Preference-Guided Adaptation for Open-Vocabulary Semantic Segmentation via Prompt Disagreement

Hyun-Kurl Jang
Visual Intelligence Lab.
KAIST
jhg0001@kaist.ac.kr

Jihun Kim
Visual Intelligence Lab.
KAIST
jihun1998@kaist.ac.kr

Kuk-Jin Yoon
Visual Intelligence Lab.
KAIST
kjyoon@kaist.ac.kr

Abstract

Open-vocabulary semantic segmentation (OVSS) enables pixel-level prediction over arbitrary text-specified vocabularies and has shown strong generalization on common benchmarks. However, OVSS performance often degrades in specialized domains such as medical imaging, remote sensing, and industrial inspection, where dense pixel-level masks for adaptation are costly to obtain and require domain-specific expertise. We propose a preference-guided adaptation framework that replaces dense mask supervision with binary preferences. We observe that different prompt templates produce systematically different segmentations for the same image, a phenomenon we call prompt disagreement, and we repurpose it as a built-in source of preference supervision. Building on this, we mine localized preference queries from regions of high cross-template uncertainty, and adapt the OVSS model with **R**egion-**L**ocalized **P**reference **O**ptimization (**RLPO**) together with consistency regularization that stabilizes updates outside the queried region. Across extensive experiments on the MESS benchmark, the proposed method achieves consistent gains across diverse OVSS backbones without any pixel-level annotation, and remains effective under noisy preferences. Our code is available at `https://github.com/blue-531/pref-ovss`.

1 Introduction

Semantic segmentation has long been studied under a closed-set assumption, where models are trained and evaluated on a fixed set of categories [1–10]. This limits their ability to recognize novel concepts and operate in open-world settings. Open-vocabulary semantic segmentation (OVSS) [11–39] addresses this limitation by allowing class names to be specified in natural language at inference time. Powered by vision-language models (VLMs) such as CLIP [40], OVSS aligns dense visual features with text embeddings and has shown strong generalization to categories unseen during training.

As OVSS moves toward real-world deployment, adapting it to specialized domains has become increasingly important. Applications such as medical imaging, remote sensing, industrial inspection, and agriculture differ substantially from web-scale pretraining data in appearance, label granularity, and vocabulary [17, 41]. Prior adaptation strategies for OVSS have explored prompt tuning [42–44] and adapter-based fine-tuning [45, 46]. However, many of these approaches often assume access to target-domain mask supervision, which are difficult to obtain in specialized domains. Unlike common-object datasets, specialized domains often involve fine-grained categories, ambiguous visual boundaries, and domain-specific terminology, making reliable mask annotation dependent on expert knowledge. Since target classes and annotation criteria often vary across specialized domains, relying on dense masks creates a recurring annotation bottleneck for OVSS adaptation.

These limitations motivate a different form of supervision that does not require annotators to construct dense ground truth yet still guides model predictions toward the intended concept. Pairwise prefer-

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

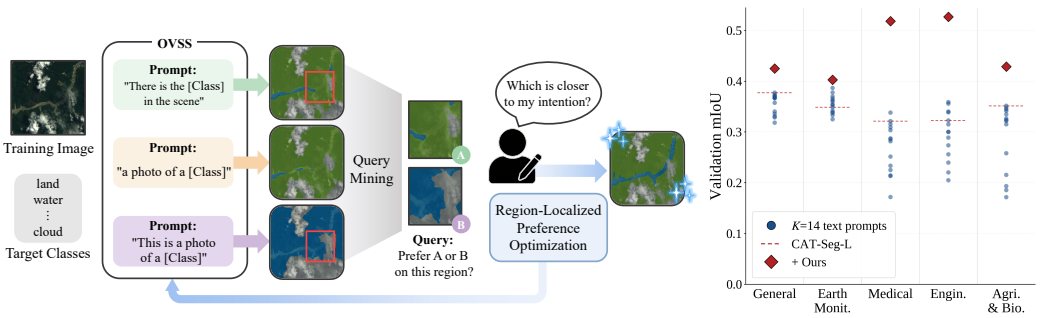


Figure 1: **Left:** For each target-domain image, our framework runs multiple prompt templates through a OVSS backbone, mines a localized preference query from regions where the templates disagree most, and uses a binary judgment ("which prediction is closer to the intended concept on this region?") to drive an update of OVSS model. **Right:** Validation mIoU across diverse specialized domains [41] when only the prompt template is varied. Across all domains, performance varies widely across templates. Our method harnesses this variance as supervision and lifts performance above the entire per-domain spread, turning prompt disagreement into training signal.

ences provide such a signal. Preference-based supervision has emerged as an effective way to align model outputs with human intent in language and vision-language generation, largely because relative judgments can convey useful training signals without requiring fully specified target outputs [47–49]. Translating this principle to OVSS leads to a lightweight protocol: given two candidate segmentations for the same image, annotators simply choose the prediction that better matches the intention. By indicating which prediction better matches the intended segmentation in a target-domain image, this relative feedback provides a natural form of supervision for adapting OVSS models without requiring pixel-level masks.

To make this supervision actionable, however, OVSS requires meaningful candidate segmentations to compare. We observe that such candidates naturally arise from the prompting interface itself. Different prompt templates can produce systematically different masks for the same image and class name [40, 50, 51]; we refer to this template-induced prediction variance as **prompt disagreement**. As shown in Fig. 1 (right), prompt disagreement is pronounced in practice: across diverse specialized domains, validation mIoU varies widely across templates for the same backbone. Rather than treating prompt disagreement as a nuisance, we repurpose it as a built-in source of pairwise supervision: prompt-induced masks serve as competing hypotheses, and binary preferences identify which hypothesis better captures the intended concept.

Building on this idea, we propose a preference-guided adaptation framework for OVSS that turns prompt disagreement into an actionable training signal, illustrated in Fig. 1 (left). The framework consists of three components. First, we mine informative preference queries from prompt ensembles. We localize high-uncertainty regions using cross-prompt entropy and, within each region, select the pair of template-induced predictions with the largest disagreement. This jointly determines where feedback should be collected and which pair of segmentation hypotheses should be compared. Second, we introduce Region-Localized Preference Optimization (RLPO) for adapting OVSS models from binary preferences. Rather than treating the preference as a single image-level signal, our objective applies it to pixel-level segmentation scores inside the selected region, enabling dense spatial supervision from a single comparison. Third, we introduce a consistency regularization to stabilize preference optimization. Because the preference objective only supervises the selected region, it resolves the queried disagreement but leaves predictions outside that region unconstrained. Our consistency regularizer therefore uses the preferred prediction as a pseudo-target for the rejected prediction outside the selected region, preventing unintended drift beyond the adapted area.

We evaluate the proposed method on the MESS benchmark [41], which covers five domain groups: general scenes, earth monitoring, medical sciences, engineering, and agriculture & biology. Across these diverse settings, our approach consistently improves OVSS performance without dense masks, demonstrating that prompt disagreement provides a practical supervision signal for adapting OVSS to specialized domains. We further show that the proposed adaptation strategy yields consistent improvements across different OVSS baselines [18, 30], suggesting that it is not specific to a single model design. The method also remains effective under noisy preference feedback, highlighting the robustness of binary preferences.

Our main contributions are as follows:

- We show that prompt disagreement can be repurposed as a source of preference supervision, enabling OVSS adaptation without human-provided pixel-level masks.
- We propose Region-Localized Preference Optimization (RLPO), which adapts OVSS models from binary preferences mined via prompt disagreement.
- We demonstrate consistent improvements on the MESS benchmark across diverse specialized domains, validating binary preference feedback as a practical supervision signal for OVSS adaptation.

2 Related Works

2.1 Open-Vocabulary Semantic Segmentation

Open-vocabulary semantic segmentation (OVSS) aims to assign pixel-level labels from an arbitrary set of text-specified classes, including those unseen during training. Building on vision-language foundations such as CLIP [40], recent methods have explored diverse strategies for bridging image-level pretraining with dense prediction. Two-stage approaches [12, 13, 15, 16, 19, 21, 23] first generate mask proposals and then classify each region via CLIP, while one-stage methods [18, 25, 34] produce masks within a side adapter during inference. Cost aggregation approaches [30, 31, 38, 52] instead refine patch-level embeddings through cost aggregation or learned decoders to directly produce per-pixel predictions.

Despite strong results on standard benchmarks dominated by everyday imagery, such as ADE20K [53] and Pascal Context [54], OVSS models often struggle in specialized domains. Such domains often exhibit domain-specific vocabulary, high inter-class similarity, ambiguous boundaries, and atypical visual appearance. Existing adaptation strategies, including prompt tuning [42–44], adapter-based methods [37, 45, 46], and personalized OVSS approaches [55], typically rely on dense mask supervision. This reliance limits scalability in specialized domains, where annotation often requires expert knowledge and must be repeated for each target vocabulary. To address this gap, we propose a preference-guided OVSS adaptation framework that learns from binary comparisons between candidate segmentations rather than dense masks.

2.2 Preference Learning

Preference learning [47, 48, 56–62] uses relative judgments between candidate outputs as supervision, instead of requiring fully specified ground-truth targets. This formulation is especially useful in settings where dense annotation is difficult to obtain but comparative feedback is easier to collect. Among recent approaches, Direct Preference Optimization (DPO) [48] has emerged as a simple objective that learns directly from pairwise preferences without requiring an explicit reward model.

While DPO was originally proposed for language model alignment, preference-based objectives have since been extended to visual tasks, including image generation [63, 49, 64, 65] and dense prediction [66, 67]. Recent work has also begun to explore preference-based objectives for segmentation [68, 69]. However, these studies have been limited to fixed-target settings, typically in medical imaging, where the task involves a single foreground structure or a small closed label space.

As a result, preference-guided adaptation remains largely unexplored for open-vocabulary semantic segmentation. We address this gap by using prompt-induced variation to construct localized binary comparisons, enabling OVSS adaptation from preference feedback without dense masks.

3 Method

Our goal is to adapt an OVSS model to a specialized target domain using binary preference feedback. Given target-domain images and a target vocabulary, we use prompt disagreement to construct localized comparison queries and obtain binary preferences over candidate segmentations. These preferences serve as the supervision signal for adaptation, without requiring dense annotations. Our overall framework is illustrated in Fig. 2, which consists of three components: preference query mining, RLPO, and consistency regularization. We first introduce the problem setting and then describe each component.

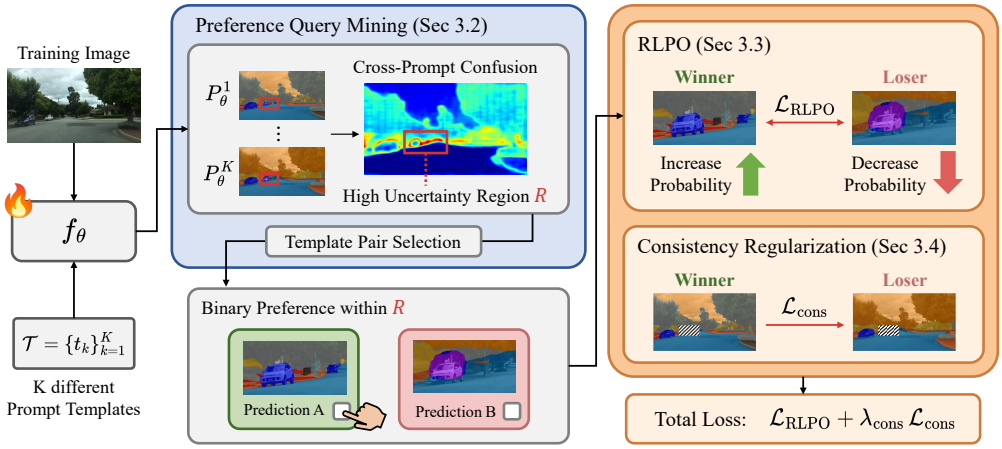


Figure 2: Given a training image, our method generates K prompt-conditioned predictions with model f_θ. It mines a preference query by selecting a high-uncertainty region R from prompt-ensemble entropy and the most disagreeing template pair within R. A binary preference identifies the winner and loser predictions. RLPO updates f_θ by increasing the winner's score relative to the loser's inside R, while the consistency regularization aligns the loser prediction with the winner pseudo-label.

3.1 Preliminaries

Open-vocabulary semantic segmentation (OVSS). OVSS aims to assign a class label to every pixel of an image given an arbitrary class vocabulary specified at test time, including categories unseen during training. Modern OVSS systems build on vision–language models such as CLIP [40], predicting segmentation maps by aligning dense visual features with text embeddings of class names prompted through templates (e.g., "`a photo of a [CLASS]`"). The choice of prompt template is known to materially affect predictions [40, 50, 51]: different templates can produce noticeably different segmentations on the same image, particularly under domain shift, where the source-trained alignment is no longer well calibrated to the target distribution.

Direct preference optimization (DPO). Standard reinforcement learning from human feedback first fits a reward model on preference data and then optimizes a policy against it; DPO [48] avoids the reward modeling stage by deriving a closed-form objective that learns directly from preference pairs. Specifically, DPO adapts a policy π_θ from a frozen reference π_{ref} using preference pairs (y_w, y_l), where y_w is preferred to y_l for the same input. It interprets the log-probability ratio

$$r_\theta(y) = \log \frac{\pi_\theta(y)}{\pi_{\text{ref}}(y)} \tag{1}$$

as an implicit reward, and optimizes the Bradley–Terry objective [70]

$$\mathcal{L}_{\text{DPO}}(\theta) = -\log \sigma\big(\beta\big[r_\theta(y_w) - r_\theta(y_l)\big]\big)\,, \tag{2}$$

which raises $r_\theta(y_w)$ for the winner while lowering $r_\theta(y_l)$ for the loser. Because the reward is defined relative to π_{ref}, the reference distribution acts as a KL anchor that prevents π_θ from drifting far from its initial behavior, with β controlling the trade-off between matching the preferences and staying close to π_{ref}. We adapt this principle to OVSS in Section 3.3 by replacing the policy log-probability ratio with a region-localized segmentation score, so that Eq. (2) takes a form directly applicable to template-conditional segmentation.

Problem setting. We adapt an OVSS model to a target domain in a streaming, single-step setting: training images from the target domain arrive sequentially, and for each image we elicit a binary preference and perform a single gradient update on the model parameters θ before moving on, without revisiting images or accumulating preferences for batch optimization. Let $x \in \mathbb{R}^{H\times W\times 3}$ denote a target-domain image with pixel domain Ω and pixel positions $\mathbf{u} \in \Omega$. Let $\mathcal{C} = \{c_1, \ldots, c_N\}$ denote the target vocabulary, and let $\mathcal{T} = \{t_k\}_{k=1}^K$ denote a fixed set of K templates with prompted vocabulary $\mathcal{C}^k = \{t_k(c) : c \in \mathcal{C}\}$. We adapt an OVSS model f_θ from a frozen reference f_{ref}.

Both share the same backbone; f_θ adds lightweight adapters on the vision and text branches as the only trainable parameters, while f_{ref} corresponds to the initial state of f_θ before adaptation. The template-k pixel-level prediction is

$$P_\theta^k(c \mid x, \mathbf{u}) = f_\theta(x, \mathcal{C}^k)_{\mathbf{u},c}, \qquad c \in \mathcal{C}, \tag{3}$$

with P_{ref}^k defined analogously. Given a small set of target-domain images, we adapt θ using binary preferences elicited between pairs of template-specific predictions $\{P_\theta^k\}_{k=1}^K$.

3.2 Preference Query Mining

We design the preference protocol around two requirements: each query should be cognitively simple to answer, and the resulting binary signal should still carry enough supervision to drive adaptation. Richer feedback formats (e.g., ratings, rankings, or multi-way selections) place a heavier cognitive load on annotators and are prone to inconsistent calibration across examples and annotators [71]. We therefore restrict each annotation to a binary choice. Even with binary feedback, whole-image preferences remain ambiguous in segmentation, because two candidates may each be better in different parts of the image. Therefore, we further localize each comparison to a small region $R \subseteq \Omega$, on which the question becomes concrete: within R, which of two template predictions P_θ^a, P_θ^b better matches the intended concept?

We construct queries by exploiting prompt disagreement, the variation across template predictions $\{P_\theta^k\}_{k=1}^K$. Template-induced predictions provide useful candidates because they vary the segmentation hypothesis while preserving the target vocabulary. This makes the resulting candidates directly comparable under the same semantic target. Concretely, prompt disagreement on a single image yields two operational signals: where the templates collectively disagree, and which pair of templates disagrees most strongly. To localize regions of high disagreement, we measure cross-prompt uncertainty at each pixel by the entropy of the ensemble distribution

$$\bar{P}(c \mid x, \mathbf{u}) = \frac{1}{K} \sum_{k=1}^{K} P_\theta^k(c \mid x, \mathbf{u}), \qquad \mathcal{H}(\mathbf{u}) = -\sum_{c \in \mathcal{C}} \bar{P}(c \mid x, \mathbf{u}) \log \bar{P}(c \mid x, \mathbf{u}). \tag{4}$$

We binarize $\mathcal{H}$ at its 0.95-quantile and take R as the bounding box of the largest connected component of the resulting high-entropy mask. Within R, we identify the most informative template pair by counting pixel-level disagreement between hard predictions and selecting the pair with the largest count:

$$(a, b) = \arg \max_{1 \le k < k' \le K} \sum_{\mathbf{u} \in R} \mathbf{1}\Big[\hat{Y}^k(\mathbf{u}) \neq \hat{Y}^{k'}(\mathbf{u})\Big], \quad \hat{Y}^k(\mathbf{u}) = \arg\max_c P_\theta^k(c \mid x, \mathbf{u}). \tag{5}$$

By targeting both the most uncertain region and the most disagreeing template pair within it, each query is visually concrete enough to be judged at a glance yet carries dense supervision for adaptation.

3.3 Region-Localized Preference Optimization (RLPO)

Given the query (R, a, b), an oracle provides a binary preference indicating which of the two predictions P_θ^a and P_θ^b is preferred on R. We denote the winner and loser template indices by w and l, respectively, with corresponding hard predictions $\hat{Y}^w$ and $\hat{Y}^l$.

Region-level score. To instantiate the DPO objective in our setting, we need a per-template scalar score that plays the role of $\log \pi_\theta(y)$ in Eq. (1). A natural choice is the log-likelihood that template k assigns to its own hard prediction, averaged over R. However, a simple pixelwise average would be dominated by classes that occupy the largest area in R, so a single large class could obscure the contribution of smaller but semantically important ones. We therefore use a class-balanced average: pixels in R are grouped by the winner's prediction $\hat{Y}^w$ to define a stable, k-independent partition, and per-pixel scores are first averaged within each class before averaging across classes. Let

$$R_c = \{\mathbf{u} \in R : \hat{Y}^w(\mathbf{u}) = c\}, \qquad U_R = \{c \in \mathcal{C} : R_c \neq \emptyset\}. \tag{6}$$

For $k \in \{w, l\}$, the score is

$$S_\theta^k(R) = \frac{1}{|U_R|} \sum_{c \in U_R} \frac{1}{|R_c|} \sum_{\mathbf{u} \in R_c} \log P_\theta^k\big(\hat{Y}^k(\mathbf{u}) \mid x, \mathbf{u}\big), \tag{7}$$

and the reference score $S_{\text{ref}}^k(R)$ is defined analogously by replacing P_θ^k with P_{ref}^k.

Region-localized preference loss. Treating the score difference $S^k_\theta(R) - S^k_{\text{ref}}(R)$ as the implicit reward r_θ of Eq. (1), we instantiate the Bradley–Terry objective of Eq. (2) for template-conditional segmentation:

$$\mathcal{L}_{\text{RLPO}}(\theta) = -\log \sigma\big(\beta\big[\big(S^w_\theta(R) - S^w_{\text{ref}}(R)\big) - \big(S^l_\theta(R) - S^l_{\text{ref}}(R)\big)\big]\big)\,. \tag{8}$$

Intuitively, minimizing $\mathcal{L}_{\text{RLPO}}$ produces the same winner-up/loser-down dynamic as standard DPO, but applied to region-localized segmentation scores: within R, f_θ becomes more confident in the winner template's prediction $\hat{Y}^w$ and less confident in the loser's prediction $\hat{Y}^l$, with f_{ref} anchoring both shifts. The full derivation of $\mathcal{L}_{\text{RLPO}}$ from the standard DPO formulation is provided in Appendix A.

3.4 Consistency Regularization

The preference loss in Eq. (8) only updates the model within R, leaving the loser branch unconstrained elsewhere. Without an additional anchor, the strong preference signal applied on R can shift the loser's predictions outside R in unintended directions. To prevent this, we add a consistency regularizer that keeps the loser-template prediction aligned with the winner pseudo-label $\hat{Y}^w$ outside R, restricted to pixels where the winner is sufficiently confident. We use the Lovász–Softmax loss [72]:

$$\mathcal{L}_{\text{cons}}(\theta) = \mathcal{L}_{\text{Lovász}}\Big(P^l_\theta\big|_{R_{\text{conf}}},\ \hat{Y}^w\big|_{R_{\text{conf}}}\Big)\,, \tag{9}$$

where R_{conf} denotes the pixels outside R at which confidence exceeds a threshold τ_{conf}.

The full per-shot objective combines preference and consistency:

$$\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_{\text{RLPO}}(\theta) + \lambda_{\text{cons}}\,\mathcal{L}_{\text{cons}}(\theta), \tag{10}$$

where λ_{cons} balances the two terms. Following the streaming protocol of Section 3.1, each incoming image produces exactly one minimization step of $\mathcal{L}_{\text{total}}(\theta)$.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on the MESS benchmark [41], a suite designed to stress-test open-vocabulary segmentation models on domains that substantially differ from web-scale image-text pretraining. MESS spans five domain groups—General [73–76], Earth Monitoring [77–80], Medical Sciences [81–83], Engineering [84–87], and Agriculture & Biology [88–90]—each containing multiple datasets with distinct visual appearance, label granularity, and vocabulary. We exclude four datasets from the original MESS benchmark—Dark Zurich [91], DRAM [92], ISPRS Potsdam [93], and CryoNuSeg [94]—as they either lack a training split or are no longer publicly accessible. Following the original benchmark protocol, we report mean intersection-over-union (mIoU, %) per domain and the overall mean across datasets; per-dataset results are provided in Appendix F. Adaptation samples are drawn from each dataset's training split, while evaluation is performed on the official held-out split.

Base OVSS models. We apply our adaptation framework to two OVSS backbones, SAN [18] and CAT-Seg [30], each in two CLIP-scale variants (ViT-B/16 and ViT-L/14). All backbones are kept frozen except for lightweight adapters: a LoRA module of rank 4 on the vision branch and a residual prompt embedding on the text branch. For adaptation, we draw $K = 14$ prompt templates from the ViLD prompt pool [50]. The full template list and an analysis of template diversity are provided in Appendix B. At evaluation time, we use each backbone's default template (the ViLD pool for SAN and `"a photo of a [CLASS] in the scene"` for CAT-Seg) to match its baseline configuration.

Preference oracle. The binary preference for each query (R, a, b) is provided by an oracle that compares the two candidate predictions to the ground-truth segmentation within R: the winner $w \in \{a, b\}$ is the template whose prediction has the higher intersection-over-union with the ground truth restricted to R. This mimics an annotator who, given two segmentations cropped to a small region, selects the one closer to the intended target. We verify that this is a faithful stand-in for a human annotator in Appendix D.

Table 1: Quantitative evaluation on the MESS benchmark (mIoU, %). Numbers for our method are mean ± standard deviation. *Dense-mask* rows use ground-truth mask supervision under the same streaming protocol, providing the dense-supervision counterpart.

VLM	Method	General	Earth Monit.	Medical Sci.	Engineering	Agri. & Biology	**Mean**
CLIP ViT-B/16	SAN-B	21.60	26.90	33.51	28.19	15.99	25.29
	+ Ours	22.09 ±0.18	27.69 ±0.60	52.97 ±1.04	33.47 ±0.21	30.36 ±0.23	32.39 ±0.09
	+ Dense-mask	*24.74 ±0.33*	*28.59 ±0.44*	*44.37 ±0.60*	*35.64 ±0.72*	*31.48 ±2.95*	*32.41 ±0.46*
	CAT-Seg-B	34.51	34.52	37.82	29.95	28.95	33.12
	+ Ours	36.06 ±1.85	35.88 ±0.60	52.66 ±2.48	43.82 ±2.16	31.04 ±0.46	39.67 ±0.08
	+ Dense-mask	*38.69 ±0.13*	*38.22 ±0.35*	*60.08 ±1.37*	*46.01 ±1.37*	*41.61 ±0.87*	*44.26 ±0.49*
CLIP ViT-L/14	SAN-L	26.23	34.51	32.00	24.15	19.24	27.40
	+ Ours	27.43 ±0.37	37.09 ±0.78	38.62 ±2.69	36.13 ±0.35	31.15 ±3.41	33.99 ±0.90
	+ Dense-mask	*31.18 ±0.31*	*35.74 ±0.73*	*50.05 ±1.30*	*38.03 ±0.26*	*35.86 ±0.64*	*37.64 ±0.15*
	CAT-Seg-L	39.36	35.64	29.52	34.22	36.41	35.26
	+ Ours	42.48 ±0.35	40.28 ±0.35	51.83 ±0.81	52.68 ±0.41	42.86 ±1.45	45.88 ±0.38
	+ Dense-mask	*44.04 ±0.41*	*41.82 ±1.69*	*55.48 ±1.98*	*49.18 ±0.69*	*44.48 ±0.78*	*46.67 ±0.74*

Baselines. We compare against two reference points: (i) the *baseline* OVSS model with no adaptation, evaluated zero-shot on each target domain; and (ii) the *Dense-mask* reference, which follows the same streaming adaptation protocol but replaces the binary preference with ground-truth mask supervision, isolating the effect of the supervision format under a matched budget rather than serving as a fully-supervised upper bound. Unless stated otherwise, both our method and the Dense-mask reference adapt each backbone with 64 target domain images per dataset[1]. The effect of varying the number of training images is studied in Section 4.3. All adaptation results are averaged over three seeds, each with independently sampled adaptation images. A broader comparison, against additional baselines, is provided in Appendix E. Additional implementation details are provided in Appendix C.

4.2 Main Results

Table 1 reports mIoU on the MESS benchmark. Our method consistently improves over the baseline across all backbones, with mean gains of $+7.10$, $+6.55$, $+6.59$, and $+10.62$ mIoU on SAN-B, CAT-Seg-B, SAN-L, and CAT-Seg-L, surpassing the dense-mask reference on several specialized domains (e.g., Medical Sciences on SAN-B, Engineering on CAT-Seg-L). The improvement is robust to backbone scale (ViT-B/16 vs ViT-L/14) and architecture, suggesting that prompt disagreement provides a generic source of supervision rather than one specific to a particular OVSS design.

Our method is most effective on domains with the largest gap from the pretrained distribution. Medical Sciences shows the largest improvement on every backbone (e.g., $+19.46$ on SAN-B and $+22.31$ on CAT-Seg-L), followed by Engineering and Agriculture & Biology. These results demonstrate the applicability of our method to specialized-domain adaptation.

Figure 3 shows qualitative comparisons for SAN and CAT-Seg across two CLIP scales (ViT-B/16 and ViT-L/14). The baseline zero-shot predictions often miss or mislabel domain-specific objects—for example, confusing fine-grained bird species in CUB-200 or hallucinating non-existent classes in aerial scenes— while our adapted predictions align much more closely with the ground truth. Improvements are consistent across both scales, indicating that the gains observed in Table 1 translate into perceptually clearer segmentations. Additional qualitative results spanning all five domain groups for each backbone are provided in Appendix I.

4.3 Ablation Studies

In this section, we ablate three components of our framework: the loss formulation, the candidate source, and the adaptation budget. All ablations are conducted on CAT-Seg-L and follow the default protocol of Section 4.2. A hyperparameter sensitivity analysis is provided in Appendix G.

Candidate generation. We use prompt disagreement as the source of comparison candidates for preference queries. Prior visual preference learning methods rely on auxiliary mechanisms such as

[1]For datasets with fewer training set images, results are reported with 8 (CHASE DB1) and 32 (CWFID) images.

Table 2: **Comparison of candidate generation strategies.** We contrast prompt disagreement against MC Dropout and test-time augmentation, with matched candidate size ($K = 14$).

Domain	Baseline	Strategies		
		MC Dropout	TTA	Prompt
General	39.36	36.48	40.95	**42.48**
Earth	35.64	32.10	37.73	**40.28**
Medical	29.52	42.64	**53.15**	51.83
Engin.	34.22	47.66	50.68	**52.68**
Agri.& Bio	36.41	36.90	42.36	**42.86**
Mean	35.26	39.09	44.66	**45.88**

Table 3: **Loss function ablation.** We ablate the two components of our objective: $\mathcal{L}_{\text{RLPO}}$ (preference loss within the queried region) and $\mathcal{L}_{\text{cons}}$ (consistency regularizer outside)

Domain	Baseline	Ablation		
		w/o $\mathcal{L}_{\text{RLPO}}$	w/o $\mathcal{L}_{\text{cons}}$	**Ours**
General	39.36	41.80	41.80	**42.48**
Earth	35.64	38.52	38.11	**40.28**
Medical	29.52	36.25	48.93	**51.83**
Engin.	34.22	44.83	47.88	**52.68**
Agri.& Bio	36.41	39.97	41.41	**42.86**
Mean	35.26	40.51	43.45	**45.88**

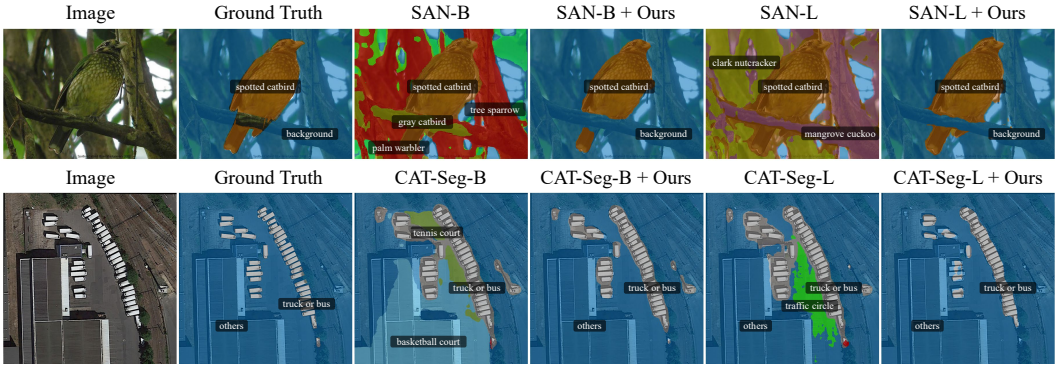


Figure 3: Qualitative comparison on the MESS benchmark. Each row shows one sample, comparing the zero-shot baseline against our adapted prediction at both CLIP scales (ViT-B/16 and ViT-L/14). Top: SAN on CUB-200 (Agriculture & Biology). Bottom: CAT-Seg on iSAID (Earth Monitoring).

stochastic sampling, input perturbation, or output sampling to produce candidate diversity [49, 63, 64, 95]. We instead exploit a source of variation already present in the prompted OVSS interface. This choice has two principled advantages. First, template-induced predictions vary the segmentation hypothesis while preserving the target vocabulary, so candidates remain comparable under the same semantic target—unlike input perturbation, which can shift the visual content itself. Second, no additional mechanism or hyperparameter (e.g., dropout rate, augmentation strength) is introduced; the diversity is obtained for free from the existing prompting interface.

To verify this design empirically, we compare prompt disagreement against two representative alternatives: MC Dropout [96] (stochastic forward) and test-time augmentation [97] (input perturbation). For MC Dropout, we apply dropout with rate 0.1 during inference. For test-time augmentation, we generate variants of each image through horizontal flips and random scaling between $0.75\times$ and $1.25\times$. Table 2 reports results with a matched ensemble size of $K = 14$. Prompt disagreement achieves the highest mean mIoU (45.88), outperforming MC Dropout (39.09) and test-time augmentation (44.66). In Appendix B, we further analyze the role of template diversity and observe that combining different types of prompt variation is essential.

Loss components. Our framework combines a region-localized preference loss, $\mathcal{L}_{\text{RLPO}}$, for supervising the queried region with a consistency regularizer, $\mathcal{L}_{\text{cons}}$, for stabilizing the loser-template prediction outside that region. Table 3 ablates these two components. Removing $\mathcal{L}_{\text{RLPO}}$ reduces the mean mIoU from 45.88 to 40.51 (-5.37), whereas removing $\mathcal{L}_{\text{cons}}$ yields 43.45 (-2.43). This indicates that RLPO is the primary driver of adaptation, while the consistency regularizer provides complementary stability by mitigating unintended drift outside the queried region.

Number of training images. Table 4 reports the effect of varying the number of target images per dataset from 4 to 128. The mean mIoU improves rapidly with more images at first, gaining $+3.00$, $+4.24$, and $+7.05$ over the zero-shot baseline at 4, 8, and 16 images, respectively, and continues improving up to 64 images ($+10.62$). Beyond this point, the mean gain saturates: 128 images yield $+10.53$, essentially the same as 64. Different domains exhibit different sample efficiency. On Medical

Table 4: **Effect of training set size.** We vary the number of target-domain training images per dataset from 0, corresponding to the zero-shot baseline, to 128. We report mIoU for each domain group and the gain over the zero-shot baseline.

	Number of training images						
Domain group	0	4	8	16	32	64	128
General	39.36	39.57 (+0.21)	40.52 (+1.16)	40.85 (+1.49)	41.96 (+2.60)	42.48 (+3.12)	**42.56** (+3.20)
Earth Monitoring	35.64	38.22 (+2.58)	37.38 (+1.74)	39.50 (+3.86)	38.69 (+3.05)	**40.28** (+4.64)	39.35 (+3.71)
Medical Sciences	29.52	41.03 (+11.51)	42.75 (+13.23)	46.30 (+16.78)	48.89 (+19.37)	**51.83** (+22.31)	51.42 (+21.90)
Engineering	34.22	36.28 (+2.06)	39.19 (+4.97)	45.20 (+10.98)	50.57 (+16.35)	52.68 (+18.46)	**54.13** (+19.91)
Agriculture & Biology	36.41	36.42 (+0.01)	38.12 (+1.71)	40.16 (+3.75)	41.36 (+4.95)	**42.86** (+6.45)	41.93 (+5.52)
Mean	35.26	38.26 (+3.00)	39.50 (+4.24)	42.31 (+7.05)	44.20 (+8.94)	**45.88** (+10.62)	45.79 (+10.53)

Sciences, our method already gains $+11.51$ mIoU at 4 images and saturates by 64 images, while on Engineering the gain grows more gradually and peaks at 128 images ($+19.91$). These results indicate that preference-guided adaptation is sample-efficient, particularly on specialized domains, where most of the gain is achieved with only a handful of preference queries.

4.4 Robustness Analysis

Our method is designed to keep annotation simple by asking the user for only a binary choice between two region-localized candidates. In practice, however, users may still occasionally make mistakes, marking the wrong candidate as preferred. To assess how sensitive our method is to such errors, we simulate noisy oracle feedback: after the oracle selects a winner w and a loser l for each query, we flip the two labels with probability p. We sweep p from 0 (clean) to 0.20 on CAT-Seg-L and report the mIoU gain over the zero-shot baseline.

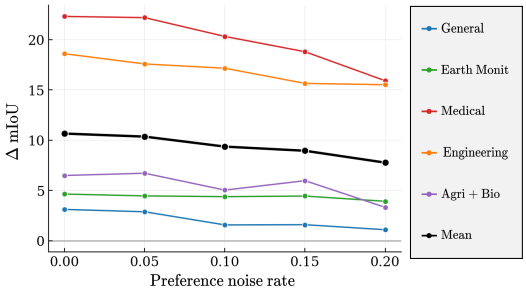


Figure 4: **Robustness to noisy preferences.** We flip a fraction p of preference labels and report the mIoU gain over the zero-shot baseline for each domain as well as the mean across datasets.

Figure 4 shows the result. Our method is robust to a substantial level of preference noise: at $p = 0.05$ the gain is essentially unchanged from the clean setting, and even at $p = 0.20$ the method still recovers roughly 8 mIoU on average. Medical Sciences and Engineering, which benefit most from adaptation, also retain the largest gains under heavy noise, indicating that the robustness extends to the regimes where our method is most useful. This robustness suggests that our framework can tolerate the imperfect feedback expected in real world applications.

5 Conclusion

We presented a preference-guided adaptation framework for open-vocabulary semantic segmentation that turns prompt disagreement into an actionable supervision signal. By mining localized comparison queries from prompt ensembles and learning from them via Region-Localized Preference Optimization with consistency regularization, our method adapts OVSS models to specialized target domains using only binary preferences over small image regions. Across the MESS benchmark, we observed consistent improvements without dense masks. The method remains sample-efficient at small adaptation budgets and robust under noisy preference feedback, suggesting that prompt-induced variation can serve as a practical supervision signal for OVSS adaptation in regimes where dense annotation is costly.

6 Limitations

Our framework assumes that the target vocabulary can be reasonably expressed through natural-language prompts and that prompt-induced candidates contain at least one useful hypothesis within the queried region. When all templates produce similarly inaccurate predictions, the preference signal becomes less informative, since selecting the "less wrong" candidate provides limited guidance toward the correct segmentation. This limitation is likely to arise for highly domain-specific concepts whose terminology, visual appearance, or label granularity is not well captured by natural-image-style templates. Extending preference-guided adaptation with richer prompt sources, such as learned, domain-specific, or expert-provided templates, is a promising direction for future work.

7 Acknowledgement

This research was supported by the Challengeable Future Defense Technology Research and Development Program through the Agency For Defense Development (ADD) funded by the Defense Acquisition Program Administration (DAPA) in 2026 (No.915102201).

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2017.

[2] Liang-Chieh Chen. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[4] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[5] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43 (10):3349–3364, 2020.

[6] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34: 17864–17875, 2021.

[7] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[8] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[9] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer, 2020.

[10] Jihun Kim, Hoyong Kwon, Hyeokjun Kweon, and Kuk-Jin Yoon. Bootstrapping video semantic segmentation model via distillation-assisted test-time adaptation. *arXiv preprint arXiv:2604.10950*, 2026.

[11] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022.

[12] 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*, pages 540–557. Springer, 2022.

[13] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European conference on computer vision*, pages 736–753. Springer, 2022.

[14] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, pages 23033–23044. PMLR, 2023.

[15] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary universal image segmentation with maskclip. *arXiv preprint arXiv:2208.08984*, 2022.

[16] 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*, pages 7061–7070, 2023.

[17] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15116–15127, 2023.

[18] 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*, pages 2945–2954, 2023.

[19] 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*, pages 2955–2966, 2023.

[20] Xin Xu, Tianyi Xiong, Zheng Ding, and Zhuowen Tu. Masqclip for open-vocabulary universal image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 887–898, 2023.

[21] Xi Chen, Shuang Li, Ser-Nam Lim, Antonio Torralba, and Hengshuang Zhao. Open-vocabulary panoptic segmentation with embedding modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1141–1150, 2023.

[22] Xudong Wang, Shufan Li, Konstantinos Kallidromitis, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Hierarchical open-vocabulary universal image segmentation. *Advances in Neural Information Processing Systems*, 36:21429–21453, 2023.

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

[24] Chaofan Ma, Yang Yuhuan, Chen Ju, Fei Zhang, Ya Zhang, and Yanfeng Wang. Attrseg: open-vocabulary semantic segmentation via attribute decomposition-aggregation. *Advances in neural information processing systems*, 36:10258–10270, 2023.

[25] 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.

[26] Jie Qin, Jie Wu, Pengxiang Yan, Ming Li, Ren Yuxi, Xuefeng Xiao, Yitong Wang, Rui Wang, Shilei Wen, Xin Pan, et al. Freeseg: Unified, universal and open-vocabulary image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19446–19455, 2023.

[27] Cong Han, Yujie Zhong, Dengjie Li, Kai Han, and Lin Ma. Open-vocabulary semantic segmentation with decoupled one-pass network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1086–1096, 2023.

[28] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1020–1031, 2023.

[29] Jingxuan Xu, Wuyang Chen, Yao Zhao, and Yunchao Wei. Transferable and principled efficiency for open-vocabulary segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15814–15824, 2024.

[30] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024.

[31] Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3426–3436, 2024.

[32] Xiangheng Shan, Dongyue Wu, Guilin Zhu, Yuanjie Shao, Nong Sang, and Changxin Gao. Open-vocabulary semantic segmentation with image embedding balancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28412–28421, 2024.

[33] Siyu Jiao, Hongguang Zhu, Jiannan Huang, Yao Zhao, Yunchao Wei, and Humphrey Shi. Collaborative vision-text representation optimizing for open-vocabulary segmentation. In *European Conference on Computer Vision*, pages 399–416. Springer, 2024.

[34] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11175–11185, 2023.

[35] 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*, pages 14998–15008, 2025.

[36] 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, pages 6254–6262, 2025.

[37] 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*, pages 9804–9815, 2025.

[38] Ziyu Zhao, Xiaoguang Li, Lingjia Shi, Nasrin Imanpour, and Song Wang. Dpseg: dual-prompt cost volume learning for open-vocabulary semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25346–25356, 2025.

[39] Sung-Hoon Yoon, Hoyong Kwon, Changgyoon Oh, and Kuk-Jin Yoon. Dinode: Continuous vision-text alignment for open-vocabulary semantic segmentation. In *European Conference on Computer Vision*, pages 258–277. Springer, 2026.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[41] Benedikt Blumenstiel, Johannes Jakubik, Hilde Kühne, and Michael Vössing. What a MESS: Multi-Domain Evaluation of Zero-shot Semantic Segmentation. *Advances in Neural Information Processing Systems*, 2023.

[42] Gonca Yilmaz, Songyou Peng, Marc Pollefeys, Francis Engelmann, and Hermann Blum. Opendas: Open-vocabulary domain adaptation for 2d and 3d segmentation. *arXiv preprint arXiv:2405.20141*, 2024.

[43] Rabin Adhikari, Safal Thapaliya, Manish Dhakal, and Bishesh Khanal. Tunevlseg: Prompt tuning benchmark for vision-language segmentation models. In *Proceedings of the Asian Conference on Computer Vision*, pages 126–144, 2024.

[44] Kanchan Poudel, Manish Dhakal, Prasiddha Bhandari, Rabin Adhikari, Safal Thapaliya, and Bishesh Khanal. Exploring transfer learning in medical image segmentation using vision-language models. *arXiv preprint arXiv:2308.07706*, 2023.

[45] Manish Dhakal, Rabin Adhikari, Safal Thapaliya, and Bishesh Khanal. Vlsm-adapter: Fine-tuning vision-language segmentation efficiently with lightweight blocks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 712–722. Springer, 2024.

[46] Ujjwal Mishra, Vinita Shukla, Praful Hambarde, and Amit Shukla. Improvise, adapt, overcome–telescopic adapters for efficient fine-tuning of vision language models in medical imaging. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7605–7615, 2026.

[47] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[48] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[49] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.

[50] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[51] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022.

[52] Aditya Gandhamal, Aniruddh Sikdar, and Suresh Sundaram. Ov-coast: Cost aggregation with optimal transport for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2506.03706*, 2025.

[53] 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*, 2017.

[54] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[55] Sunghyun Park, Jungsoo Lee, Shubhankar Borse, Munawar Hayat, Sungha Choi, Kyuwoong Hwang, and Fatih Porikli. Personalized ovss: Understanding personal concept in open-vocabulary semantic segmentation. *arXiv preprint arXiv:2507.11030*, 2025.

[56] Ashesh Jain, Brian Wojcik, Thorsten Joachims, and Ashutosh Saxena. Learning trajectory preferences for manipulators via iterative improvement. *Advances in neural information processing systems*, 26, 2013.

[57] Róbert Busa-Fekete, Balázs Szörényi, Paul Weng, Weiwei Cheng, and Eyke Hüllermeier. Preference-based reinforcement learning: evolutionary direct policy search using a preference-based racing algorithm. *Machine learning*, 97(3):327–351, 2014.

[58] Andras Kupcsik, David Hsu, and Wee Sun Lee. Learning dynamic robot-to-human object handover from human feedback. In *Robotics Research: Volume 1*, pages 161–176. Springer, 2017.

[59] Dorsa Sadigh, Anca Dragan, Shankar Sastry, and Sanjit Seshia. Active preference-based learning of reward functions. 2017.

[60] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR, 2024.

[61] Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024.

[62] Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *arXiv preprint arXiv:2410.07163*, 2024.

[63] Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Mingxi Cheng, Ji Li, and Liang Zheng. Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13199–13208, 2025.

[64] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8941–8951, 2024.

[65] Shentao Yang, Tianqi Chen, and Mingyuan Zhou. A dense reward view on aligning text-to-image diffusion with preference. *arXiv preprint arXiv:2402.08265*, 2024.

[66] Miaomiao Cai, Simiao Li, Wei Li, Xudong Huang, Hanting Chen, Jie Hu, and Yunhe Wang. Dspo: Direct semantic preference optimization for real-world image super-resolution. *arXiv preprint arXiv:2504.15176*, 2025.

[67] Rongyuan Wu, Lingchen Sun, Zhengqiang Zhang, Shihao Wang, Tianhe Wu, Qiaosi Yi, Shuai Li, and Lei Zhang. DP^{2}O-SR: Direct perceptual preference optimization for real-world image super-resolution. *arXiv preprint arXiv:2510.18851*, 2025.

[68] Aishik Konwer, Zhijian Yang, Erhan Bas, Cao Xiao, Prateek Prasanna, Parminder Bhatia, and Taha Kass-Hout. Enhancing sam with efficient prompting and preference optimization for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20990–21000, 2025.

[69] Yonghuang Wu, Wenwen Zeng, Xuan Xie, Chengqian Zhao, Guoqing Wu, and Jinhua Yu. Sampo: Visual preference optimization for intent-aware segmentation with vision foundation models. *arXiv preprint arXiv:2508.02464*, 2025.

[70] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[71] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

[72] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4413–4421, 2018.

[73] Seyed Mohammad Hassan Erfani, Zhenyao Wu, Xinyi Wu, Song Wang, and Erfan Goharian. Atlantis: A benchmark for semantic segmentation of waterbody images. *Environmental Modelling & Software*, 149:105333, 2022.

[74] Xiongwei Wu, Xin Fu, Ying Liu, Ee-Peng Lim, Steven CH Hoi, and Qianru Sun. A large-scale benchmark for food image segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 506–515, 2021.

[75] Jianshu Li, Jian Zhao, Yunchao Wei, Congyan Lang, Yidong Li, Terence Sim, Shuicheng Yan, and Jiashi Feng. Multiple-human parsing in the wild. *arXiv preprint arXiv:1705.07206*, 2017.

[76] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020.

[77] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. UAVid: A semantic segmentation dataset for UAV imagery. *ISPRS journal of photogrammetry and remote sensing*, 165:108–119, 2020.

[78] Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Roberson Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021.

[79] Gonzalo Mateo-Garcia, Joshua Veitch-Michaelis, Lewis Smith, Silviu Vlad Oprea, Guy Schumann, Yarin Gal, Atılım Güneş Baydin, and Dietmar Backes. Towards global flood mapping onboard low cost satellites with machine learning. *Scientific reports*, 11(1):1–12, 2021.

[80] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. iSAID: A large-scale dataset for instance segmentation in aerial images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019.

[81] Constantin Seibold, Simon Reiß, Saquib Sarfraz, Matthias A. Fink, Victoria Mayer, Jan Sellner, Moon Sung Kim, Klaus H. Maier-Hein, Jens Kleesiek, and Rainer Stiefelhagen. Detailed Annotations of Chest X-Rays via CT Projection for Report Understanding. In *Proceedings of the 33th British Machine Vision Conference (BMVC)*, 2022.

[82] Muhammad Moazam Fraz, Paolo Remagnino, Andreas Hoppe, Bunyarit Uyyanonvara, Alicja R Rudnicka, Christopher G Owen, and Sarah A Barman. An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Transactions on Biomedical Engineering*, 59(9):2538–2548, 2012.

[83] Debesh Jha, Sharib Ali, Krister Emanuelsen, Steven A Hicks, Vajira Thambawita, Enrique Garcia-Ceja, Michael A Riegler, Thomas de Lange, Peter T Schmidt, Håvard D Johansen, et al. Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. *MultiMedia Modeling: 27th International Conference, MMM 2021*, pages 218–229, 2021.

[84] Dina Bashkirova, Mohamed Abdelfattah, Ziliang Zhu, James Akl, Fadi Alladkani, Ping Hu, Vitaly Ablavsky, Berk Calli, Sarah Adel Bargal, and Kate Saenko. Zerowaste dataset: Towards deformable object segmentation in cluttered scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21147–21157, 2022.

[85] Shreyas S Shivakumar, Neil Rodrigues, Alex Zhou, Ian D Miller, Vijay Kumar, and Camillo J Taylor. PST900: RGB-thermal calibration, dataset and segmentation network. *2020 IEEE international conference on robotics and automation (ICRA)*, pages 9441–9447, 2020.

[86] Yahui Liu, Jian Yao, Xiaohu Lu, Renping Xie, and Li Li. Deepcrack: A deep hierarchical feature learning architecture for crack segmentation. *Neurocomputing*, 338:139–153, 2019. doi: 10.1016/j.neucom.2019.01.036.

[87] Eric Bianchi and Matthew Hebdon. Corrosion condition state semantic segmentation dataset. *University Libraries, Virginia Tech: Blacksburg, VA, USA*, 2021.

[88] Sebastian Haug and Jörn Ostermann. A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks. In *Computer Vision - ECCV 2014 Workshops*, pages 105–116. Springer, 2015.

[89] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. Caltech-UCSD Birds 200. *California Institute of Technology*, 2011.

[90] Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic segmentation of underwater imagery: Dataset and benchmark. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1769–1776, 2020.

[91] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7374–7383, 2019.

[92] Nadav Cohen, Yael Newman, and Ariel Shamir. Semantic Segmentation in Art Paintings. *Computer Graphics Forum*, 41(2):261–275, 2022. ISSN 1467-8659. doi: 10.1111/cgf.14473.

[93] BSF Swissphoto. Isprs potsdam dataset within the isprs test project on urban classification, 3d building reconstruction and semantic labeling. https://www.isprs.org/education/ benchmarks/UrbanSemLab/default.aspx, 2012. URL `https://www.isprs.org/education/benchmarks/UrbanSemLab/default.aspx`.

[94] Amirreza Mahbod, Gerald Schaefer, Benjamin Bancher, Christine Löw, Georg Dorffner, Rupert Ecker, and Isabella Ellinger. CryoNuSeg: A dataset for nuclei instance segmentation of cryosectioned H&E-stained histological images. *Computers in biology and medicine*, 132: 104349, 2021.

[95] Shamil Ayupov, Maksim Nakhodnov, Anastasia Yaschenko, Andrey Kuznetsov, and Aibek Alanov. Dreamboothdpo: Improving personalized generation using direct preference optimization, 2025. URL `https://arxiv.org/abs/2505.20975`.

[96] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[97] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

[98] Daniele Calandriello, Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko, Tianqi Liu, et al. Human alignment of large language models through online preference optimisation. *arXiv preprint arXiv:2403.08635*, 2024.

[99] Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. *arXiv preprint arXiv:2312.11456*, 2023.

[100] 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 *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 18061–18070. IEEE, 2022.

A Derivation of the Region-Localized Preference Optimization Loss

We derive the RLPO loss by specializing the standard DPO framework [48] to OVSS. Section A.1 compresses the standard derivation; Section A.2 carries out the specialization, including the cross-template Bradley–Terry comparison, the class-balanced regional score, and the online-DPO interpretation.

A.1 DPO Recap

We briefly recall the standard DPO derivation [48]. Given a generative policy $\pi(y \mid c)$ over responses $y \in \mathcal{Y}$ conditioned on a context c, the KL-regularized objective

$$\max_{\pi} \; \mathbb{E}_{c\sim\mathcal{D}}\big[\mathbb{E}_{y\sim\pi(\cdot|c)}[\, r(c,y)\,] \;-\; \beta\, D_{\mathrm{KL}}(\pi(\cdot \mid c)\,\|\,\pi_{\mathrm{ref}}(\cdot \mid c))\big] \tag{11}$$

admits the closed-form per-context optimum

$$\pi^*(y \mid c) = \frac{1}{Z(c)}\,\pi_{\mathrm{ref}}(y \mid c)\exp\Big(\tfrac{1}{\beta} r(c,y)\Big)\,, \tag{12}$$

which can be equivalently rearranged as a log-ratio expression for the reward,

$$r(c,y) = \beta \log \frac{\pi^*(y \mid c)}{\pi_{\mathrm{ref}}(y \mid c)} + \beta \log Z(c). \tag{13}$$

Substituting Eq. (13) into the Bradley–Terry preference $P(y_w \succ y_l \mid c) = \sigma(r(c,y_w) - r(c,y_l))$ [70], the context-dependent normalizers $\beta \log Z(c)$ cancel, yielding the DPO loss

$$\mathcal{L}_{\mathrm{DPO}}(\theta) = -\mathbb{E}_{(c,y_w,y_l)\sim\mathcal{D}_{\mathrm{pref}}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid c)}{\pi_{\mathrm{ref}}(y_w \mid c)} - \beta\log\frac{\pi_\theta(y_l \mid c)}{\pi_{\mathrm{ref}}(y_l \mid c)}\right)\right]. \tag{14}$$

The remainder of this appendix specializes Eq. (14) to our setting.

A.2 Adaptation to Template-Conditional Segmentation

Setting. We instantiate the abstract context c of Section A.1 as $c = (x, t_k)$, an input image $x \in \mathbb{R}^{H\times W\times 3}$ paired with one of K prompt templates $t_k \in \mathcal{T}$; the term *prompt* hereafter refers exclusively to these OVSS templates, while *context* denotes the DPO input as in Section A.1. The response of template k is the segmentation it induces:

$$\hat{Y}^k = \big\{\hat{Y}^k(\mathbf{u})\big\}_{\mathbf{u}\in\Omega}, \qquad \hat{Y}^k(\mathbf{u}) = \arg\max_{c'\in\mathcal{C}} P_\theta^k(c' \mid x, \mathbf{u}). \tag{15}$$

Under a pixelwise-independence factorization $\pi_\theta(\hat{Y}^k \mid x, t_k) = \prod_{\mathbf{u}\in\Omega} P_\theta^k(\hat{Y}^k(\mathbf{u}) \mid x, \mathbf{u})$, which is standard for OVSS heads producing per-pixel softmax outputs, the response log-likelihood decomposes additively across pixels,

$$\log \pi_\theta(\hat{Y}^k \mid x, t_k) = \sum_{\mathbf{u}\in\Omega} \log \max_{c'\in\mathcal{C}} P_\theta^k(c' \mid x, \mathbf{u}), \tag{16}$$

where $\log P_\theta^k(\hat{Y}^k(\mathbf{u}) \mid x, \mathbf{u}) = \log \max_{c'} P_\theta^k(c' \mid x, \mathbf{u})$ holds by the definition of $\hat{Y}^k(\mathbf{u})$.

Cross-template Bradley–Terry comparison. Standard DPO compares two responses under a *shared* context, so that the context-dependent normalizers $\beta \log Z(c)$ in Eq. (13) cancel inside the BT difference. In our setting the winner and loser responses arise from *sibling* contexts $c_w = (x, t_w)$ and $c_l = (x, t_l)$ that share the underlying image x and the target vocabulary $\mathcal{C}$ but differ in the prompt template. Substituting Eq. (13) into the BT preference therefore yields

$$\begin{aligned} &r(c_w, \hat{Y}^w) - r(c_l, \hat{Y}^l) \\ &\quad= \beta\left[\log\frac{\pi_\theta(\hat{Y}^w \mid c_w)}{\pi_{\mathrm{ref}}(\hat{Y}^w \mid c_w)} - \log\frac{\pi_\theta(\hat{Y}^l \mid c_l)}{\pi_{\mathrm{ref}}(\hat{Y}^l \mid c_l)}\right] \\ &\qquad+ \underbrace{\beta[\log Z(c_w) - \log Z(c_l)]}_{=:\,\Delta_Z(c_w,c_l),\ \text{constant in } \theta}. \end{aligned} \tag{17}$$

The residual term $\Delta_Z(c_w, c_l)$ does not vanish in general, but it depends only on the ground-truth reward and the reference policy, not on the trainable parameters θ. As an additive offset inside the BT log-sigmoid loss it contributes no gradient to the optimization, so the closed form of Eq. (14) extends to our cross-template setting up to a θ-independent constant. The KL regularization of Eq. (11) is thus inherited per template: each context $c_k = (x, t_k)$ is anchored to its own $\pi_{\text{ref}}(\cdot \mid c_k)$, which is precisely the regularization our adaptation needs.

Region restriction. Preferences in our setting are elicited over a query region $R \subseteq \Omega$ rather than the full image. Restricting the sum in Eq. (16) to R yields a region-localized log-likelihood,

$$\bar{S}^k_\theta(R) = \sum_{\mathbf{u} \in R} \log \max_{c' \in \mathcal{C}} P^k_\theta(c' \mid x, \mathbf{u}), \tag{18}$$

which is the strict consequence of Eq. (16) under region restriction.

Class-balanced averaging. Eq. (18) weights each pixel uniformly, so a class that occupies most of R dominates the score and can suppress contributions from smaller but semantically important classes. To prevent this, we replace the uniform sum with a class-balanced average that equalizes contributions across present classes:

$$S^k_\theta(R) = \frac{1}{|\mathcal{U}_R|} \sum_{c \in \mathcal{U}_R} \frac{1}{|R_c|} \sum_{\mathbf{u} \in R_c} \log \max_{c' \in \mathcal{C}} P^k_\theta(c' \mid x, \mathbf{u}), \tag{19}$$

with buckets $R_c = \{\mathbf{u} \in R : \hat{Y}^w(\mathbf{u}) = c\}$ and present-class set $\mathcal{U}_R = \{c \in \mathcal{C} : R_c \neq \emptyset\}$. We use the winner's hard prediction $\hat{Y}^w$ to define a stable, k-independent partition: the same buckets $\{R_c\}$ are used in all four score evaluations $\{(\theta, w), (\theta, l), (\text{ref}, w), (\text{ref}, l)\}$, so the BT difference depends only on the per-template logits, not on which template's prediction is used to bucket. We view this class-balanced averaging as a deliberate departure from the strict derivation, motivated by class imbalance within R, rather than a derived consequence of Eq. (14). The same construction defines $S^k_{\text{ref}}(R)$ by replacing P^k_θ with P^k_{ref}.

Online-DPO interpretation. Within each optimization step, $\hat{Y}^k$ is a fixed discrete label, so the BT closed form of Eq. (14) applies to $(y_w, y_l) = (\hat{Y}^w, \hat{Y}^l)$. Across steps, $\hat{Y}^k$ evolves with θ, placing RLPO within the family of online/iterative DPO methods [98, 99]; the $\arg\max$ corresponds to a deterministic mode response equivalent to the zero-temperature limit of softmax sampling.

The RLPO loss. Combining the cross-template BT extension with the class-balanced region score, and defining the region-localized implicit reward $r^k(R) = \beta\big[S^k_\theta(R) - S^k_{\text{ref}}(R)\big]$, the BT closed form of Eq. (14) applied to $(\hat{Y}^w, \hat{Y}^l)$ yields

$$\begin{aligned} \mathcal{L}_{\text{RLPO}}(\theta) &= -\log \sigma\big(r^w(R) - r^l(R)\big) \\ &= -\log \sigma\big(\beta\big[\big(S^w_\theta(R) - S^w_{\text{ref}}(R)\big) - \big(S^l_\theta(R) - S^l_{\text{ref}}(R)\big)\big]\big)\,. \end{aligned} \tag{20}$$

B Template Analysis

We further analyze how different prompt-template variations affect adaptation. Table 12 lists the full ViLD template pool used in our main experiments. Table 5 compares the full ViLD pool with two diagnostic template sets—a sentence-only ViLD subset and a controlled scale-only set—to isolate the roles of sentence and scale variation. The *sentence-only* subset consists of the five ViLD templates without scale modifiers (rows 1–5). The *scale-only* subset fixes the sentence structure to `a photo of a [scale] {} in the scene` and varies only the scale modifier over {`none`, `small`, `medium`, `large`}.

The sentence-only and scale-only subsets achieve similar mean mIoU, 43.79 and 43.65, respectively, suggesting that scale modifiers alone do not consistently improve segmentation quality. However, their effects are complementary across domains: scale-only templates improve over sentence-only templates in some agriculture and engineering datasets, whereas sentence-only templates are stronger in several medical datasets. Using the full 14-template set achieves the best mean mIoU of 45.88,

Table 5: Per-dataset adaptation performance for three template subsets that vary along different diversity axes. The best per dataset across the three subsets is highlighted in bold; the zero-shot baseline is shown for reference.

	General				Earth Monit.				Medical			Engineering				Agri. & Bio.			
Subset	BDD100K	MHP v1	FoodSeg103	ATLANTIS	iSAID	WorldFloods	FloodNet	UAVid	Kvasir-Inst.	CHASE_DB1	PAXRay-4	Corrosion CS	DeepCrack	PST900	ZeroWaste-f	SUIM	CUB-200	CWFID	*Mean*
Baseline	48.23	30.77	32.92	45.51	19.67	39.94	41.05	41.90	65.49	3.32	19.75	7.47	25.27	78.73	25.42	49.75	21.89	37.58	*35.26*
Sentence-only	50.76	**36.95**	36.73	**45.70**	35.74	39.90	**41.19**	43.76	80.21	19.65	46.05	21.40	60.22	80.75	24.33	55.74	22.38	46.69	*43.79*
Scale-only	**50.89**	34.67	**37.41**	45.41	32.59	40.33	41.16	**46.19**	76.12	20.14	47.09	20.94	61.44	79.98	24.76	57.71	**22.47**	46.48	*43.65*
Full pool	50.58	36.41	**37.41**	45.50	**36.17**	**41.13**	41.07	42.74	**87.41**	**20.74**	**47.34**	**22.90**	**74.39**	**80.96**	**32.47**	**58.49**	22.22	**47.87**	***45.88***

Table 6: Detailed Implementation configuration.

CAT-Seg [30]		SAN [18]	
optimizer	AdamW	optimizer	AdamW
weight decay	1e-4	weight decay	1e-4
learning rate	3e-3	learning rate	1e-3
β	0.1	β	0.1
λ_{cons}	0.1	λ_{cons}	0.1
τ_{conf}	0.8	τ_{conf}	0.8
q	0.95	q	0.95
vision LoRA rank, α	$r = 4$, $\alpha = 1.0$	vision LoRA rank, α	$r = 4$, $\alpha = 1.0$
text adapter rank	$r = 4$	text adapter rank	$r = 4$
exception lr	1e-3†	exception lr	3e-4†
†Applied to `cub_200`, `atlantis`, `isaid`, `pst900`.		†Applied to `cub_200`, `atlantis`, `pst900`.	

outperforming both subsets. This indicates that the benefit of scale-aware prompts is not simply higher standalone accuracy, but the additional candidate diversity they introduce around the same target vocabulary. By combining sentence and scale variations, the full prompt ensemble provides more diverse yet semantically aligned segmentation hypotheses, which improves preference-query mining.

C Implementation Details

Adapter configuration. The adapter consists of two lightweight modules: a vision LoRA on the CLIP image encoder and a residual adapter on the text classifier embedding. Both modules are rank 4. The vision LoRA is attached to the Q and V projections of the last four transformer blocks of the CLIP image encoder. The text adapter is a low-rank residual on the mean-of-K classifier embedding. All other backbone parameters are kept frozen.

Training protocol. We adapt each OVSS backbone with AdamW (weight decay 1e-4) and a constant learning rate, kept fixed throughout the streaming run. Across all backbones we use the same DPO temperature $\beta = 0.1$, consistency weight $\lambda_{\text{cons}} = 0.1$, confidence threshold $\tau_{\text{conf}} = 0.8$, and entropy quantile $q = 0.95$. Each adaptation step processes a single image (batch size 1), consistent with the streaming protocol of Section 3.1. The default learning rate differs by backbone. For a small subset of datasets, we additionally use a lower learning rate. The full configuration, including backbone-specific learning rates and dataset-specific exceptions, is summarized in Table 6.

D Validation of the Preference Oracle

Throughout the main experiments, each binary preference is supplied by an oracle that compares the two candidate predictions against the ground-truth mask within the queried region (Section 4.1). While this enables large-scale evaluation, it raises the question of how well the oracle reflects human judgments. We examine this in two ways: whether humans agree with the oracle (Appendix D.1) and whether the method remains robust to errors concentrated on cases where humans and the oracle disagree (Appendix D.2).

Table 7: **Agreement between human annotators and the preference oracle.** Mined preference pairs are split into three difficulty tiers by the IoU margin between the two candidates, with boundaries set to the terciles of the margin distribution observed during actual adaptation runs. Each of the 20 participants judged 30 pairs (10 per tier), for 600 judgments in total.

Tier	IoU margin	#Judgments	**Agreement**
Easy	≥ 0.18	200	0.967
Medium	$[0.04,\ 0.18)$	200	0.953
Hard	< 0.04	200	0.840
Overall	—	600	**0.920**

Table 8: **Robustness to preference noise placed where humans actually disagree** (CAT-Seg-L, mIoU %). Noise is injected only into near-tie queries whose IoU margin falls below τ. *Random* replaces the oracle choice with a coin flip; *Flip* actively inverts it, an adversarial upper bound on annotator error.

Condition	τ	General	Earth Monit.	Medical Sci.	Engineering	Agri. & Biology	**Mean**
Zero-shot	—	39.36	35.64	29.52	34.22	36.41	35.26
Clean (oracle)	—	42.48	40.28	51.83	52.68	42.86	**45.88**
Random	0.05	42.39	39.83	51.60	49.38	43.74	45.13
	0.10	42.08	39.18	50.66	51.95	40.85	44.85
	0.15	42.21	37.52	50.57	48.68	42.93	44.12
Flip	0.05	42.74	39.25	51.30	51.78	41.62	45.21
	0.10	41.62	38.93	48.86	48.40	37.33	43.02
	0.15	40.80	39.32	46.93	49.10	35.34	42.43

D.1 Human–Oracle Agreement

Protocol. We construct a preference-pair database by running Preference Query Mining (Section 3.2), the same procedure that generates comparisons during adaptation, so that the pairs shown to participants follow the same distribution as those encountered in an actual run. Pairs are grouped into *easy*, *medium*, and *hard* tiers by the IoU margin between the two candidates inside the queried region. The tier boundaries, 0.04 and 0.18, are set to the terciles of the margin distribution over all queries answered during our adaptation runs, so each tier reflects a third of the comparisons the method actually issues. Each participant judged 30 pairs, 10 sampled from each tier, which allows agreement to be estimated separately at every difficulty level rather than only in aggregate. In total, 20 participants provided 600 judgments.

Results. Table 7 reports agreement with the oracle. Humans select the same candidate as the oracle on 0.920 of all pairs, taking 9.4 seconds per judgment on average. Agreement is near-ceiling on easy and medium pairs (0.967 and 0.953) and drops on hard pairs (0.840), which is the expected pattern: when the two candidates are separated by a very small IoU margin they are close to equally good, and the choice becomes genuinely ambiguous rather than incorrect. The oracle is therefore a close proxy for human preference over the range of comparisons our method issues, with the residual disagreement concentrated in near-tie queries.

D.2 Noise Concentrated on Near-Tie Queries

Section 4.4 studies robustness under preference labels flipped uniformly at random. The agreement study above suggests a more targeted stress test, since human error is not uniform but concentrated on hard, near-tie pairs. We therefore inject noise only into queries whose IoU margin falls below a threshold τ, under two models: *Random* replaces the oracle choice with a coin flip, and *Flip* actively inverts it, which is adversarial rather than merely noisy and upper-bounds any realistic annotator.

Table 9: **Comparison with alternative adaptation strategies on CAT-Seg-L** (mIoU, %). **Annotation** denotes target-domain supervision as *type / scope*, with rows ordered by annotation cost. *Grayed* rows require stronger supervision than binary preferences and are included as references. Results are mean ± standard deviation over three runs.

Method	**Annotation**	General	Earth Monit.	Medical Sci.	Engineering	Agri. & Biology	**Mean**
CAT-Seg-L	—	39.36	35.64	29.52	34.22	36.41	35.26
+ Prompt ens.	—	39.53	37.66	25.11	34.97	33.78	34.74
+ Ours	binary preference / image	42.48 ±0.35	40.28 ±0.35	51.83 ±0.81	52.68 ±0.41	42.86 ±1.45	45.88 ±0.38
+ Weakly-sup. (point)	*click / object*	*41.79 ±0.62*	*34.20 ±0.86*	*51.86 ±2.70*	*44.15 ±2.07*	*42.43 ±0.97*	*42.41 ±0.68*
+ Prompt selection	*dense mask / image*	*39.61 ±0.29*	*40.15 ±0.73*	*35.05 ±0.10*	*39.13 ±0.04*	*35.67 ±0.26*	*38.20 ±0.08*
+ Dense-mask	*dense mask / image*	*44.04 ±0.41*	*41.82 ±1.69*	*55.48 ±1.98*	*49.18 ±0.69*	*44.48 ±0.78*	*46.67 ±0.74*
+ Supervised	*dense mask / image*	*45.83 ±0.23*	*46.96 ±1.36*	*73.03 ±0.50*	*55.91 ±0.86*	*47.73 ±0.51*	*53.17 ±0.50*

Table 8 reports the result. Performance degrades gracefully in both models. Even in the most severe setting, where every near-tie query with margin below 0.15 is actively inverted, the mean remains at 42.43, well above the 35.26 zero-shot baseline. Since humans disagree with the oracle on only 16% of hard pairs, the realistic error regime sits comfortably inside the range the method tolerates.

E Comparison with Alternative Adaptation Strategies

Section 4.1 compares our method with the zero-shot baseline and a matched-budget Dense-mask reference while keeping the adaptation protocol fixed. Because no prior method uses an equivalent preference-supervision budget, we further compare against alternatives ranging from no annotation to stronger point and dense supervision. All methods use the same CAT-Seg-L backbone and the same 64 target-domain images, with results summarized in Table 9.

Strategies compared. *Prompt ensemble* averages predictions from the K templates in our candidate pool without training or annotation. *Weakly-sup. (point)* supervises one point at the distance-transform maximum of each connected component, requiring instance-level localization. *Prompt selection* uses dense ground-truth masks to select the best template for each dataset, without updating model parameters. *Dense-mask*, introduced in Section 4.1, follows our single-step adaptation protocol but replaces the binary preference with a ground-truth mask. *Supervised* is a fully supervised upper reference, using GT-based prompt tuning following [51] and [100] on dense masks from the same 64 images for 200 epochs. Together, these baselines span supervision from annotation-free inference to full dense supervision.

Results. Prompt ensembling does not improve over the zero-shot baseline (34.74 versus 35.26). Prompt selection reaches 38.20 despite requiring per-dataset dense masks, so template choice alone does not account for the gain, and the point baseline reaches 42.41 while requiring the annotator to click every object instance. Our method surpasses both at 45.88 with a single binary judgment per image. The two mask-supervised references bracket it from above: at a matched single-step budget dense masks reach 46.67, and multi-epoch fully-supervised prompt tuning reaches 53.17, which we do not claim to match.

F Detailed Results on MESS Benchmark

We provide details of the MESS datasets used in this work and report per-dataset performance for each of the four base OVSS backbones, as well as per-dataset breakdowns of the ablation studies in Section 4.3. Table 13 lists the datasets grouped by domain, along with class count, license, and a sample of class labels.

Table 14 expands the group-level numbers in Table 1 of the main paper with per-dataset results for all four backbones (SAN-B, CAT-Seg-B, SAN-L, and CAT-Seg-L). Within each backbone, the better of the zero-shot baseline and our method is highlighted in bold; the dense-mask reference is shown for reference and excluded from the comparison.

Tables 15–17 expand the three ablation studies of Section 4.3 with per-dataset results on CAT-Seg-L. Table 15 breaks down the loss ablation of Table 3, comparing the full method against ablating $\mathcal{L}_{\text{RLPO}}$

Table 10: Sensitivity to the hyperparameters of our method. Default values and the best per domain group within each hyperparameter group are shown in bold.

Param	Value	General	Earth Monit.	Medical Sci.	Engineering	Agri. & Bio.	Mean
β	0.05	**42.50**	39.95	**52.02**	50.27	43.01	*45.33*
	0.1	42.48	40.28	51.83	**52.68**	42.86	***45.88***
	0.2	41.77	**40.31**	51.61	51.84	**43.24**	*45.57*
λ_{cons}	0.05	41.60	40.21	50.98	52.06	42.42	*45.31*
	0.1	**42.48**	**40.28**	51.83	**52.68**	**42.86**	***45.88***
	0.2	41.90	39.83	**51.96**	50.41	41.80	*44.99*
q	0.90	42.31	39.84	50.00	**53.80**	41.53	*45.47*
	0.95	**42.48**	**40.28**	51.83	52.68	**42.86**	***45.88***
	0.98	41.70	39.30	**52.65**	52.48	41.95	*45.43*
τ_{conf}	0.70	41.95	40.02	51.39	51.25	**43.43**	*45.41*
	0.80	**42.48**	40.28	**51.83**	**52.68**	42.86	***45.88***
	0.90	41.82	**40.68**	51.29	50.34	42.72	*45.19*

Table 11: Compute and memory profile of our adaptation on single NVIDIA H200 GPU.

Metric	CAT-Seg-L	SAN-L
Full model parameters	433.7 M	436.7 M
Trainable parameters	71,681	71,681
Vision LoRA	65,536	65,536
Text residual adapter	6,145	6,145
Trainable fraction	0.017 %	0.016 %
Inference cost		
Inference time (ms)	72.61	53.41
Inference peak memory (MiB)	5,869	2,918
Adaptation cost (per training step)		
Step time (ms)	1,616	418
Step peak memory (MiB)	7,335	8,392

and $\mathcal{L}_{\text{cons}}$. Table 16 contrasts prompt disagreement against MC Dropout and test-time augmentation under matched candidate size, expanding Table 2. Finally, Table 17 reports the effect of varying the number of target-domain training images per dataset from 0 (zero-shot baseline) to 128, expanding Table 4. The best per dataset is highlighted in bold across all three ablation tables.

G Hyperparameter Sensitivity

We analyze the sensitivity of our method to four hyperparameters on CAT-Seg-L: the DPO temperature β, the consistency weight λ_{cons}, the entropy quantile q used for query-region selection, and the confidence threshold τ_{conf} for winner-pseudo-label pixels. Each ablation varies one hyperparameter while keeping the others fixed to their default values. Table 10 reports mIoU for each domain group and the mean across groups. Across all four hyperparameters, the mean mIoU remains within approximately ± 1 point of the default setting, and every variation stays substantially above the zero-shot baseline of 35.26. The default values achieve the best overall mean and are competitive across domain groups, indicating that our method is robust to hyperparameter choices.

H Compute Resources and Efficiency Analysis

In Table 11, we report the compute and memory profile of our adaptation framework on a single NVIDIA H200 GPU, using BDD100K (998 validation images, 19 classes) as a representative target.

H.1 Trainable Parameters

Our adaptation introduces only two trainable modules per backbone: a vision LoRA on the last four CLIP-ViT blocks (Q and V projections, rank 4) and a rank-4 residual text adapter shared across all

classes and templates. The trainable footprint is **71,681 parameters** on both CAT-Seg-L and SAN-L (65,536 vision LoRA + 6,145 text adapter), corresponding to roughly 0.017% of the full model. The footprint is essentially identical because both adapter modules attach to shared CLIP components, independent of the surrounding OVSS architecture.

H.2 Adaptation Cost

We measure the additional cost incurred by our adaptation as the gap between a base inference forward and a single training step, where each step performs one update on a support image including the $K = 14$ candidate forwards used for winner/loser selection. The adaptation overhead is $+1{,}466$ MiB and $+1{,}543$ ms on CAT-Seg-L, and $+5{,}474$ MiB and $+365$ ms on SAN-L. CAT-Seg-L incurs a smaller memory overhead but a longer per-step time because its IoU-based winner/loser selection requires one full sliding-window forward per template, whereas SAN-L scores all $K = 14$ candidates with native single-pass forwards.

I Additional Qualitative Results

We present additional qualitative results. Figures 5–8 show per-backbone adaptation results across the five MESS domain groups: for each backbone (SAN-B, CAT-Seg-B, SAN-L, CAT-Seg-L), we visualize one sample per domain group, comparing the input image, the zero-shot baseline prediction, our adapted prediction, and the ground-truth segmentation.

Figure 9 visualizes the preference query mining process described in Section 3.2. For each example, we show the input image, the cross-prompt uncertainty map (per-pixel ensemble entropy across the $K = 14$ templates), and the selected query region $\mathcal{R}$ overlaid on the prediction as a bounding box. The high-entropy regions concentrate on object boundaries and parts of the image where templates disagree most strongly, illustrating how prompt disagreement provides a localized, informative signal for preference query mining.

Table 12: The $K = 14$ prompt templates from the ViLD prompt pool used throughout this work.

#	Template
1	a photo of a {}.
2	This is a photo of a {}.
3	There is a {} in the scene.
4	There is the {} in the scene.
5	a photo of a {} in the scene.
6	a photo of a small {}.
7	a photo of a medium {}.
8	a photo of a large {}.
9	This is a photo of a small {}.
10	This is a photo of a medium {}.
11	This is a photo of a large {}.
12	There is a small {} in the scene.
13	There is a medium {} in the scene.
14	There is a large {} in the scene.

Table 13: Datasets in the MESS benchmark used in this work, grouped by domain.

Dataset	License	# classes	Classes
General Scenes			
BDD100K [76]	custom	19	[road; sidewalk; building; wall; fence; pole; traffic light; ...]
MHP v1 [75]	custom	19	[others; hat; hair; sunglasses; upper clothes; skirt; pants; ...]
FoodSeg103 [74]	Apache 2.0	104	[background; candy; egg tart; french fries; chocolate; biscuit; ...]
ATLANTIS [73]	Flickr (images)	56	[bicycle; boat; breakwater; bridge; building; bus; canal; ...]
Earth Monitoring			
iSAID [80]	Google Earth (images)	16	[others; boat; storage tank; baseball diamond; tennis court; ...]
WorldFloods [79]	CC NC 4.0	3	[land; water and flood; cloud]
FloodNet [78]	custom	10	[building-flooded; building-non-flooded; road-flooded; water; ...]
UAVid [77]	CC BY-NC-SA 4.0	8	[others; building; road; tree; grass; moving car; parked car; humans]
Medical Sciences			
Kvasir-Instrument [83]	custom	2	[others; tool]
CHASE_DB1 [82]	CC BY 4.0	2	[others; blood vessels]
PAXRay-4 [81]	custom	4×2	[others, lungs], [others, bones], [others, mediastinum], [others, diaphragm]
Engineering			
Corrosion CS [87]	CC0	4	[others; steel with fair, poor, severe corrosion]
DeepCrack [86]	custom	2	[concrete or asphalt; crack]
PST900 [85]	GPL-3.0	5	[background; fire extinguisher; backpack; drill; human]
ZeroWaste-f [84]	CC-BY-NC 4.0	5	[background or trash; rigid plastic; cardboard; metal; soft plastic]
Agriculture & Biology			
SUIM [90]	MIT	8	[human diver; reefs and invertebrates; fish and vertebrates; ...]
CUB-200 [89]	custom	201	[background; Laysan Albatross; Sooty Albatross; Crested Auklet; ...]
CWFID [88]	custom	3	[ground; crop seedling; weed]

Table 14: Per-dataset results on the MESS benchmark across four OVSS backbones (mIoU, %). Within each backbone, the better of the zero-shot baseline and our method (+ *Ours*) is highlighted in bold. (+ *Dense.*) indicates the dense-mask reference.

		General				Earth Monit.				Medical			Engineering				Agri. & Bio.			
Backbone	Method	BDD100K	MHP v1	FoodSeg103	ATLANTIS	iSAID	WorldFloods	FloodNet	UAVid	Kvasir-Inst.	CHASE_DB1	PAXRay-4	Corrosion CS	DeepCrack	PST900	ZeroWaste-f	SUIM	CUB-200	CWFID	*Mean*
SAN-B	base	35.36	9.39	**8.40**	33.25	4.18	30.15	**33.95**	39.31	62.38	18.47	19.69	4.53	**49.27**	40.69	**18.25**	36.54	5.79	5.63	*25.29*
	+ Ours	**37.78**	**10.69**	6.59	**33.31**	**4.19**	**32.59**	33.73	**40.25**	**73.30**	**46.43**	**39.18**	**21.01**	48.23	**46.49**	18.16	**40.64**	**6.77**	**43.68**	***32.39***
	+ Dense.	40.04	12.40	12.77	33.75	6.70	33.30	34.70	39.66	47.71	46.68	38.71	21.01	61.59	42.20	17.76	49.58	6.82	38.05	*32.41*
CAT-Seg-B	base	47.03	23.89	26.67	**40.43**	19.34	**38.52**	37.16	43.04	48.20	**23.99**	41.26	12.46	32.71	57.13	17.51	44.82	**10.41**	31.61	*33.12*
	+ Ours	**48.12**	**29.00**	**27.78**	39.35	**24.70**	36.60	**37.91**	**44.29**	**84.05**	23.16	**50.77**	**22.33**	**61.49**	**73.79**	**17.65**	**47.08**	9.43	**36.60**	***39.67***
	+ Dense.	48.93	34.69	29.67	41.48	24.14	39.57	40.18	49.00	77.39	46.65	56.21	26.67	56.11	75.31	25.94	63.58	14.98	46.26	*44.26*
SAN-L	base	42.65	9.16	14.50	38.61	10.31	**48.89**	37.42	41.42	62.25	4.43	29.33	8.17	19.65	53.73	15.03	**48.03**	8.27	1.43	*27.40*
	+ Ours	**44.68**	**11.17**	**15.23**	**38.63**	**19.38**	47.95	**38.56**	**42.45**	**67.50**	**5.97**	**42.40**	**20.88**	**47.84**	**59.38**	**16.41**	47.07	**8.84**	**37.54**	***33.99***
	+ Dense.	46.75	14.91	23.86	39.19	22.77	37.75	38.51	43.91	85.10	15.47	49.59	21.04	47.84	63.29	19.96	49.13	9.78	48.68	*37.64*
CAT-Seg-L	base	48.23	30.77	32.92	**45.51**	19.67	39.94	41.05	41.90	65.49	3.32	19.75	7.47	25.27	78.73	25.42	49.75	21.89	37.58	*35.26*
	+ Ours	**50.58**	**36.41**	**37.41**	45.50	**36.17**	**41.13**	**41.07**	**42.74**	**87.41**	**20.74**	**47.34**	**22.90**	**74.39**	**80.96**	**32.47**	**58.49**	**22.22**	**47.87**	***45.88***
	+ Dense.	51.59	39.03	38.57	46.97	31.51	40.63	43.63	51.49	85.13	21.45	59.85	22.74	62.14	78.68	33.17	59.01	26.93	47.49	*46.67*

Table 15: Per-dataset loss ablation. The best performance is highlighted in bold.

	General				Earth Monit.				Medical			Engineering				Agri. & Bio.			
Method	BDD100K	MHP v1	FoodSeg103	ATLANTIS	iSAID	WorldFloods	FloodNet	UAVid	Kvasir-Inst.	CHASE_DB1	PAXRay-4	Corrosion CS	DeepCrack	PST900	ZeroWaste-f	SUIM	CUB-200	CWFID	*Mean*
Baseline	48.23	30.77	32.92	45.51	19.67	39.94	41.05	41.90	65.49	3.32	19.75	7.47	25.27	78.73	25.42	49.75	21.89	37.58	*35.26*
w/o $\mathcal{L}_{\text{RLPO}}$	50.08	35.41	36.65	45.05	31.67	40.83	40.01	41.56	70.68	3.32	34.76	20.88	61.25	80.62	16.57	55.04	21.54	43.32	*40.51*
w/o $\mathcal{L}_{\text{cons}}$	50.36	34.48	36.79	**45.56**	25.45	**41.29**	**42.07**	**43.64**	84.86	20.52	41.40	**23.78**	59.33	80.22	28.18	54.90	21.15	**48.17**	*43.45*
Ours	**50.58**	**36.41**	**37.41**	45.50	**36.17**	41.13	41.07	42.74	**87.41**	**20.74**	**47.34**	22.90	**74.39**	**80.96**	**32.47**	**58.49**	**22.22**	47.87	***45.88***

Table 16: Per-dataset candidate generation comparison. With matched candidate size ($K = 14$), we contrast prompt disagreement (Ours) against MC Dropout and test-time augmentation (TTA). The best performance is highlighted in bold.

	General				Earth Monit.				Medical			Engineering				Agri. & Bio.			
Method	BDD100K	MHP v1	FoodSeg103	ATLANTIS	iSAID	WorldFloods	FloodNet	UAVid	Kvasir-Inst.	CHASE_DB1	PAXRay-4	Corrosion CS	DeepCrack	PST900	ZeroWaste-f	SUIM	CUB-200	CWFID	*Mean*
Baseline	48.23	30.77	32.92	45.51	19.67	39.94	41.05	41.90	65.49	3.32	19.75	7.47	25.27	78.73	25.42	49.75	21.89	37.58	*35.26*
MC Dropout	47.59	17.67	35.41	45.23	18.63	29.23	38.28	42.27	81.48	15.01	31.43	**23.22**	70.32	77.16	19.93	45.85	21.33	43.52	*39.09*
TTA	47.59	34.41	35.78	**46.01**	27.12	40.02	**41.11**	42.66	**88.12**	**21.06**	**50.28**	22.30	72.84	74.21	**33.36**	57.35	**22.56**	47.17	*44.66*
Ours	**50.58**	**36.41**	**37.41**	45.50	**36.17**	**41.13**	41.07	**42.74**	87.41	20.74	47.34	22.90	**74.39**	**80.96**	32.47	**58.49**	22.22	**47.87**	***45.88***

Table 17: Per-dataset effect of training set size. The best performance is highlighted in bold.

	General				Earth Monit.				Medical			Engineering				Agri. & Bio.			
# images	BDD100K	MHP v1	FoodSeg103	ATLANTIS	iSAID	WorldFloods	FloodNet	UAVid	Kvasir-Inst.	CHASE_DB1	PAXRay-4	Corrosion CS	DeepCrack	PST900	ZeroWaste-f	SUIM	CUB-200	CWFID	*Mean*
0	48.23	30.77	32.92	**45.51**	19.67	39.94	41.05	41.90	65.49	3.32	19.75	7.47	25.27	78.73	25.42	49.75	21.89	37.58	*35.26*
4	49.59	31.71	33.30	43.69	29.30	39.96	41.04	42.57	68.67	20.24	34.17	15.21	26.02	79.57	24.33	52.39	18.64	38.24	*38.26*
8	50.24	32.36	34.10	45.40	26.51	39.78	41.14	42.10	69.34	20.74	38.19	23.04	29.24	79.86	24.63	52.54	21.08	40.75	*39.50*
16	50.73	31.89	35.67	45.13	34.12	40.47	41.22	42.20	74.79	20.74	43.38	22.89	53.77	80.58	23.53	54.73	19.25	46.51	*42.31*
32	50.78	34.43	37.20	45.43	30.62	40.27	**41.25**	42.63	80.26	20.74	45.68	22.24	72.25	80.31	27.46	53.58	22.64	**47.85**	*44.20*
64	50.58	36.41	**37.41**	45.50	**36.17**	41.13	41.07	**42.74**	**87.41**	20.74	47.34	22.90	74.39	**80.96**	32.47	**58.49**	22.22	**47.85**	***45.88***
128	**51.24**	**36.71**	36.83	45.47	33.85	**41.34**	40.76	41.44	85.21	20.74	**48.30**	**25.47**	**75.99**	80.49	**34.59**	54.53	**23.41**	**47.85**	*45.79*

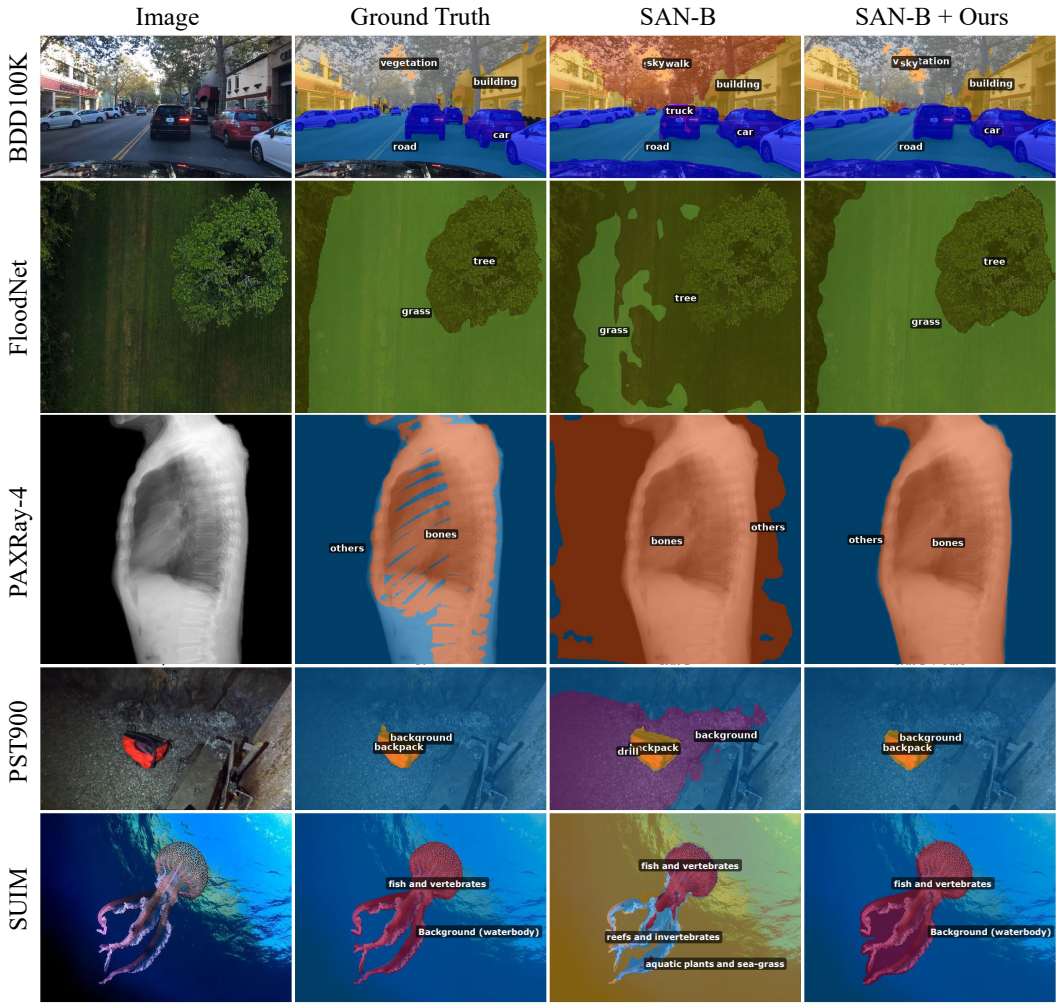


Figure 5: Qualitative results on SAN-B. Each row shows one sample from one of the five MESS domain groups (top to bottom: General, Earth Monitoring, Medical Sciences, Engineering, Agriculture & Biology). Columns show the input image, ground-truth segmentation, zero-shot baseline prediction, and our adapted prediction.

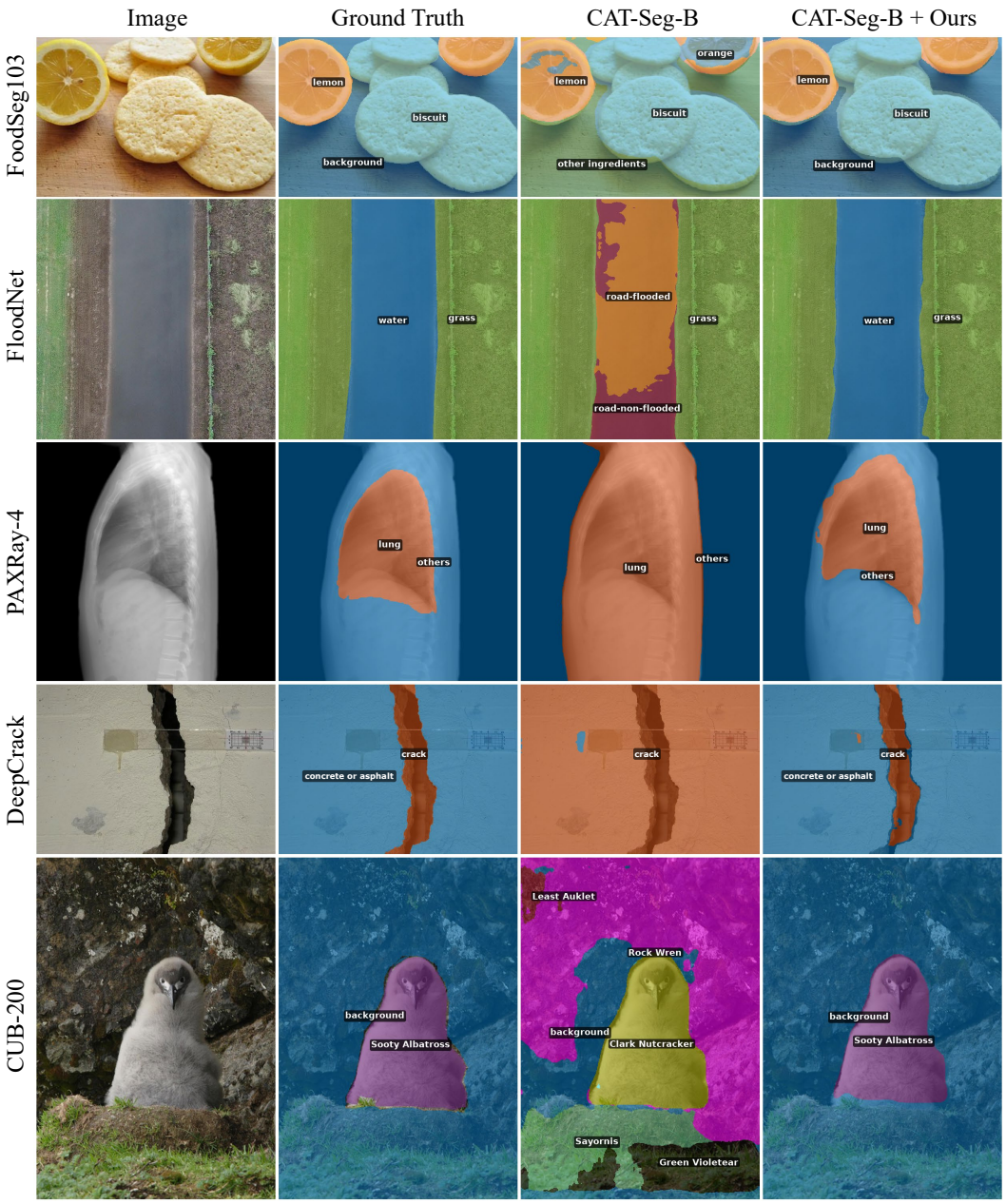


Figure 6: Qualitative results on CAT-Seg-B. Each row shows one sample from one of the five MESS domain groups (top to bottom: General, Earth Monitoring, Medical Sciences, Engineering, Agriculture & Biology). Columns show the input image, ground-truth segmentation, zero-shot baseline prediction, and our adapted prediction.

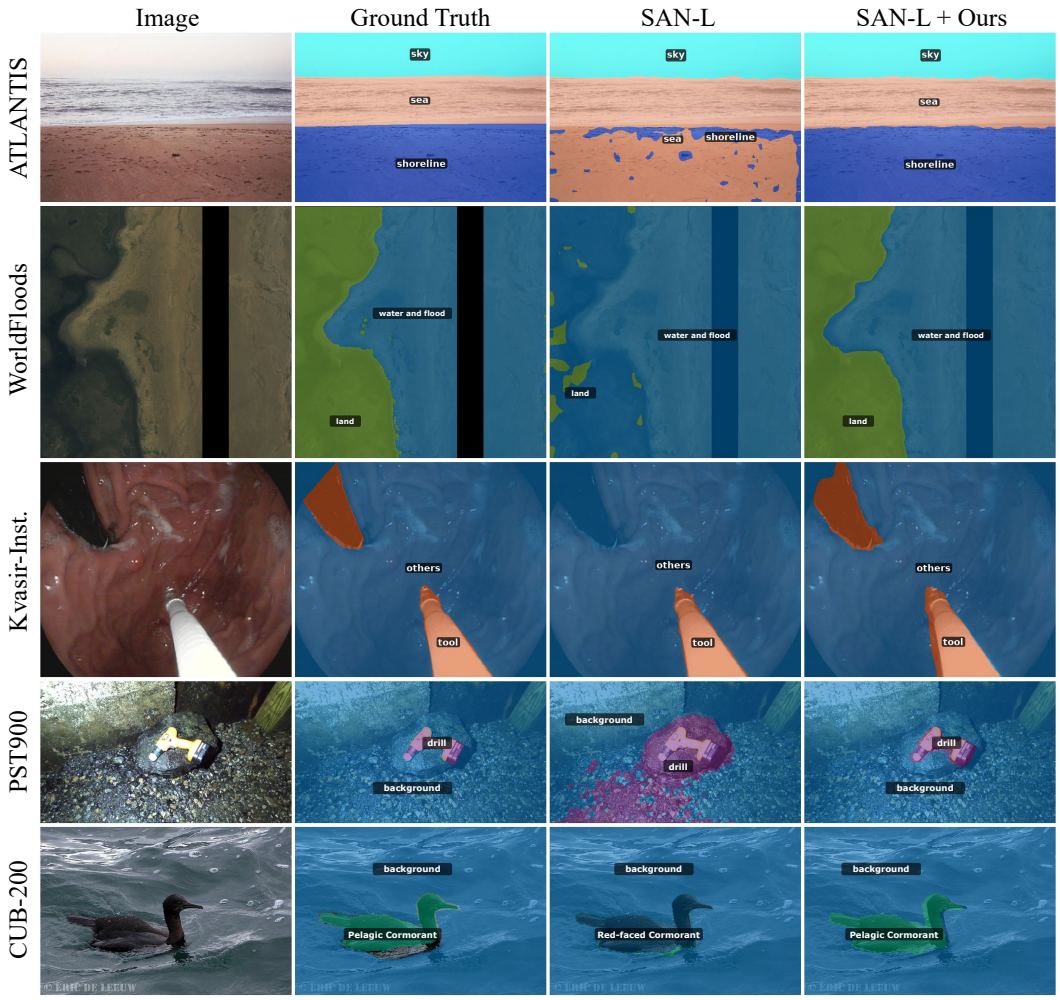


Figure 7: Qualitative results on SAN-L. Each row shows one sample from one of the five MESS domain groups (top to bottom: General, Earth Monitoring, Medical Sciences, Engineering, Agriculture & Biology). Columns show the input image, ground-truth segmentation, zero-shot baseline prediction, and our adapted prediction.

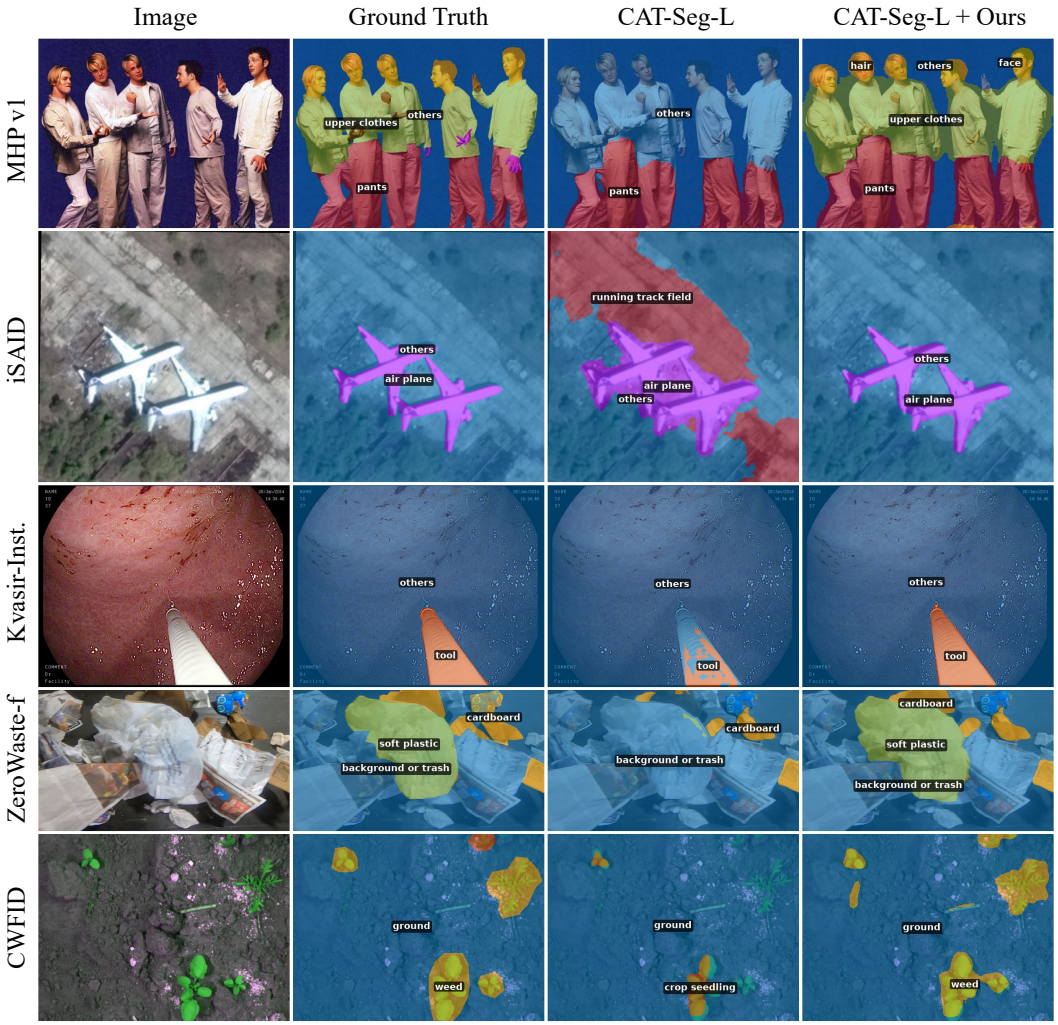


Figure 8: Qualitative results on CAT-Seg-L. Each row shows one sample from one of the five MESS domain groups (top to bottom: General, Earth Monitoring, Medical Sciences, Engineering, Agriculture & Biology). Columns show the input image, ground-truth segmentation, zero-shot baseline prediction, and our adapted prediction.

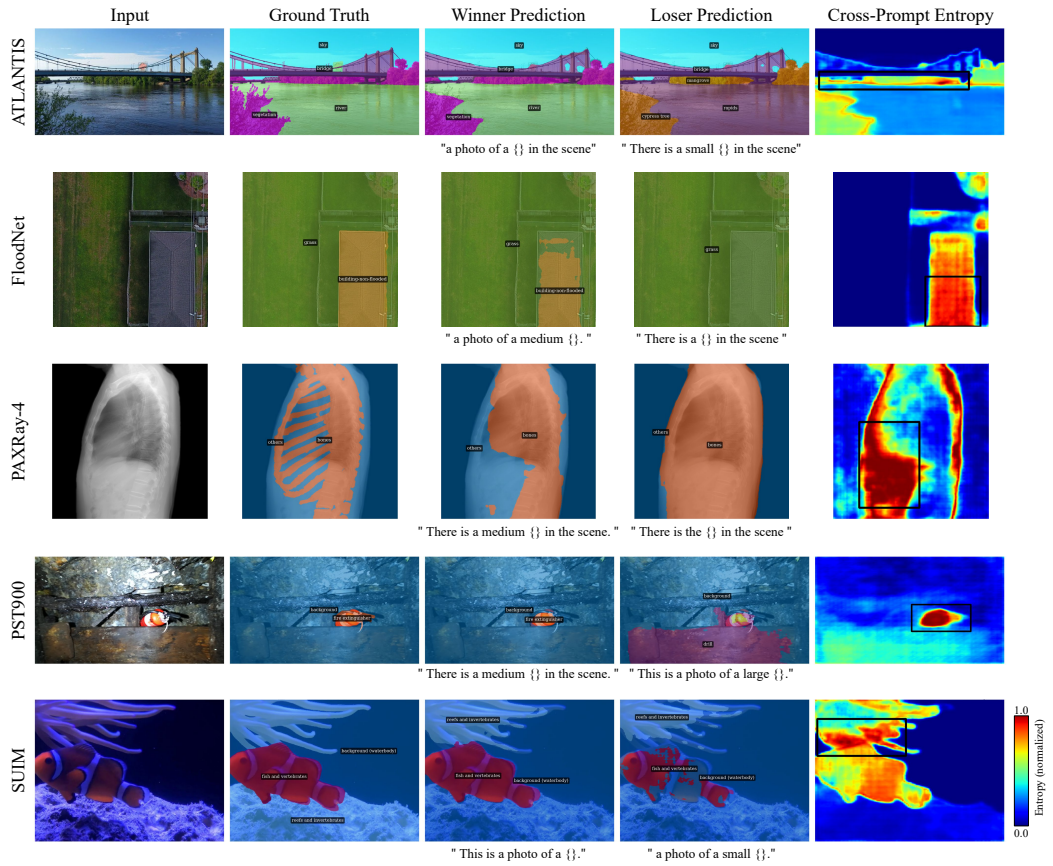


Figure 9: Visualization of preference query mining. For each example (one per row), we show the input image, ground truth, winner prediction, loser prediction, and cross-template entropy ($K = 14$ ViLD templates). The selected query region $\mathcal{R}$ is overlaid on the entropy map as a bounding box.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions—prompt disagreement as preference supervision, RLPO for OVSS adaptation, and consistent improvements on MESS—each supported by Section 4 (Tables 1–3) and Section 4.4 (noisy-preference robustness).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 discusses the main limitation: when all templates are uniformly inaccurate on a region, preference selection cannot recover a useful supervisory signal.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The RLPO objective is derived by specializing the standard DPO formulation to template-conditional segmentation. The full derivation is provided in Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 4.1 and Appendix describe the datasets, backbones, adapter configuration, prompt template pool, oracle definition, and training protocol in sufficient detail to reproduce the main results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide open access to the code and dataset at `https://github.com/blue-531/pref-ovss`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 and Appendix specify all training details, including optimizer, learning rates, hyperparameters.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 1 reports the mean $\pm$ standard deviation over three independent runs, each with independently sampled adaptation images. All experiments in this manuscript are conducted over three independent runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix reports trainable parameter counts, wall-clock time, and peak memory all measured on a single NVIDIA H200 GPU.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:The research uses publicly available benchmarks and pretrained models.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We believe that this research does not present any notable societal impact.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release assets with high misuse risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and their licenses are listed in Appendix. Base OVSS models, CLIP, and the ViLD prompt pool are properly credited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new datasets or pretrained models are released at submission time.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The user study in Appendix D.1 involved 20 voluntary participants. The instructions are described in Appendix D.1.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: We did not obtain IRB approval for user study in Appendix D.1. Participants only made binary choices between segmentation predictions on public benchmark images, and no personal information was collected.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: This paper does not involve the use of large language models (LLMs) as a core methodological component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.