AWSAM: ONE-SHOT VOCABULARY-FREE SEGMENTATION OF ANYTHING WITH SAM-3

Anonymous authors
Paper under double-blind review

ABSTRACT

Promptable vision foundation models (VFMs) such as SAM-3 can segment targets from text or geometric prompts, but each input creates distinct limitations for domain-specific applications. Text prompts fail for concepts outside the model's learned semantic space, while geometric prompts incorporating points and boxes require specifying the target in every query image. Few-shot semantic segmentation (FSS) addresses this challenge by specifying a novel category at inference from a small set of annotated support examples. Existing FSS methods largely follow one of two routes: (i) converting support examples into geometric prompts by matching support and query patch features in a frozen VFM, which improves generalization but underperforms on fine-grained targets due to coarse patch features; or (ii) learning an explicit concept representation through task-specific training, tens of labeled images, or learned text-token conditioning, which improves performance but limits generalization. Given these limitations, we revisit FSS and SAM-3 prompt conditioning approaches and ask *whether a single annotated example can be used to define a transferable concept directly, without training, a target dataset, or a class name to improve generalization?* In our study, we demonstrate that a frozen SAM-3 can hold such a concept in its prompt space, a shared token sequence into which its text, geometric, and visual-exemplar encoders write. Fitting concept tokens in this space bypasses the text encoder entirely, and thus the language prior that limits generalization. We introduce AwSAM, a vocabulary-free FSS framework that learns a set of concept tokens from a single annotated example. We further study the contribution of external localization cues by pairing the fitted tokens with point prompts derived from a simple closed-form discriminant on VFM features. AwSAM sets a new state-of-the-art in one-shot vocabulary-free FSS across 12 diverse benchmarks, outperforming the strongest prior method by 11.3 mIoU while providing $2.9\times$ faster query inference.

1 INTRODUCTION

Despite success in fine-grained visual understanding, supervised semantic segmentation models remain fundamentally constrained by the concepts (classes) represented in their training data. Extending a model to new concepts requires additional pixel-level annotation and another round of training, and in specialized domains that annotation requires experts. To reduce this dependence on task-specific supervision, recent approaches built on vision foundation models (VFMs) have explored zero-shot, open-vocabulary, and promptable segmentation such as SAM-3(Carion et al., 2026), enabling models to segment previously unseen categories specified at inference time, typically through text prompts or geometric prompts. However, text-prompted segmentation also underperforms on rare or domain-specific concepts underrepresented in the model's learned vocabulary and cross-modal vision-language semantic space.

Few-shot segmentation (FSS) offers another route for specifying novel concepts at inference time, using a small number of annotated examples rather than relying on a predefined class name or textual description. Existing exemplar-based FSS methods largely use these examples for two complementary functions: localizing *where* the target appears and representing *what* the target is. *Feature-matching* methods primarily perform localization by comparing spatial patch features between the support and query images. This enables transfer to novel and out-of-domain targets, but requires the support example for every query. Moreover, coarse patch features can miss fine-grained details and

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.

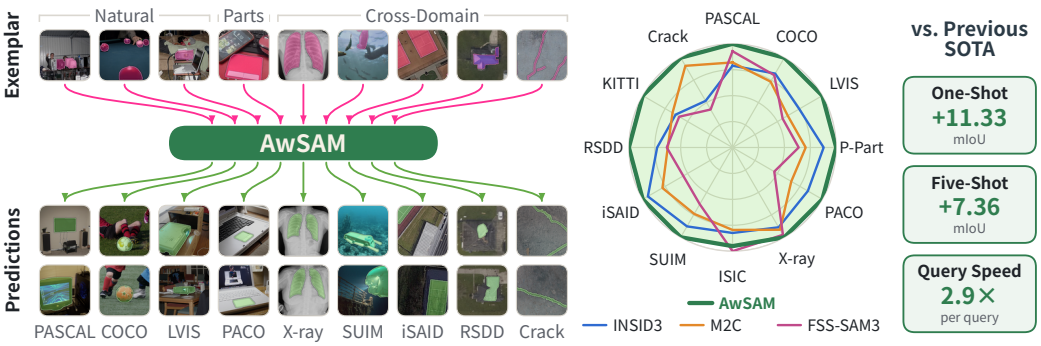

Figure 1: **One exemplar, any concept.** Given a single support Example, AwSAM segments the same concept in new images across natural, part and cross-domain benchmarks. AwSAM outperforms prior SOTA vocabulary-free method across all benchmarks.

provide no explicit representation of the target concept. *Prompt-conditioning* methods instead learn a representation of the target concept that directly conditions a pretrained segmentation model. Once learned, this representation can be reused across query images without requiring the support example at inference. Existing approaches, however, typically require prior training (Sun et al., 2024), tens of labeled images (Piater et al., 2025), or textual description in addition to the exemplar (Zhou et al., 2026). The latter reintroduces a semantic prior from SAM-3's pretrained vocabulary, so the target is no longer defined by the exemplar alone, limiting generalization to concepts outside the pretrained vocabulary. We therefore ask whether a single annotated image can directly define a reusable, transferable concept representation without prior training, additional target-specific data, or language input.

In this study, we show that a frozen SAM-3 can represent a concept in its native prompt space, which is shared across its text, geometric, and exemplar inputs. We introduce AwSAM, a vocabulary-free FSS framework that fits concept tokens in this space from a single annotated example, bypassing the text encoder and its language prior. We further study the contribution of external localization by pairing the fitted tokens with point prompts generated from a closed-form discriminant on frozen C-RADIOv4 (Ranzinger et al., 2026) features. The support therefore specifies *what* to segment, while the external VFM provides a query-specific cue for *where* it appears, without text supervision, a fixed vocabulary, or a learned support-to-prompt mapping. Our main contributions are:

- We introduce AwSAM, a vocabulary-free FSS dual-branch framework that fits concept tokens from a single annotated example directly in the native prompt space of a frozen SAM-3 via a concept branch, bypassing the text encoder. We combine these tokens with point prompts generated by a closed-form discriminant on frozen VFM features via a localization branch, explicitly separating *what* to segment from *where* it appears in the query.
- A controlled study of exemplar, vocabulary, concept prompt, and localization. We evaluate all methods on identical episodes across 12 benchmarks spanning natural objects, object parts, and cross-domain targets. For SAM-3-based methods, we measure the contribution of vocabulary priors for in-distribution and out-of-distribution concepts and perform controlled ablations to isolate the contributions of concept and localization branches.
- We achieve an 11.33-point improvement in one-shot mIoU over the strongest prior method with a single configuration, while providing 2.9× faster per-query inference, and a 7.36-point improvement in five-shot mIoU (Fig. 1).

2 Related Work

Prompting SAM with Feature Matching. Frozen SAM models rely on geometric prompts, such as points or boxes, to specify *where* a mask should be produced, but these prompts do not explicitly encode *what* concept should be segmented. Few-shot methods built on SAM therefore infer the target from support-query feature correspondence and convert the resulting localization into geometric

prompts. PerSAM (Zhang et al., 2024b) builds a support-query similarity map from SAM features and takes its most and least similar locations as positive and negative point prompts. Matcher (Liu et al., 2024) instead matches DINOv2 (Oquab et al., 2024) features before converting the correspondences into point prompts. GF-SAM (Zhang et al., 2024a) groups the matched points with a graph built from the coverage of SAM's masks to obtain more reliable prompts, while PR-MaGIC (Lee et al., 2026) refines the prompts at test time with gradients from SAM's mask decoder. These feature-matching methods rely on estimated coarse correspondences, so matching errors can misplace geometric prompts and cause SAM to segment the surrounding object rather than the intended concept. Other approaches match support and query features without SAM. INSID3 (Cuttano et al., 2026), for example, removes positional bias from frozen DINOv3-L (Siméoni et al., 2025) features through SVD, clusters and matches the debiased features, and refines the prediction with a dense CRF (Krähenbühl & Koltun, 2011). However, its localization remains constrained by the spatial resolution of the underlying patch features. More recently, SAM-3 (Carion et al., 2026) introduced concept prompts, given as short noun phrases or image exemplars, so the model can condition on *what* to segment as well as *where*. FSS-SAM3 (Tsai et al., 2026) exploits this capability by placing the support and query images on a shared canvas and marking the support object with a box prompt. This provides direct concept-aware conditioning, but each added support leaves less resolution for every image on the canvas.

Learned and fitted concept representations. A second line of work conditions the segmentation model on a learned or fitted representation of the target concept rather than on spatial correspondence alone. VRP-SAM (Sun et al., 2024) learns a visual-reference prompt encoder on base classes that converts annotated reference images into prompt embeddings for SAM. SegIC (Meng et al., 2024) trains a mask decoder conditioned on geometric, visual, and meta instructions extracted from a frozen vision foundation model. SANSA (Cuttano et al., 2025) trains lightweight adapters on base classes to make the features of a frozen SAM-2 more semantically aligned for few-shot segmentation. PTSAM (Piater et al., 2025) learns task-specific prompt parameters for a frozen SAM from as few as sixteen annotated training images. M2C (Zhou et al., 2026) instead optimizes a reusable concept embedding inside a frozen SAM-3, starting from the embedding of a text prompt and refining it on a few labeled images. These methods represent *what* to segment more explicitly, but they obtain that representation through training on external labeled data, optimization over several labeled target images, or a language-derived starting point.

3 METHOD

3.1 PRELIMINARIES: SAM-3 WITH CONCEPT

SAM-3 keeps *what* to segment separate from *where* to segment through a shared prompt space. SAM-3 has three encoders: an image encoder reads the image, a text encoder reads the class name, and a geometry encoder reads points and boxes, including the boxes that mark image exemplars. The image and text encoders operate in a 1024-dimensional space, while the geometry encoder produces prompt features from point or box information and the corresponding image features. The outputs of these encoders are projected into a shared 256-dimensional prompt-token space, which SAM-3 passes to the decoder as a single concatenated sequence without any positional information. The rest of the network receives prompt information through this sequence, without explicitly distinguishing the encoder that produced each token providing a common prompt interface through which tokens from different encoders can condition the decoder. Existing FSS methods reach this space through SAM-3's geometric or text encoders. AwSAM instead conditions tokens directly into this shared projected prompt space, letting the support mask define the concept without text or clicks.

3.2 OVERVIEW OF AWSAM

We introduce AwSAM with two complementary branches: a Concept Branch that represents *what* to segment and an optional Localization Branch that identifies *where* the target appears in the query (Fig. 2). Given a single support image $\mathbf{I}_s$ with mask M_s, the Concept Branch obtains a support-derived representation P_s through Visual Prompt Inversion (VPI). VPI keeps SAM-3 frozen and optimizes noise-initialized prompt tokens so that its semantic segmentation head reconstructs the annotated support mask M_s. We decompose $P_s = [P_s^{\text{co}}; P_s^{\text{sc}}]$ into concept and scene tokens, which

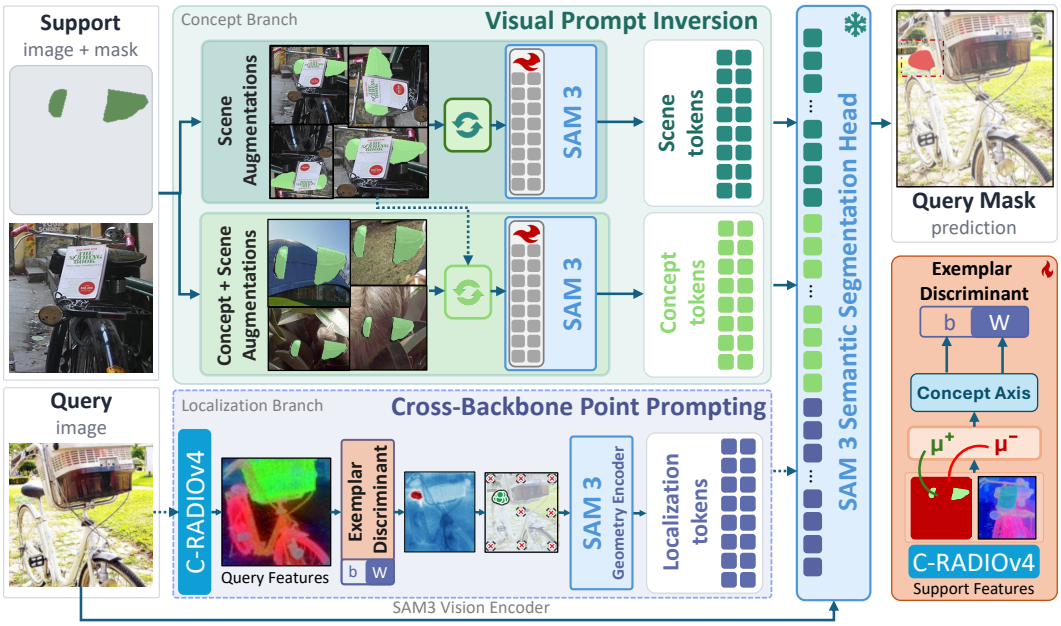

Figure 2: **Overview of AwSAM.** We perform few-shot segmentation via two complementary branches. The *Concept Branch* fits two sets of prompt tokens through frozen SAM 3 to reproduce the support mask: scene tokens on scene augmentations of the support, and concept tokens on scene augmentations plus composites of the object on unrelated background. The *Localization Branch* fits a Exemplar Discriminant on C-RADIOv4 features, applies it to the query, and encodes the resulting points (+/-) with SAM 3's geometry encoder. The three token sets are once concatenated and decoded by SAM 3's for each query image.

are optimized under complementary augmentation regimes. Scene tokens P_s^{sc} are fitted on views that preserve the target within its original scene and therefore retain contextual cues associated with the support concept. Concept tokens P_s^{co} are additionally fitted on views that place the target on unrelated backgrounds, which reduces dependence on scene context and encourages a target concept representation. The Localization Branch complements this representation with query-specific information about *where* the target appears. A closed-form discriminant is fitted to separate the foreground and background C-RADIOv4 features of the support image and then applied to the query image $\mathbf{I}_q$ to identify the likely target and background regions. Positive and negative points sampled from these regions are converted by SAM-3's geometry encoder into localization tokens L_q. SAM-3 finally encodes the query image and predicts its mask using P_s and localization tokens L_q.

3.3 Concept Branch: Visual Prompt Inversion

The Concept Branch learns *what* the support concept is through two complementary augmentation regimes that differ in how they preserve scene context. The scene-view set $\mathcal{V}_{\text{sc}}$ preserves the target within its original surroundings, whereas the concept-view set $\mathcal{V}_{\text{co}}$ extends these views with composites that place the support foreground on unrelated backgrounds. This separation exposes P_s^{sc} to support-specific context while requiring P_s^{co} to remain effective when that context changes. We construct the scene-view set $\mathcal{V}_{\text{sc}}$ by applying the same transformation τ_v to the support image and mask, $(I_v, M_v) = \big(\tau_v(I_s), \tau_v(M_s)\big)$. These views include the full support image, mask-centered crops under four rotations and one horizontal flip, and wide and tight crops around the support mask. To construct a composite view, we extract the masked foreground from the support image and paste it at a random position and scale onto a randomly selected unlabeled ImageNet image, with M_v given by the corresponding pasted mask. We generate four such composite views, denoted $\mathcal{V}_{\text{comp}}$, and define the concept-view set as $\mathcal{V}_{\text{co}} = \mathcal{V}_{\text{sc}} \cup \mathcal{V}_{\text{comp}}$.

Using these two view sets, we obtain the scene and concept tokens by optimizing prompt tokens against the corresponding support-view masks while keeping SAM-3 frozen. For each fit, we ini-

tialize $P \in \mathbb{R}^{16\times256}$ from $\mathcal{N}(0, I)$ and optimize only P, with no additional prompts, so that SAM-3 reproduces the corresponding view mask:

$$\mathcal{L}_{\text{VPI}} = \mathcal{L}_{\text{Dice}}(\hat{M}_v, M_v) + \mathcal{L}_{\text{BCE}}(\hat{M}_v, M_v), \qquad \hat{M}_v = f(I_v; P), \tag{1}$$

where f denotes the frozen SAM-3 segmentation pipeline. We optimize the prompt tokens with Adam optimizer while sampling one view at each step. From the same noise initialization, two optimizations over $\mathcal{V}_{\text{sc}}$ and $\mathcal{V}_{\text{co}}$ produce P_s^{sc} and P_s^{co}, respectively. We concatenate the two optimized sets of token to form $P_s = [\, P_s^{\text{co}}; P_s^{\text{sc}} \,] \in \mathbb{R}^{32\times256}$, and reused across all query images.

3.4 Localization Branch: Cross-Backbone Point Prompting

The Localization Branch identifies *where* the support concept appears in the query by transferring a support-derived *exemplar discriminant* to the query features. Frozen C-RADIOv4 encodes the support and query images, each resized to 1152×1152 with patch size 16, into dense 72×72 feature grids, where $\mathbf{f}_s(u), \mathbf{f}_q(u) \in \mathbb{R}^C$, $C = 1152$, is the feature at grid cell u. We ℓ_2-normalize each cell, so the discriminant depends on feature direction rather than magnitude.

With the normalized support features, we max-pool the support mask onto the support feature grid to split its cells into foreground and background sets, pooling each set across supports when multiple examples are given. The set means $\boldsymbol{\mu}_s^+$ and $\boldsymbol{\mu}_s^-$ define the *concept axis* $\mathbf{a}_s = \boldsymbol{\mu}_s^+ - \boldsymbol{\mu}_s^-$. Centering each set on its own mean gives the within-group covariance $\boldsymbol{\Sigma}_s \in \mathbb{R}^{C\times C}$ over all support features. As in regularized discriminant analysis (Friedman, 1989), we shrink $\boldsymbol{\Sigma}_s$ toward a scaled identity with strength $\lambda = 0.8$. Applying this inverse to the concept axis gives the weight $\mathbf{w}_s$, which down-weights feature directions that vary strongly within either set. The decision threshold b_s completes the exemplar discriminant in Eq. 2, where $\mathbf{I}_C$ is the $C \times C$ identity matrix.

$$\boldsymbol{\Sigma}_{s,\lambda} = (1-\lambda)\boldsymbol{\Sigma}_s + \lambda\frac{\operatorname{tr}(\boldsymbol{\Sigma}_s)}{C}\mathbf{I}_C, \qquad \mathbf{w}_s = \boldsymbol{\Sigma}_{s,\lambda}^{-1}\mathbf{a}_s, \qquad b_s = \frac{1}{2}\mathbf{w}_s^\top\left(\boldsymbol{\mu}_s^+ + \boldsymbol{\mu}_s^-\right) \tag{2}$$

The discriminant is then transfers to the query unchanged, scoring each query feature as $D_q(u) = \mathbf{w}_s^\top \mathbf{f}_q(u) - b_s$. Cells with $D_q(u) > 0$ and $D_q(u) \le 0$ form the positive and negative regions, which localize the support concept on the query grid. Farthest-point sampling then picks up to $K = 8$ spatially distributed points per region from cells at least two grid cells inside its boundary. SAM-3's frozen geometry encoder then maps these labeled query coordinates to up to 16 point tokens and appends its learned CLS token, giving up to 17 query-specific localization tokens L_q, which spatially guide the segmentation head.

3.5 Inference

For query inference, AwSAM combines the Concept and Localization branch outputs, P_s and L_q, which provide complementary information about *what* to segment and *where* it appears in the query. For a query image $\mathbf{I}_q$, we concatenate the two prompt sequences as $[P_s; L_q]$ and provide them along with the query image features from SAM-3's vision encoder to the frozen semantic segmentation head, which predicts the query mask. AwSAM-lite follows the same inference procedure but uses only the concept representation P_s, without localization tokens L_q.

4 Experiments

We evaluate AwSAM for vocabulary-free one-shot and five-shot segmentation across 12 benchmarks, and analyze the contributions of the localization and concept branches and the effect of text prompts on in-distribution and out-of-distribution concepts. Each episode contains a query image and either one or five annotated support image-mask pairs of the same target concept.

Benchmarks and Baselines. The benchmarks span natural images, object parts, and cross-domain scenes. Natural-image benchmarks are COCO-20^i (80 classes, 4 folds), PASCAL-5^i (20, 4), KITTI (19), and LVIS-92^i (920, 10) (Nguyen & Todorovic, 2019; Shaban et al., 2017; Geiger et al., 2012; Gupta et al., 2019). Object-part benchmarks are PASCAL-Part (56 parts, 4 folds) and PACO-Part (303, 4) (Chen et al., 2014; Ramanathan et al., 2023). Cross-domain benchmarks are Chest X-ray

(lung fields), ISIC-2018 (3 lesion classes), SUIM (7 underwater classes), iSAID (15 remote-sensing classes, 3 folds), RSDD (7 roof-defect classes), and OmniCrack, which contains 16 crack datasets treated as separate single-class folds (Candemir et al., 2014; Codella et al., 2019; Islam et al., 2020; Waqas Zamir et al., 2019; Hamburger & Kijewski-Correa, 2025; Benz & Rodehorst, 2024). We compare with PerSAM, PerSAM-F, Matcher, GF-SAM, PR-MaGIC, INSID3, M2C, and FSS-SAM3 using their released code and published benchmark-specific settings.

Evaluation Metric and Protocol. We report mean intersection-over-union (mIoU), accumulating intersections and unions over episodes for each class before averaging across classes and folds. All methods are evaluated on the same fixed episodes and random seed and receive only the support image(s) and mask(s), without class names or text descriptions. For text-capable baselines, M2C uses SAM-3's default "visual" prompt and FSS-SAM3 is evaluated without text, avoiding additional supervision unavailable to the other methods. Text-conditioned comparisons, including AwSAM, are reported separately in Sec.4.3. Additional details are available in supplementary.

Implementation Details. We use frozen SAM-3 at 1008×1008 as the segmenter and frozen C-RADIO-v4 (SO400M) at 1152×1152 as the localization backbone. The scene and concept branches each optimize 16 noise-initialized tokens for 100 Adam steps with learning rate 0.05. For localization, we set $\lambda = 0.8$ and sample eight positive and eight negative point prompts. The same configuration is used across all benchmarks without dataset-specific tuning. All experiments run on a single NVIDIA L40S GPU.

4.1 Main Results

One- and five-shot results. AwSAM is the only method that leads consistently across domains, at both one and five shots (Table 1). At one shot, it outperforms INSID3, the strongest prior, on all twelve benchmarks by 11.33 mIoU on average and beats the best prior of each benchmark on eleven, trailing only FSS-SAM3 on ISIC. The priors instead split by regime, so the best prior changes across benchmarks (INSID3 on five, FSS-SAM3 on three, GF-SAM and M2C on two each). The SAM-3-based FSS-SAM3 is best on PASCAL-5^i but reaches only 20.63 on PACO-Part, whereas the feature-matching INSID3 leads on both part benchmarks while trailing FSS-SAM3 by 12.09 on PASCAL-5^i and M2C by 12.92 on Crack. AwSAM, which pairs a fitted concept with feature-based localization, leads in both regimes. At one shot, its margin widens with target concept complexity on natural images, from 6.22 on PASCAL-5^i to 17.16 on KITTI, and is smallest on large, centered targets such as X-ray (2.51). Even without localization, AwSAM-lite exceeds the strongest prior by 8.48 mIoU. With five shots, M2C gains the most (+15.26) and becomes the strongest prior, yet AwSAM still outperforms it on all twelve benchmarks by 7.36 mIoU, leads the best prior on eleven, and cuts its ISIC deficit from 3.32 to 0.91. One-shot AwSAM already exceeds every five-shot prior except M2C, which it trails by only 0.93 mIoU.

The accuracy gains of AwSAM come at a lower amortized cost than the strongest prior at both shot counts. Because a reference is enrolled once and reused across queries, Table 2 reports the amortized cost per query over 25 queries, query $+$ reference$/25$. At one shot, AwSAM answers each query $2.9\times$ faster than INSID3 ($1.2\times$ amortized), and at five shots it enrolls a reference $5.8\times$ faster than M2C ($4.1\times$ amortized).

Parts, and Cross-domain. Part-level benchmarks evaluate fine-grained, non-salient targets within larger objects, exposing a limitation of SAM-based FSS methods that rely on sparse point prompts to specify the target. FSS-SAM3 reaches only 20.63 on PACO-Part, while localization-based INSID3 performs substantially better by identifying targets from feature clusters. AwSAM further improves over INSID3 by 6.94 on PASCAL-Part and 7.15 on PACO-Part, while also outperforming *AwSAM-Lite*, which uses only the concept branch. These gains highlight the value of combining concept branch with localization branch to identify the target part within its surrounding object.

The six cross-domain benchmarks reveal three regimes. For large, centered targets (Chest X-ray, ISIC), support and query images are highly similar, leaving limited room for improvement; the added localization cue can even be detrimental. FSS-SAM3 leads on ISIC, the only benchmark where AwSAM trails. Notebly, FSS-SAM3 outperforms the other prior methods only on the three benchmarks whose support and query images are most similar (Chest X-ray, ISIC and PASCAL-5^i),

		Natural				*Part*		*Cross-domain*						
	Method	PASCAL 5^i	COCO 20^i	LVIS 92^i	KITTI	PASCAL	PACO	X-ray	ISIC	SUIM	iSAID	RSDD	Crack	Mean
One-shot	PerSAM	49.21	22.02	11.65	16.18	31.64	20.35	32.83	22.01	28.55	20.34	18.16	2.94	22.99
	PerSAM-F	47.43	22.66	12.43	14.05	30.63	20.17	43.47	22.20	28.25	20.56	20.04	2.35	23.69
	PR-MaGIC	65.39	44.80	24.05	19.86	42.82	30.34	76.42	39.90	40.69	33.62	28.44	9.33	37.97
	Matcher	67.90	51.85	32.65	19.53	45.14	33.47	72.70	38.88	39.27	33.14	24.65	6.89	38.84
	GF-SAM	71.81	58.49	35.05	27.99	46.12	36.01	54.47	48.87	55.20	48.54	25.51	12.70	43.40
	FSS-SAM3	80.57	55.53	31.25	30.74	36.48	20.63	87.90	**68.08**	39.29	30.20	29.63	14.00	43.69
	M2C	71.21	50.59	34.43	34.58	40.33	29.05	82.82	54.13	45.98	43.00	29.19	30.11	45.45
	INSID3	68.48	56.74	41.62	32.81	50.23	37.25	80.36	56.12	54.62	51.95	33.97	17.19	48.45
	AwSAM-lite	87.97	66.12	47.59	49.51	50.67	36.33	89.40	66.11	61.21	51.47	44.70	32.04	56.93
	AwSAM	**86.79**	**68.95**	**56.03**	**51.74**	**57.17**	**44.40**	**90.42**	64.76	**62.18**	**55.08**	**46.72**	**33.10**	**59.78**
	Δ *vs. best prior*	**+6.22**	**+10.45**	**+14.41**	**+17.16**	**+6.94**	**+7.15**	**+2.51**	**-3.32**	**+6.98**	**+3.12**	**+12.75**	**+2.99**	**+7.28**
	Δ *vs. INSID3*	**+18.31**	**+12.20**	**+14.41**	**+18.94**	**+6.94**	**+7.15**	**+10.05**	**+8.64**	**+7.56**	**+3.12**	**+12.75**	**+15.92**	**+11.33**
Five-shot	GF-SAM	82.29	67.00	44.39	37.70	53.85	41.02	53.11	59.03	61.93	54.91	25.50	14.48	49.60
	FSS-SAM3	87.06	68.99	47.07	42.63	56.32	38.37	71.96	**75.06**	56.93	45.20	37.30	20.89	53.98
	INSID3	80.04	64.66	45.89	41.24	58.92	44.86	78.90	66.12	63.22	58.33	35.74	18.06	54.66
	M2C	83.35	65.59	54.94	59.37	61.35	46.31	87.97	70.60	60.80	57.94	47.30	33.03	60.71
	AwSAM-lite	91.87	73.96	60.43	60.39	60.63	47.67	91.09	76.32	70.35	63.09	51.86	38.41	65.51
	AwSAM	**90.45**	**75.67**	**65.71**	**64.51**	**65.38**	**54.16**	**91.62**	74.15	**72.77**	**67.23**	**55.39**	**39.85**	**68.07**
	Δ *vs. best prior*	**+3.39**	**+6.68**	**+10.77**	**+5.14**	**+4.03**	**+7.86**	**+3.65**	**-0.91**	**+9.55**	**+8.90**	**+8.10**	**+6.82**	**+6.16**
	Δ *vs. M2C*	**+7.10**	**+10.08**	**+10.77**	**+5.14**	**+4.03**	**+7.86**	**+3.65**	**+3.55**	**+11.97**	**+9.29**	**+8.10**	**+6.82**	**+7.36**

Table 1: **One-shot and five-shot mean IoU (%, ↑)** across twelve benchmarks; **bold** best and underline second best, with AwSAM-lite shown but not ranked.

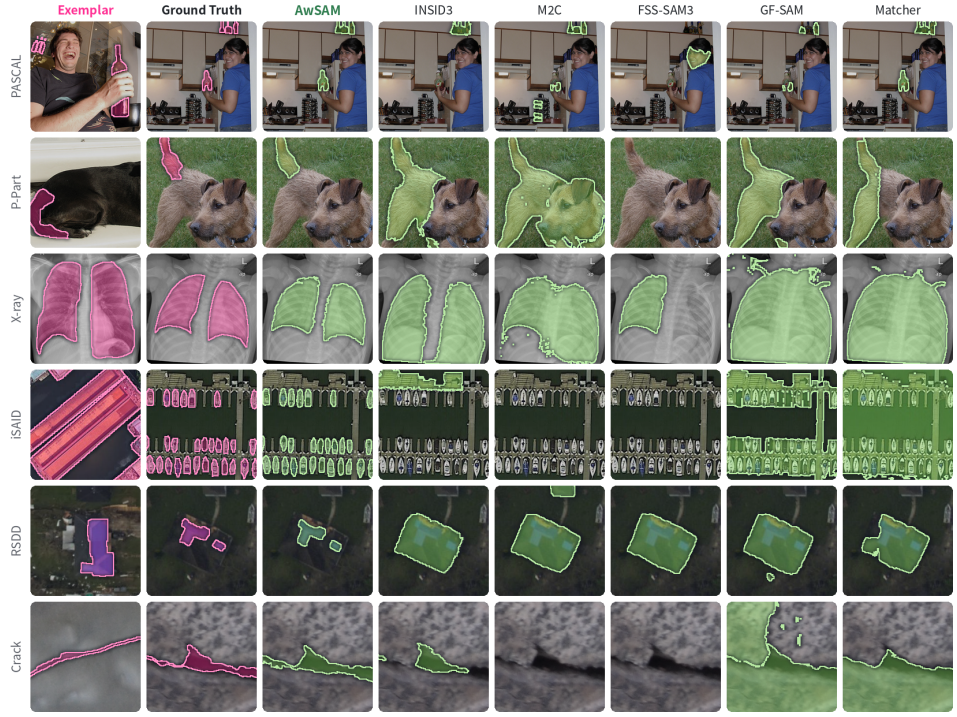

Figure 3: **Qualitative one-shot results.** Exemplar and ground truth in pink, predictions in green. AwSAM isolates the annotated target, including parts and thin structures, where the priors struggle.

consistent with its shared canvas, which lets SAM 3 compare the two images directly. However, AwSAM still improves over INSID3 by 8.64 points. For scene-level targets, AwSAM gains 6.98 on SUIM and 3.12 on iSAID over the best prior method. For fine-grained or small targets, AwSAM leads by 12.75 on RSDD, where both SAM-based and localization-based methods struggle. On

		GF-SAM	FSS-SAM3	M2C	INSID3	AwSAM-lite	Δ_{prev}	AwSAM	Δ_{prev}
One-shot	reference, once (ms) ↓	–	–	11,235	**52**	5,278	101.0× more	5,456	104.3× more
	per query (ms) ↓	471	163	86	417	**76**	5.5× faster	143	2.9× faster
	amortised (ms) ↓	471	**163**	535	419	287	1.5× faster	361	1.2× faster
	mIoU ↑	43.40	43.69	45.45	48.45	56.93	**+8.48**	**59.78**	**+11.33**
Five-shot	reference, once (ms) ↓	–	–	45,838	**321**	7,248	6.3× less	7,913	5.8× less
	per query (ms) ↓	698	321	87	424	**76**	1.1× faster	147	1.7× more
	amortised (ms) ↓	698	**321**	1,920	436	366	5.2× faster	463	4.1× faster
	mIoU ↑	49.60	53.98	60.71	54.66	65.51	**+4.79**	**68.07**	**+7.36**

Table 2: **Cost and accuracy at one and five shots**; **bold** best and underline second best per row, and Δ_{prev} relative to the strongest prior at that shot count (INSID3 at one shot, M2C at five).

		Natural				*Part*		*Cross-domain*						
	Method	PASCAL 5^i	COCO 20^i	LVIS 92^i	KITTI	PASCAL	PACO	X-ray	ISIC	SUIM	iSAID	RSDD	Crack	Mean
One-shot	**AwSAM**	86.79	68.95	56.03	51.74	57.17	44.40	90.42	64.76	62.18	55.08	46.72	33.10	**59.78**
	Only Points	72.05	51.32	39.05	32.61	49.35	34.28	84.88	38.11	53.05	49.23	36.16	12.80	46.07
	Only Concept	87.97	66.12	47.59	49.51	50.67	36.33	89.40	66.11	61.21	51.47	44.70	32.04	56.93
	Δ_{points}	**-1.18**	**+2.83**	**+8.44**	**+2.23**	**+6.50**	**+8.07**	**+1.02**	**-1.35**	**+0.96**	**+3.61**	**+2.02**	**+1.07**	**+2.85**
	$\Delta_{concept}$	**+14.74**	**+17.63**	**+16.98**	**+19.13**	**+7.82**	**+10.12**	**+5.53**	**+26.66**	**+9.12**	**+5.85**	**+10.56**	**+20.31**	**+13.70**
Five-shot	**AwSAM**	90.45	75.67	65.71	64.51	65.38	54.16	91.62	74.15	72.77	67.23	55.39	39.85	**68.07**
	Only points	84.11	64.04	44.79	40.49	55.48	40.55	86.50	41.84	65.91	59.03	40.71	16.95	53.37
	Only Concept	91.87	73.96	60.43	60.39	60.63	47.67	91.09	76.32	70.35	63.09	51.86	38.41	65.51
	Δ_{points}	**-1.41**	**+1.71**	**+5.28**	**+4.12**	**+4.75**	**+6.50**	**+0.53**	**-2.17**	**+2.42**	**+4.14**	**+3.53**	**+1.44**	**+2.57**
	$\Delta_{concept}$	**+6.34**	**+11.63**	**+20.92**	**+24.02**	**+9.90**	**+13.61**	**+5.12**	**+32.31**	**+6.86**	**+8.20**	**+14.68**	**+22.90**	**+14.71**

Table 3: **Complementary contributions of the concept and localization branches** in one- and five-shot segmentation.

OmniCrack, localization methods such as INSID3 are limited by coarse patch-level features when segmenting thin cracks, highlighting the benefit of combining concept representation with query-specific localization.

Qualitative results. Figure 3 compares one-shot predictions across methods. On PASCAL, the small bottle in a cluttered kitchen is missed or only partially recovered by the priors, while AwSAM captures its full extent. Rows 2 and 5 show the part-level challenge: when the exemplar marks a dog's tail or a tarp covering part of a damaged roof, the priors segment the entire object or extended target region, whereas AwSAM isolates the annotated region. Other cross-domain examples reveal a similar limitation: on Chest X-ray, the priors either recover only one lung field or leak into the ribcage. On iSAID, most priors fail or segment the dock instead of the individual boats. On the thin crack, insufficient reliable correspondences produce fragmented or empty predictions, while AwSAM follows the target across its full extent. Across all six examples, AwSAM recovers the target, consistent with its fitted prompt encoding the concept itself rather than relying on sparse coordinates or feature similarity.

4.2 Contributions of the Concept and Localization Branches

Table 3 isolates the two branches of AwSAM. In both one-shot and five-shot setting, removing the concept branch reduces mean IoU by 13.70 and 14.71 points, whereas removing query-specific localization reduces it by 2.85 and 2.57 points. Thus, visual prompt inversion provides the primary support-defined representation of *what* to segment, while the localization branch supplies complementary query-specific evidence of *where* it appears. The concept branch improves every benchmark, while localization improves 10 of 12 benchmarks except PASCAL-5^i and ISIC, where concept-only prompting is slightly stronger.

4.3 Class Name vs. Exemplar

We evaluate FSS-SAM3, M2C, and AwSAM under controlled text access to distinguish exemplar-driven performance from gains inherited from SAM-3's in-distribution vocabulary (Table 4). Since all three methods decode through SAM-3, their performance may partly reflect its learned vocab-

		Name	*Exemplar*					*Exemplar + name*				
	Benchmark	SAM-3	FSS-SAM3	M2C	AwSAM-lite	AwSAM	vs. M2C	FSS-SAM3	M2C	AwSAM-lite	AwSAM	vs. M2C
Known vocabulary	PASCAL-5^i	90.92	80.57	71.21	**87.97**	86.79	**+15.58**	89.11	85.65	**90.23**	89.39	**+3.73**
	COCO-20^i	78.81	55.53	50.59	66.12	**68.95**	**+18.35**	74.31	71.44	**74.47**	74.15	**+2.70**
	LVIS-92^i	70.80	31.25	34.43	47.59	**56.03**	**+21.60**	52.35	57.68	59.58	**61.67**	**+3.99**
	Mean	80.17	55.78	52.08	67.22	**70.59**	**+18.51**	71.92	71.59	74.76	**75.07**	**+3.47**
	$\Delta_{\text{SAM-3}}$	0.00	-24.39	-28.10	-12.95	-9.59		-8.25	-8.58	-5.41	-5.11	
Unknown vocabulary	P-Part	26.76	36.48	40.33	50.67	**57.17**	**+16.84**	6.51	46.13	50.11	**57.02**	**+10.89**
	PACO	35.77	20.63	29.05	36.33	**44.40**	**+15.35**	11.48	40.44	38.73	**45.90**	**+5.46**
	RSDD	29.61	29.63	29.19	44.70	**46.72**	**+17.53**	26.82	34.92	45.54	**47.29**	**+12.37**
	Mean	30.71	28.91	32.86	43.90	**49.43**	**+16.57**	14.94	40.50	44.79	**50.07**	**+9.57**
	$\Delta_{\text{SAM-3}}$	0.00	-1.80	+2.14	+13.18	+18.72		-15.78	+9.78	+14.08	+19.36	
Mean (all six)		55.44	42.35	42.47	55.56	**60.01**	**+17.54**	43.43	56.05	59.78	**62.57**	**+6.52**

Table 4: **Class name vs. exemplar at one shot (mIoU, %, ↑)**; **bold** best and underline second best per row within each prompt group, $\Delta_{\text{SAM-3}}$ relative to SAM-3 given the name alone, and *vs. M2C* is AwSAM minus M2C under the same prompt.

ulary. We therefore evaluate each method with the exemplar alone and with the exemplar plus the class name, and compare both against name-only SAM-3. We group benchmarks as known or unknown vocabulary by name-only SAM 3 accuracy and whether class names are common object categories. Two findings emerge. First, AwSAM extracts substantially more information from a single exemplar regardless of vocabulary coverage: without the class name, it leads M2C by 18.51 mIoU on known benchmarks and 16.57 on unknown ones. On the unknown benchmarks at one shot, AwSAM also exceeds name-only SAM-3 (30.71 mIoU) by 18.72, whereas M2C and FSS-SAM3 remain within 2.2 mIoU of that baseline. Second, prior methods benefit much more from SAM-3's vocabulary. Adding the class name improves M2C by 19.51 mIoU on known benchmarks, reducing AwSAM's lead to 3.47, but by only 7.64 on unknown benchmarks, where AwSAM remains 9.57 ahead; FSS-SAM3 instead drops by 13.97. In contrast, AwSAM changes by only 4.48 mIoU on known and 0.64 on unknown benchmarks. These results indicate that AwSAM derives the target concept primarily from the visual exemplar and can therefore exploit frozen SAM-3 even when the concept is not reliably accessible through its text vocabulary. The supplementary reports the full comparison across all twelve benchmarks and at the class name level.

5 Conclusion

We introduced AwSAM, a vocabulary-free few-shot segmentation (FSS) framework that learns transferable concept tokens from one or more annotated support examples and segments the corresponding concept in unseen images without class names, text descriptions, or task-specific training. AwSAM optimizes a concept token in the shared prompt space of frozen SAM-3 to encode *what* to segment via a concept branch, while an optional localization branch transfers a foreground-background discriminant from frozen C-RADIOv4 support features to the query to indicate *where* the target appears. We show that combining both branches enables better generalization across domains, with the concept branch driving most of the accuracy and the localization branch providing complementary gains, particularly for object parts and fine-grained, non-salient targets. We further investigate whether each method learns the concept from the exemplar or borrows it from SAM-3's vocabulary, by scoring it with and without the class name. The class name closes most of the gap between prior methods and AwSAM on concepts SAM-3 already knows, but not on concepts outside its vocabulary, where, at one shot, prior methods given only the exemplar stay within 2.2 mIoU of SAM-3 given the name alone, and AwSAM surpasses it by 18.72. Across twelve natural-image, part-segmentation and cross-domain benchmarks, AwSAM uses a single configuration and improves over the strongest vocabulary-free FSS prior by 11.33 and 7.36 mIoU at one and five shots, respectively. AwSAM ranks first on 11 of the 12 benchmarks at both one and five shots and outperforms the strongest prior method on every benchmark, despite variation in the best-performing baseline.

REFERENCES

Christian Benz and Volker Rodehorst. Omni-Crack30k: A benchmark for crack segmentation and the reasonable effectiveness of transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 3876–3886, 2024. doi: 10.1109/cvprw63382.2024.00392.

Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P. Musco, Rahul K. Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J. McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Transactions on Medical Imaging*, 33(2):577–590, 2014. doi: 10.1109/tmi.2013.2290491.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, 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 Rädle, 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 Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=r35clVtGzw.

Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1979–1986, 2014. doi: 10.1109/cvpr.2014.254.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC). *arXiv preprint arXiv:1902.03368*, 2019.

Claudia Cuttano, Gabriele Trivigno, Giuseppe Averta, and Carlo Masone. SANSA: Unleashing the hidden semantics in SAM2 for few-shot segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 131857–131891, 2025. doi: 10.52202/085713-3969.

Claudia Cuttano, Gabriele Trivigno, Christoph Reich, Daniel Cremers, Carlo Masone, and Stefan Roth. INSID3: Training-free in-context segmentation with DINOv3. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21638–21648, 2026.

Jerome H Friedman. Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175, 1989.

A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3354–3361, 2012. doi: 10.1109/cvpr.2012.6248074.

Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5351–5359, 2019. doi: 10.1109/cvpr.2019.00550.

Rachel Hamburger and Tracy Kijewski-Correa. Computer vision-enabled roof subassembly damage detection from hurricanes using aerial reconnaissance imagery. *Natural Hazards Review*, 26(1): 04024057, 2025. doi: 10.1061/NHREFO.NHENG-2278. URL https://ascelibrary.org/doi/abs/10.1061/NHREFO.NHENG-2278.

Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic segmentation of underwater imagery: Dataset and benchmark. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1769–1776, 2020. doi: 10.1109/iros45743.2020.9340821.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 24, 2011.

Minjae Lee, Sungwoo Hur, Soojin Hwang, and Won Hwa Kim. PR-MaGIC: Prompt refinement via mask decoder gradient flow for in-context segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21659–21668, 2026.

Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=yzRXdhk2he.

Lingchen Meng, Shiyi Lan, Hengduo Li, Jose M. Alvarez, Zuxuan Wu, and Yu-Gang Jiang. SegIC: Unleashing the emergent correspondence for in-context segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 203–220, 2024. doi: 10.1007/978-3-031-72920-1_12.

Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 622–631, 2019. doi: 10.1109/iccv.2019.00071.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Tristan Piater, Björn Barz, and Alexander Freytag. Prompt-tuning SAM: From generalist to specialist with only 2,048 parameters and 16 training images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4688–4698, 2025. doi: 10.1109/cvprw67362.2025.00455.

Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, Amir Mousavi, Yiwen Song, Abhimanyu Dubey, and Dhruv Mahajan. PACO: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7141–7151, 2023. doi: 10.1109/cvpr52729.2023.00690.

Mike Ranzinger, Greg Heinrich, Collin McCarthy, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. C-RADIOv4 (tech report). *arXiv preprint arXiv:2601.17237*, 2026.

Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017. doi: 10.5244/c.31.167.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Yanpeng Sun, Jiahui Chen, Shan Zhang, Xinyu Zhang, Qiang Chen, Gang Zhang, Errui Ding, Jingdong Wang, and Zechao Li. VRP-SAM: SAM with visual reference prompt. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23565–23574, 2024. doi: 10.1109/cvpr52733.2024.02224.

Yi-Jen Tsai, Yen-Yu Lin, and Chien-Yao Wang. Few-shot semantic segmentation meets SAM3. *arXiv preprint arXiv:2604.05433*, 2026.

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

Anqi Zhang, Guangyu Gao, Jianbo Jiao, Chi Harold Liu, and Yunchao Wei. Bridge the points: Graph-based few-shot segment anything semantically. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a. URL https://openreview.net/forum?id=jYypS5VIPj.

Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Yu Qiao, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. In *International Conference on Learning Representations (ICLR)*, 2024b. URL https://openreview.net/forum?id=6Gzkhoc6YS.

Quan Zhou, Shaoqing Zhai, Qiang Hu, Jia Chen, Qiang Li, and Zhiwei Wang. Mask to Concept: Auto-Promptable SAM3 via Efficient Test-Time Concept Embedding Search for Few-Shot Annotation . In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2026*, volume LNCS 16879. Springer Nature Switzerland, September 2026.