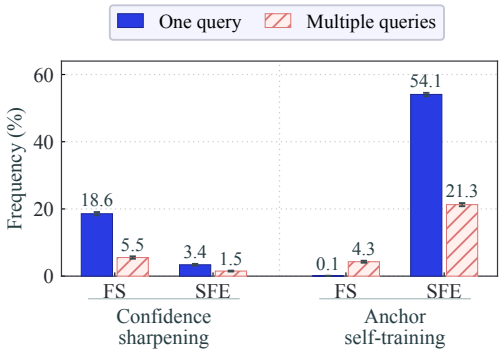


Figure 2: **Failure frequencies.** FS denotes foreground suppression; SFE denotes spurious foreground expansion, measured here on absent classes. Solid blue and red hatched bars denote one and multiple queries, respectively; whiskers show 95% confidence intervals.

Spurious foreground expansion under anchor self-training. To mitigate foreground suppression in one-query OVS, BAPA (Wang et al., 2026b) assigns hard pseudo-labels and refreshes the top $K = 20\%$ of each predicted class as dynamic anchors. At adaptation step t, let $c \in \{\mathrm{bg}, \mathrm{fg}\}$ denote the background or foreground class, F_c^t its predicted pixel set, $|F_c^t|$ the number of pixels in that set, m_c^t the selected anchor mass, and $\bar{\ell}_c^t$ the mean anchor loss. The class weight B_c^t weights this loss in the binary objective

$$\begin{aligned} \mathcal{L}_{\mathrm{bal}}^t &= \sum_{c\in\{\mathrm{bg},\mathrm{fg}\}} B_c^t \bar{\ell}_c^t, \\ B_c^t &= \frac{1}{2}\frac{m_c^t}{\max(1, m_c^t)}, \qquad m_c^t = K|F_c^t|, \quad K = 0.2. \end{aligned} \tag{3}$$

The class-loss term is zero when $F_c^t = \varnothing$. Otherwise, classwise normalization allows even a small false foreground region to receive the same maximum loss weight as a supported object. Updating on these false anchors can enlarge the region, which then supplies further erroneous supervision at the next step.

BAPA effectively mitigates foreground suppression through balanced anchor selection, but its original formulation considers only a single foreground query. Multi-BAPA denotes its direct multi-query extension: each predicted class refreshes its own top-20% anchors, while all classes share one softmax and parameter update. Building on this anchor formulation, SFE-Guard controls the resulting multi-query feedback. Reinforcing an incorrect class also affects its competitors, and dynamic anchor refresh can propagate these errors across steps. To control the resulting spurious foreground expansion, SFE-Guard introduces CDI to limit supervision from unreliable predictions and HRC to correct candidate updates before they reinforce these errors.

3.2 Method Overview

SFE-Guard applies CDI and HRC at every adaptation step. CDI derives a fixed credit γ_k from the zero-shot prediction for each query and uses it to bound supervision from dynamically refreshed anchors. The optimizer then proposes a candidate step u_t, producing $\hat{\delta}_{t+1} = \delta_t + u_t$. This proposal is termed the *candidate update*. HRC evaluates its aggregate and background risks against the protected reference and applies the smallest required correction before commitment.

$$p^0 \longrightarrow \{\gamma_k\} \xrightarrow{\text{CDI}} \mathcal{L}_t \xrightarrow{\text{optimizer}} \hat{\delta}_{t+1} \xrightarrow{\text{HRC}} \delta_{t+1}. \tag{4}$$

Only the committed state determines the prediction and anchors used in the next step. Figure 3 summarizes the adaptation loop, and Section 3.5 defines its reference distributions.

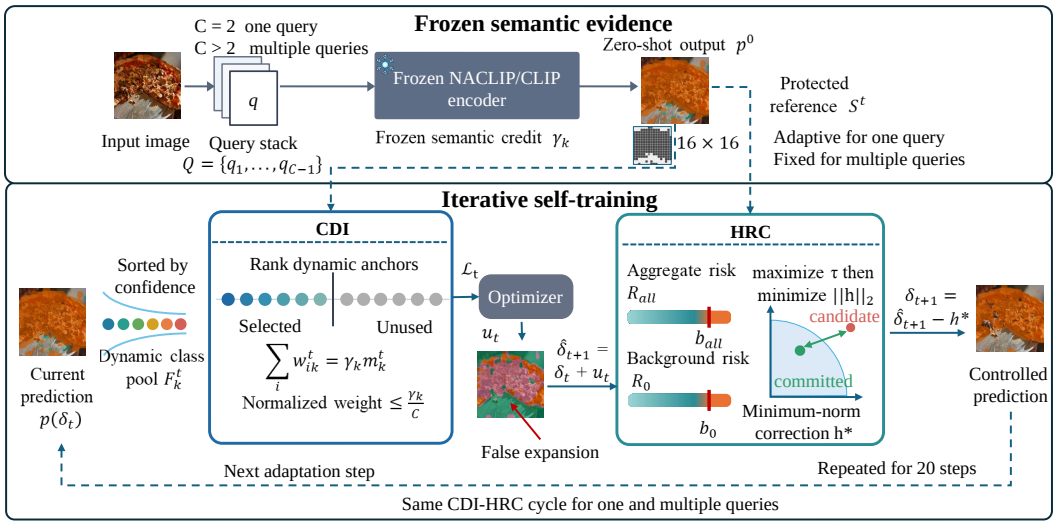


Figure 3: **SFE-Guard architecture.** Initial predictions provide frozen semantic credit for CDI and initialize the protected reference for HRC. CDI bounds supervision from dynamic anchors, and HRC corrects candidate updates before commitment. The process repeats for 20 adaptation steps.

3.3 Credit-Denominated Issuance

CDI first asks whether the zero-shot prediction contains spatially concentrated evidence for a query. It converts this fixed evidence into a limit on the supervision that the query can issue during adaptation. For class k, let $p_{ik} = p_{ik}^0$ be its zero-shot probability at pixel i. CDI compares that class with all non-k classes through

$$p_i^{(k)} = \big(1 - p_{ik},\ p_{ik}\big) \qquad \text{(other classes, class } k\text{)}. \tag{5}$$

Here, $1 - p_{ik}$ combines all other classes as the alternative. This provides the same binary comparison for one- and multi-query predictions. The two probabilities are represented by their logarithms and averaged by area on a 16×16 grid, producing $\bar{z}^0 \in \mathbb{R}^{2\times16\times16}$. Its two channels are the log probabilities of the non-k and k entries in Eq. (5). At each of the $N = 256$ grid positions, evidence for class k is the class–alternative margin $d_i = \bar{z}_{i,k}^0 - \bar{z}_{i,\neg k}^0$. The calculation is applied once to each query, and k is omitted below. The semantic credit is

$$\gamma = \sigma\left(\log \frac{1}{N} \sum_{i=1}^{N} e^{d_i}\right). \tag{6}$$

Here σ is the logistic sigmoid, so $0 \leq \gamma \leq 1$ and stronger frozen query evidence receives more credit. The aggregation in Eq. (6) assigns credit when a compact region provides strong evidence, without requiring a minimum foreground area. The sigmoid maps this evidence to the unit interval used to scale issued supervision. Area averaging uses the native 16×16 visual token layout and prevents dense interpolation from counting replicated evidence as additional support. Equation (6) is computed once from the initial zero-shot reference and then frozen. At step t, let F_k^t contain the pixels whose current winner is class k, and set $m_k^t = 0.2|F_k^t|$. These pixels are sorted by their current probability for class k and assign weight to the most confident pixels until the total reaches $\gamma_k m_k^t$. Selected pixels receive weight 1, with the last pixel assigned any remaining fractional weight to match this issued amount exactly. Thus $w_{ik}^t \in [0, 1]$ and $\sum_i w_{ik}^t = \gamma_k m_k^t$; weights outside F_k^t are zero. The foreground loss for query k is

$$\mathcal{L}_k^t = -\frac{\sum_i w_{ik}^t \log p_{ik}(\delta_t)}{C \max(1, m_k^t)}. \tag{7}$$

Each foreground class uses its semantic credit γ_k, while the background credit is fixed at $\gamma_0 = 1$. Background otherwise follows the same normalized classwise rule. The total loss is $\mathcal{L}_t = \sum_{k=0}^{C-1} \mathcal{L}_k^t$.

A class with no predicted pixels provides no anchors and has zero loss at that step. Because its normalized issued mass is $\sum_i w_{ik}^t/[C\max(1,m_k^t)] = \gamma_k m_k^t/[C\max(1,m_k^t)] \le \gamma_k/C$. Thus expanding a predicted region cannot create additional credit. This loss advances the optimizer exactly once and produces a candidate step $u_t \in \mathbb{R}^{C\times d}$.

3.4 Hierarchical Risk Correction

HRC uses the initial zero-shot prediction to identify the pixels associated with each class. Let $I_c^0 = \{i : \arg\max_r p_{ir}^0 = c\}$ be the pixels assigned to class c at the initial prediction. Each pixel in this set receives the same weight, $w_{ic}^0 = \mathbb{I}[i \in I_c^0]/|I_c^0|$ when I_c^0 is nonempty and zero otherwise, where $\mathbb{I}$ is the indicator function. The weights therefore sum to one for every initially predicted class, so a large region does not outweigh a small region. Let S^t be the probability reference that HRC protects; its construction for one and multiple queries is given in Section 3.5. For any state δ, HRC measures how much probability has been removed from the initial class assignments. This class-specific quantity is called the *risk* $R_c(\delta)$. It is the average squared drop in log probability below the reference,

$$R_c(\delta) = \sum_i w_{ic}^0 [\log S_{ic}^t - \log p_{ic}(\delta)]_+^2. \tag{8}$$

Here $[x]_+ = \max(x, 0)$, so a pixel contributes only when its current probability for class c falls below the protected reference. At step t, the optimizer proposes u_t, giving $\hat{\delta}_{t+1} = \delta_t + u_t$. HRC reviews this proposed state before it is committed. The correction is constrained by two risks. The aggregate risk $R_{\text{all}} = \sum_{c=0}^{C-1} R_c$ summarizes probability losses across all initially predicted classes. The background risk R_0 checks the background pixels separately, which prevents reductions in other class risks from offsetting an increase in background risk. For each checked quantity $j \in \{0, \text{all}\}$, the proposed update introduces the nonnegative risk increase

$$b_j = \left[R_j(\hat{\delta}_{t+1}) - R_j(\delta_t)\right]_+, \qquad j \in \{0, \text{all}\}. \tag{9}$$

Let $g_j = \nabla_\delta R_j(\delta)|_{\hat{\delta}_{t+1}}$ be the gradient of this risk at the proposed state. Near that state, $g_j^\top h$ estimates how much subtracting a correction h will reduce the risk. HRC subtracts $h \in \mathbb{R}^{C\times d}$ from the proposed update, producing $\delta_t + (u_t - h) = \hat{\delta}_{t+1} - h$. The dot product and norms treat the $C \times d$ entries of g_j, h, and u_t as vectors. The scalar τ is the fraction of both risk increases removed by the correction. Feasible pairs (τ, h) satisfy

$$\mathcal{F}_t = \left\{(\tau, h) \,\middle|\, 0 \le \tau \le 1,\ \|h\|_2 \le \|u_t\|_2,\ g_j^\top h \ge \tau b_j\ \forall j \in \{0, \text{all}\}\right\}. \tag{10}$$

The correction is bounded by the size of the proposed update. The last constraint requires the linearized reduction to cover at least a fraction τ of both risk increases. HRC first maximizes feasible τ, then chooses the correction with the smallest norm among the solutions that attain it.

$$\tau^* = \max\{\tau : (\tau, h) \in \mathcal{F}_t\}, \quad h^* = \arg\min\{\|h\|_2 : (\tau^*, h) \in \mathcal{F}_t\}. \tag{11}$$

The selected correction is h^*. HRC applies $\tilde{u}_t = u_t - h^*$ and moves to $\delta_{t+1} = \delta_t + \tilde{u}_t = \hat{\delta}_{t+1} - h^*$. When the candidate does not increase either risk, $h^* = 0$. Thus, HRC maximizes the shared first-order reduction of the aggregate and background risks, and the committed state is evaluated by the full segmentation model at the next adaptation step.

Proposition 1. *Assume that each active R_j has an L_j-Lipschitz gradient on the segment from $\hat{\delta}_{t+1}$ to $\hat{\delta}_{t+1} - h^*$. The HRC solution is the lexicographic optimum that maximizes common linearized risk reduction and then minimizes correction norm. The corrected state satisfies*

$$\begin{aligned} \epsilon_j^t &= R_j(\hat{\delta}_{t+1} - h^*) - R_j(\hat{\delta}_{t+1}) + g_j^\top h^*, \\ R_j(\delta_{t+1}) - R_j(\delta_t) &\le (1 - \tau^*) b_j + \epsilon_j^t, \qquad \epsilon_j^t \le \frac{L_j}{2}\|h^*\|_2^2. \end{aligned} \tag{12}$$

Consequently, $g_j^\top h^* \ge R_j(\hat{\delta}_{t+1}) - R_j(\delta_t) + \frac{L_j}{2}\|h^*\|_2^2$ is a sufficient condition for the risk after correction not to exceed its value before the step.

Proof. The feasible set is compact and convex. Maximizing τ followed by Euclidean projection of the origin onto the maximizing correction set gives the stated lexicographic solution. The descent lemma at $\hat{\delta}_{t+1}$ gives the bound on ϵ_j^t; substituting $g_j^\top h^* \ge \tau^* b_j$ and $R_j(\hat{\delta}_{t+1}) - R_j(\delta_t) \le b_j$ gives Eq. (12). □

The proposition characterizes the local calculation. The term ϵ_j^t is the difference between the linearized prediction and the exact risk after correction; Eq. (12) bounds its size. If no joint correction is feasible, HRC retains the candidate produced under CDI.

3.5 Reference Distributions

The protected reference S^t follows the probability structure of the query setting. With one foreground query ($|\mathcal{Q}| = 1$, $C = 2$), each pixel has only foreground and background outcomes. Increasing its foreground probability decreases its background probability by the same amount. Each image–query pair starts independently from its zero-shot prediction. Let $p_{i\mathrm{fg}}^t$ be the accepted foreground probability before step t, and let $\Delta_i^t = p_{i\mathrm{fg}}^t - p_{i\mathrm{fg}}^0$ denote its change from the zero-shot value. The reference is

$$S_{i\mathrm{fg}}^t = p_{i\mathrm{fg}}^0 + \min(\Delta_i^t, 0) + \frac{1}{2}\max(\Delta_i^t, 0), \qquad S_{i\mathrm{bg}}^t = 1 - S_{i\mathrm{fg}}^t. \tag{13}$$

When $p_{i\mathrm{fg}}^t$ has decreased, $S_{i\mathrm{fg}}^t = p_{i\mathrm{fg}}^t$. When it has increased, $S_{i\mathrm{fg}}^t$ is halfway between $p_{i\mathrm{fg}}^0$ and $p_{i\mathrm{fg}}^t$. Thus, only half of an accepted foreground expansion becomes protected at the next step.

With multiple foreground queries ($|\mathcal{Q}| > 1$, $C = |\mathcal{Q}| + 1$), all queried classes compete in one C-class softmax and share one $C \times d$ residual update. The reference remains fixed at the initial prediction, $S^t = p^0$, throughout adaptation.

4 Experiments

4.1 Experimental Setup

Settings and protocol. One-query OVS evaluates each text query against background, whereas multi-query OVS jointly predicts all dataset classes through one shared softmax and update. During adaptation and prediction, methods receive only the test image and text queries, not annotations or information about which queried classes occur in the image. Evaluation covers 1,449 VOC images, 5,000 COCO images, and 500 Cityscapes images, with 20, 80, and 19 foreground classes, respectively. All dataset classes are queried for each image, including those absent from it. All methods use the same NACLIP ViT-L/14 backbone, preprocessing, class names, evaluated images and queries.

Implementation. SFE-Guard uses three text templates per class in both query settings and 20 adaptation steps with a learning rate of 10^{-3}. Only the text prototype residual is updated (Wang et al., 2026b); the encoders remain frozen. Adapted parameters and optimizer states are reset for each episode. Baselines follow their respective adaptation procedures.

Metrics and comparisons. Present Dice first averages Dice over images in which each class is present, then averages across classes within each dataset. Absent FG and absent class area measure the pixel fraction assigned to absent queries in the respective settings; PA is categorical pixel accuracy. In one-query OVS, prediction collapse denotes a mask for a present query whose predicted foreground occupies less than 1% or more than 99% of the image. In multi-query OVS, it denotes a prediction in which one class occupies more than 95% of the image. Cross-dataset scores give each dataset equal weight. Further protocol and implementation details are provided in Appendix A.

4.2 Results

Comparison with existing methods. SFE-Guard achieves the highest mean Present Dice in both query settings and the highest overall mean (Table 1). For one-query OVS, SFE-Guard attains the strongest average across datasets, improving BAPA by 3.066 points and leading on every dataset. For multi-query OVS, it leads on VOC and COCO, improving Multi-BAPA by 8.469 points and MLMP by 3.177 points. Averaged across both settings, SFE-Guard attains 37.258% Present Dice with a 3.864% collapse rate.

Table 1 also shows that MLMP is 2.174 points below zero-shot with one query but 0.749 points above it with multiple queries. BAPA/Multi-BAPA shows the opposite pattern: it exceeds zero-shot by 0.936 points with one query but falls 4.543 points below it with multiple queries. SFE-Guard leads in both settings. Extended false-foreground and pixel accuracy results are reported in the appendix.

Table 1: **One-query and multi-query OVS results (%).** Dice is Present Dice averaged over classes, and Coll. is prediction collapse. "Avg." equally weights the three datasets; "Overall" averages the two query settings. † denotes a dense OVS port. Best and second-best results are bold and underlined.

	One-query OVS ($\|\mathcal{Q}\| = 1$)					Multi-query OVS ($\|\mathcal{Q}\| > 1$)					Overall	
Method	VOC	COCO	City	Avg. Dice↑	Avg. Coll.↓	VOC	COCO	City	Avg. Dice↑	Avg. Coll.↓	Avg. Dice↑	Avg. Coll.↓
Reference without adaptation												
Zero-shot (Hajimiri et al., 2025)	51.592	36.175	15.072	34.280	17.208	46.298	33.742	16.886	32.309	0.589	33.294	8.898
Confidence Sharpening												
TENT† (Wang et al., 2021)	52.434	36.796	14.790	34.673	19.998	46.050	33.681	16.690	32.141	1.304	33.407	10.651
ETA† (Niu et al., 2022)	51.526	36.128	15.140	34.265	14.398	46.294	33.744	16.885	32.308	0.589	33.286	7.494
TPT† (Shu et al., 2022)	24.516	9.963	5.025	13.168	87.198	41.265	25.519	5.563	24.116	33.674	18.642	60.436
SAR† (Niu et al., 2023)	51.526	36.126	15.139	34.264	14.404	46.287	33.741	16.881	32.303	0.589	33.283	7.497
CLIPArTT (Vargas Hakim et al., 2025)	30.338	14.713	6.945	17.332	39.884	37.935	25.271	13.739	25.648	15.509	21.490	27.696
MLMP (Noori et al., 2025)	50.465	33.634	12.219	32.106	27.899	47.308	32.537	**19.329**	33.058	5.496	32.582	16.698
O-TPT† (Sharifdeen et al., 2025)	29.442	13.969	5.429	16.280	81.659	41.507	25.703	5.841	24.350	32.330	20.315	56.994
Anchor Self-Training												
Pseudo-labeling (PL) (Lee, 2013)	43.550	31.755	13.180	29.495	40.743	39.900	30.312	13.941	28.051	16.177	28.773	28.460
BAPA / Multi-BAPA (Wang et al., 2026b)	55.941	35.362	14.345	35.216	8.291	38.188	27.112	17.998	27.766	**0.000**	31.491	4.145
SFE-Guard	**58.622**	**39.074**	**17.148**	**38.281**	**7.729**	**53.389**	**36.317**	18.998	**36.235**	**0.000**	**37.258**	**3.864**

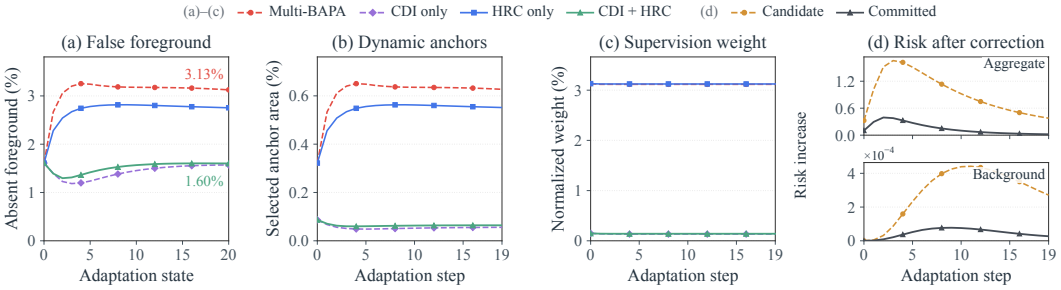


Figure 4: **Adaptation dynamics.** (a) Absent foreground. (b) Dynamic anchors. (c) Supervision weight. (d) Candidate and committed risk under CDI+HRC. Aggregate and background risk use separate scales.

Component ablation. Table 2 isolates CDI and HRC under the same prompts, optimization, and evaluation protocol. CDI+HRC attains the highest Present Dice in the one-query, multi-query, and averaged evaluations. CDI-only attains the lowest multi-query absent area and the highest multi-query and averaged PA, showing that credit-bounded supervision directly limits false foreground. Adding HRC to CDI raises Present Dice. CDI limits the false supervision entering each update, while HRC corrects the candidate update before it refreshes the next anchor set. Their effects over the adaptation trajectory are examined in Figure 4.

Qualitative results. Figure 5 illustrates both query settings. For the couch query, MLMP suppresses the foreground from 4.55% of the image at its initialization to zero. SFE-Guard preserves the object and attains a Dice of 0.82, compared with 0.51 for BAPA and 0.00 for MLMP. In the multi-query example, Multi-BAPA expands incorrectly labeled foreground from 71.81% of the image at initialization to 89.81% after adaptation. SFE-Guard reduces this area to 22.43% and achieves a present-class macro Dice of 0.79, compared with 0.48 for Multi-BAPA and 0.67 for MLMP.

Effect of HRC on candidate updates. HRC modifies 56.697% of one-query candidate updates and 96.461% of multi-query candidate updates. Relative to the uncorrected candidate, the correction reduces mean excess aggregate risk by 84.0%. On 99.983% of steps, the committed update has no greater aggregate risk than the candidate.

Spurious foreground expansion during adaptation. Figure 4 traces Multi-BAPA, CDI-only, HRC-only, and CDI+HRC throughout adaptation. Panels (a)–(c) show the SFE feedback loop. Under Multi-BAPA, absent foreground rises from 1.610% to 3.126%, while dynamic anchor area grows from 0.322% to 0.627% of the image. The enlarged false region therefore produces more anchors and more supervision at later steps. CDI-only reduces the final absent area to 1.570% and lowers issued supervision by 95.7% relative to Multi-BAPA. HRC-only limits the expansion of absent foreground

Table 2: **Component ablation (%).** Checkmarks indicate active components; the empty row is BAPA for one query and Multi-BAPA for multiple queries. PD, Abs., and PA denote Present Dice, absent foreground or absent class area, and pixel accuracy. All rows share the same prompts and optimization protocol. Best results are bold and second-best results are underlined.

Components		One-query			Multi-query			Mean		
CDI	HRC	PD↑	Abs.↓	PA↑	PD↑	Abs.↓	PA↑	PD↑	Abs.↓	PA↑
		35.216	26.827	73.901	27.766	3.179	20.833	31.491	15.003	47.367
✓		37.374	14.751	85.540	34.516	**1.570**	**50.336**	35.945	8.160	**67.938**
	✓	36.728	19.163	81.272	33.389	2.754	26.801	35.058	10.959	54.037
✓	✓	**38.281**	**13.757**	**86.496**	**36.235**	1.602	46.042	**37.258**	**7.679**	66.269

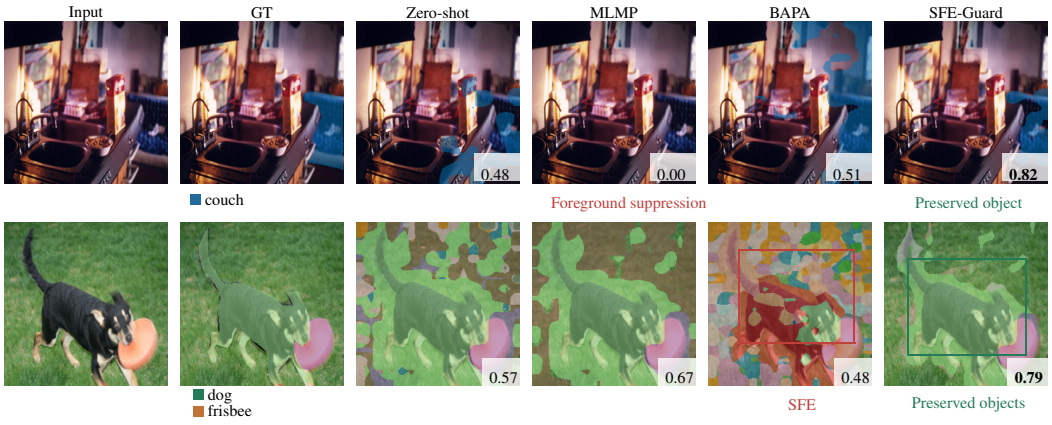


Figure 5: **Representative segmentation results.** Top: one-query couch segmentation. Bottom: multi-query segmentation with dog and frisbee present. Columns show the input, ground truth, Zero-shot, MLMP, BAPA or Multi-BAPA, and SFE-Guard. Numbers report foreground Dice (top) and present-class macro Dice (bottom); annotations mark foreground suppression and SFE.

to $+1.144$ points. Panel (d) shows how HRC changes the candidate before it becomes the next state; CDI+HRC ends with 1.602% absent area. The trajectory separates CDI's control of pseudo supervision from HRC's correction of the candidate update.

Computational cost. Both methods reuse one visual encoding per RGB. Multi-BAPA uses 22 dense callbacks; HRC evaluates 20 additional candidates before commitment, giving 42 callbacks. Under the matched GH200 protocol in Table B.2, cached adaptation takes 0.346 versus 0.363 seconds per RGB, so HRC adds about 5.0% runtime. The full method therefore incurs only a small additional computational cost.

5 Conclusion

Open-vocabulary segmentation must process text-defined classes that may be absent from an image. Anchor based adaptation can then repeatedly enlarge unsupported foreground through its own pseudo supervision. SFE-Guard is the first test-time adaptation method to control this recursive SFE in both one-query and multi-query OVS. CDI bounds the supervision issued from each prediction, and HRC corrects the candidate update before it becomes the next prediction.

Across all three datasets, SFE-Guard achieves the strongest Present Dice among evaluated methods in both settings. It removes 77.9% of the excess false foreground introduced by self training with one query and reduces multi-query absent class area to the zero-shot level. Future work will combine stronger evidence that a queried class occurs in the image with class specific and nonlinear risk constraints, and extend the framework to additional OVS architectures, domains, and streaming inputs.

AI USE STATEMENT

Generative AI assisted with figure preparation and manuscript editing. The authors verified all claims, citations, figures, code, and reported results and take full responsibility for the submission.

Reproducibility statement. The paper and appendix specify the datasets, evaluation protocol, implementation details, baseline configurations, and statistical analyses. An anonymous core implementation of CDI and HRC accompanies the submission; the complete evaluation implementation will be released with the accepted version.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 22–31, 2017.

Luca Barsellotti, Lorenzo Bianchi, Nicola Messina, Fabio Carrara, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, and Rita Cucchiara. Talking to DINO: Bridging self-supervised vision backbones with language for open-vocabulary segmentation. In *IEEE/CVF International Conference on Computer Vision*, pp. 22025–22035, 2025. doi: 10.1109/ICCV51701.2025.02045.

Dimitri P. Bertsekas. On the Goldstein–Levitin–Polyak gradient projection method. *IEEE Transactions on Automatic Control*, 21(2):174–184, 1976. doi: 10.1109/TAC.1976.1101194.

Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. CAT-Seg: Cost aggregation for open-vocabulary semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4113–4123, 2024.

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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11583–11592, 2022.

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.

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, 2022.

Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5061–5071, 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ProxyCLIP: Proxy attention improves CLIP for open-vocabulary segmentation. In *European Conference on Computer Vision*, 2024.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*, 2013.

Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022.

Xinzhuo Li, Adheesh Juvekar, Jiaxun Zhang, Xingyou Liu, Muntasir Wahed, Kiet A. Nguyen, Yifan Shen, Tianjiao Yu, and Ismini Lourentzou. Counterfactual segmentation reasoning: Diagnosing and mitigating pixel-grounding hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 7450–7460, 2026.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7061–7070, 2023.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 6028–6039, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 2014.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning*, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023.

Mehrdad Noori, David Osowiechi, Gustavo Vargas Hakim, Ali Bahri, Moslem Yazdanpanah, Sahar Dastani, Farzad Beizaee, Ismail Ben Ayed, and Christian Desrosiers. Test-time adaptation of vision-language models for open-vocabulary semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. DenseCLIP: Language-guided dense prediction with context-aware prompting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Ashshak Sharifdeen et al. O-TPT: Orthogonality constraints for calibrating test-time prompt tuning in vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19942–19951, 2025.

Gyungin Shin, Weidi Xie, and Samuel Albanie. ReCo: Retrieve and co-segment for zero-shot transfer. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Heeseong Shin, Chaehyun Kim, Sunghwan Hong, Seokju Cho, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Towards open-vocabulary semantic segmentation without semantic labels. In *Advances in Neural Information Processing Systems*, volume 37, pp. 9153–9177, 2024. doi: 10.52202/079017-0290.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, 2022.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, pp. 9229–9248, 2020.

Gustavo A. Vargas Hakim et al. CLIPArTT: Adaptation of CLIP to new domains at test time. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7092–7101, 2025.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Feng Wang, Jieru Mei, and Alan Yuille. SCLIP: Rethinking self-attention for dense vision-language inference. In *European Conference on Computer Vision*, 2024.

Haozhe Wang, Jintao Cheng, Weibin Li, and Xiaoyu Tang. Test-time prototype adaptation for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2608.08290*, 2026a.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022.

Zhengshan Wang, Joshua Charles Webster-Ford, Yifei Tian, Xinxin Wang, Long Chen, and Weiping Ding. Balanced prompt adaptation against entropy-induced collapse for test-time binary segmentation. *arXiv preprint arXiv:2609.21743*, 2026b. URL https://arxiv.org/abs/2609.21743.

Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. CLIP-DINOiser: Teaching CLIP a few DINO tricks for open-vocabulary semantic segmentation. In *European Conference on Computer Vision*, 2024.

Monika Wysoczańska, Antonin Vobecky, Amaia Cardiel, Tomasz Trzciński, Renaud Marlet, Andrei Bursuc, and Oriane Siméoni. Test-time contrastive concepts for open-world semantic segmentation with vision-language models. *Transactions on Machine Learning Research*, 2025.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. GroupViT: Semantic segmentation emerges from text supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18134–18144, 2022.

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 *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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2945–2954, 2023b.

Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang Yoo. C-TPT: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *International Conference on Learning Representations*, 2024.

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

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836, 2020.

Marvin Zhang, Sergey Levine, and Chelsea Finn. MEMO: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In *European Conference on Computer Vision*, 2022.

Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. ZegCLIP: Towards adapting CLIP for zero-shot semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11175–11185, 2023. doi: 10.1109/CVPR52729.2023.01075.

Yang Zou, Zhiding Yu, B. V. K. Vijaya Kumar, and Jinsong Wang. Domain adaptation for semantic segmentation via class-balanced self-training. In *European Conference on Computer Vision*, 2018.

A Reproducibility Details

This appendix specifies the evaluation protocol, reports extended quantitative results and HRC diagnostics, documents implementation details, and then presents generalization tests, qualitative comparisons, per-dataset results, and paired statistical analyses.

Table A.1: **OVS evaluation sets.** One-query OVS evaluates every RGB–class pair; multi-query OVS evaluates all classes jointly per RGB.

Dataset	RGBs	FG queries	1Q episodes	Present pairs
VOC2012 val	1,449	20	28,980	2,148
COCO val	5,000	80	400,000	14,631
Cityscapes val	500	19	9,500	6,005
Total	6,949	119	438,480	22,784

Table B.1: **Accuracy and false foreground results (%).** PD, Abs., and PA denote Present Dice averaged over classes, absent foreground or absent class area, and pixel accuracy. Each setting equally weights the three datasets; Mean averages both settings. † denotes a dense OVS port. Best results are bold and second-best results are underlined.

	One-query OVS			Multi-query OVS			Mean		
Method	PD↑	Abs.↓	PA↑	PD↑	Abs.↓	PA↑	PD↑	Abs.↓	PA↑
Reference without adaptation									
Zero-shot	34.280	10.049	89.313	32.309	1.604	33.819	33.294	5.826	61.566
Confidence Sharpening									
TENT†	34.673	8.782	90.678	32.141	1.462	34.150	33.407	5.122	62.414
ETA†	34.265	10.056	89.312	32.308	1.602	33.828	33.286	5.829	61.570
TPT†	13.168	5.948	**93.479**	24.116	1.134	28.901	18.642	3.541	61.190
SAR†	34.264	10.054	89.314	32.303	1.603	33.820	33.283	5.828	61.567
CLIPArTT	17.332	24.947	72.858	25.648	1.448	30.670	21.490	13.198	51.764
MLMP	32.106	**4.647**	92.125	33.058	0.953	45.855	32.582	**2.800**	**68.990**
O-TPT†	16.280	6.002	93.455	24.350	1.141	28.956	20.315	3.571	61.206
Anchor Self-Training									
Pseudo-labeling	29.495	8.921	90.041	28.051	**0.938**	31.475	28.773	4.930	60.758
BAPA / Multi-BAPA	35.216	26.827	73.901	27.766	3.179	20.833	31.491	15.003	47.367
SFE-Guard	**38.281**	13.757	86.496	**36.235**	1.602	**46.042**	**37.258**	7.679	66.269

A.1 Evaluation Protocol

Evaluation uses VOC2012 val (Everingham et al., 2010), COCO val2017 (Lin et al., 2014), and Cityscapes val (Cordts et al., 2016) without filtering images or queries according to whether a class occurs in an image. One-query OVS evaluates every RGB–class pair independently. Multi-query OVS supplies all native dataset classes and returns one categorical prediction per RGB.

Metrics and aggregation. Present Dice first averages each class over images where that class is present and then averages classes. Absent FG is the fraction of pixels assigned to a queried class with no ground-truth pixels. In multi-query OVS, absent class area applies the same definition to each absent class in the shared categorical mask. PA is categorical accuracy over all pixels. Each dataset score averages the eligible native classes, and the reported mean across datasets gives each dataset equal weight. The model never receives annotations indicating whether a queried class occurs in an image. Collapse measures a one-query present mask whose predicted foreground occupies below 1% or above 99% of the image. In multi-query OVS, dominant collapse denotes an image for which one categorical class occupies above 95% of the prediction.

B Extended Quantitative Results

B.1 Accuracy and False Foreground Results

Table B.1 reports false foreground and pixel accuracy metrics for every method in the primary comparison. Because a a prediction dominated by background can obtain low absent foreground while missing present objects, Abs. is interpreted together with PD and PA.

The comparison in Table B.1 measures false foreground added by adaptation for absent queries, relative to the initial zero-shot prediction. BAPA and Multi-BAPA add 16.778 points of absent foreground in one-query OVS and 1.575 points of absent class area in multi-query OVS. SFE-Guard limits the one-query addition to 3.708 points, removing 77.9% of the adaptation-induced excess.

Table B.2: **Multi-query compute cost on one GH200.** Time is seconds per RGB, reported as mean $\pm$ standard deviation over three repetitions of the same 300 images. TPT and O-TPT use one batch of 64 augmented global views and one final dense prediction; other entries report dense callbacks per RGB.

Method	Model evaluations	Time (s/RGB)
Pseudo-labeling	21 dense	0.856 ± 0.001
TENT	2 dense	0.079 ± 0.001
ETA	2 dense	0.072 ± 0.000
TPT	64 views + 1 dense	0.677 ± 0.001
SAR	3 dense	0.112 ± 0.000
CLIPArTT	21 dense	2.128 ± 0.004
MLMP	11 dense	0.475 ± 0.001
O-TPT	64 views + 1 dense	0.679 ± 0.001
Multi-BAPA	22 dense	0.346 ± 0.002
SFE-Guard	42 dense	0.363 ± 0.000

Table C.1: **HRC diagnostics over all update steps.** The maximum KKT residual is 5.03×10^{-15}, with no radius violation.

Statistic	One-query	Multi-query
Reviewed update steps	8,769,600	138,980
HRC activation (%)	56.697	96.461
Mean correction norm	0.001551	0.006748
Changed update steps (%)	–	96.451
Hard prediction disagreement (%)	–	5.161
Mean L_1 change in soft probabilities (%)	–	0.205

Its one-query Absent FG and PA are 13.757% and 86.496%, respectively. In multi-query OVS, its absent class area is 1.602%, near the 1.604% zero-shot level and removing the Multi-BAPA-induced increase.

B.2 Computational Cost

Table B.2 measures every adaptation method on the same GH200. The benchmark uses 300 fixed RGBs, with 100 from each dataset, nine untimed warm-up images, and three timed repetitions. Decoded inputs are resident before timing, which covers method specific adaptation and final prediction. The reported values are the mean and sample standard deviation across repetition means. Under this matched protocol, HRC increases the anchor pipeline runtime from 0.346 to 0.363 seconds per RGB, a $1.05\times$ cost.

C HRC Mechanism Diagnostics

C.1 Activation and Numerical Checks

Table C.1 reports how often HRC corrects a candidate update and verifies the numerical conditions of the correction step.

Exact nonlinear behavior after HRC. No step with a positive risk increase and $\tau^* = 0$ was observed in either the one-query or multi-query evaluation. Table C.2 evaluates the exact account values before the candidate, at the candidate, and after HRC commitment. The multi-query result bank contains all 138,980 update steps over all three datasets. Aggregate risk is almost always increased by the candidate, yet HRC returns an aggregate risk no larger than that candidate on 99.983% of steps. For background, HRC produces positive changes on more steps but makes their average magnitude much smaller. The following tables report both the frequency and magnitude of these changes.

HRC controls the risk carried by a candidate update. The increase rate of the background account rises from 8.647% to 12.574% because this rate counts every positive change, including changes close to zero. Over the same adaptation runs, the mean positive excess falls from 0.000258 to 0.000041, an 84.1% reduction. The joint local correction can therefore allow small positive changes on more steps

Table C.2: **Exact multi-query risk changes after HRC (%).** Increases are measured from the accepted state before the step; the final column compares committed and candidate risks.

Account	Candidate increase	Committed increase	Committed $\leq$ candidate
Aggregate	96.443	96.430	99.983
Senior background	8.647	12.574	91.238

Table C.3: **Distribution of exact risk excess over 138,980 multi-query updates.** Values are measured from the accepted state before the step.

	Raw candidate			Committed state			Mean reduction
Account	Mean	Median	95th pct.	Mean	Median	95th pct.	$\downarrow$
Aggregate	1.152581	0.732	3.737	0.184271	0.081	0.751	84.0%
Senior background	0.000258	0.000000	0.000031	0.000041	0.000000	0.000011	84.1%

while substantially reducing their average size. This behavior preserves useful foreground recovery while limiting the magnitude of background degradation.

Comparison of commitment rules. Table C.4 evaluates a stricter commitment rule that recomputes both exact nonlinear accounts and rejects a step whenever either account increases. Exact monotonicity rejects 99.999% of candidate updates and returns within 0.001 Present Dice points of zero-shot prediction. Proportional HRC lowers absent area by 0.008 points while gaining 3.927 Present Dice points and 12.053 points in pixel accuracy over exact rejection. The result supports HRC's proportional risk reduction rule as the mechanism that retains useful adaptation while reducing candidate risk.

Robustness of the protected reference. Table C.5 changes only the HRC reference while keeping semantic CDI, prompts, optimizer, update count, and complete evaluation population fixed. For one query, the proposed half-update reference gives the highest PD, lowest absent foreground, and highest PA. It follows decreases in the accepted foreground while protecting half of each increase. A fixed step-zero reference is slightly less accurate, and a fully rolling reference protects newly expanded foreground too quickly. For multiple queries, the fixed reference gives the highest PD, while rolling improves absent area and PA at the cost of PD. These matched results support the references used in the main evaluation.

For one-query OVS, the logs record HRC activation, correction norm, risk reduction, KKT residual, and radius compliance over all 8,769,600 steps. Differences between the candidate and committed predictions require both dense outputs and are therefore reported from the multi-query update log, which stores them at every step. The one-query Present Dice, Absent FG, PA, and correction statistics use all evaluated episodes.

D Implementation Details

D.1 Model and Adaptation Settings

Backbone and preprocessing. All new open-vocabulary segmentation experiments use the same NACLIP ViT-L/14 checkpoint (SHA-256 b8cca3fd41ae0c99ba7e8951adf17d26 7cdb84cd88be6f7c2e0eca1737a03836) using 32-bit floating-point arithmetic. RGB inputs are resized to 224×224, scaled to $[0, 1]$, and normalized by mean $(.48145466, .4578275, .40821073)$ and standard deviation $(.26862954, .26130258, .27577711)$. Dense features use the official NACLIP mean of outputs $-1, \ldots, -12$ and are bilinearly interpolated to 224^2. Each RGB is visually encoded once and cached.

Prompts and trainable state. For one-query OVS, SFE-Guard uses three foreground templates *a photo of a q.*, *an image of a q.*, and *a picture of a q.* The paired background templates are *a scene without a q.*, *a background photo.*, and *an image with no q.* Multi-query OVS uses the same three foreground templates and the background templates *a background scene.*, *a background photo.*, and

Table C.4: **Multi-query commitment rules (%).** Both variants use identical CDI, candidate updates, prompts, and evaluation images.

Commitment rule	PD↑	Abs.↓	PA↑	Rejected steps↓
Zero-shot	32.309	1.604	33.819	–
Exact nonlinear rejection	32.308	1.610	33.989	99.999
HRC common risk reduction	**36.235**	**1.602**	**46.042**	0.000

Table C.5: **Protected-reference robustness on complete runs (%).** Values are equal-domain means across all three datasets.

Setting	Protected reference	PD↑	Abs.↓	PA↑
One query	Fixed at step zero	38.184	14.397	85.896
	Fully rolling	37.923	17.314	83.244
	Half of foreground increase	**38.281**	**13.757**	**86.496**
Multiple queries	Fixed at step zero	**36.235**	1.602	46.042
	Fully rolling	35.748	**1.554**	**49.266**

an image background. Each template embedding is normalized, the embeddings are averaged per class, and the result is normalized again. All component and step-by-step studies use this three-template construction in both settings. For methods that adapt prompts, the trainable tensor is a residual initialized at zero, added to the text prototypes, and normalized by row before every dense evaluation. Each one-query image–text episode or multi-query RGB run starts with a fresh optimizer and seed 2027.

Credit. CDI computes each semantic credit on the 16×16 grid averaged by area using Eq. (6). No target statistic, area threshold, learned presence classifier, or branch specific to a dataset enters the credit. Equation (5) constructs the binary view for each foreground query, and Eq. (7) uses its credit to weight pseudo supervision. For $C = 2$ this is the original binary query; for $C > 2$, each γ_k is computed independently and receives fractional quota $0.2\gamma_k|F_k^t|$ with denominator $C \max(1, 0.2|F_k^t|)$.

Semantic credit calibration. Table D.1 compares semantic credit with query presence after all prediction files have been verified. Across all 438,480 one-query episodes, present queries receive consistently higher credit than absent queries. The macro AUROC over classes, with equal weight for each dataset, is 0.827, directly confirming that CDI ranks supported supervision ahead of unsupported supervision without target access during adaptation.

Semantic credit design sensitivity. Table D.2 evaluates the credit construction on all 9,500 Cityscapes image–query pairs. Labels are accessed only after the credit files have been completed and authenticated. Changing the spatial grid from 8×8 to 32×32 changes the macro AUROC over classes by at most 0.004, showing that the ranking is insensitive to the chosen resolution. All three aggregation rules assign higher mean credit to present queries. Averaging the margin or its sigmoid improves presence ranking, whereas the aggregation in Eq. (6) retains more credit for localized positive evidence. The latter property motivates the default 16×16 construction used by CDI.

Confident errors in frozen credit. The calibration result does not assume that every high-credit query is present. Table D.3 isolates all 415,696 absent image–query pairs and groups them by frozen semantic credit. Higher credit is associated with stronger zero-shot false foreground, confirming the difficult case in which the initial model confidently supports an absent class. SFE-Guard reduces false foreground produced by anchor self-training in every credit range, including the two ranges above 0.5. CDI prevents these queries from issuing more supervision than their frozen evidence allows, and HRC reviews the update produced from that bounded supervision.

Optimization and correction. BAPA and SFE-Guard use 20 Adam steps, learning rate 10^{-3}, weight decay 10^{-2}, $\beta = (0.9, 0.999)$, and $\epsilon = 10^{-8}$ (Kingma & Ba, 2015). Adam advances once per step. The HRC active-set system considers three cases in 64-bit floating-point arithmetic according to whether zero, one, or both risks are active. If a full correction exceeds the radius set

Table D.1: **Semantic credit calibration.** AUROC is computed per class from present and absent query credits and then averaged by dataset.

Dataset	Episodes	Pres. credit↑	Abs. credit↓	AUROC↑
VOC	28,980	0.769	0.229	0.961
COCO	400,000	0.557	0.123	0.883
Cityscapes	9,500	0.458	0.372	0.636
Equal-domain mean	438,480	0.595	0.241	0.827

Table D.2: **Semantic credit design sensitivity on complete Cityscapes.** Pres. and Abs. report mean credit for present and absent queries; AUROC is averaged over the 19 classes.

Aggregation	Grid	Pres.↑	Abs.↓	AUROC↑
$\sigma(\mathrm{LME}(d))$	8×8	0.425	0.330	0.638
$\sigma(\mathrm{LME}(d))$	16×16	<u>0.458</u>	0.372	0.636
$\sigma(\mathrm{LME}(d))$	32×32	**0.465**	0.382	0.634
$\sigma(\mathrm{mean}(d))$	16×16	0.319	**0.170**	<u>0.650</u>
$\mathrm{mean}(\sigma(d))$	16×16	0.315	<u>0.201</u>	**0.652**

by the candidate step, HRC uses the largest common reduction fraction that fits the radius. When both risks are zero, HRC commits the candidate unchanged. A zero-gradient or exactly conflicting positive risk system records $\tau = 0$ and also commits the candidate unchanged. Non-finite inputs or failed KKT or radius checks terminate without a valid artifact. HRC replaces the candidate parameter with its corrected value, while the Adam moments retain the history of the candidate update. The matched alternatives in Table F.3 validate this state policy empirically. For either query-set size, each adaptation run performs one dense evaluation for the reference, one before and one after each of 20 candidate updates, and one final evaluation, giving 42 evaluations in total.

D.2 Baseline Reproduction and Fairness

Baseline implementations and dense OVS ports. Every one-query baseline receives the binary list $[\mathrm{background}, q]$, resets its model and optimizer for each RGB–query pair, and returns one binary mask. Every multi-query baseline receives the background and all dataset foreground classes (21/81/20 classes on VOC/COCO/Cityscapes), resets before each RGB, and returns one categorical mask from a shared softmax. No multi-query method is evaluated as a set of independent binary predictions.

BAPA and Multi-BAPA use the shared three-template class construction, refresh the classwise top-20% anchors, and assign equal loss mass to the classes currently predicted. PL performs 20 steps of refreshed hard pixel pseudo-label self-training on the visual LayerNorm affine parameters. TPT retains the released 64 augmented views, lowest-10% entropy selection, four tunable context tokens, and one AdamW step. O-TPT uses the same protocol with its released Householder orthogonality term. The adapted TPT and O-TPT prompts are then used for dense prediction without further optimization.

TENT, ETA, and SAR update only visual LayerNorm affine parameters and apply one released update per episode. TENT minimizes dense entropy with Adam at learning rate 10^{-3}. ETA applies the released reliable-entropy weighting with SGD at learning rate 2.5×10^{-4} and momentum .9; its entropy margin preserves the released threshold as a fraction of the maximum entropy after changing the active class count to C. The source Fisher term is unavailable under source free adaptation, so this comparator is reported as ETA. SAR uses the released SAM–SGD update with learning rate 10^{-3}, momentum .9, $\rho = .05$, margin $0.4 \log C$, EMA, and reset rule. CLIPArTT uses its official implementation with learning rate 10^{-3}, the top $\min(3, C)$ classes, and 10 refinement steps. MLMP uses its official seven-template prompt bank, 18 visual outputs, and 10-step visual-normalization adaptation. Zero-shot performs no update. These implementations share the checkpoint, RGB preprocessing, class names, evaluated images, and episodic reset specified above. Each method retains its original objective and internal selection rule wherever the released dense interface permits; every necessary OVS change is enumerated below.

Table D.3: **Foreground for absent queries by semantic credit range (%).** Counts cover the complete one-query evaluation; values are equal-domain means.

Frozen credit	Episodes	Zero-shot	Anchor Self-Training	SFE-Guard
[0, .25)	350,165	**1.201**	20.795	7.958
[.25, .50)	44,022	**14.167**	42.638	24.098
[.50, .75)	14,256	36.313	39.687	**31.053**
[.75, 1]	7,253	44.296	35.005	**30.680**

Table D.4: **Baseline implementation map.** The final column lists the changes required for one- and multi-query OVS.

Method	Released implementation used	OVS port	Interface change
TENT	Official TENT release	Binary and multi-query dense ports	Image batches become dense pixels; visual LayerNorm affine parameters remain the only trainable state.
ETA	Official EATA release	Binary and multi-query dense ports	Released entropy and reliability operators are retained; selection acts on pixels and the source Fisher term is omitted under the source free protocol.
SAR	Official SAR release	Binary and multi-query dense ports	Released SAM, reliable-entropy selection, EMA, and reset operate on dense pixels.
TPT	Released prompt learner, objective, and confidence-selection rule (Shu et al., 2022)	Prompt adaptation followed by dense OVS evaluation	The released 64 augmented views, four tunable context tokens, and one-step update are retained; the adapted prompt is evaluated by the fixed dense NACLIP model.
O-TPT	Released O-TPT orthogonality objective (Sharifdeen et al., 2025)	Prompt adaptation followed by dense OVS evaluation	The released TPT protocol and Householder orthogonality term are retained before dense evaluation.
CLIPArTT	Official CLIPArTT release	Official wrapper in both OVS runners	The released class-refinement and LayerNorm update are wrapped around the fixed NACLIP model and native OVS class list.

Baseline source and port map. Table D.4 identifies the released implementation used for each of the six external baselines that requires a dense OVS port and the only code in which that port is implemented. Zero-shot requires no port; PL, BAPA, and MLMP are specified in the preceding paragraph. TENT and SAR call released update operators directly, ETA reuses the released entropy and reliability equations, and CLIPArTT wraps the released implementation. TPT and O-TPT retain their released classification-view prompt updates; only the final application of the adapted prompts is changed to obtain a categorical dense prediction. The released implementations are wrapped by a common OVS evaluation interface. The binary and multi-query adapters are included in the anonymous code release, and each result shard records the corresponding source-version hashes.

Published defaults and matched OVS settings. Table D.5 records the source-task defaults and the exact dense settings used here. The OVS settings were fixed before evaluation. No target mask, Present Dice, absent class statistic, or pixel accuracy entered method or parameter selection. All methods use the same seed, evaluation set, checkpoint, class names, and episodic reset; entropy margins depend only on the number of queried classes, C.

Trainable state and runtime. The NACLIP visual encoder contains 102,400 trainable LayerNorm affine parameters. A class-residual method trains $768C$ parameters, giving 1,536 in one-query OVS and 15,360/16,128/62,208 in Cityscapes/VOC/COCO multi-query OVS; the released TPT prompt learner instead trains four shared context tokens, or 3,072 parameters. Table D.6 reports trainable state, algorithmic evaluations, and the matched multi-query GH200 timing described in Table B.2.

Target isolation and aggregation. Prediction files contain only public RGB identifiers, hard predictions, and update diagnostics that do not use target labels. Dataset targets are read only after all prediction shards and source hashes are checked. One-query and multi-query reports retain their respective evaluation units. Present Dice is averaged by class within each dataset and then across datasets. An undefined class-specific metric is omitted rather than replaced with zero, and episodes are not pooled across datasets.

Table D.5: **Published and OVS hyperparameters.** LN denotes visual LayerNorm affine parameters; C includes background.

Method	Released source-task setting	Setting in this work
TENT	Norm affine parameters; entropy minimization; one update per incoming batch; example learning rate 10^{-3}.	LN; Adam; one released update; learning rate 10^{-3}; mean dense entropy.
ETA	Reliable sample selection from EATA with $e_0 = 0.4 \log 1000$; EATA additionally uses diversity filtering and source Fisher regularization.	LN; SGD with momentum .9; one update; learning rate 2.5×10^{-4}; class-count normalized reliability margin; no source Fisher term.
SAR	SAM–SGD with $\rho = .05$, momentum .9, reliability margin $0.4 \log 1000$, EMA recovery, and online reset.	LN; one released SAM–SGD update; learning rate 10^{-3}; $\rho = .05$; momentum .9; margin $0.4 \log C$; released EMA and reset threshold.
TPT	64 augmented views, lowest-10% entropy selection, one AdamW step, learning rate 5×10^{-3}, and four tunable context tokens.	The same released setting; the adapted prompts are used for the final dense prediction.
O-TPT	TPT classification protocol with Householder orthogonality calibration.	The same TPT setting with the released Householder term weighted by 18; the adapted prompts are used for the final dense prediction.
CLIPArTT	Released wrapper defaults to top-3 class refinement and 10 LN steps; the learning rate is supplied by the caller.	Released wrapper; top-$\min(3, C)$ refinement; 10 LN–Adam steps; learning rate 10^{-3}; one reset per RGB–query pair or multi-query RGB.

Table D.6: **Trainable parameters, model evaluations, and runtime.** Parameter ranges reflect dataset class counts. Multi-query time is measured on one GH200 and reported as mean ± standard deviation over three repetitions.

	Trainable parameters		Model evaluations		GH200 time
Method	One query	Multiple queries	One query	Multiple queries	Multiple queries
PL	102,400	102,400	21 dense	21 dense	0.856 ± 0.001
TENT	102,400	102,400	2 dense	2 dense	0.079 ± 0.001
ETA	102,400	102,400	2 dense	2 dense	0.072 ± 0.000
TPT	3,072	3,072	64 views + 1 dense	64 views + 1 dense	0.677 ± 0.001
SAR	102,400	102,400	3 dense	3 dense	0.112 ± 0.000
CLIPArTT	102,400	102,400	21 dense	21 dense	2.128 ± 0.004
MLMP	102,400	102,400	11 dense	11 dense	0.475 ± 0.001
O-TPT	3,072	3,072	64 views + 1 dense	64 views + 1 dense	0.679 ± 0.001
Multi-BAPA	–	15,360–62,208	–	22 dense	0.346 ± 0.002
SFE-Guard	1,536	15,360–62,208	42 dense	42 dense	0.363 ± 0.000

E Direct Comparison with Hallucination-Oriented OVS

Test-Time Contrastive Concepts (TTCC) (Wysoczańska et al., 2025), Counterfactual Segmentation Reasoning (CSR) (Li et al., 2026), and SFE-Guard address consecutive stages of unsupported foreground prediction. TTCC introduces separate alternatives for each query so that visually related content can be assigned to background at inference. CSR learns abstention from paired factual and counterfactual examples. SFE-Guard controls the subsequent self-training loop in which an unsupported prediction becomes supervision and is enlarged by repeated parameter updates. Table E.1 summarizes these distinct intervention points.

Matched TTCC comparison. The Cityscapes comparison fixes the NACLIP ViT-L/14 backbone, input preprocessing, three-template class prototypes, image population, and query order. TTCC uses the authors' published class-specific contrastive concepts and their multi-query similarity filter. Its predictions are evaluated before adaptation with PD, absent foreground area, and PA. Table E.2 compares this inference-time intervention with matched anchor self-training and SFE-Guard under identical images and queries.

TTCC provides the strongest absent-query suppression and one-query PA by introducing alternative concepts before a prediction is formed. SFE-Guard leads PD in both query settings and multi-query PA while substantially reducing the absent error produced by anchor self-training. The comparison isolates two intervention stages. Inference-time contrast changes the initial mask, whereas CDI and HRC control how that mask is reused and amplified during adaptation.

Table E.1: **Direct comparison with hallucination-oriented OVS.** The three methods intervene at consecutive stages of unsupported foreground prediction.

Method	Prediction setting	Main intervention	Additional information	Controlled failure
TTCC	One- and multi-query OVS	Adds separate contrastive concepts for each query to the inference vocabulary	Concept co-occurrences extracted from captions or concept lists produced by a language model	Unsupported foreground in the initial prediction
CSR	Referring and reasoning segmentation	Learns to abstain on counterfactual image–text pairs	Supervised factual and counterfactual training pairs	Pixel-grounding hallucination at inference
SFE-Guard	One- and multi-query OVS with test-time adaptation	Bounds pseudo supervision and reviews every candidate update before commitment	No source samples, target labels, information about whether a queried class occurs, or external language model	Recursive amplification of unsupported foreground during adaptation

Table E.2: **Matched comparison with TTCC on complete Cityscapes (%).** TTCC changes the inference vocabulary without adaptation; the other methods perform 20 test-time updates.

Setting	Method	PD↑	Abs.↓	PA↑
One query	Anchor Self-Training	14.345	33.047	68.434
	TTCC	16.371	**3.382**	**92.715**
	SFE-Guard	**17.148**	17.329	83.357
Multiple queries	Anchor Self-Training	17.998	4.115	36.208
	TTCC	14.251	**1.398**	34.853
	SFE-Guard	**18.998**	2.257	**48.629**

Counterfactual comparison. The CSR comparison uses all 1,340 pairs in the official HalluSeg-Bench referring split and its factual and counterfactual image–text combinations. RobustSeg is reported with the published CSR results. The remaining methods use the same frozen NACLIP backbone and three-template query prototypes; the two iterative methods perform 20 updates. Table E.3 reports the official counterfactual mask score (CMS) and text and visual ΔIoU, together with PD and absent-query foreground for the OVS methods.

Table E.3: **Direct comparison on the complete HalluSegBench referring split.** CMS and ΔIoU follow the official evaluator. PD and absent foreground are percentages. RobustSeg denotes the published supervised CSR model; the other rows use frozen NACLIP.

Method	CMS fact.↓	CMS cf.↓	Text ΔIoU↑	Visual ΔIoU↑	PD↑	Abs. FG↓
RobustSeg (Li et al., 2026)	**.1062**	**.4049**	**.6105**	**.5278**	–	–
Zero-shot	.4085	1.6762	.1581	.1145	51.764	**20.036**
Anchor Self-Training	1.2017	1.6492	.1033	.1151	54.666	30.009
SFE-Guard	.7500	1.2791	.1213	.1326	**56.236**	20.628

RobustSeg obtains the strongest static counterfactual abstention through supervised factual–counterfactual training. Within iterative OVS, SFE-Guard improves every reported measure over anchor self-training. It raises PD by 1.570 points, removes 9.381 points of absent foreground, lowers factual and counterfactual CMS, and increases both counterfactual ΔIoU scores. The result directly shows that CDI and HRC suppress the additional hallucination introduced when an initial prediction is repeatedly reused as self-supervision.

F Generalization and Robustness

F.1 Alternative Anchor Policies

Table F.1 evaluates representative direct, confidence-filtered, and teacher-smoothed anchor policies on all three evaluation sets. Each iterative policy raises absent class area above zero-shot. The EMA policy improves PD to 34.064, yet its absent area remains 1.334 points higher than zero-shot, showing that SFE also persists under teacher-smoothed pseudo-label refresh.

Table F.1: **Multi-query anchor policies (%).** Results equally weight all three datasets.

Policy	Anchor fraction	Steps	PD↑	Abs.↓	PA↑
Zero-shot	–	0	32.309	**1.604**	**33.819**
Dynamic anchors	0.20	20	27.766	3.179	20.833
Confidence ≥ 0.80	0.20	20	29.346	1.936	26.993
EMA anchors	0.20	20	**34.064**	2.938	27.427

F.2 Transfer to a Second OVS Architecture

Table F.2 transfers the multi-query study to the released SPAR MaskCLIP ViT-B/16 model on all 1,449 VOC validation images. BAPA raises absent area from 0.234 to 1.752 and lowers PA by 23.811 points, reproducing SFE under a different backbone, feature scale, prompt ensemble, and class-probability structure. CDI+HRC improves BAPA by 3.314 PD points, reduces absent area by 1.307 points, and recovers 21.962 PA points. CDI-only provides the highest PD, while CDI+HRC supplies the lowest absent area and highest PA among the adapted methods.

Table F.2: **SPAR MaskCLIP ViT-B/16 transfer on VOC (%).** Adapted methods share the prompt ensemble, trainable state, 20 updates, and reset policy.

Method	PD↑	Abs.↓	PA↑
Zero-shot	55.311	**0.234**	**80.124**
BAPA	54.118	1.752	56.313
CDI-only	**58.369**	0.641	75.694
CDI+HRC	57.432	0.445	78.275

F.3 Optimizer State after HRC Commitment

Table F.3 changes only the optimizer state used after HRC commits its corrected parameter displacement. Retaining the proposal moments gives the highest PD and PA and improves the rollback alternative by 0.310 PD, 0.010 absent area, and 1.949 PA points. Removing momentum further reduces absent area while losing PD and PA. The retained state therefore carries useful information from the candidate update into the next reviewed step.

Table F.3: **Multi-query optimizer-state comparison (%).** All rows use the same CDI–HRC configuration and evaluation images.

State after HRC commitment	PD↑	Abs.↓	PA↑
Retain proposal moments	**36.235**	1.602	**46.042**
Rollback moments	35.925	1.612	44.093
SGD without momentum	32.929	**1.349**	41.583

G Qualitative Results

Figures G.1 and G.2 show representative examples from the evaluated predictions. Selection requires a visible ground-truth foreground and excludes excessively large predictions; the multi-query examples additionally contain at least two present classes. Each row compares Zero-shot, MLMP, BAPA, and SFE-Guard on the same input and ground truth. Together with Figure 5, these examples span all three datasets and complement the per-dataset results in Table H.1.

H Per-Dataset Results and Statistical Analysis

H.1 Component Results by Dataset

Table H.1 reports the component comparison by dataset. The four arms use the same RGBs, queries, prompts, optimization budget, and reset policy within each setting. CDI+HRC gives the strongest PD on all three one-query datasets and on VOC and COCO in multi-query OVS; HRC-only gives the

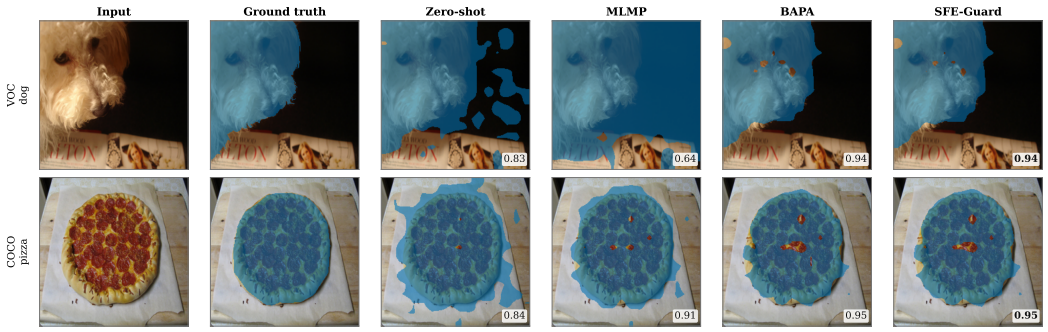


Figure G.1: **One-query qualitative comparison.** Rows show the input, ground truth, and four predictions. Blue denotes the queried foreground; numbers report Dice.

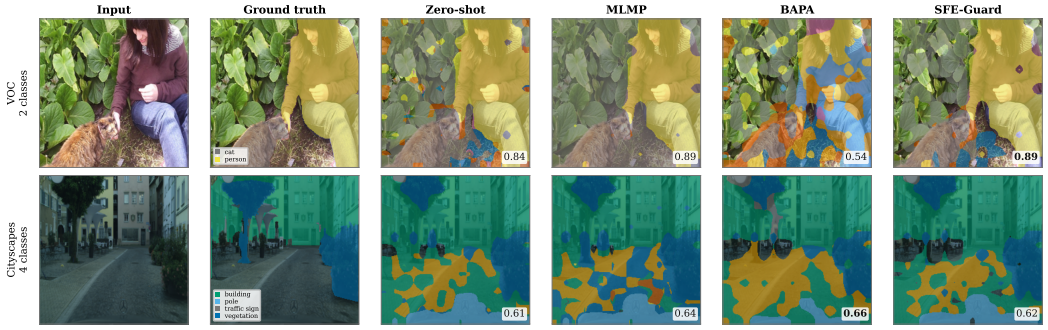


Figure G.2: **Multi-query qualitative comparison.** The same class-color palette is used for ground truth and all predictions; numbers report present-class macro Dice.

strongest multi-query Cityscapes PD. CDI+HRC also gives the strongest one-query absent-foreground control and PA on all three datasets. CDI-only gives the lowest Abs. and highest PA on multi-query VOC and COCO, whereas CDI+HRC gives both on Cityscapes.

H.2 Primary Comparisons

The same image-level paired bootstrap is applied to the central comparisons in Table 1. Each resample uses a common image multiplicity for every method and both query settings, and the reported PD remains an macro average over classes with equal weight for each dataset. Table H.2 reports 2,000 support-preserving resamples with seed 2027.

Every PD interval in Panel A lies above zero. BAPA is the one-query runner-up, MLMP is the multi-query runner-up, and TENT is the runner-up after averaging the two settings. From Table B.1, BAPA and Multi-BAPA add 16.778 points above zero-shot in one-query OVS and 1.575 points in multi-query OVS, respectively. Panel B shows that SFE-Guard removes 13.070 and 1.577 points, respectively, returning multi-query absent area to the zero-shot level.

H.3 Failure Frequencies

Figure 2 uses all 6,949 images, with 22,784 present and 415,696 absent image–class pairs in each query setting. For each dataset, the event count is divided by the number of eligible pairs; the three dataset rates are then averaged equally. An image is the bootstrap unit, so all its queries retain the same resampling multiplicity. Percentile intervals use 2,000 paired resamples within each dataset and seed 2027.

Figure 2 counts newly suppressed present classes, whose predicted area falls from at least 1% at initialization to below 1% after adaptation. The denominator includes all present pairs, preserving the same evaluation population across methods. The appendix tables additionally report endpoint

Table H.1: **Component results by dataset (%).** All rows use the same primary protocol. PD, Abs., and PA denote Present Dice, absent foreground or absent class area, and pixel accuracy. Best results are bold and second-best results are underlined.

Setting	Dataset	Method	PD↑	Abs.↓	PA↑
One-query	VOC	BAPA	55.941	25.820	74.826
		CDI-only	57.206	16.119	84.058
		HRC-only	57.744	19.085	81.285
		CDI+HRC	**58.622**	**14.824**	**85.339**
	COCO	BAPA	35.362	21.614	78.443
		CDI-only	38.270	9.274	90.606
		HRC-only	37.167	14.470	85.509
		CDI+HRC	**39.074**	**9.117**	**90.792**
	Cityscapes	BAPA	14.345	33.047	68.434
		CDI-only	16.645	18.859	81.957
		HRC-only	15.274	23.934	77.021
		CDI+HRC	**17.148**	**17.329**	**83.357**
Multi-query	VOC	Multi-BAPA	38.188	4.311	16.202
		CDI-only	50.047	**1.829**	**55.458**
		HRC-only	48.438	3.734	23.029
		CDI+HRC	**53.389**	1.966	49.888
	COCO	Multi-BAPA	27.112	1.110	10.089
		CDI-only	35.187	**0.477**	**48.594**
		HRC-only	31.980	1.013	14.468
		CDI+HRC	**36.317**	0.582	39.610
	Cityscapes	Multi-BAPA	17.998	4.115	36.208
		CDI-only	18.315	2.405	46.956
		HRC-only	**19.748**	3.515	42.906
		CDI+HRC	18.998	**2.257**	**48.629**

Table H.2: **Paired uncertainty for the primary comparisons.** Panel A reports Present Dice gains; Panel B reports reductions in BAPA/Multi-BAPA-induced absent error. Brackets give 95% confidence intervals from paired image resampling.

Setting	Comparator	ΔPD	95% CI
A. Equal-domain Present Dice			
One-query	BAPA (runner-up)	+3.066	[+2.880, +3.259]
Multi-query	Multi-BAPA	+8.469	[+8.044, +8.902]
Multi-query	MLMP (runner-up)	+3.177	[+2.841, +3.513]
Overall	BAPA / Multi-BAPA	+5.767	[+5.548, +5.989]
Overall	TENT (runner-up)	+3.851	[+3.654, +4.052]
B. Reduction of BAPA/Multi-BAPA-induced absent error			
Setting	Quantity	Estimate	95% CI
One-query	Excess removed (points)	13.070	[12.623, 13.485]
	Relative reduction	77.9%	[76.0, 79.9]%
Multi-query	Excess removed (points)	1.577	[1.361, 1.784]
	Relative reduction	100.2%	[93.3, 106.8]%

suppression, defined as final area below 1% regardless of initial area. This includes initial misses and small objects. Both differ from the whole-prediction collapse measures in Table 1. For an absent class, expansion means its final area exceeds its initial area by more than one percentage point. Expansion rates use all absent image–class pairs as the denominator. The reported frequencies give each of the three datasets equal weight in both query settings. The two statistics have different eligible populations and are not added together.

In Figure 2, confidence sharpening and anchor self-training are represented by MLMP and BAPA/Multi-BAPA, respectively, in the one- and multi-query settings. BAPA and SFE-Guard use the matched three-template diagnostic protocol. The one-query calculation combines the authenticated initial predictions with the corresponding factor and full-model prediction banks. For multiple queries, initial and final areas are taken from the same recorded adaptation trajectories. MLMP retains its recorded official prediction protocol. Its initial predictions are recomputed with the matching checkpoint, prompts, preprocessing, and visual aggregation, so expansion is measured against its own initialization. Ground truth is used only to identify the present and absent evaluation

groups. Table H.3 reports the endpoint suppression statistic by dataset. Table H.4 reports thresholds of 0.5%, 1%, and 2%.

Table H.3: **Endpoint foreground suppression and SFE by dataset (%).** Foreground suppression is measured at the final prediction; SFE is measured relative to initialization. BAPA denotes Multi-BAPA in multiple-query rows.

		Endpoint foreground suppression				SFE		
Setting	Dataset	Zero-shot	MLMP	BAPA	SFE-Guard	MLMP	BAPA	SFE-Guard
One query	VOC	4.84	6.42	1.35	1.77	3.24	52.44	44.35
One query	COCO	14.85	24.71	5.82	7.81	1.83	46.38	39.87
One query	Cityscapes	23.08	45.51	10.39	12.81	5.06	63.40	50.99
Multiple queries	VOC	5.31	6.84	4.66	5.91	1.90	29.79	13.21
Multiple queries	COCO	26.68	28.19	25.19	27.67	1.15	13.26	5.62
Multiple queries	Cityscapes	58.10	52.31	41.63	51.06	1.46	20.77	11.93

Table H.4: **Threshold sensitivity of endpoint foreground suppression and SFE (%).** BAPA denotes Multi-BAPA in multiple-query rows. Rates equally weight the three datasets. Brackets give 95% image-bootstrap intervals.

		Endpoint foreground suppression		SFE	
Setting	Threshold	MLMP	SFE-Guard	BAPA	SFE-Guard
One query	0.5%	23.25 [22.71, 23.80]	6.05 [5.75, 6.36]	54.67 [54.14, 55.14]	46.66 [46.18, 47.16]
One query	1%	25.55 [25.02, 26.08]	7.46 [7.13, 7.81]	54.07 [53.55, 54.55]	45.07 [44.57, 45.56]
One query	2%	28.89 [28.30, 29.46]	11.37 [10.92, 11.84]	53.13 [52.62, 53.63]	41.95 [41.44, 42.46]
Multiple queries	0.5%	25.38 [24.88, 25.91]	23.70 [23.22, 24.21]	22.76 [22.27, 23.25]	12.84 [12.43, 13.27]
Multiple queries	1%	29.11 [28.60, 29.63]	28.21 [27.72, 28.77]	21.27 [20.81, 21.74]	10.25 [9.89, 10.64]
Multiple queries	2%	34.22 [33.66, 34.80]	35.59 [35.06, 36.14]	18.59 [18.18, 19.01]	6.92 [6.61, 7.22]

H.4 Component-Level Paired Uncertainty

Table H.5 reports the paired difference between CDI+HRC and CDI-only. Source images are resampled with replacement 2,000 times using seed 2027. A draw is repeated if it removes all support for a defined class, preserving the macro average over classes without filling a missing stratum with zero. All queries and metrics from a sampled image receive the same bootstrap multiplicity. Entries are percentage-point differences with percentile 95% confidence intervals.

Table H.5: **Paired bootstrap comparison of CDI+HRC with CDI-only.** Δ denotes CDI+HRC minus CDI-only; brackets give 95% confidence intervals over 2,000 paired image resamples.

Setting	Dataset	ΔPD	ΔAbs.	ΔPA
One-query	VOC	$+1.416\,[+1.131, +1.689]$	$-1.295\,[-1.410, -1.174]$	$+1.281\,[+1.169, +1.392]$
	COCO	$+0.804\,[+0.641, +0.959]$	$-0.157\,[-0.202, -0.114]$	$+0.186\,[+0.143, +0.230]$
	Cityscapes	$+0.503\,[+0.221, +0.752]$	$-1.530\,[-2.357, -0.623]$	$+1.400\,[+1.216, +1.575]$
Multi-query	VOC	$+3.342\,[+2.701, +3.996]$	$+0.137\,[+0.103, +0.171]$	$-5.570\,[-6.230, -4.955]$
	COCO	$+1.130\,[+0.959, +1.303]$	$+0.105\,[+0.101, +0.109]$	$-8.984\,[-9.338, -8.635]$
	Cityscapes	$+0.683\,[+0.511, +0.866]$	$-0.148\,[-0.406, +0.102]$	$+1.673\,[+1.257, +2.096]$

The paired intervals confirm that CDI+HRC improves PD over CDI-only on all three datasets in both settings. In one-query OVS, absent foreground and PA also improve on all three datasets. In multi-query OVS, the additional correction trades absent area and PA on VOC and COCO. The Cityscapes absent-area difference is unresolved from zero, while its PA improves by 1.673 points.

prediction artifacts will accompany the submission.