ACTIVESAM: FAST AND ACCURATE OPEN-VOCABULARY SEMANTIC SEGMENTATION WITH FROZEN SAM 3

Anonymous authors
Paper under double-blind review

ABSTRACT

Segment Anything Model 3 (SAM 3) provides a strong frozen backbone for concept-prompted segmentation, but applying it directly to open-vocabulary semantic segmentation (OVSS) is inefficient: full-resolution decoding is typically run over the entire dataset vocabulary, whereas each image contains only a small active subset of classes. We introduce ActiveSAM, a training-free inference framework that turns SAM 3 into an active-vocabulary segmenter. ActiveSAM first canonicalizes and expands class prompts, then uses evidence-proportional grounding to estimate an image-conditioned active set from a low-resolution presence preview. Only retained prompts receive full-resolution mask prediction, using bucketed prompt multiplexing with the frozen SAM 3 decoder. The preview stage uses only class-presence evidence and skips unnecessary segmentation-head computation. To resolve overlapping concept responses, exclusive concept decoding compares each pixel's joint score vector with class signatures estimated once per vocabulary from unlabeled images. ActiveSAM requires no weight updates, no oracle class-presence labels and no per-dataset hyperparameter tuning. Across eight OVSS benchmarks, ActiveSAM improves the speed-accuracy trade-off of training-free open-vocabulary semantic segmentation, outperforming the current state-of-the-art SegEarth-OV3 by $+2.1$ mIoU on average while running much faster, with 7.3–12.2$\times$ speedups on large-vocabulary datasets. ActiveSAM also achieves the highest accuracy under image corruptions that simulate real-world distribution shift, making it well-suited for deployment in noisy-input domains such as autonomous driving and embodied AI.

1 INTRODUCTION

Real perception systems rarely operate with a fixed label set. Autonomous-driving stacks encounter new traffic objects, embodied agents receive task-specific object names, and image-analysis workflows revise their taxonomies as the task changes. In these scenarios, collecting new pixel-level annotations and retraining a segmentation model for every vocabulary update is impractical. Open-vocabulary semantic segmentation (OVSS) targets this setting: the model receives a vocabulary at test time and should produce dense masks for those classes without collecting new pixel annotations or retraining for every category update. This capability is particularly valuable as segmentation models move from closed environments to open-world deployment, where the deployment vocabulary can be large, dynamic, and only partially relevant to an individual image.

Vision-language models have made this possible by adapting CLIP-style (Radford et al., 2021; Jia et al., 2021) image-text alignment to dense prediction through region classification (Li et al., 2022; Cho et al., 2024), mask proposals (Ghiasi et al., 2022; Ding et al., 2021), side adapters (Xu et al., 2023b), diffusion features (Xu et al., 2023a), and universal segmentation frameworks (Zhang et al., 2023). Recent methods have increasingly focused on training-free OVSS, where foundation models are kept frozen and composed at inference time (Zhou et al., 2022; Bousselham et al., 2024; Wang et al., 2024a; Hajimiri et al., 2025; Lan et al., 2024b; Zhang et al., 2025; Yang et al., 2025; Chen et al., 2025).

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.

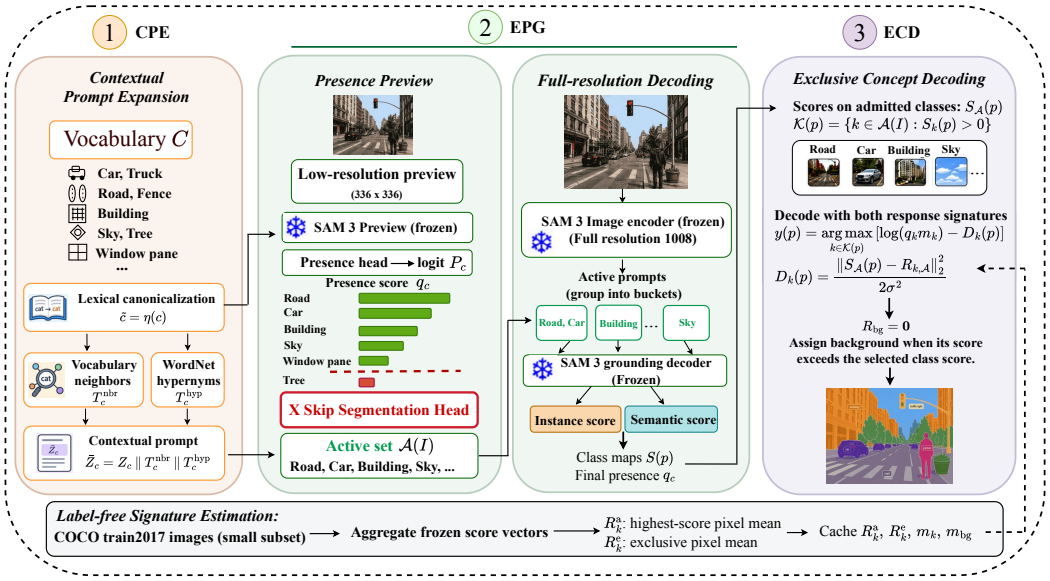

Figure 1: Overview of ActiveSAM, which decodes only the prompts relevant to each image and resolves overlapping concept responses, with all SAM 3 weights frozen. (1) **Contextual Prompt Expansion (CPE)** canonicalizes class names and expands their prompts. (2) A low-resolution **presence preview** skips the segmentation head and admits only the classes likely to be present. (3) **Full-resolution decoding** encodes the image once and decodes the active prompts in buckets. Steps (2) and (3) form **Evidence-Proportional Grounding (EPG)**. (4) **Exclusive Concept Decoding (ECD)** labels each pixel by comparing its score vector with class signatures estimated once per vocabulary from unlabeled images.

However, directly applying image-level recognition to pixel-level tasks often yields coarse results, as OVSS demands precise boundaries, local consistency, and fine-grained separation in complex scenes, and recent methods import spatial correspondence (Lan et al., 2024b), reconstruct CLIP patch correlations (Zhang et al., 2025), or modify attention (Wang et al., 2024a; Hajimiri et al., 2025; Yang et al., 2025) to close this gap. This limitation has motivated a shift toward segmentation foundation models. SAM (Kirillov et al., 2023) and SAM 2 (Ravi et al., 2024) provide strong spatial masks, and recent work extends promptable segmentation with language prompts or VLM embeddings (Xiao et al., 2026). SAM 3 (Carion et al., 2026) makes the connection even more direct: it accepts concept prompts such as noun phrases, returns masks for matching instances, and decouples recognition from localization with a presence head. Recent SAM 3-based OVSS pipelines therefore convert the evaluation vocabulary into text prompts, obtain class-conditioned semantic and instance evidence, and use presence scores to suppress absent categories (Li et al., 2025). While this approach improves mask quality, it creates a new systems bottleneck: full-resolution prompt-conditioned decoding scales with the vocabulary, although any individual image contains only a small active subset of classes.

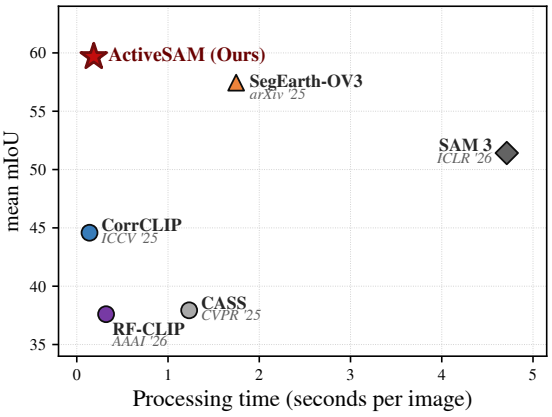

Figure 2: ActiveSAM (⋆ Ours) achieves state-of-the-art mIoU while running significantly faster than SegEarth-OV3 (Li et al., 2025). Reported values are averaged over six benchmarks: Context59, Context60, COCO-Object, COCO-Stuff, ADE20K, and Pascal VOC21.

A second limitation concerns how concept masks are converted into semantic labels. SAM 3 evaluates each prompt separately and can give high scores to related concepts on the same pixels. A rider can match both *rider* and *person* because both descriptions are valid. Semantic segmentation instead requires one label per pixel. Selecting the highest score can label rider pixels as *person* even when the model also responds strongly to *rider*. Filtering absent classes cannot resolve this ambiguity when both classes are present. The response to *rider* still provides useful evidence even when its score is lower. This motivates a decision that considers the class scores together rather than relying only on the highest score.

We propose **ActiveSAM**, a training-free inference framework that turns frozen SAM 3 (Carion et al., 2026) into an active-vocabulary segmenter. The key idea is to use SAM 3's presence signal to prune absent classes before full-resolution decoding, rather than only as a filtering score after decoding. As shown in Figure 1, Contextual Prompt Expansion (CPE) first canonicalizes class names and adds cached semantic context. Evidence-Proportional Grounding (EPG) then selects prompts through a low-resolution presence preview that skips the segmentation head. The same presence threshold screens the selected prompts again at full resolution before mask prediction. Retained prompts are decoded in buckets that reuse image features. The number of decoded prompts therefore adapts to the content of each image, without a fixed class budget.

To distinguish overlapping concepts, we introduce Exclusive Concept Decoding (ECD). Each class is represented by *response signatures* that summarize the model's scores across concepts on pixels selected for that class by the frozen model. ECD compares each pixel's score vector with these signatures using a nearest-signature classifier with a class prior. It first revises the highest-score label only when two signature estimates agree on the replacement and both the original and proposed classes have enough calibration pixels. For vocabularies with a background label, ECD then compares the selected class against a *null signature*, a zero score vector representing background.

The signatures are estimated once per vocabulary from one shared pool of unlabeled COCO training images, without semantic annotations or evaluation images. ActiveSAM keeps all model weights frozen, requires no oracle class-presence labels, and uses the global set of hyperparameters across all eight benchmarks.

Our main contributions are summarized as follows:

- We introduce ActiveSAM, a training-free inference framework that uses SAM 3's presence predictions to decide which prompts receive full-resolution mask decoding for each image, replacing the standard but inefficient practice of decoding the entire dataset vocabulary on every image. The framework combines three proposed components: *Contextual Prompt Expansion* (CPE) for richer class prompts, *Evidence-Proportional Grounding* (EPG) for prompt selection at two resolutions, and *Exclusive Concept Decoding* (ECD) for assigning semantic labels from joint concept responses. No SAM 3 weights are updated.
- ActiveSAM achieves the highest mean mIoU among the compared methods across eight standard OVSS benchmarks. It reaches 66.02% mean mIoU, outperforms the current state-of-the-art method SegEarth-OV3 (Li et al., 2025) by 2.1 points while dramatically running faster. On large vocabulary datasets, ActiveSAM achieves $7.3\times$ to $12.2\times$ faster while improving segmentation accuracy on each benchmark (Figure 2). Table 2 provides the per-dataset comparison.
- ActiveSAM demonstrates strong robustness to distribution shifts. Through rigorous testing on common image corruptions (noise, blur, JPEG compression, fog), we show that ActiveSAM maintains high performance where traditional CLIP-based baselines suffer severe accuracy degradation, a crucial property for real-world applications such as autonomous driving and embodied AI.

2 Related Work

Open-vocabulary semantic segmentation. Early OVSS methods such as LSeg (Li et al., 2022) align pixel embeddings with text-label embeddings. GroupViT (Xu et al., 2022) shows that semantic grouping can emerge from text supervision, while OpenSeg (Ghiasi et al., 2022) scales open-vocabulary segmentation using image/mask-level alignment. ZegFormer (Ding et al., 2021) decouples zero-shot segmentation into class-agnostic grouping and segment-level classification. Seg-

CLIP (Luo et al., 2023) learns patch aggregation with learnable centers, and TCL (Cha et al., 2023) and CODE (Wu et al., 2024) learn text-grounded masks by decomposing paired images and texts into region and word components. CLIP-DINOISER (Wysoczańska et al., 2024) uses a DINO (Caron et al., 2021) backbone to strengthen CLIP dense features. TALK2DINO (Barsellotti et al., 2025) bridges DINOv2 (Oquab et al., 2023) features with CLIP language embeddings by training a lightweight alignment layer on an external dataset. Despite their strong performance, these architectures require additional training or weak supervision, making them unsuitable for deployment where the foundation model must remain strictly frozen at test time.

Training-free open-vocabulary semantic segmentation. Training-free OVSS aims to preserve segmentation capability without requiring learnable parameters or additional training. MASKCLIP (Zhou et al., 2022) adapts frozen CLIP for dense prediction without additional training or pixel-level annotations. Later training-free methods try to modify CLIP attention at inference time: SCLIP (Wang et al., 2024a) explores deeper self-attention layers, PROXYCLIP (Lan et al., 2024b) uses proxy attention, NACLIP (Hajimiri et al., 2025) enforces neighbor-aware localization, CLIPTRASE (Shao et al., 2024) recalibrates patch self-correlations, RESCLIP (Yang et al., 2025) uses residual cross-correlation self-attention with semantic feedback refinement, and DIH-CLIP (Duan et al., 2025) exploits diversity across multi-head self-attention.

Several approaches combine the CLIP-based model with other strong foundation models. FREEDA (Barsellotti et al., 2024) builds offline diffusion-augmented prototypes for open-vocabulary segmentation. CORRCLIP (Zhang et al., 2025) reconstructs CLIP patch correlations using multiple foundation models such as DINO (Caron et al., 2021) and SAM 2 (Ravi et al., 2024), and TRIDENT (Shi et al., 2025) combines CLIP, DINO, and SAM (Kirillov et al., 2023) for high-resolution training-free segmentation.

SAM-based open-vocabulary semantic segmentation. The SEGMENT ANYTHING MODEL (SAM) (Kirillov et al., 2023) introduced promptable segmentation with strong spatial masks, and SAM 2 (Ravi et al., 2024) extended this idea to images and videos, but neither solves semantic labeling over an arbitrary vocabulary. SAM 3 (Carion et al., 2026) moves closer to semantic open-vocabulary segmentation by associating segmentation with concepts. SEGEARTH-OV3 (Li et al., 2025) explores SAM 3 for remote-sensing OVSS and natural OVSS benchmarks. This method uses dual-head mask fusion and the presence head for decoder phase and generates the final labels. MM-OVSEG (Wei et al., 2026) extends open-vocabulary segmentation to multimodal remote sensing. Our ActiveSAM is related to SAM 3 (Carion et al., 2026) and SEGEARTH-OV3 (Li et al., 2025), but it aims at two different important questions: How can we exploit SAM 3's strong backbone while reducing the computational cost of standard SAM 3-based inference, and how can its independent concept responses be turned into one label per pixel? Instead of applying the conventional encoder-decoder pass of SAM 3, our method uses a cheap presence-only preview to select an image-specific active vocabulary before full-resolution decoding, then decodes only active class prompts with bucketed class-prompt multiplexing and return the predicted class maps jointly instead of taking their pixelwise maximum.

3 ACTIVESAM

Problem setup. Given an image I and a semantic vocabulary $\mathcal{C} = \{c_1, \ldots, c_C\}$ containing C classes, our goal is to predict a dense label map $\hat{y} \in (\mathcal{C} \cup \{\mathrm{bg}\})^{H \times W}$, where $\hat{y}$ assigns one semantic label to each pixel, H and W are the output height and width, and bg denotes background for benchmarks that define an explicit background class. We build on a frozen SAM 3 model, which performs concept-prompted segmentation and decouples recognition from localization with a presence head (Carion et al., 2026). ActiveSAM is training-free and keeps all SAM 3 weights frozen.

Overview. ActiveSAM consists of three components. First, Contextual Prompt Expansion (CPE) constructs and caches a contextual prompt $\bar{Z}_c$ for each class $c \in \mathcal{C}$. Second, Evidence-Proportional Grounding (EPG) uses a low-resolution presence preview to remove unlikely prompts before full-resolution decoding. The full-resolution image encoder is evaluated once, and the prompts of the active set are processed in buckets. After one decoder layer, their presence is checked again using the same global threshold τ, and only prompts that still have sufficient evidence continue through the full decoder layers and the segmentation head. Third, Exclusive Concept Decoding (ECD) converts

the resulting class score maps into one semantic label per pixel. Rather than selecting only the class with the highest independent score, ECD compares the full vector of class scores with cached class signatures estimated once per vocabulary from unlabeled images. As a result, expensive mask decoding is performed only for prompts supported by the image, while the final prediction uses the joint response across classes.

3.1 Contextual Prompt Expansion

Class names in open-vocabulary segmentation are often short category labels rather than natural text prompts. Compound labels such as *windowpane* and *trafficlight*, for example, may be harder for the text encoder to interpret than their natural forms, *window pane* and *traffic light*. To obtain clearer class prompts, our *Contextual Prompt Expansion* (CPE) first canonicalizes each class name and then adds two complementary sources of context: related classes from the vocabulary and WordNet hypernyms. For each class $c \in \mathcal{C}$, a fixed lexical canonicalizer η converts the original class name into a canonicalized form $\tilde{c} = \eta(c)$, such as *windowpane*$\mapsto$*window pane*. The frozen SAM 3 text encoder (Carion et al., 2026) then maps $\tilde{c}$ to its prompt tokens Z_c and pooled text embedding e_c. Using e_c, we find the $M_s = 2$ other classes in the vocabulary with the highest cosine similarity and project their pooled embeddings into prompt tokens, denoted by T_c^{nbr}. We also obtain up to $M_t = 2$ noun hypernyms of $\tilde{c}$ from WordNet (Miller, 1995) and encode them in the same way to obtain T_c^{hyp}. The final contextual prompt is

$$\bar{Z}_c = Z_c \parallel T_c^{\mathrm{nbr}} \parallel T_c^{\mathrm{hyp}}, \tag{1}$$

where $\bar{Z}_c$ is the contextual prompt used for class c, Z_c contains the tokens of its canonicalized class name, T_c^{nbr} contains the $M_s = 2$ vocabulary-neighbor tokens, and T_c^{hyp} contains up to $M_t = 2$ WordNet hypernym tokens. The operator $\parallel$ denotes concatenation along the token sequence. All three sources depend only on the class vocabulary and fixed lexical resources: the canonicalization rules and WordNet lookup are fixed before evaluation and use no images or annotations. The resulting prompts are constructed once for each vocabulary and cached for both full-resolution inference and the signature-estimation pass in Section 3.4. The low-resolution presence preview in Section 3.2 uses only the canonicalized class tokens Z_c, without the additional context tokens, to keep the preview inexpensive. Appendix A gives further details.

3.2 Evidence-Proportional Grounding

Most images contain only a small subset of the classes in the full vocabulary. ActiveSAM therefore uses the SAM 3 presence head to remove unlikely classes before expensive full-resolution decoding. A single global presence threshold τ is used at both the preview and full-resolution stages.

Presence preview. We run frozen SAM 3 at a preview resolution of 336 over all C class prompts, using the canonicalized class tokens Z_c without the additional CPE context tokens. Since this stage only estimates class presence, it uses the image–text fusion encoder and a shortened object decoder while skipping the segmentation head. This shortened preview gives the same class-admission decisions as the full decoder on all tested benchmarks (Appendix A). Let $q_c^{\mathrm{p}} = \mathrm{sigmoid}(\rho_c)$ denote the preview presence score of class c, where ρ_c is the corresponding SAM 3 presence logit. We admit

$$\mathcal{A}(I) = \{c \in \mathcal{C} : q_c^{\mathrm{p}} \geq \tau\} \cup \{\mathrm{bg}\}, \tag{2}$$

where $\mathcal{A}(I)$ is the active set of image I, the classes admitted by the preview, and τ is the single global presence threshold shared across all datasets. For benchmarks with a background class, the background prompt bg is always admitted, and for other benchmarks, bg is omitted. If no semantic class reaches τ, we retain the class with the highest presence score. Thus, only classes with sufficient presence evidence proceed to the full-resolution stage. Classes rejected by the preview require no full-resolution decoding, whereas SAM 3 (Carion et al., 2026) and SegEarth-OV3 (Li et al., 2025) decode the full vocabulary before applying presence scores. We use a preview resolution of 336 throughout, and performance is stable across a wide range of preview resolutions (Table 12).

Full-resolution decoding. ActiveSAM runs the full-resolution SAM 3 image encoder once at resolution 1008 and divides the active classes into buckets of at most $K = 32$. Classes in the same

bucket are processed together in one decoder forward pass. For each bucket, the image–text fusion encoder and the first decoder layer produce a full-resolution presence score $q_c^{(1)}$ for class c. If $q_c^{(1)} < \tau$, the class is stopped early and assigned an all-zero score map. Otherwise, it is processed by the full decoder, segmentation head, and post-processing. Following SegEarth-OV3 (Li et al., 2025), we obtain instance and semantic score maps and fuse them with the final presence score q_c,

$$S_c(p) = q_c \max\big(S_c^{\text{inst}}(p), S_c^{\text{sem}}(p)\big), \tag{3}$$

where $S_c(p)$ is the final score of class c at pixel p, $S_c^{\text{inst}}(p)$ and $S_c^{\text{sem}}(p)$ are its instance and semantic scores, and q_c is its final full-resolution presence score. Classes rejected before full decoding are assigned zero scores, i.e., $S_c(p) = 0$ for all pixels p. We denote the full class-score vector at pixel p by $S(p) \in [0,1]^C$, and the standard one-versus-rest prediction by $y_0(p) = \arg\max_c S_c(p)$.

3.3 Exclusive Concept Decoding

SAM 3 predicts a score map for each class prompt independently, so related classes can respond to the same pixels. For example, a *rider* can also receive a high *person* score, and an *armchair* can also receive a high *chair* score. A pixelwise argmax y_0 considers only the largest score and can therefore favor the broader class. Exclusive Concept Decoding (ECD) instead uses the joint score pattern across classes.

Response signatures. For each class k, we define a response signature $R_k \in [0,1]^C$ as the expected class-score vector on pixels selected for that class by the frozen model. For example, a *rider* signature can contain high responses for both *rider* and *person*, while a *person* signature need not respond strongly to *rider*. The signatures are estimated once per vocabulary from unlabeled images, together with the pixel share m_k of each class (Section 3.4).

Decoding. For each image, let $S_{\mathcal{A}}(p)$ denote the score vector $S(p)$ restricted to the active classes $\mathcal{A}(I)$, and let $R_{k,\mathcal{A}}$ denote the corresponding entries of class signature R_k. We consider classes with a positive score at pixel p, $\mathcal{K}(p) = \{k \in \mathcal{A}(I) : S_k(p) > 0\}$. ECD combines the class prior with its distance to the response signature and selects the class with the highest score

$$y(p) = \arg\max_{k \in \mathcal{K}(p)} \left[\log(q_k m_k) - \frac{\|S_{\mathcal{A}}(p) - R_{k,\mathcal{A}}\|_2^2}{2\sigma^2} \right], \tag{4}$$

where $y(p)$ is the ECD prediction, $\|S_{\mathcal{A}}(p) - R_{k,\mathcal{A}}\|_2^2$ is the squared Euclidean distance between the pixel score vector and the class signature, q_k is the final full-resolution presence score of class k, and m_k is its pixel share on the shared unlabeled pool. The term $\log(q_k m_k)$ accounts for support from the current image and the unlabeled pool, while the distance term measures how closely the pixel score pattern matches each class signature. The scale σ controls the balance between these two terms and is shared across all datasets.

Agreement and support. We evaluate Eq. 4 using the two signature estimates R_k^{a} and R_k^{e} defined in Section 3.4. We accept the ECD prediction only when both signature estimates give the same class, otherwise the pixel keeps $y_0(p)$. Classes with insufficient support in the unlabeled pool also keep the baseline prediction.

Null-signature background. For benchmarks with an explicit background class, we additionally represent background by a null signature $R_{\text{bg}} = \mathbf{0}$. Its prior is determined by the background mass m_{bg} estimated from the shared unlabeled pool and a single global offset b shared across datasets. A pixel assigned to class y is changed to background when the null signature receives the higher score,

$$\log m_{\text{bg}} + b - \frac{\|S_{\mathcal{A}}(p)\|_2^2}{2\sigma^2} > \log(q_y m_y) - \frac{\|S_{\mathcal{A}}(p) - R_{y,\mathcal{A}}^{\text{a}}\|_2^2}{2\sigma^2}, \tag{5}$$

where $\|S_{\mathcal{A}}(p)\|_2^2$ is the squared Euclidean distance from the pixel score vector to the null background signature, while $\|S_{\mathcal{A}}(p) - R_{y,\mathcal{A}}^{\text{a}}\|_2^2$ is its squared Euclidean distance to the signature of the predicted class y. The pixel is therefore assigned to background only when the null signature gives the higher score. This uses the same signature comparison as the semantic classes and avoids a dataset-specific background threshold.

3.4 LABEL-FREE SIGNATURE ESTIMATION

We estimate the response signatures from a shared unlabeled subset of COCO train2017 images (Lin et al., 2014). The same image pool is used across all benchmarks in this work, without using semantic annotations or any evaluation image. Each pool image is processed by CPE and EPG without ECD, producing the class-score maps used to estimate the signatures. This pass only aggregates outputs from the frozen model and does not update any model parameters.

For each class k, we collect three groups of pixels: pixels where k has the highest score, pixels where k responds strongly, and pixels where k responds without a competing class. Let μ_k^{a}, μ_k^{p}, and μ_k^{e} denote the mean class-score vectors over these three groups, respectively. The first ECD signature is $R_k^{\mathrm{a}} = \mu_k^{\mathrm{a}}$. The second signature R_k^{e} uses the exclusive estimate μ_k^{e} when it has sufficient support, and otherwise falls back to μ_k^{p} and then μ_k^{a}. We also estimate the class mass m_k from the fraction of pixels where class k has the highest score.

These two signatures capture complementary views of the same class. R_k^{a} reflects the response pattern where the class wins against other classes, while R_k^{e} aims to capture its response when competing classes are weak. Their agreement is used by ECD to avoid uncertain reassignments. All signatures and class masses are estimated once per vocabulary from the shared unlabeled pool. The exact pixel-selection thresholds, support rule, pool construction, and implementation details are given in Appendix B.

4 EXPERIMENTAL RESULTS

Experimental Setup. We evaluate ActiveSAM on eight standard OVSS benchmarks: Pascal VOC 20/21 (with a background class) (Everingham et al., 2010), Pascal Context 59/60 (Mottaghi et al., 2014), COCO-Object (Lin et al., 2014), COCO-Stuff (Caesar et al., 2018), Cityscapes (Cordts et al., 2016), and ADE20K (Zhou et al., 2017). COCO-Object evaluates the 80 MS COCO classes and treats "stuff" as background. Following the standard practice in training-free OVSS, we use no target-dataset annotations such as ground-truth class presence or pixel-level labels, and no evaluation image is used before inference. All experiments are conducted on a single NVIDIA RTX 5090, and performance is reported using mean Intersection over Union (mIoU). We use the same hyperparameters across all datasets without per-benchmark tuning: $\tau = 0.1$, $\sigma = 0.15$, pixel-selection thresholds 0.5 and 0.3, minimum support $N = 2000$, and $b = 1$. The preview resolution is 336. Our hyperparameter analysis (Table 6) and preview-resolution analysis (Table 12) support these choices: each default is within 0.04 mean mIoU of the best tested value, and performance remains stable across all tested values. Further implementation details are provided in Appendix A.

4.1 MAIN RESULTS

Comparison with state-of-the-art methods. As summarized in Table 1, ActiveSAM achieves the highest mean mIoU of 66.0 across eight OVSS benchmarks, outperforming the current state-of-the-art method, SegEarth-OV3 (Li et al., 2025), by 2.1 points and SAM 3 (Carion et al., 2026) by 8.5 points. The largest gains over SAM 3 are on COCO-Stuff ($+11.6$), Cityscapes ($+10.7$), and Pascal Context-60/59 ($+10.5$/$+10.4$). Compared with SegEarth-OV3, which uses the same SAM 3 backbone, ActiveSAM improves VOC21 by 3.6 points, Cityscapes by 3.3, Context60 by 3.2, ADE20K by 2.4, and COCO-Stuff by 2.1. Together with the efficiency results in Table 2, these gains support our motivation to avoid full-vocabulary decoding when most classes are absent from an image. Qualitative comparisons with SegEarth-OV3 are shown in Appendix I.

Speed–accuracy advantage. ActiveSAM achieves a substantial efficiency gain across all eight benchmarks. As shown in Table 2, it runs much faster than both SAM 3 (Carion et al., 2026) and SegEarth-OV3 (Li et al., 2025) on every benchmark. Compared with the current state-of-the-art SegEarth-OV3, ActiveSAM runs 7.3–12.2$\times$ faster on Context59/60, COCO-Object, COCO-Stuff, and ADE20K, while also improving mIoU on all five datasets. Across all eight benchmarks, ActiveSAM is $7.5\times$ faster on average. These results highlight a key strength of ActiveSAM: it achieves both goals at once, substantially faster inference and improved segmentation accuracy.

Robustness under severe image corruptions. Practical segmentation systems should remain reliable under input degradation. Autonomous vehicles encounter fog and motion blur. We therefore

Table 1: **Open-vocabulary semantic segmentation on 8 standard benchmarks.** (mIoU, %). The best score in each column is shown in **bold**. Further baselines are listed in Table 8 (Appendix F). DI and DI2 denote DINO and DINOv2, respectively. S/16, B/8, B/14, B/16, L/14, and L/16 denote ViT-S/16, ViT-B/8, ViT-B/14, ViT-B/16, ViT-L/14, and ViT-L/16, respectively. † indicates DINOv2 with registers. $\diamond$ = DI (B/8) + SAM2 (Hiera-L), ‡ = DI (B/16) + SAM (B/16), and § = DI2 (L/14) + SAM (L/16). P59/60: Pascal Context 59/60, C-O/C-S: COCO-Object/COCO-Stuff.

Method	Venue	Backbone (Size)	*with background*			*without background*					Avg.
			V21	P60	C-O	V20	P59	C-S	City	ADE	
Training-based											
GroupViT (Xu et al., 2022)	CVPR'22	GroupViT (S/16)	50.4	18.7	27.5	79.7	23.4	15.3	11.1	9.2	29.4
TCL (Cha et al., 2023)	CVPR'23	CLIP	51.2	24.3	30.4	77.5	30.3	19.6	23.1	14.9	33.9
CLIP-DINOiser (Wysoczańska et al., 2024)	ECCV'24	CLIP + DI (B/16)	62.1	32.4	34.8	80.9	35.9	24.6	31.7	20.0	40.3
Talk2DINO (Barsellotti et al., 2025)	ICCV'25	CLIP + DI2† (B/14)	65.8	37.7	45.1	88.5	42.4	30.2	38.1	22.5	46.3
Training-free											
FreeDA (Barsellotti et al., 2024)	CVPR'24	CLIP + DI2 (B/14)	51.8	35.3	36.3	84.3	39.7	25.7	34.1	20.8	41.0
ProxyCLIP (Lan et al., 2024b)	ECCV'24	CLIP + DI2† (B/14)	58.6	33.8	37.4	83.0	37.2	25.4	33.9	19.7	41.1
CASS (Kim et al., 2025)	CVPR'25	CLIP + DI (B/8)	65.8	36.7	37.8	87.8	40.2	26.7	39.4	20.4	44.4
CorrCLIP (Zhang et al., 2025)	ICCV'25	CLIP + $\diamond$	76.7	44.9	49.4	91.5	50.8	34.0	51.1	30.7	53.6
Trident (Shi et al., 2025)	ICCV'25	CLIP + ‡	67.1	38.6	41.1	84.5	42.2	28.3	42.9	21.9	45.8
ReME (Xuan et al., 2025)	ICCV'25	CLIP + §	82.2	44.6	48.2	93.2	53.1	33.3	59.0	28.2	55.2
SegEarth-OV3 (Li et al., 2025)	arXiv'25	SAM 3 (PE-L+/14)	79.8	53.4	72.0	96.8	59.2	42.8	69.7	37.6	63.9
RF-CLIP (Li et al., 2026)	AAAI'26	CLIP	67.2	37.9	39.1	87.0	41.4	27.5	43.0	21.0	45.5
SAM 3 (Carion et al., 2026)	ICLR'26	SAM 3 (PE-L+/14)	81.9	46.1	65.4	88.9	50.0	33.3	62.3	31.8	57.5
ActiveSAM	Ours	SAM 3 (PE-L+/14)	**83.4**	**56.6**	**72.6**	**97.1**	**60.4**	**44.9**	**73.0**	**40.0**	**66.0**

Table 2: **Per-dataset mIoU and FPS comparison** on the eight OVSS benchmarks where "$+\Delta$" is the mIoU improvement. "$N\times$" means ActiveSAM is N times faster than the respective method on the corresponding dataset. Best score per column is in **bold**.

Method	Metric	ADE20K	C59	C60	CO-O	CO-S	V21	V20	City	Mean
CASS (Kim et al., 2025)	mIoU	20.4 (+19.6)	40.2 (+20.2)	36.7 (+19.9)	37.8 (+34.8)	26.7 (+18.2)	65.8 (+17.6)	87.8 (+9.3)	39.4 (+33.6)	44.4 (+21.6)
	FPS	0.70 (5.0×)	0.81 (7.8×)	0.81 (7.6×)	0.68 (10.5×)	0.63 (6.0×)	2.30 (4.1×)	0.82 (11.6×)	0.11 (36.2×)	0.86 (11.1×)
SegEarth-OV3 (Li et al., 2025)	mIoU	37.6 (+2.4)	59.2 (+1.2)	53.4 (+3.2)	72.0 (+0.6)	42.8 (+2.1)	79.8 (+3.6)	96.8 (+0.3)	69.7 (+3.3)	63.9 (+2.1)
	FPS	0.35 (10.0×)	0.86 (7.4×)	0.84 (7.3×)	0.63 (11.4×)	0.31 (12.2×)	2.19 (4.3×)	1.92 (5.0×)	1.81 (2.2×)	1.11 (7.5×)
SAM 3 (Carion et al., 2026)	mIoU	31.8 (+8.2)	50.0 (+10.4)	46.1 (+10.5)	65.4 (+7.2)	33.3 (+11.6)	81.9 (+1.5)	88.9 (+8.2)	62.3 (+10.7)	57.5 (+8.5)
	FPS	0.13 (27.0×)	0.32 (19.8×)	0.32 (19.1×)	0.24 (29.8×)	0.11 (34.5×)	0.93 (10.1×)	0.93 (10.3×)	0.76 (5.2×)	0.47 (19.5×)
ActiveSAM	mIoU	**40.0**	**60.4**	**56.6**	**72.6**	**44.9**	**83.4**	**97.1**	**73.0**	**66.0**
	FPS	**3.51**	**6.33**	**6.12**	**7.16**	**3.79**	**9.37**	**9.54**	**3.98**	**6.23**

evaluate robustness using four representative ImageNet-C corruptions (Hendrycks & Dietterich, 2019). All corruptions are evaluated at severity 5. ActiveSAM achieves the highest mean mIoU under all four corruptions, outperforming the current state-of-the-art SegEarth-OV3 (Li et al., 2025) by 1.6-3.6 points across six benchmarks (Table 9). These results show that ActiveSAM's accuracy advantage extends beyond clean images and remains consistent under severe input corruption.

4.2 Ablation Studies and Analysis

Component ablations. We ablate each component separately: CPE and ECD in Table 4, and EPG in Table 10. CPE and ECD mainly improve accuracy, whereas EPG mainly reduces computation. Removing CPE lowers mean mIoU by 1.4 points, with the largest drop on VOC21 (83.4 to 79.0). Among the CPE components, lexical canonicalization contributes the most (Table 7), showing that clear class names matter for concept prompting. Replacing ECD with a pixelwise argmax lowers mean mIoU by 0.7 points. The gain is concentrated on Cityscapes ($+3.3$), where *rider* overlaps *person*, and on Context60 ($+1.5$). On the remaining benchmarks, ECD changes mIoU by at most 0.6 points. With $\tau = 0.1$, the preview keeps on average only 7.3 of 81 prompts per image on COCO-Object and 30.8 of 150 on ADE20K (Table 11). Decoding only these prompts changes mIoU by at most 0.34 on any benchmark, yet makes inference 1.8–4.1$\times$ faster on seven of the eight benchmarks (Table 10). Cityscapes is the exception ($1.25\times$): its images contain most of its classes, so the preview keeps 18.0 of 23 prompts and there is little to skip.

Prompt selection and bucketing. Table 3 removes prompt selection and bucketing one at a time. Prompt selection provides the largest share of the speedup: decoding every prompt slows ActiveSAM by $2.73\times$ on average and by 3.3–$4.1\times$ on COCO-Object, COCO-Stuff and ADE20K. Bucketed decoding also contributes substantially: removing it slows ActiveSAM by $1.74\times$ on

Table 3: **Prompt selection and bucketing.** We remove one component at a time to measure its contribution to the speedup. Without bucketing, each active prompt is decoded in its own decoder pass instead of in buckets, which leaves mIoU unchanged. Without prompt selection, every prompt is decoded with full decoding layers.

Dataset	Slowdown vs. ActiveSAM		Speedup over SegEarth-OV3	
	w/o bucketing	w/o prompt selection	ActiveSAM	w/o bucketing
VOC21	1.28×	1.84×	4.3×	3.3×
VOC20	1.22×	2.17×	5.0×	4.1×
Cityscapes	1.74×	1.25×	2.2×	1.3×
Context59	1.87×	2.65×	7.4×	3.9×
Context60	1.92×	2.55×	7.3×	3.8×
COCO-Object	1.53×	3.96×	11.4×	7.4×
COCO-Stuff	2.15×	4.12×	12.2×	5.7×
ADE20K	2.25×	3.34×	10.0×	4.5×
Mean	1.74×	2.73×	7.5×	4.3×

Table 4: **Component ablation.** mIoU (%) after removing ECD or CPE. The EPG ablation is reported in Table 10.

Configuration	*with background*			*without background*					Mean	Δ vs Full
	V21	P60	C-O	V20	P59	C-S	City	ADE		
Full ActiveSAM (CPE + EPG + ECD)	83.4	56.6	72.6	97.1	60.4	44.9	73.0	40.0	**66.0**	—
− ECD (pixelwise argmax)	83.3	55.1	72.2	97.1	60.5	44.3	69.7	40.0	65.3	−0.7
− CPE (plain prompts, signatures re-estimated)	79.0	56.1	72.8	96.7	60.3	43.4	70.1	38.8	64.6	−1.4
SegEarth-OV3 (Li et al., 2025) (for reference)	79.8	53.4	72.0	96.8	59.2	42.8	69.7	37.6	63.9	−2.1

average without changing mIoU. Even without bucketing, ActiveSAM remains 4.3× faster than SegEarth-OV3 (Li et al., 2025) on average, keeping most of its 7.5× faster inference.

Shared calibration pool. ActiveSAM does not select calibration images per benchmark: the signatures for all eight vocabularies are estimated from the same pool of unlabeled COCO training images, and no evaluation image is used before inference. To measure what this stricter setting costs, we also estimate the signatures from each benchmark's own evaluation images, without labels (Table 5). This variant raises mean mIoU only slightly, from 66.02 to 66.40, so the shared pool remains competitive with calibrating on each benchmark's own images. We keep the shared pool for all other results because it follows the stricter protocol.

One configuration for all benchmarks. ActiveSAM uses the same hyperparameters on all eight benchmarks, whereas SegEarth-OV3 (Li et al., 2025) adjusts its confidence and background thresholds per dataset. A shared configuration matters in open-vocabulary use, where a new vocabulary rarely comes with labeled data for tuning. Table 6 tests whether this configuration depends on its exact values: it varies six hyperparameters one at a time and disables the agreement rule and the class-mass term. All 14 variants keep mean mIoU at 65.1 or higher, above SegEarth-OV3's 63.9, and 13 of them stay within 0.5 points of the default. The only larger drop, 0.9 points, comes from raising the signature-distance scale σ to 0.20, which gives more weight to the class prior.

5 Conclusion

We presented ActiveSAM, a training-free framework for open-vocabulary semantic segmentation with frozen SAM 3. CPE clarifies class prompts, EPG selects which prompts receive full-resolution mask prediction, and ECD uses joint class responses to distinguish overlapping concepts. The response signatures are estimated once per vocabulary from a shared unlabeled image pool. Across eight benchmarks, ActiveSAM improves mean mIoU by 2.1 over SegEarth-OV3 (Li et al., 2025) and runs much faster on every benchmark, by 7.3× to 12.2× on the large-vocabulary benchmarks. It also achieves the highest segmentation accuracy under severe image corruptions, which is desirable in real-world conditions. These results show that selecting relevant prompts and resolving their overlap can improve both the efficiency and accuracy of a frozen segmentation model.

AI USE STATEMENT

In this work, we used generative AI tools to polish writing and improve grammar. We also used them to refine the format of tables and figures. All research ideas, methods, experiments, analyses, and substantive writing were carried out by the authors without AI/LLM assistance.

ETHICS STATEMENT

ActiveSAM may reduce the computational cost of deploying segmentation models, but it could also be misused for surveillance or tracking. Its English lexical resources and calibration pool may underrepresent some concepts. Prompt filtering can also omit small or occluded objects. We use public benchmark datasets under their respective licenses and collected no new personal data. These limitations require evaluation in the intended domain before any safety-critical use.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of our work. Section 3 and Appendix A describe the method and detailed implementation settings. Appendix B details the signature estimation. Appendices C and D examine calibration sources and hyperparameter sensitivity. We commit to providing our source code of implementation upon publication to facilitate reproduction of our results.

REFERENCES

Luca Barsellotti, Roberto Amoroso, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3689–3698, 2024.

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

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025.

Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3828–3837, 2024.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1209–1218, 2018.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *ICLR 2026*, 2026.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11165–11174, 2023.

Qi Chen, Lingxiao Yang, Yun Chen, Nailong Zhao, Jianhuang Lai, Jie Shao, and Xiaohua Xie. Training-free class purification for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23124–23134, 2025.

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.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Jian Ding, Nan Xue, Guisong Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. 2022 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11573–11582, 2021.

Songsong Duan, Xi Yang, and Nannan Wang. Dih-clip: Unleashing the diversity of multi-head self-attention for training-free open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22794–22803, 2025.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European conference on computer vision*, pp. 540–557. Springer, 2022.

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

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Shuo Jin, Siyue Yu, Bingfeng Zhang, Mingjie Sun, Yi Dong, and Jimin Xiao. Feature purification matters: Suppressing outlier propagation for training-free open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20291–20300, 2025.

Dahyun Kang and Minsu Cho. In defense of lazy visual grounding for open-vocabulary semantic segmentation. In *European Conference on Computer Vision*, pp. 143–164. Springer, 2024.

Chanyoung Kim, Dayun Ju, Woojung Han, Ming-Hsuan Yang, and Seong Jae Hwang. Distilling spectral graph for object-context aware open-vocabulary semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15033–15042, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *European Conference on Computer Vision*, pp. 143–160. Springer, 2024a.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *European Conference on Computer Vision*, pp. 70–88. Springer, 2024b.

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

Jiahao Li, Yang Lu, Yachao Zhang, Yong Xie, Fangyong Wang, Yuan Xie, and Yanyun Qu. Target refocusing via attention redistribution for open-vocabulary semantic segmentation: An explainability perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6181–6189, 2026.

Kaiyu Li, Shengqi Zhang, Yupeng Deng, Zhi Wang, Deyu Meng, and Xiangyong Cao. Segearth-ov3: Exploring sam 3 for open-vocabulary semantic segmentation in remote sensing images. *arXiv preprint arXiv:2512.08730*, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

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*, pp. 23033–23044. PMLR, 2023.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 891–898, 2014.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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*, pp. 8748–8763. PmLR, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Tong Shao, Zhuotao Tian, Hang Zhao, and Jingyong Su. Explore the potential of clip for training-free open vocabulary semantic segmentation. In *European Conference on Computer Vision*, pp. 139–156. Springer, 2024.

Yuheng Shi, Minjing Dong, and Chang Xu. Harnessing vision foundation models for high-performance, training-free open vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23487–23497, 2025.

Vladan Stojnić, Yannis Kalantidis, Jiří Matas, and Giorgos Tolias. Lposs: Label propagation over patches and pixels for open-vocabulary semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9794–9803, 2025.

Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. Clip as rnn: Segment countless visual concepts without training endeavor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13171–13182, 2024.

Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *European conference on computer vision*, pp. 315–332. Springer, 2024a.

Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3635–3647, 2024b.

Yimin Wei, Aoran Xiao, Hongruixuan Chen, Junshi Xia, and Naoto Yokoya. Mm-ovseg: Multimodal optical-sar fusion for open-vocabulary segmentation in remote sensing. *arXiv preprint arXiv:2603.17528*, 2026.

Ji-Jia Wu, Andy Chia-Hao Chang, Chieh-Yu Chuang, Chun-Pei Chen, Yu-Lun Liu, Min-Hung Chen, Hou-Ning Hu, Yung-Yu Chuang, and Yen-Yu Lin. Image-text co-decomposition for text-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26794–26803, 2024.

Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation. In *European Conference on Computer Vision*, pp. 320–337. Springer, 2024.

Shiting Ginny Xiao, Rishabh Kabra, Yuhang Li, Donghyun Lee, Joao Carreira, and Priyadarshini Panda. Openworldsam: Extending sam2 for universal image segmentation with language prompts. *Advances in Neural Information Processing Systems*, 38:35507–35532, 2026.

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.

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

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

Xiwei Xuan, Ziquan Deng, and Kwan-Liu Ma. Reme: A data-centric framework for training-free open-vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20954–20965, 2025.

Yuhang Yang, Jinhong Deng, Wen Li, and Lixin Duan. Resclip: Residual attention for training-free dense vision-language inference. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29968–29978, 2025.

Dengke Zhang, Fagui Liu, and Quan Tang. Corrclip: Reconstructing patch correlations in clip for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24677–24687, 2025.

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*, pp. 1020–1031, 2023.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.

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

APPENDIX

A IMPLEMENTATION DETAILS

We use the official SAM 3 model (Carion et al., 2026) with the Perception Encoder-Large+ backbone (Bolya et al., 2025) and keep all weights frozen. Images are resized to 1008×1008 for full-resolution inference and 336×336 for the preview. Together with the bucket size $K = 32$, $M_s = M_t = 2$ context tokens and the instance confidence threshold of 0.3, these settings are shared by all eight benchmarks. Class names pass through the fixed canonicalizer η before text encoding. We use no test-time augmentation or additional mask refinement. For the speed–accuracy comparison, all methods are run from their official implementations on one NVIDIA RTX 5090 under the same evaluation setup.

Contextual prompt expansion. CPE adds two vocabulary-neighbor tokens and up to two WordNet hypernym tokens to each canonicalized class prompt. Each added token is the pooled text embedding of the neighbor class or hypernym, mapped by a frozen projection from the 1024-dimensional pooled space to the 256-dimensional prompt-token space. For WordNet, we take the first noun synset of the canonicalized name and its first two immediate hypernyms; if none exist, no hypernym tokens are added. The canonicalizer η is a fixed set of string rules that rewrite class names into clearer forms, such as splitting compound words; it is applied identically to all datasets. The contextual prompts $\bar{Z}_c$ are cached once per vocabulary, so prompt construction is not repeated for each image.

Evidence-proportional grounding. The preview uses the canonicalized class tokens Z_c, the first two of SAM 3's six decoder layers, and the first 50 of its 200 object queries. It skips the segmentation head. On all tested benchmarks, this shortened preview admits the same classes as the full preview decoder. Prompts with presence at least $\tau = 0.1$ are retained, along with the explicit background prompt where applicable. If no semantic prompt passes, the highest-scoring one is retained.

At full resolution, the active prompts are processed in vocabulary order in buckets of at most $K = 32$, reusing the image features. After one decoder layer, prompts with presence below τ stop and receive zero score maps while retaining their one-layer presence scores. The remaining prompts use the full six-layer decoder, segmentation head, and post-processing. When a class has several names, each name is processed separately and the class score map and presence are the maxima over its names. This explains why VOC20 uses 26 prompts for 20 classes and Cityscapes uses 23 for 19.

Dual-head fusion and presence gating. Following SegEarth-OV3 (Li et al., 2025), we score each instance mask by the product of its class probability and prompt-presence score. Masks below the shared confidence threshold of 0.3 are discarded, and a pixelwise maximum over the remaining masks gives S_c^{inst}. We fuse this map with the semantic map S_c^{sem} and multiply by the final full-resolution presence q_c, as in Eq. 3. Logits are upsampled to the original image size in float32 before applying the sigmoid. The final presence q_c weights the score map, while preview presence q_c^{p} and first-layer presence $q_c^{(1)}$ control prompt selection.

Exclusive concept decoding. ECD loads the two signature matrices R^{a} and R^{e}, class masses, background mass where applicable, and the list of classes with insufficient calibration support. It restricts score vectors and signatures to the active classes and evaluates Eq. 4 with both signature estimates. A pixel keeps the baseline label $y_0(p)$ if the estimates disagree, either the original or proposed class lacks support, or no active class has a positive score. For background benchmarks, Eq. 5 then uses R^{a} on pixels not protected by the support rule.

B SIGNATURE ESTIMATION DETAILS

Calibration images. The shared pool consists of 9,858 COCO train2017 images (Lin et al., 2014), evenly spaced in the file list and selected without regard to image content. Each image is processed with CPE and EPG, without ECD. The presence-weighted score maps are average-pooled to one quarter of the original height and width. No annotation file is used.

Table 5: **Calibration source.** mIoU (%) with signatures estimated from different unlabeled image sets. The final row uses evaluation images without labels and is shown for reference only.

Unlabeled pool	V21	P60	C-O	V20	P59	C-S	City	ADE	Mean
Shared COCO pool, 10k images (default)	83.42	56.59	72.63	97.08	60.44	44.90	73.04	40.03	66.02
Shared COCO pool, 2k images	83.49	56.66	72.58	97.07	60.57	44.74	72.86	40.02	66.00
Evaluation images, no labels	83.63	56.52	72.75	97.10	60.43	44.83	74.89	41.02	66.40

Pixel groups and signatures. For each class k, we collect pixels where it has the highest positive score, where its score is high, and where its score is high while competing scores are low:

$$\begin{aligned}\Omega_k^{\mathrm{a}} &= \{p : \arg\max_j S_j(p) = k,\ \max_j S_j(p) > 0\},\\ \Omega_k^{\mathrm{p}} &= \{p : S_k(p) > 0.5\},\\ \Omega_k^{\mathrm{e}} &= \{p : S_k(p) > 0.5,\ \max_{j\neq k} S_j(p) < 0.3\}.\end{aligned} \tag{6}$$

Here, p indexes pooled pixels across the calibration images, and $S_j(p)$ is the score of class j. Let μ_k^x be the mean score vector over Ω_k^x and $n_k^x = |\Omega_k^x|$ for $x \in \{\mathrm{a}, \mathrm{p}, \mathrm{e}\}$. We define the two signatures and class mass as

$$\begin{aligned}R_k^{\mathrm{a}} &= \mu_k^{\mathrm{a}}, \qquad m_k = \frac{n_k^{\mathrm{a}}}{\sum_j n_j^{\mathrm{a}}},\\ R_k^{\mathrm{e}} &= \begin{cases}\mu_k^{\mathrm{e}} & n_k^{\mathrm{e}} \geq N,\\ \mu_k^{\mathrm{p}} & n_k^{\mathrm{e}} < N \leq n_k^{\mathrm{p}},\\ \mu_k^{\mathrm{a}} & \text{otherwise.}\end{cases}\end{aligned} \tag{7}$$

The minimum support is $N = 2000$ pooled pixels. A class with $n_k^{\mathrm{a}} < N$ is marked as uncalibrated, and ECD preserves the baseline label when either the original or proposed class is uncalibrated.

C Calibration Source

Table 5 changes only the unlabeled images used to estimate the signatures. Besides the default pool of about 10k images (Appendix B), we test a smaller 2k-image pool, which changes mean mIoU by only 0.02. For reference, we also estimate the signatures from each benchmark's own evaluation images, without labels. This variant raises mean mIoU from 66.02 to 66.40, but we do not use it, because our protocol never accesses evaluation images before inference. Apart from this reference variant, every experiment in the paper uses a shared COCO pool, whose train2017 images are disjoint from the val2017 images used to evaluate COCO-Object and COCO-Stuff.

D Hyperparameter Sensitivity

Table 6 varies one setting at a time. Changing the presence threshold τ, the two pixel-selection thresholds, or the minimum support N changes mean mIoU by at most 0.42, and changing the background offset b changes the mean over the three background benchmarks by at most 0.04. For the signature-distance scale σ, 0.10 and 0.15 give similar accuracy, while 0.20 gives more weight to the class prior and lowers mean mIoU to 65.13. Disabling the agreement rule or the class-mass term lowers mean mIoU by 0.22 and 0.27, respectively. Every variant still outperforms SegEarth-OV3 (63.9 mean mIoU, or 68.4 on the three background benchmarks), so the gains of ActiveSAM are stable and a single configuration can serve all eight benchmarks.

E CPE Components

Table 7 removes one CPE component at a time on the 2k pool and re-estimates the signatures with the resulting prompts. Lexical canonicalization contributes the most: removing it lowers mean mIoU by 1.06, with the largest drop on VOC21. Removing the neighbor or hypernym tokens lowers mean mIoU by 0.41 and 0.45, respectively. On the default 10k pool, replacing full CPE with plain prompts lowers mean mIoU by 1.37.

Table 6: **Hyperparameter sensitivity.** One setting is varied at a time. Values are mean mIoU (%) across eight benchmarks, except for b, which only affects the three benchmarks with a background class and is averaged over them (VOC21, Context60 and COCO-Object). The default setting is the middle value, or "on" for the two rules.

Setting	Role	Values tested	mIoU result
τ	prompt selection	0.05 / 0.1 / 0.2	66.03 / 66.02 / 65.81
σ	signature distance versus prior	0.10 / 0.15 / 0.20	65.92 / 66.02 / 65.13
agreement rule	decode with R^{a} and R^{e}	on / off	66.02 / 65.80
mass term	prior $q_k m_k$ versus q_k	on / off	66.02 / 65.75
confident-score threshold	confident and exclusive pixel groups	0.4 / 0.5 / 0.6	65.62 / 66.02 / 66.05
competitor-score threshold	competing scores in the exclusive group	0.2 / 0.3 / 0.4	66.05 / 66.02 / 65.96
N	minimum support (pooled pixels)	1000 / 2000 / 4000	65.60 / 66.02 / 66.03
b	background offset (three background benchmarks)	0.5 / 1 / 1.5	70.86 / 70.88 / 70.92

Table 7: **CPE components.** mIoU (%) after removing each prompt component and re-estimating signatures. Each Δ is relative to full CPE using the same pool. The upper block uses the 2k calibration pool, and the lower block uses the default 10k pool.

Prompts	V21	P60	C-O	V20	P59	C-S	City	ADE	Mean	Δ
Full CPE (2k pool)	83.49	56.66	72.58	97.07	60.57	44.74	72.86	40.02	66.00	—
without lexical canonicalization	79.07	55.94	72.50	96.69	60.02	43.40	72.88	39.01	64.94	−1.06
without neighbor tokens	83.02	56.78	72.80	97.10	60.76	44.58	70.19	39.52	65.59	−0.41
without hypernym tokens	83.27	56.72	72.66	97.08	60.64	44.66	69.57	39.75	65.54	−0.45
Full CPE (10k pool, default)	83.42	56.59	72.63	97.08	60.44	44.90	73.04	40.03	66.02	—
without CPE (plain prompts)	79.02	56.07	72.84	96.70	60.25	43.40	70.09	38.82	64.65	−1.37

F Additional Baselines

Table 8 reports further baselines under the protocol of Table 1. A dash marks an unreported result, and means are given only when all eight results are available. Among the baselines with all eight results, the strongest is SFP (Jin et al., 2025) at 44.0 mean mIoU, compared with 66.0 for ActiveSAM.

G Robustness to Image Corruptions

Table 9 reports mIoU under four ImageNet-C corruptions (Hendrycks & Dietterich, 2019) at severity 5. ActiveSAM has the highest mean mIoU under every corruption and leads in 23 of the 24 dataset–corruption pairs; the exception is motion blur on Context60, where it reaches 42.5 mIoU versus 43.1 for SegEarth-OV3. Both CLIP-based methods stay far below the two SAM 3-based methods under every corruption, and under Gaussian noise their mean mIoU falls below 20.

Table 8: **Additional baselines.** mIoU (%) on eight OVSS benchmarks.

Method	Venue	Backbone (Size)	V21	P60	C-O	V20	P59	C-S	City	ADE	Mean
Training-based											
SegCLIP (Luo et al., 2023)	ICML'23	CLIP	52.6	24.7	27.5	-	-	-	-	-	-
CoDe (Wu et al., 2024)	CVPR'24	CLIP	57.7	30.5	32.3	-	-	23.9	28.9	17.7	-
SAM-CLIP (Wang et al., 2024b)	CVPR'24	CLIP + SAM (ViT-B/16)	60.6	29.2	-	-	-	31.5	-	17.1	-
Training-free											
CLIP (Radford et al., 2021)	ICML'21	CLIP	18.6	7.8	6.5	49.1	11.2	7.2	6.7	3.2	13.8
GEM (Bousselham et al., 2024)	CVPR'24	CLIP	46.2	-	-	-	32.6	15.7	-	-	-
CaR (Sun et al., 2024)	CVPR'24	CLIP	48.6	13.6	15.4	73.7	18.4	-	-	5.4	-
LaVG (Kang & Cho, 2024)	ECCV'24	CLIP + DINO (ViT-B/8)	62.1	31.6	34.2	82.5	34.7	23.2	26.2	15.8	38.8
CLIPtrase (Shao et al., 2024)	ECCV'24	CLIP	50.9	29.9	43.6	81.0	33.8	22.8	21.3	16.4	37.5
ClearCLIP (Lan et al., 2024a)	ECCV'24	CLIP	51.8	32.6	33.0	80.9	35.9	23.9	30.0	16.7	38.1
SCLIP (Wang et al., 2024a)	ECCV'24	CLIP	59.1	30.4	30.5	80.4	34.1	22.4	32.2	16.1	38.2
NACLIP (Hajimiri et al., 2025)	WACV'25	CLIP	58.9	32.2	33.2	79.7	35.2	23.3	35.5	17.4	39.4
LPOSS (Stojnić et al., 2025)	CVPR'25	CLIP + DINO (ViT-B/16)	61.1	34.6	33.4	78.8	37.8	25.9	37.3	21.8	41.3
DIH-CLIP (Duan et al., 2025)	ICCV'25	CLIP	64.2	36.0	37.4	84.9	39.7	26.7	40.2	19.6	43.6
SFP (Jin et al., 2025)	ICCV'25	CLIP	63.9	37.2	37.9	84.5	39.9	26.4	41.1	20.8	44.0
ActiveSAM	Ours	SAM 3 (PE-L+/14)	**83.4**	**56.6**	**72.6**	**97.1**	**60.4**	**44.9**	**73.0**	**40.0**	**66.0**

Table 9: **Robustness to image corruptions.** mIoU (%) on clean inputs and four corruptions at severity 5. Means are over the six benchmarks shown. The best result for each condition and benchmark is bold. P60 and C-S denote Context60 and COCO-Stuff.

Condition	Method	V20	V21	City	ADE	P60	C-S	Mean
Clean	CorrCLIP (Zhang et al., 2025)	91.5	76.7	51.1	30.7	44.9	34.0	54.8
	SegEarth-OV3 (Li et al., 2025)	96.83	79.77	69.69	37.55	53.43	42.76	63.34
	RF-CLIP (Li et al., 2026)	87.0	67.2	43.0	21.0	37.9	27.5	47.3
	ActiveSAM (Ours)	**97.08**	**83.42**	**73.04**	**40.03**	**56.59**	**44.90**	**65.84**
Gaussian noise	CorrCLIP (Zhang et al., 2025)	35.66	25.25	8.99	5.51	11.55	9.15	16.02
	SegEarth-OV3 (Li et al., 2025)	77.40	60.25	47.15	22.44	34.94	27.34	44.92
	RF-CLIP (Li et al., 2026)	42.47	27.39	8.86	8.56	11.87	10.85	18.33
	ActiveSAM (Ours)	**84.22**	**66.21**	**47.72**	**24.32**	**38.29**	**30.27**	**48.51**
Motion blur	CorrCLIP (Zhang et al., 2025)	62.64	47.91	37.04	15.69	28.39	19.62	35.21
	SegEarth-OV3 (Li et al., 2025)	88.38	69.49	55.67	28.07	**43.12**	32.53	52.88
	RF-CLIP (Li et al., 2026)	69.21	50.81	28.28	14.37	25.87	18.52	34.51
	ActiveSAM (Ours)	**90.62**	**70.00**	**59.21**	**30.15**	42.50	**34.40**	**54.48**
JPEG compression	CorrCLIP (Zhang et al., 2025)	78.50	62.92	38.68	19.53	36.27	25.52	43.57
	SegEarth-OV3 (Li et al., 2025)	92.30	74.96	58.40	31.55	45.10	35.46	56.29
	RF-CLIP (Li et al., 2026)	71.71	52.39	26.21	15.22	26.70	19.19	35.24
	ActiveSAM (Ours)	**92.48**	**77.13**	**60.76**	**33.44**	**47.52**	**37.82**	**58.19**
Fog	CorrCLIP (Zhang et al., 2025)	77.66	62.13	34.33	22.17	33.99	25.48	42.63
	SegEarth-OV3 (Li et al., 2025)	94.74	79.44	64.18	34.98	50.15	40.63	60.69
	RF-CLIP (Li et al., 2026)	76.55	53.43	25.66	16.55	29.82	22.15	37.36
	ActiveSAM (Ours)	**95.02**	**81.94**	**67.37**	**37.47**	**52.29**	**41.70**	**62.63**

Table 10: **EPG ablation.** Full ActiveSAM (w/ preview) versus decoding every prompt at full resolution (w/o preview), on the full validation sets. FPS is end-to-end throughput on one RTX 5090 GPU, and speedup is the ratio of the two FPS values.

Dataset	Images	mIoU		FPS		Speedup
		w/ preview	w/o preview	w/ preview	w/o preview	
VOC21	1,449	83.42	83.35	9.37	5.09	1.84×
VOC20	1,449	97.08	97.05	9.54	4.40	2.17×
Cityscapes	500	73.04	72.99	3.98	3.19	1.25×
Context59	5,105	60.44	60.51	6.33	2.39	2.65×
Context60	5,105	56.59	56.50	6.12	2.40	2.55×
COCO-Object	5,000	72.63	72.55	7.16	1.81	3.96×
COCO-Stuff	5,000	44.90	45.24	3.79	0.92	4.12×
ADE20K	2,000	40.03	40.05	3.51	1.05	3.34×

H Ablation of Evidence-Proportional Grounding

Table 10 compares ActiveSAM with decoding every prompt at full resolution. EPG changes mIoU by at most 0.34 on any benchmark and by -0.01 on average, while making inference 1.8–4.1× faster on seven benchmarks. Cityscapes gains only $1.25\times$ because most of its prompts are admitted. Averaged over the eight benchmarks, EPG reduces the time per image from 0.521 to 0.184 seconds, and most of this saving comes from the preview.

Class availability after filtering. Table 11 reports the number of prompts admitted by the preview and the percentage of ground-truth image–class pairs that remain available after both EPG stages, that is, pairs whose class can still be predicted. Across the eight benchmarks, 93.8% of these pairs remain available; the fraction is 93.9% for connected components smaller than 1,024 pixels and 90.6% for pairs in the least frequent third of each vocabulary. Retention is lowest on COCO-Stuff (79.8%), yet decoding every prompt improves its mIoU by only 0.34, which suggests that many of the filtered classes were already difficult for the full pipeline.

Table 12 varies only the preview resolution while other parameters are fixed. From 336 to 672 pixels, mean mIoU stays within 0.04 while the time per image grows by up to 30%, so we use 336 pixels, the fastest setting in this range. ActiveSAM outperforms SegEarth-OV3 (63.9 mean mIoU) at every

Table 11: **Prompt admission and class retention.** Mean active and total prompt counts use the preview threshold $\tau = 0.1$. Retention is the percentage of ground-truth image–class pairs that remain available after both EPG stages. The total count prompts, and a class with several names contributes one prompt per name (VOC20: 26 prompts for 20 classes, Cityscapes: 23 for 19).

Dataset	Mean active / total	Retention (%)
VOC21	3.7 / 21	97.4
VOC20	3.1 / 26	99.4
Cityscapes	18.0 / 23	99.0
Context59	12.9 / 59	93.4
Context60	13.9 / 60	93.4
COCO-Object	7.3 / 81	93.9
COCO-Stuff	25.7 / 171	79.8
ADE20K	30.8 / 150	93.6

Table 12: **Preview resolution.** Signatures estimated at a 672-pixel preview are fixed across all rows. mIoU (%) is reported on eight benchmarks. Time is averaged over VOC21, Context60, COCO-Object, VOC20, and Context59, with the ratio to 336 in parentheses. Bold marks the default setting.

Preview (px)	V21	P60	C-O	V20	P59	C-S	City	ADE	Mean	Time (s/img)
224	83.11	56.41	70.66	96.20	60.03	43.58	72.49	39.57	65.26	0.131 (0.97×)
336	83.44	56.57	72.64	97.08	60.45	44.88	72.94	40.24	**66.03**	**0.135** (1.00×)
448	83.37	56.47	72.53	97.32	60.46	44.75	73.34	40.26	66.06	0.145 (1.08×)
560	83.39	56.51	72.62	97.37	60.49	44.81	72.98	40.23	66.05	0.158 (1.17×)
672	83.33	56.52	72.52	97.31	60.89	44.76	73.02	40.17	66.07	0.175 (1.30×)

tested resolution, including 224 pixels (65.26). These results show that the preview step gives us the flexibility in controlling the preference on inference time or segmentation accuracy while showing the stability across different preview resolution.

I Qualitative Comparison

Figures 3–7 compare the segmentation outputs of ActiveSAM and SegEarth-OV3 (Li et al., 2025) across five OVSS benchmarks (GT denotes ground truth). For each example, we provide the input image, ground-truth annotations, and the respective predictions from each method. ActiveSAM consistently produces masks with fewer spurious detections, particularly in complex large-vocabulary scenes and background-heavy regions. For instance, ActiveSAM suppresses false positives in the VOC21 monitor scene and the Context60 computer-monitor scene. These visualizations demonstrate that our proposed method yields more robust semantic grounding in diverse environments.

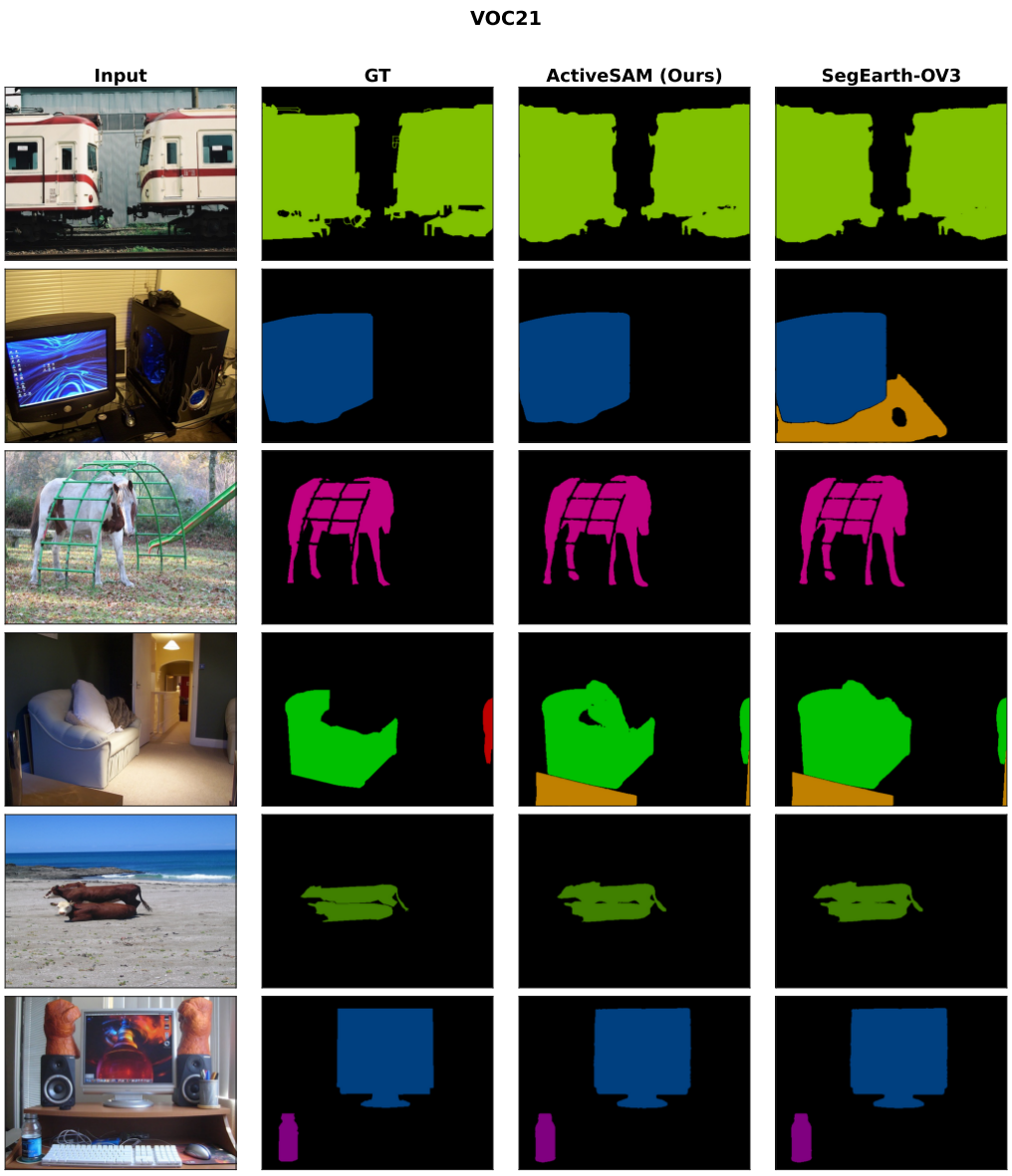

Figure 3: Qualitative results on Pascal VOC21 comparing ActiveSAM (Ours) to SegEarth-OV3 (Li et al., 2025).

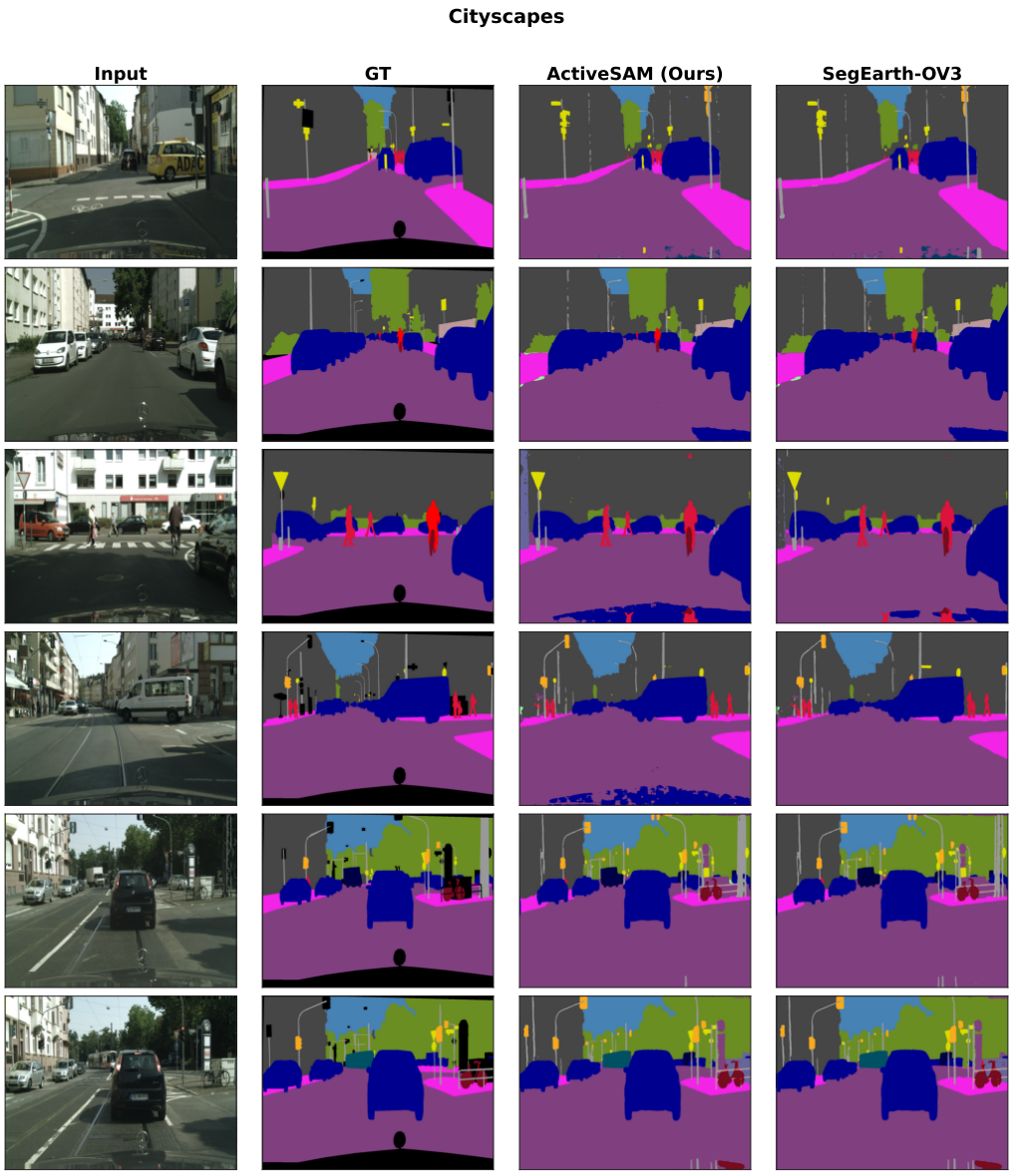

Figure 4: Qualitative results on Cityscapes comparing ActiveSAM (Ours) to SegEarth-OV3 (Li et al., 2025).

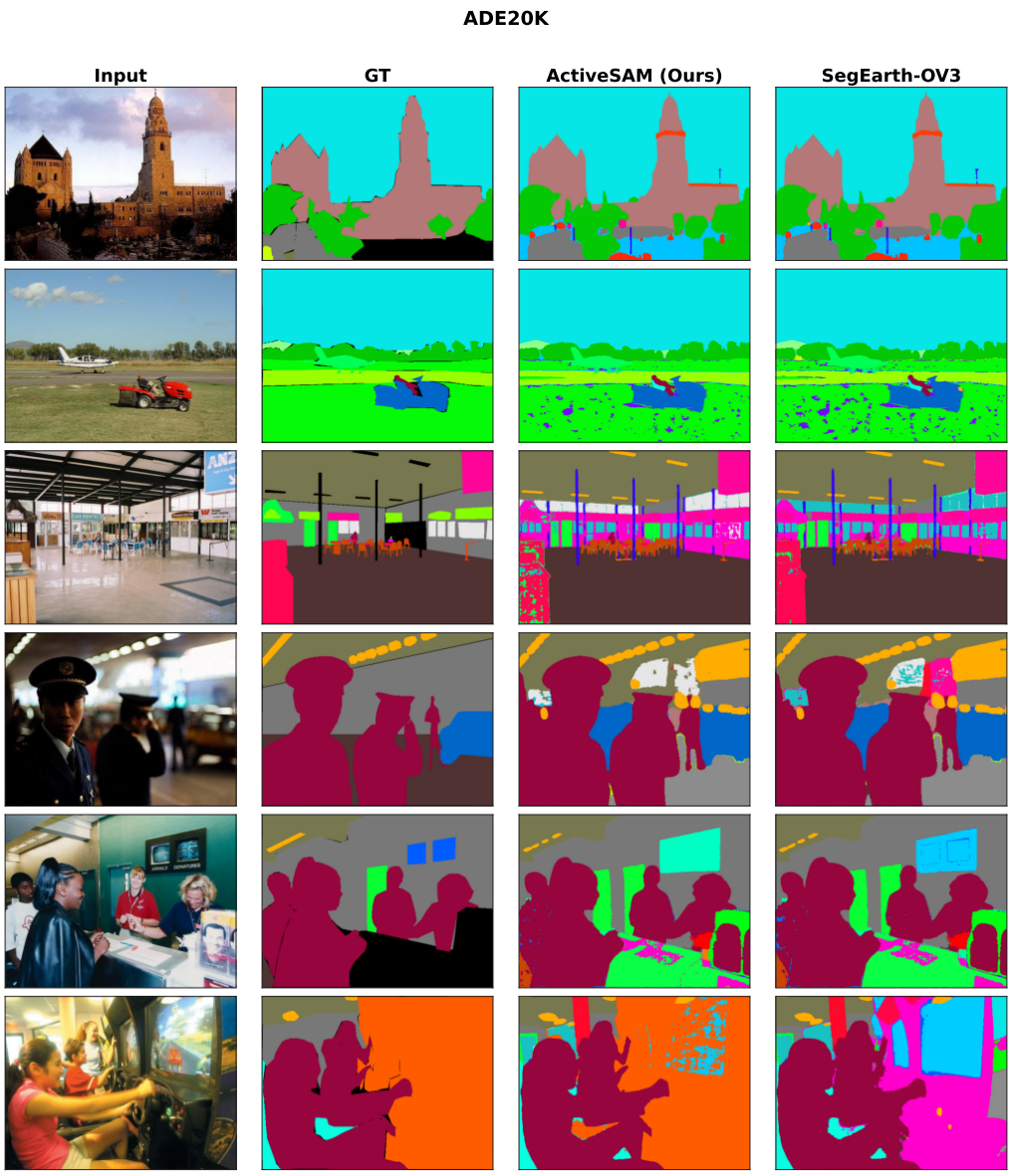

Figure 5: Qualitative results on ADE20K comparing ActiveSAM (Ours) to SegEarth-OV3 (Li et al., 2025).

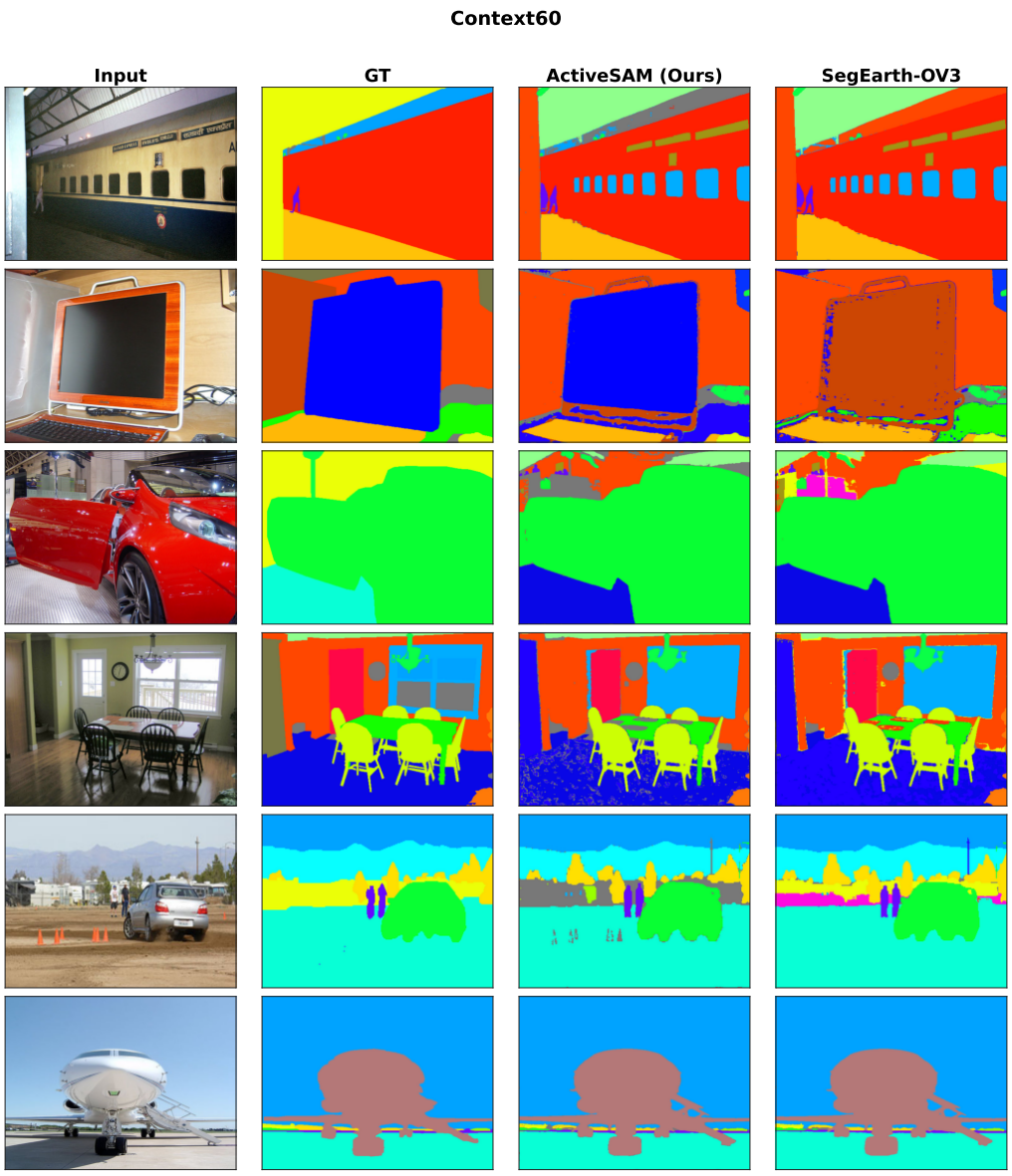

Figure 6: Qualitative results on Context60 comparing ActiveSAM (Ours) to SegEarth-OV3 (Li et al., 2025).

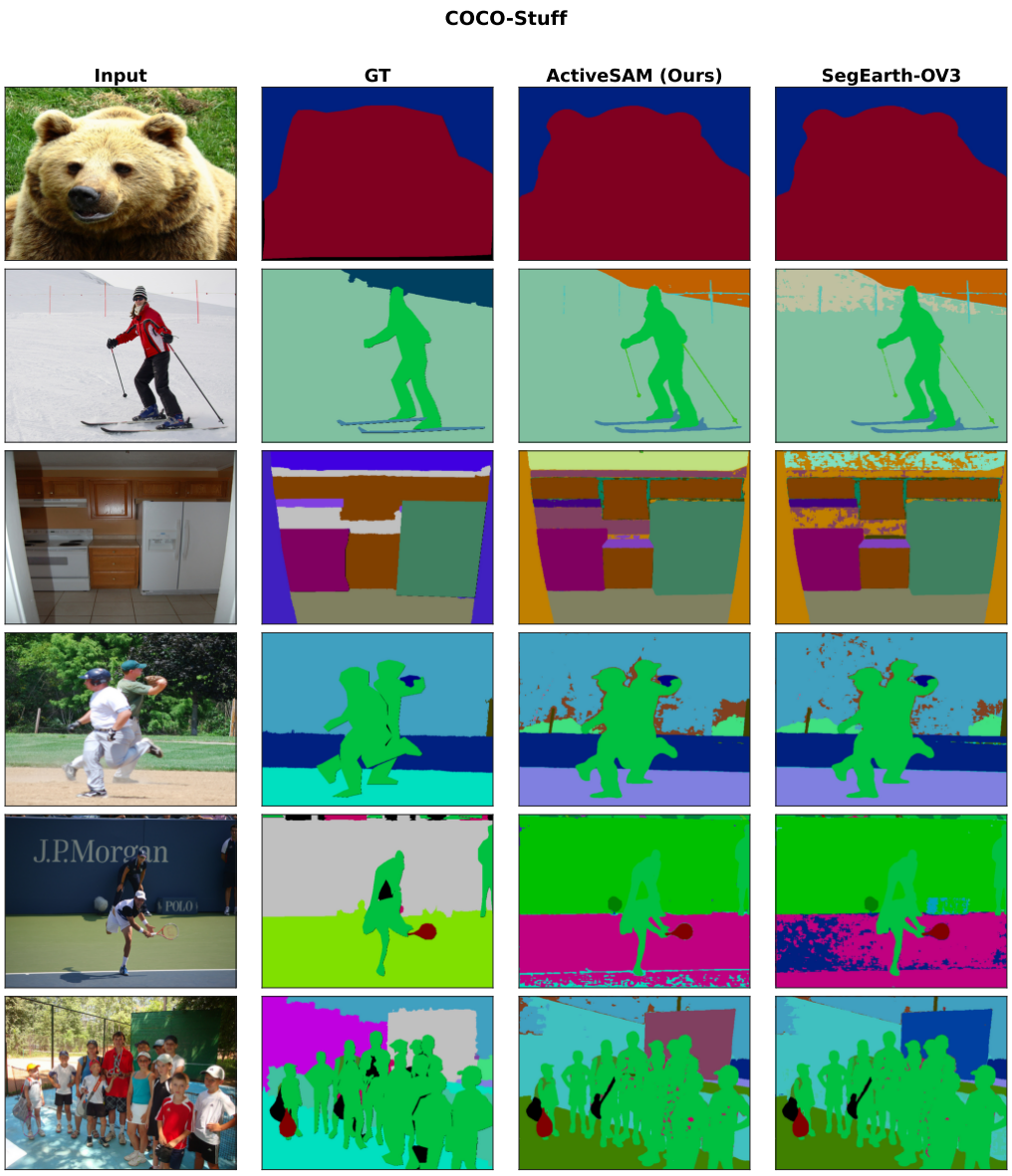

Figure 7: Qualitative results on COCO-Stuff comparing ActiveSAM (Ours) to SegEarth-OV3 (Li et al., 2025).