PROMPTCORESET: BEHAVIOR-AWARE TEXT-PROMPT SELECTION FOR REMOTE-SENSING SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-conditioned segmentation is sensitive to prompt wording, but exhaustive prompt selection requires dense annotations and repeated inference over the target dataset. We study *calibration-set design*: which target images should be labeled to select a prompt under a limited annotation budget? We introduce **PromptCoreset**, a training-free method that represents unlabeled images by foreground extent and pairwise agreement of candidate masks. Deterministic k-medoids selects representative images for annotation, and cluster-weighted true-positive and union counts estimate global-IoU utility. Unlike purely visual sampling, the chosen images represent variation in foreground extent and mask agreement among prompts. The method selects one deployment prompt without changing the segmentation model or prompt bank. Across six remote-sensing datasets and three frozen segmentation systems, PromptCoreset improves mean segmentation IoU and F1 and lowers mean Prompt Regret relative to random calibration on each system using only a small labeled calibration set. Prompt-conditioned calls decrease by approximately 90% relative to exhaustive search. These results show that behavior-aware calibration can make text-prompt selection more annotation- and compute-efficient without retraining the segmentation model.

1 INTRODUCTION

Text-conditioned segmentation allows categories to be specified at inference time. This flexibility is useful in remote sensing, where semantic classes, sensors, spatial resolution, and scene composition vary across datasets. Alongside advances in spatially prompted segmentation such as SAM (Kirillov et al., 2023) and SAM 2 (Ravi et al., 2024), systems including SAM 3 (Carion et al., 2025), CAT-Seg (Cho et al., 2024), and SegEarth-OV (Li et al., 2025) support language-conditioned recognition and segmentation.

Yet semantically similar descriptions of the same target can yield markedly different masks. Figure 1 illustrates this sensitivity across three segmentation systems. Candidate masks reveal prompt-specific success and failure modes across the target population before any labels are requested; calibration should use these patterns to select one reusable prompt.

An exhaustive Oracle search evaluates every prompt on every target image against dense ground truth. A practical alternative estimates prompt utility from a small labeled calibration set. Its reliability depends on both the utility estimator and the images selected for annotation. Visual diversity need not track prompt behavior: similar-looking images can yield different candidate masks, whereas visually distinct images can be redundant when all prompts respond similarly. This motivates selecting calibration images in a representation induced by candidate predictions.

We propose **PromptCoreset**, a behavior-aware calibration method. Candidate prompts first produce masks on a small unlabeled screening pool. Foreground extent and pairwise mask agreement form an image-level behavior signature, and deterministic k-medoids selects representative images for annotation. Cluster-weighted true-positive and union counts then estimate each prompt's global IoU. We fix the deployment prompt before full-population ground-truth evaluation. PromptCoreset trains no model but requires labels for the selected calibration images.

We evaluate six remote-sensing datasets and three frozen segmentation systems at matched annotation budgets, using every tile in the pre-specified evaluation manifests. Our contributions are:

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

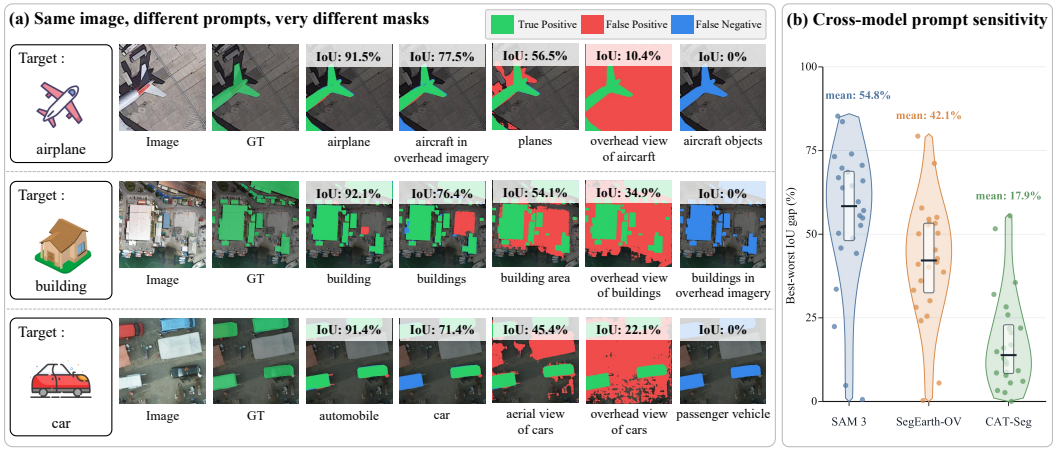

Figure 1: **Prompt sensitivity on the current benchmark. (a)** Semantically related prompts produce different masks for the same image and class. **(b)** Best-minus-worst full-population IoU differences across 24 dataset–class units per model, shown on a 0–100% scale (numerically equal to percentage-point gaps). Dots mark units, violins show distributions, boxes the interquartile range, black bars the median, and text the mean. These post-hoc gaps never enter selection.

- We cast fixed-bank prompt selection as calibration-set design, targeting low Prompt Regret under annotation and inference constraints rather than exact recovery of the Oracle prompt.
- We introduce a training-free selector based on candidate-mask behavior, representative annotation, and a cluster-weighted global-IoU estimator.
- We provide a frozen-method evaluation over 24 dataset–class units and three segmentation systems, with matched-budget comparisons, component ablations, and separate accounting for inference calls and GPU time.

2 Related Work

Text-conditioned segmentation. CLIP enabled transferable image–text representations from natural-language supervision (Radford et al., 2021). Subsequent dense prediction methods introduced pixel-level alignment, grouping, region–text matching, and mask-level recognition, including LSeg (Li et al., 2022), GroupViT (Xu et al., 2022), OpenSeg (Ghiasi et al., 2022), DenseCLIP (Rao et al., 2022), OVSeg (Liang et al., 2023), SAN (Xu et al., 2023), and CAT-Seg (Cho et al., 2024). Remote-sensing extensions address scale variation, dense repetition, orientation, and texture-dominated appearance through SAM-generated supervision (Wang et al., 2023), learned prompting (Chen et al., 2024), multi-scale modeling (Zhang et al., 2024), and domain-specific vision–language representations such as SegEarth-OV (Li et al., 2025). We keep the segmentation systems frozen and study the text condition used for deployment.

Prompt optimization and selection. CoOp (Zhou et al., 2022b), CoCoOp (Zhou et al., 2022a), and MaPLe (Khattak et al., 2023) learn continuous prompts, while test-time prompt tuning adapts prompts on unlabeled target samples (Shu et al., 2022). CuPL generates category descriptions with a language model for zero-shot classification (Pratt et al., 2023); segmentation methods also explore prompt ensembles, synonym fusion, and filtering (Lin et al., 2023; Yang et al., 2026). These approaches optimize, generate, or aggregate text conditions. We instead hold a human-readable prompt bank fixed and select one prompt for deployment. This separates bank construction from the downstream decision: which target images should be annotated to identify a strong prompt from the existing candidates?

Representative subset selection. Batch active learning selects labels to improve a trainable predictor. Core-set methods cover an embedding space (Sener & Savarese, 2018), BADGE combines diversity and uncertainty in gradient embeddings (Ash et al., 2020), and CEREALS accounts for

region-annotation cost in semantic segmentation (Mackowiak et al., 2018). Predictor disagreement has long been used to identify informative queries (Seung et al., 1992). Our use of mask agreement differs from selecting the most disputed images: agreement and foreground extent jointly describe how the *candidate prompts* behave, and medoids cover that behavior space. The acquired labels compare frozen predictors rather than update the segmentation backbone. Cluster weights represent the screening images assigned to each annotation, while aggregation of true-positive and union counts preserves the ratio-of-sums structure of global IoU.

Label-efficient selection among frozen predictors. Active model selection is closer to our setting than conventional active learning: it queries a few labels to choose among trained predictors. Model Picker studies online selection from pretrained classifiers (Reza Karimi et al., 2021); Model Selector and CODA use unlabeled candidate predictions to guide label acquisition for pool-based classifier selection (Okanovic et al., 2025; Kay et al., 2025). We adopt this decision-oriented framing. Here, however, each text prompt induces a candidate *dense segmenter* within one frozen model, and utility is global IoU rather than per-image classification accuracy. Because global IoU is a ratio of population-level counts, averaging image scores need not preserve the prompt ranking. We specialize mask-behavior representation and weighted count estimation to this setting and report inference alongside annotation cost.

3 Behavior-Aware Prompt Calibration

3.1 Problem Setting

A benchmark unit is a dataset–class pair $u = (\mathcal{D}, c)$. Let $\mathcal{D}_u = \{(x_n, y_n^c)\}_{n=1}^{N_u}$ denote a fixed target population, where labels are unavailable to the selector except for the requested calibration images. Let

$$\mathcal{P}_c = \{p_1, \ldots, p_M\} \tag{1}$$

be a fixed prompt bank. For a frozen segmentation system F, prompt p produces mask $\hat{y}_n(p) = F(x_n, p)$. Its full-dataset utility is global IoU,

$$J_u(p) = \frac{\sum_n \mathrm{TP}_n(p)}{\sum_n [\mathrm{TP}_n(p) + \mathrm{FP}_n(p) + \mathrm{FN}_n(p)]}. \tag{2}$$

The Oracle prompt is $p_u^* = \arg\max_{p \in \mathcal{P}_c} J_u(p)$. A selector returning $\hat{p}_u$ incurs

$$\mathcal{R}_u = J_u(p_u^*) - J_u(\hat{p}_u), \tag{3}$$

which we call *Prompt Regret*. The objective is to minimize regret using at most B labeled target images and model outputs from a small unlabeled screening pool. This is transductive calibration: calibration images belong to the evaluated target population. We do not interpret the resulting utility as performance on an independent held-out test set.

3.2 Framework Overview

PromptCoreset has three stages (Figure 2). First, we evaluate all M prompts on a fixed unlabeled screening pool $\mathcal{S}_u = \{x_i\}_{i=1}^K$. Second, the masks define prompt-behavior signatures from which we select B representative images without ground truth. Third, we annotate those images and estimate prompt utility. Model parameters, prompt text, screening images, and coreset assignments remain fixed throughout calibration.

3.3 Prompt-Behavior Representation

For image x_i with pixel domain Ω_i and prompt p_m, let $\hat{y}_{im} = F(x_i, p_m)$ denote the predicted foreground set. The foreground-area feature is

$$a_{im} = \frac{|\hat{y}_{im}|}{|\Omega_i|}, \tag{4}$$

and pairwise prompt agreement is

$$g_{im\ell} = \mathrm{IoU}(\hat{y}_{im}, \hat{y}_{i\ell}), \qquad m < \ell. \tag{5}$$

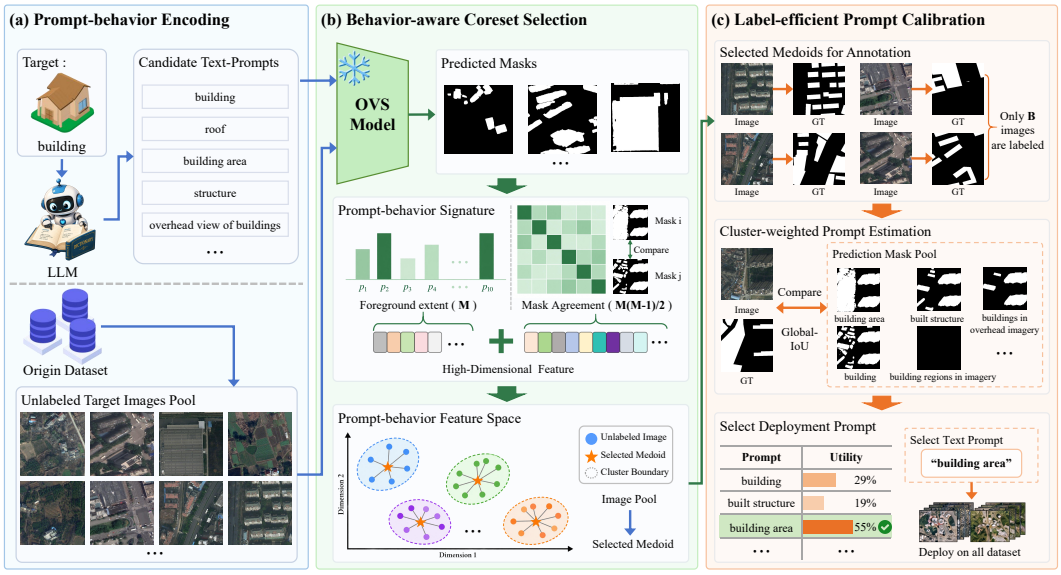

Figure 2: **PromptCoreset overview.** **(a)** Evaluate a fixed prompt bank on unlabeled screening images. **(b)** Encode foreground extent and pairwise mask agreement; select representative behaviors by deterministic k-medoids. **(c)** Annotate only the medoids, estimate cluster-weighted global IoU, and select one deployment prompt. Illustrative utilities use the unit scale of Equation 7.

When both masks are empty, agreement is one; when only one is empty, it is zero. For $M = 10$, the image signature contains ten area values and 45 pairwise agreements:

$$z_i = \left[\frac{a_i}{\sqrt{10}}, \frac{g_i}{\sqrt{45}} \right] \in \mathbb{R}^{55}. \tag{6}$$

The fixed block normalization prevents either feature family from dominating solely because of dimensionality.

3.4 Coreset Selection and Utility Estimation

We apply deterministic PAM-style k-medoids (Kaufman & Rousseeuw, 1990) with BUILD and SWAP steps to $\{z_i\}_{i=1}^{K}$ using Euclidean distance. Let S_B be the selected medoid indices. Each screening image is assigned to its nearest medoid; ties use the frozen image order. If C_j is the cluster represented by medoid j, its weight is $w_j = |C_j|$, with $\sum_{j \in S_B} w_j = K$. Ground truth is requested only for these B images.

For prompt p, define $U_j(p) = \mathrm{TP}_j(p) + \mathrm{FP}_j(p) + \mathrm{FN}_j(p)$. PromptCoreset estimates global IoU with the cluster-weighted ratio

$$\hat{J}_{\mathrm{PC}}(p) = \frac{\sum_{j \in S_B} w_j \, \mathrm{TP}_j(p)}{\sum_{j \in S_B} w_j \, U_j(p)}, \qquad \hat{p}_{\mathrm{PC}} = \arg\max_{p \in \mathcal{P}_c} \hat{J}_{\mathrm{PC}}(p). \tag{7}$$

Each labeled medoid stands in for the screening images assigned to its cluster. We therefore sum weighted true-positive and union counts before taking their ratio; averaging image-level IoUs would not preserve the global-IoU objective. A zero union receives utility zero, as in evaluation, and score ties follow the frozen prompt order.

Scope of the coreset approximation. The estimator directly approximates utility on the screening pool. If normalized true-positive and union contributions are prompt-wise Lipschitz in the behavior metric and the mean union is bounded away from zero, its error is bounded by a constant times the average medoid distortion. A uniform utility error of at most γ then implies selection regret of at most 2γ *on that pool*: the estimated maximizer can lose at most γ on either side of the comparison with the true maximizer. These are conditional assumptions, not properties established by the representation.

Extending this argument to the full target population additionally requires control of screening-pool sampling error; we assess that transfer empirically.

3.5 Annotation and Inference Cost

PromptCoreset requires B class masks and no model training. For one class and one model on a dataset with N_d evaluation images, its prompt-conditioned inference cost is

$$C_{\mathrm{PC}} = MK + N_d, \tag{8}$$

whereas exhaustive prompt search costs $C_{\mathrm{Oracle}} = MN_d$. The nominal saving is

$$1 - \frac{MK + N_d}{MN_d}. \tag{9}$$

The count includes screening and single-prompt deployment, without crediting reuse of screening predictions during deployment. It describes the intended selection workflow, not the additional all-prompt inference used to audit it. We report annotation count, prompt-conditioned calls, and GPU-time estimates separately; the latter use measured encoder and prompt-head timings with model-specific feature reuse.

4 Benchmark

4.1 Datasets and Evaluation Populations

The benchmark comprises iSAID, LoveDA, ISPRS Potsdam, UAVid, UDD5, and OpenEarthMap: four fixed classes per dataset and 24 dataset–class units. We merge instance annotations into class-level binary masks where necessary and partition large scenes into deterministic, non-overlapping 512×512 RGB tiles. The frozen manifests contain 6,363, 10,088, 7,406, 10,800, 1,824, and 1,515 tiles, respectively (37,996 total). Evaluation uses every valid entry, without a fixed-size cap.

We use the labeled validation splits of iSAID and UDD5, the repository-defined validation scenes of Potsdam, and the locally available labeled pools of LoveDA and UAVid. OpenEarthMap uses the locally paired validation subset, excluding entries whose RGB images are unavailable. Thus, *full-split* below means the complete frozen evaluation manifest, not necessarily the complete official validation split. All methods use identical populations; none trains the segmentation backbone on these data.

4.2 Models, Prompts, and Protocol

We evaluate frozen SAM 3, SegEarth-OV (Li et al., 2025), and CAT-Seg (Cho et al., 2024). Each class has $M = 10$ fixed, human-readable prompts spanning its canonical name, morphological variants, synonyms, and visual descriptions. A class-independent pool of $K = 32$ images per dataset is shared across all four classes and three models. We fix its image IDs by deterministic hash-based sampling of the seed, dataset identifier, and image key; full-split evaluation reuses these IDs rather than resampling from the expanded population.

We use annotation budgets $B \in \{4, 8, 16\}$, with $B = 8$ primary. Selection of calibration images and cluster weights uses only predictions and unlabeled metadata and is frozen before calibration labels are accessed. After utility estimation, the selected prompt is frozen before full-population ground-truth evaluation. Exhaustive predictions outside the screening pool do not enter selection. This is a frozen-method validation on the existing benchmark, not an independent external evaluation of a method developed without access to these datasets.

4.3 Baselines and Metrics

Table 1 summarizes the selectors. All calibration methods select from the same screening pool at matched label budgets. Image k-Center uses frozen image features, while Top Disagreement prioritizes disagreement among candidate masks. The unweighted PromptCoreset baseline uses the same medoids but sets all weights to one. Random Calibration averages pre-specified sampling episodes within each unit. All utility-based methods use global IoU and the same prompt-order tie breaking.

Table 1: Compared prompt-selection methods.

Method	Selection signal	Labels	Cluster weights
Class Name	Canonical prompt	0	–
Random Calibration	Random images	B	No
Image k-Center	Image features	B	No
Top Disagreement	Prompt disagreement	B	No
PromptCoreset (Unweighted)	Prompt behavior	B	No
PromptCoreset	Prompt behavior	B	Yes
Oracle	Full-split ground truth	Full	–

Prompt Regret is the primary endpoint. We also report IoU, F1, Exact Top-1 (selection of the tie-broken Oracle), Top-3 Recall (Oracle inclusion), and Near-Oracle@1 pp (regret ≤ 0.01), together with Pairwise Accuracy and Spearman correlation. IoU, F1, and recovery rates are percentages; regret and utility MAE are percentage points (pp). Equations retain the unit scale. We macro-average over units. Confidence intervals use 10,000 dataset-cluster bootstrap resamples, retaining all four classes per dataset and averaging random episodes within each unit before resampling.

5 Results

We compare selectors at the primary budget before examining inference cost, budget scaling, and component ablations. We distinguish lower mean regret from exact Oracle recovery and from consistent gains across datasets.

5.1 Main Comparison

Table 2 reports the primary $B = 8$ comparison over the same 24 units for each model.

Table 2: Full-split prompt selection at $B = 8$. Values are macro averages over 24 units per model; Regret is in percentage points (pp), and all other metrics are in percent. Bold denotes the best non-Oracle value in each column. Regret is the primary endpoint.

Model	Method	Regret ↓ (pp)	IoU ↑ (%)	F1 ↑ (%)	Top-1 ↑ (%)	Near@1 pp ↑ (%)
SAM 3	Class Name	5.36	49.49	62.69	**37.5**	41.7
	Random Calibration	5.99	48.86	62.15	35.4	**50.1**
	Image k-Center	7.89	46.96	60.48	**37.5**	45.8
	Top Disagreement	4.22	50.63	64.19	25.0	41.7
	PromptCoreset (Unweighted)	4.43	50.42	63.87	20.8	45.8
	PromptCoreset	**3.90**	**50.96**	**64.53**	29.2	50.0
	Oracle	0.00	54.85	67.81	100.0	100.0
SegEarth-OV	Class Name	5.21	38.60	52.95	16.7	33.3
	Random Calibration	2.86	40.96	55.50	38.9	50.8
	Image k-Center	2.45	41.36	56.02	33.3	45.8
	Top Disagreement	2.60	41.22	55.79	50.0	54.2
	PromptCoreset (Unweighted)	2.85	40.96	55.58	**54.2**	**58.3**
	PromptCoreset	**2.30**	**41.51**	**56.14**	50.0	**58.3**
	Oracle	0.00	43.82	58.50	100.0	100.0
CAT-Seg	Class Name	2.77	46.94	60.59	12.5	50.0
	Random Calibration	2.21	47.50	61.18	34.2	56.6
	Image k-Center	2.17	47.54	61.24	25.0	50.0
	Top Disagreement	2.07	47.64	61.28	25.0	54.2
	PromptCoreset (Unweighted)	**1.09**	**48.62**	**62.27**	45.8	**75.0**
	PromptCoreset	1.24	48.47	62.12	**50.0**	70.8
	Oracle	0.00	49.71	63.38	100.0	100.0

At $B = 8$, PromptCoreset lowers mean regret relative to Random Calibration by 2.09 pp on SAM 3, 0.55 pp on SegEarth-OV, and 0.97 pp on CAT-Seg. It improves 17, 19, and 18 of 24 units, respectively:

Table 3: Prompt-ranking fidelity at $B = 8$. Each entry is Pairwise Accuracy (%) / Spearman correlation, macro-averaged over the same 24 units per model as Table 2. All calibration methods use the same label budget; bold marks the best value for each model and metric.

Method	SAM 3	SegEarth-OV	CAT-Seg
Random Calibration	80.9 / 0.671	79.0 / 0.667	71.1 / 0.503
Image k-Center	77.9 / 0.617	78.0 / 0.636	71.3 / 0.515
Top Disagreement	84.7 / 0.764	78.9 / 0.666	64.9 / 0.370
PromptCoreset (Unweighted)	83.6 / 0.754	80.7 / 0.699	77.7 / 0.651
PromptCoreset	**85.3 / 0.788**	**81.0 / 0.705**	**78.9 / 0.672**

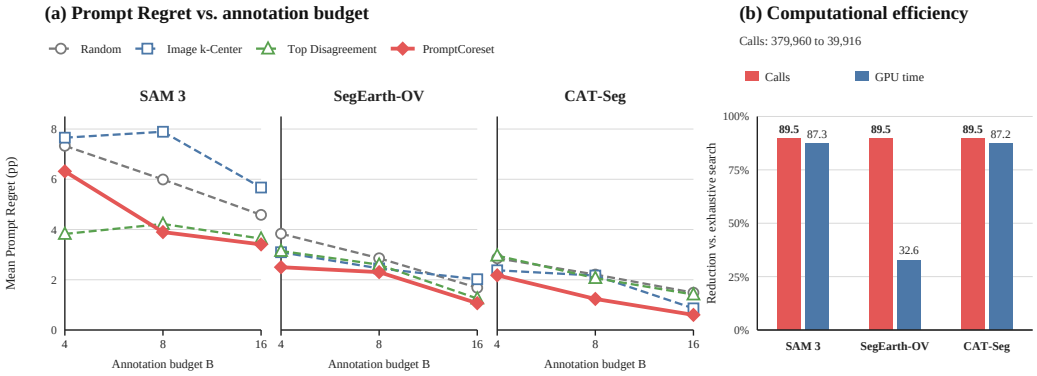

Figure 3: **Cost–quality trade-off. (a)** Mean Prompt Regret (pp) at matched annotation budgets. PromptCoreset improves with budget on all three systems, although Top Disagreement is stronger on SAM 3 at $B = 4$. **(b)** Reduction relative to exhaustive search: 89.5% in prompt-conditioned calls, and estimated GPU-time reductions of 87.3%, 32.6%, and 87.2% for SAM 3, SegEarth-OV, and CAT-Seg. GPU estimates use measured component timings; call totals follow the per-class accounting in Table 4.

54 improvements, three ties, and 15 regressions across 72 model–unit pairs. These pairs are not independent dataset replications. Dataset-cluster bootstrap 95% intervals for PromptCoreset regret are $[2.50, 5.68]$, $[0.98, 3.96]$, and $[0.57, 1.95]$ pp. For paired gain (Random minus PromptCoreset), the 95% interval excludes zero on SAM 3 ($[0.80, 3.54]$ pp) but includes zero on SegEarth-OV ($[-0.96, 1.79]$ pp) and CAT-Seg ($[-0.001, 1.86]$ pp). Paired-gain intervals use 20,000 resamples of the six dataset clusters; evidence for the smaller gains remains inconclusive.

Table 3 assesses the full prompt ordering. PromptCoreset has the highest Pairwise Accuracy and Spearman correlation among matched-budget methods on all three systems. Relative to Random Calibration, both metrics improve in six of six dataset clusters for SAM 3 and CAT-Seg and five of six for SegEarth-OV; these descriptive counts are not significance tests.

Lower regret need not imply more exact Oracle selections. On SAM 3, PromptCoreset has the lowest non-Oracle regret in Table 2 despite an Exact Top-1 rate of 29.2% and no improvement in Near-Oracle frequency over Random Calibration. Near-tied alternatives incur little regret, whereas a few poor selections can dominate the mean. On CAT-Seg, the unweighted variant has lower regret than the weighted method; the component study examines this exception.

5.2 Inference Efficiency

Across the six datasets, PromptCoreset requires 39,916 prompt-conditioned calls per class and model versus 379,960 for exhaustive search, a reduction of 89.5% (Table 4). Evaluating all four classes multiplies both counts by four without changing the percentage. These deployment costs exclude the additional exhaustive inference used to construct the benchmark reference.

Figure 3(b) estimates GPU time from measured component timings with model-specific encoder reuse. Per class, PromptCoreset uses 1.40 versus 11.00 GPU-hours for SAM 3 (87.3% saving), 0.88 versus 6.87 for CAT-Seg (87.2%), and 4.40 versus 6.54 for SegEarth-OV (32.6%). These are component-based estimates, not end-to-end wall-clock measurements. SegEarth-OV saves less GPU time because encoder cost dominates and deployment images still require encoding.

Table 4: Prompt-conditioned calls for one class and one model per dataset at $K = 32$, $M = 10$. PromptCoreset costs $MK + N_d$; exhaustive search costs MN_d. Calls do not depend on B once screening predictions are cached.

Dataset	Exhaustive	PromptCoreset	Reduction
iSAID	63,630	6,683	89.50%
LoveDA	100,880	10,408	89.68%
Potsdam	74,060	7,726	89.57%
UAVid	108,000	11,120	89.70%
UDD5	18,240	2,144	88.25%
OpenEarthMap	15,150	1,835	87.89%
Total	379,960	39,916	89.49%

5.3 Ablation Studies

Label budget. Figure 3(a) reports the regret–budget trade-off under matched calibration budgets; Appendix Table 7 retains the exact values.

Across SAM 3, SegEarth-OV, and CAT-Seg, PromptCoreset regret falls from 6.32/2.50/2.18 pp at $B = 4$ to 3.41/1.07/0.60 pp at $B = 16$. It has the lowest regret among the four selectors at $B = 8$ and $B = 16$. At $B = 4$, Top Disagreement instead leads on SAM 3 (3.82 versus 6.32 pp). All budgets reuse the same $K = 32$ screening predictions, so increasing B changes annotation rather than prompt-conditioned inference cost.

Sampling strategy and utility weighting. Table 5 crosses sampling and weighting post hoc. Paired estimators use identical images and the frozen behavior metric for cluster assignments; Random reuses 256 frozen subsets. Selections and weights were fixed before evaluating the additional cells, leaving the primary method unchanged.

Table 5: Sampling $\times$ weighting ablation at $B = 8$, $K = 32$, and $M = 10$. Mean full-split Prompt Regret (pp, lower is better) over 24 units per model. All six post-hoc factorial cells are reported; bold marks the lowest regret per model.

Sampling	Cluster weighting	SAM 3	SegEarth-OV	CAT-Seg
Random	No	5.99	2.86	2.21
	Yes	6.12	2.99	2.24
Image k-Center	No	7.89	2.45	2.17
	Yes	8.16	2.70	2.17
PromptCoreset	No	4.43	2.85	**1.09**
	Yes	**3.90**	**2.30**	1.24

Weighting alone does not improve mean regret for Random or Image k-Center. With PromptCoreset sampling, weighting reduces SAM 3 regret by 0.53 pp (paired 95% CI for weighted minus unweighted: $[-1.11, -0.02]$ pp) and SegEarth-OV regret by 0.55 pp, but raises CAT-Seg regret by 0.14 pp; the latter two intervals include zero. Weighting therefore interacts with representative selection rather than providing a universal gain. Appendix A reports the paired effects.

5.4 Screening-Pool Error and Selection Error

The $K = 32$ screening pool introduces two sources of utility error. For prompt p, let $J_{\mathcal{S}}(p)$ be global IoU on all 32 screening images, $\widehat{J}(p)$ its estimate from the B labeled medoids, and $J_u(p)$ utility on

the full frozen population. Then

$$\left|\widehat{J}(p) - J_u(p)\right| \le \left|\widehat{J}(p) - J_{\mathcal{S}}(p)\right| + \left|J_{\mathcal{S}}(p) - J_u(p)\right|. \tag{10}$$

The first term measures how well labeled medoids approximate their screening pool; the second measures how well that pool represents the evaluation population. The conditional coreset argument in Section 3 concerns the first term, not the second. Appendix Table 6 evaluates both errors retrospectively with ground truth; neither is available during prompt selection.

For both estimators on every model, full-population MAE exceeds screening-pool MAE, consistent with transfer error beyond the 32-image pool. The difference between these MAEs is *not* an estimate of the second term in Equation 10: absolute errors are not additive. Nor does lower bank-wide MAE ensure a better prompt choice. On CAT-Seg, weighting lowers full-population MAE from 13.57 to 11.66 pp, yet regret rises from 1.09 to 1.24 pp and Near-Oracle@1 pp falls from 75.00% to 70.83%. SegEarth-OV exhibits the reverse mismatch: regret improves with little change in full-population MAE. Selection depends on errors among leading prompts, not only mean error across the bank.

The iSAID storage-tank unit on SAM 3 illustrates the coverage limit. All ten prompt utilities on its frozen screening pool are zero. The medoid estimator matches them exactly, yielding zero screening MAE, and the tie rule selects the canonical prompt. On the full population, that prompt reaches 50.41% IoU versus 66.89% for the best candidate, or 16.48 pp regret. Reweighting medoids in the same pool cannot distinguish these prompts. This case was selected for analysis after evaluation and did not alter screening or selection.

Limitations. The study contains only six dataset clusters, and its frozen evaluation populations are not uniformly official test splits. It evaluates calibration on the target population, not generalization to unseen datasets or images. Performance depends on the fixed screening pool covering relevant target behavior; a coreset cannot recover behavior absent from that pool. Annotation budgets count class masks rather than human effort, which varies with scene complexity. Finally, the conditional approximation argument concerns screening-pool utility, and timing estimates do not replace an end-to-end deployment benchmark.

6 Conclusion

PromptCoreset uses candidate-mask behavior to select calibration images and estimate global IoU without model training. Across 24 dataset–class units per system, it lowers mean regret relative to random calibration on three frozen segmenters while reducing prompt-conditioned calls by 89.5%. Post-hoc factorial analysis shows that weighting depends on the chosen representatives and model, with the clearest benefit on SAM 3. Behavior-aware calibration improves the observed cost–quality trade-off, although the fixed signature and its weighting are not uniformly optimal.

With fixed $K \times M$ screening predictions, increasing B changes annotation rather than inference cost. The matched-budget gains at $B = 8$ and $B = 16$ on all three systems, together with the unweighted coreset results, show that behavior-aware representative selection contributes beyond cluster weighting. This training-free design supports prompt choice when only a small number of target images can be annotated.

The benefit is visible in the deployed masks, not only in the ordering of candidate texts. At $B = 8$, PromptCoreset improves mean IoU over the canonical Class Name by 1.47, 2.91, and 1.53 pp on SAM 3, SegEarth-OV, and CAT-Seg, respectively; the corresponding mean F1 gains are 1.84, 3.19, and 1.53 pp (Table 2). These improvements require no backbone update and only eight labeled class masks per unit. Relative to Random Calibration at the same budget, the method also lowers mean Prompt Regret on each system. Behavior-aware calibration therefore improves the final one-prompt segmentation decision, rather than merely a surrogate ranking statistic, while retaining substantially lower prompt-conditioned inference cost than exhaustive search.

AI USE STATEMENT

In this work, generative AI tools were used solely for language editing and polishing. All AI-assisted text was carefully reviewed by the authors, who take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

Sections 3–4 describe the utility, behavior representation, medoid selection, estimator, evaluation populations, label budgets, and statistical units. Exact prompt banks, class mappings, image manifests, inference configurations, tie-breaking rules, random seeds, and selection audit hashes are maintained with the experiment artifacts and are intended for the anonymous supplementary release.

REFERENCES

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2020.

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

Keyan Chen, Chenyang Liu, Hao Chen, Haotian Zhang, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17, 2024. doi: 10.1109/TGRS.2024.3356074.

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*, pp. 4113–4123, 2024.

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

Leonard Kaufman and Peter J. Rousseeuw. Partitioning around medoids (program PAM). In *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, 1990. doi: 10.1002/9780470316801.ch2.

Justin Kay, Grant Van Horn, Subhransu Maji, Daniel Sheldon, and Sara Beery. Consensus-driven active model selection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4594–4604, 2025.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023.

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

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

Kaiyu Li, Ruixun Liu, Xiangyong Cao, Xueru Bai, Feng Zhou, Deyu Meng, and Zhi Wang. SegEarth-OV: Towards training-free open-vocabulary segmentation for remote sensing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10545–10556, 2025.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7061–7070, 2023.

Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. CLIP is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15305–15314, 2023.

Radek Mackowiak, Philip Lenz, Omair Ghori, Ferran Diego, Oliver Lange, and Carsten Rother. CEREALS: Cost-effective region-based active learning for semantic segmentation. In *British Machine Vision Conference*, 2018.

Patrik Okanovic, Andreas Kirsch, Jannes Kasper, Torsten Hoefler, Andreas Krause, and Nezihe Merve Gürel. All models are wrong, some are useful: Model selection with limited labels. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2035–2043, 2025.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15691–15701, 2023.

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

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. DenseCLIP: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18082–18091, 2022.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman R"adle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Mohammad Reza Karimi, Nezihe Merve Gürel, Bojan Karlaš, Johannes Rausch, Ce Zhang, and Andreas Krause. Online active model selection for pre-trained classifiers. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 307–315, 2021.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

H. Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pp. 287–294, 1992. doi: 10.1145/130385.130417.

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

Di Wang, Jing Zhang, Bo Du, Minqiang Xu, Lin Liu, Dacheng Tao, and Liangpei Zhang. SAMRS: Scaling-up remote sensing segmentation dataset with segment anything model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8815–8827, 2023.

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

Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2945–2954, 2023.

Yi Yang, Min Chen, Yu Zhang, Xueyan Hu, and Xiaochun Cao. Test-time multi-prompt adaptation for open-vocabulary remote sensing image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Enkai Zhang, Jingjing Liu, Anda Cao, Zhen Sun, Haofei Zhang, Huiqiong Wang, Li Sun, and Mingli Song. RS-SAM: Integrating multi-scale information for enhanced remote sensing image segmentation. In *Proceedings of the Asian Conference on Computer Vision*, pp. 994–1010, 2024.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022a.

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