Training-Free Few-Shot Segmentation via Vision-Language Guided Prompting

Euihyun Yoon, Taejin Park and Jaekoo Lee
College of Computer Science, Kookmin University
Seoul, Korea

Abstract

Object segmentation relies heavily on costly pixel-level annotations and struggles to generalize to unseen domains. The recent introduction of the Segment Anything Model (SAM), a foundation model for segmentation, offers a prompt-driven, zero-shot capability that has been applied in various domains (e.g., autonomous driving, satellite imagery, medical imaging) and extended to Few-Shot Segmentation (FSS) tasks. However, existing SAM-based FSS methods typically generate prompts by using a vision encoder to measure support–query image similarity, which often biases towards the support images and fails when there are significant support–query context shifts. To address this limitation, we propose a training-free FSS approach that combines visual and textual cues to generate effective prompts for the target class. By leveraging both vision and language information, our approach bridges the support–query gap and guides SAM to segment novel objects more reliably. Without any additional training, our method outperforms previous state-of-the-art FSS methods on established benchmarks ($COCO$-20^i, $Pascal$-5^i), demonstrating its effectiveness and robust generalization. Our code is publicly available on GitHub.

1. Introduction

Segmentation is foundational to modern computer vision, powering applications as diverse as autonomous driving [24], satellite remote sensing [36], and computer-aided diagnosis [1]. Although today's segmentation methods achieve near-perfect accuracy on their source distribution, they deteriorate sharply when confronted with out-of-domain imagery or novel object categories, limiting their real-world utility [4].

As a vision foundation model, the Segment Anything Model (SAM) [15] approaches the generalization challenge from a prompt-driven, category-agnostic perspective. Trained on more than one billion masks, SAM [15] converts coarse geometric prompts (points, boxes, or polygons) into high-quality instance masks without relying on a fixed label set. This category-agnostic approach confers impressive domain transfer [15], but it also strips the model of semantic focus: SAM [15] predicts "something", not "the thing of interest". Recent analyses confirm that SAM's internal representations exhibit only limited class separability, a weakness that becomes acute in few-shot semantic segmentation (FSS) [38], where the goal is to segment all instances of a novel class after seeing only a handful of labeled exemplars.

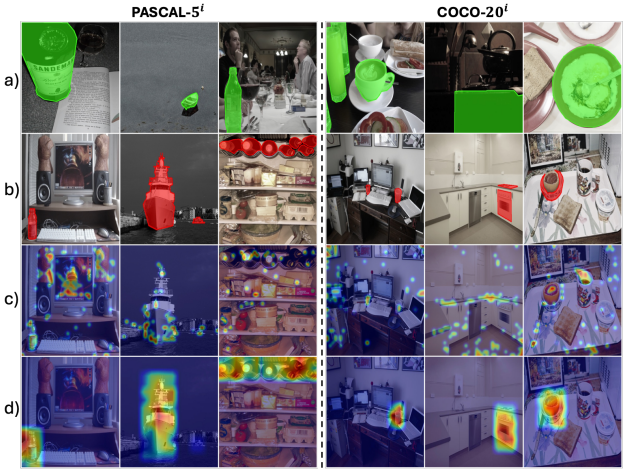

Figure 1. Illustration of the support-query context shift problem in few-shot segmentation. Qualitative comparison of attention maps highlighting segmentation result differences between prior FSS methods and our vision-language guided approach. a) Support images with ground-truth masks, b) Query images with ground-truth masks, c) Well-known Matcher [20] illustrating problematic segmentation results due to these shifts, and d) Ours.

To avoid retraining, several methods [20, 38] generate SAM [15] prompts by matching low-level visual features between support masks and a query image. Matcher [20], for example, transfers class identity through feature similarity, partially overcoming SAM [15]'s semantic shortfall. However, as shown in Figure 1 c), shifts in appearance or scene context between the support and query images cause purely visual matching to latch onto background clutter or visually similar distractors, resulting in misplaced prompts.

Other methods use a CLIP [27] encoder that fuses

class names with image features before feeding the embeddings into SAM's decoder [32]. Although they enhance cross-category generalization, their reliance on the limited support masks induces bias and, ultimately, yields only modest performance gains.

Our work poses the following question: Can we impart semantic guidance to SAM while leaving every parameter frozen?

We answer in the affirmative by introducing a three-stage, training-free pipeline that composes two frozen foundation models. a) *Text-Conditioned Region Proposal (TCRP)*: given a class name, OWL-ViT [21] predicts candidate bounding boxes that likely enclose objects of that class. b) *Vision–Language Alignment Selection (VLAS)*: for each candidate, we compute the similarity between its visual embedding and the text embedding, retaining only boxes with strong semantic alignment; these serve as class-aware prompts for SAM [15]. c) *Semantic Mask Refinement (SMR)*: SAM [15] segments each retained box, after which we prune over-segmented or off-class pixels by re-evaluating mask embeddings against the text embedding, thereby removing residual distractors.

We validate our approach on COCO-20^i [22] and Pascal-5^i [29] FSS benchmarks. In publicly available FSS benchmarks, our approach surpasses alternative SOTA FSS approaches.

We summarize our main contributions as follows:

- **Training-free prompting for SAM.** We present a novel pipeline that enables SAM to focus on "the thing of interest" by pairing two frozen foundation models.
- **Vision-language guided mask selection and refinement for FSS.** To mitigate support–query context shifts in few-shot segmentation, we introduce an efficient scheme that jointly filters and refines visual and textual features.
- **Competitive FSS performance.** Our approach clearly outperforms the other SOTA FSS approaches with negligible computational overhead.

2. Related Work

2.1. Segmentation Foundation Models

The Segment Anything Model (SAM) [15] has recently emerged as a powerful foundation model, trained on over one billion segmentation masks. By generating high-quality masks from simple prompts (e.g., points or bounding boxes), SAM [15] has demonstrated strong zero-shot generalization across various segmentation tasks [15].

Early work leveraging SAM [15] includes image editing, such as Edit Everything [34], which utilizes SAM-generated masks for object-level text-driven editing. Similarly, Track Anything Model (TAM) [35] employs SAM [15] to track selected objects across video frames. Segment-to-Cluster (S2C) [16] further uses SAM-generated segments for weakly supervised learning via contrastive training.

However, these methods typically require explicit class labels or user interactions to specify the object of interest, highlighting a key limitation—SAM's inherent category agnosticism. To address this, recent approaches [20, 38] apply SAM [15] within FSS frameworks, aiming to segment novel objects given minimal visual examples.

2.2. Few-Shot Segmentation (FSS)

FSS aims to generalize segmentation models to novel classes using only a few annotated examples. Classical FSS methods [17, 25] are often meta-trained on base classes, but their generalization to unseen classes can be limited due to base-class bias. For instance, BAM [17] employs a base-class classifier to explicitly suppress base-class predictions during inference, helping to segment novel classes effectively. HDMNet [25] utilizes hierarchical self-attention modules for fine-grained matching between support and query images, significantly enhancing segmentation resolution. However, both approaches require task-specific retraining and remain vulnerable to base-class bias.

To overcome retraining and enhance generalization, recent work integrates SAM [15] into training-free FSS frameworks [20, 38]. PerSAM [38] proposes minimalistic prompting with one positive and one negative point derived from support-query similarity. While effective, this approach struggles to segment complex objects or scenes involving multiple instances. Matcher [20] extends this idea by generating multiple point prompts via comprehensive bidirectional feature matching, improving mask completeness and accuracy. Despite these advances, such vision-only approaches heavily rely on visual similarity, limiting their robustness when encountering significant appearance or contextual variations.

To mitigate this, we propose an approach that goes beyond classical FSS by incorporating language cues alongside visual information, thereby providing stronger semantic guidance and enhancing robustness in few-shot scenarios without additional training.

2.3. Language-Guided Visual Understanding

The emergence of vision-language models, notably CLIP [27], has enabled significant advances across diverse computer vision tasks, including segmentation [39], retrieval [18], and detection [21]. CLIP aligns images and textual descriptions in a shared embedding space, enabling powerful zero-shot capabilities. However, as CLIP [27] relies on global image-caption matching, it often struggles with spatially precise tasks such as detection and segmentation [19].

Recent adaptations of CLIP [27] address these localization issues [2, 21]. OWL-ViT [21] integrates open-

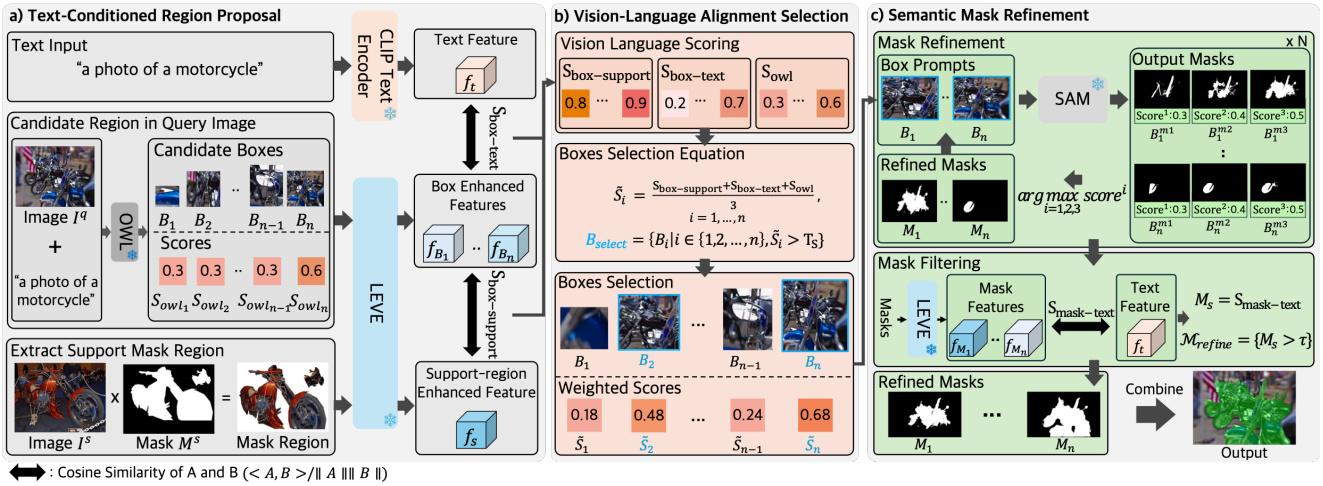

Figure 2. Overview of the proposed approach. The approach consists of three stages —*TCRP*, *VLAS*, and *SMR*—which operate entirely training-free by leveraging publicly available foundation models.

vocabulary detection capabilities by fine-tuning vision transformers on text-guided bounding box prediction tasks. SC-CLIP [2] improves zero-shot segmentation by extracting localized features from CLIP's intermediate transformer layers, demonstrating significantly enhanced spatial precision.

Inspired by these advances, we propose a novel FSS approach combining OWL-ViT [21] and CLIP [27]. Specifically, OWL-ViT [21] provides text-conditioned region proposals, which are refined through CLIP's localized embeddings. This integration of semantic textual guidance and robust visual matching facilitates accurate segmentation even under challenging conditions, maintaining the training-free advantage and significantly boosting generalization in few-shot segmentation.

3. Method

3.1. Overview

Few-shot semantic segmentation (FSS) aims to segment a query image given only a few annotated support examples of a novel class. Formally, we have a support set $S = \{(I_i^s, M_i^s)\}_{i=1}^K$ of K images $I_i^s \in \mathbb{R}^{H\times W\times 3}$ with corresponding binary masks $M_i^s \in \mathbb{R}^{H\times W}$ marking the target class. The goal is to predict the mask M^q for a query image I^q of the same class. During episodic training, both support and query masks are available for supervision, whereas at test time the model must infer the query mask using only the K support pairs (i.e. without seeing M^q).

Existing FSS approaches [20, 25, 38] follow a prototype matching paradigm: they extract the feature representation of the support foreground (using M^s to mask out the support image) and then compare it to query image features to produce a segmentation mask. While effective, this strategy can be overly biased to the specific support example. In practice, if there is a significant appearance or context shift between the support and query images, the support prototype may not align well with the query features, leading to poor segmentation results – a phenomenon often referred to as the inter-image gap. This limitation has been noted in prior work [3], where large feature discrepancies between support and query hinder effective matching and yield inaccurate segmentation priors.

To address the above issue, we propose a training-free approach that leverages both visual and textual cues for more robust query segmentation. Recent studies [20, 38] have begun exploring the use of vision-language models and promptable segmenters to improve FSS generalization.For example, the SAM [15] can segment arbitrary objects given appropriate prompts, and PerSAM [38] demonstrated a one-shot personalization of SAM [15] without additional training. Inspired by such advances, our approach introduces textual information (the class name) alongside visual features, enabling the model to localize the target object in the query image based on high-level semantic cues. By forgoing any fine-tuning and instead harnessing powerful pre-trained vision-language models, our approach remains training-free and adaptable to any new class out of the box.

Figure 2 illustrates an overview of our approach, which consists of three key modules. (a) *Text-Conditioned Region Proposal (TCRP)* (Section 3.2) uses OWL-ViT [21] with the given class name as a text prompt to identify candidate regions in the query image that are likely to contain the target object. This can be achieved by computing image–text alignment scores over the query image (e.g., via a CLIP-based model) to generate class-specific region proposals. We then extract visual features from these candidate regions for further stages.

(b) *Vision–Language Alignment Selection (VLAS)* (Section 3.3) takes the set of candidate region features and the

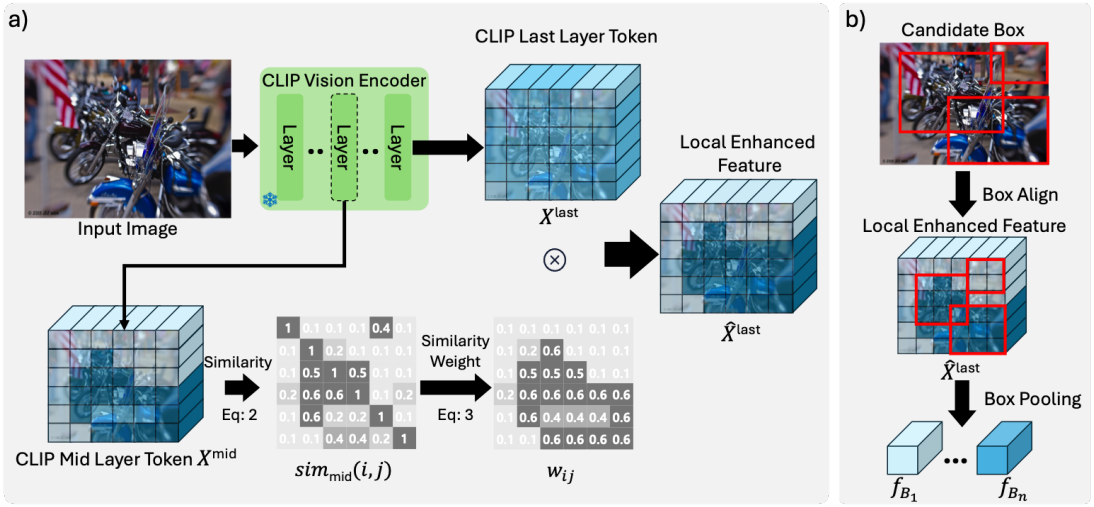

Figure 3. Overview of LEVE operation. a) semantic local feature enhancement leveraging semantic similarities computed from mid-layer tokens. b) Region feature aggregation through box aligning and pooling.

text embedding of the class name, and computes similarity scores to determine which region(s) best match the target class. In essence, *VLAS* performs a semantic alignment between each proposal's visual feature and the class description, selecting the top-scoring boxes that are most likely to correspond to the object of interest.

(c) *Semantic Mask Refinement (SMR)* (Section 3.4) then leverages the SAM [15] to obtain precise segmentation masks for the selected region proposals. Each chosen box is fed to SAM [15] as a bounding box prompt, which yields a corresponding object mask. We refine these masks (for example, by removing any small stray regions or merging multiple masks if necessary) to produce the final predicted mask M^q for the query.

3.2. Text-Conditioned Region Proposal

Text-Conditioned Region Proposal (TCRP) (Figure 2 a)) aims to propose candidate regions in the query image I^q that are likely to contain the target class (given its name), by enhancing the visual features of these regions in a text-guided manner. We first obtain a text feature f_t by feeding the prompt "a photo of a {class}" into a pretrained CLIP text encoder [27]. This yields an embedding that represents the target class in the joint vision-language space.

For visual region proposals, we leverage a pretrained open-vocabulary object detection model, OWL-ViT [21]. Given the query image I^q and the same text prompt "a photo of a {class}", OWL-ViT predicts a set of n bounding boxes that likely contain an instance of the target class, along with confidence scores for each. Formally, we obtain:

$$(B_i,\ S_{owl_i})_{i=1}^{n} = \text{OWL}(I^q,\ \text{"a photo of a \{class\}"}) \quad (1)$$

where B_i is the i-th predicted box and S_{owl_i} is its confidence score.

To improve local feature representations, we introduce a *Local Enhanced Vision Encoder (LEVE)* built on a pretrained CLIP vision encoder. The key idea is to leverage CLIP's intermediate features, which carry rich semantic information, to enhance its final-layer patch embeddings for better localization. CLIP's original vision encoder tends to focus on global image features (using a [CLS] token for the whole image) at the expense of local details. This can weaken the representation of fine-grained regions (e.g., individual objects) in the feature map. To address this, *LEVE* uses CLIP's mid-layer patch features to guide and refine the last-layer patch features, thereby strengthening local region representations in the query image.

As shown in Figure 3, given a query image I^q, we first feed it into the CLIP vision encoder. Let $X^{\text{last}} \in \mathbb{R}^{P\times D}$ denote the matrix of patch token features from CLIP's final layer, where $P = G \times G$ is the total number of image patches (for a $G \times G$ patch grid) and D is the feature dimension. We also extract the patch token features from an intermediate layer of the CLIP vision encoder, denoted $X^{\text{mid}} \in \mathbb{R}^{P\times D}$. Prior work [2] has observed that these mid-layer features retain strong local semantic signals, making them well-suited for enhancing patch-level representations. Next, we compute a patch-wise similarity matrix using the intermediate features X^{mid}. For each pair of patches i and j (where $i, j \in \{1, \ldots, P\}$), we define the semantic similarity $sim_{\text{mid}}(i,j)$ as the cosine similarity between their mid-layer features:

$$sim_{\text{mid}}(i,j) = \langle X_i^{\text{mid}},\ X_j^{\text{mid}} \rangle / \|X_i^{\text{mid}}\|\,\|X_j^{\text{mid}}\| \quad (2)$$

where $\langle\ \rangle$ denotes inner product, $\|\ \|$ denotes L2 norm. This measure $sim_{\text{mid}}(i,j)$ captures the semantic affinity between patch i and patch j using CLIP's mid-level representations. If two patches belong to the same object or region, their

mid-layer features will be highly similar, yielding a large $sim_{\text{mid}}(i,j)$. In this way, the similarity matrix highlights groups of patches that are semantically related (likely part of the same object or context), guiding the model to focus on consistent regions.

We then use the mid-layer similarity matrix to reweight and aggregate the final-layer features. For each patch i, we compute a set of weights $\{w_{ij}\}_{j=1}^{P}$ that emphasize patches similar to i (and de-emphasize unrelated patches). Specifically, we normalize the similarities for patch i across all $j \in \{1,\dots,P\}$ to obtain weights:

$$w_{ij} = \frac{sim_{\text{mid}}(i,j)}{\sum_{k=1}^{P} sim_{\text{mid}}(i,k)} \tag{3}$$

These weights w_{ij} form a soft affinity distribution over all patches relative to patch i, grounded in mid-level semantics. Finally, we derive the local-enhanced feature for patch $i \in \{1,\dots,P\}$ by a weighted sum of all final-layer patch features, using w_{ij} as importance scores:

$$\hat{X}_i^{\text{last}} = \sum_{j=1}^{P} w_{ij}\, X_j^{\text{last}} \tag{4}$$

Here $\hat{X}_i^{\text{last}}$ is the enriched feature for patch i, which integrates information from other patches that are semantically similar to i (according to the mid-layer features). This operation enhances the representation of patch i by infusing context from its related patches, effectively reinforcing the features of local object regions.

Given the local-enhanced patch features $\hat{X}^{\text{last}}$, we obtain region-level features for the query image's candidate boxes. We align the candidate boxes like RoIAlign [13] and aggregate the enhanced patch features falling inside each box. For the n-th predicted box B_n, let $\mathcal{P}_n$ be the set of patch indices that lie within that box. We compute the box-enhanced feature f_{B_n} as follows:

$$f_{B_n} = \{\sum_{i\in\mathcal{P}_n}\Big(\sum_{j\in\mathcal{P}_n} w_{ij}\Big)\hat{X}_i^{\text{last}}\}/\{\sum_{i\in\mathcal{P}_n}\sum_{j\in\mathcal{P}_n} w_{ij}\} \tag{5}$$

f_{B_n} represents the feature of region B_n by pooling the patch features inside that region.

We apply the same *LEVE* procedure to the support image's mask region, producing a support-region feature f_S. These semantically and locally enhanced features (f_{B_n} and f_S) provide improved semantic alignment between query proposals and the support example, facilitating more accurate matching in subsequent stages.

3.3. Vision-Language Alignment Selection

While OWL-ViT [21] provides candidate bounding boxes conditioned on text prompts in the *TCRP* stage, these boxes may include false positives. To mitigate this, we introduce a *Vision-Language Alignment Selection (VLAS)*, illustrated in Figure 2 b), which filters candidate boxes by evaluating their semantic alignment with both visual and textual cues.

Specifically, we first calculate the visual similarity ($S_{\text{box-support}}$) between each candidate box feature f_{B_i} and the support-region enhanced feature f_S. This similarity measures the visual alignment of each predicted region with the provided support region. Additionally, we compute textual similarity ($S_{\text{box-text}}$) between each box feature f_{B_i} and the text feature f_t extracted from the class-name prompt, capturing their general semantic correspondence.

Combining these visual and textual similarities with OWL-ViT's original box class confidence S_{owl}, we derive a unified alignment score $\tilde{S}_i$ for each candidate box as follows:

$$\tilde{S}_i = (S_{\text{box–support}} + S_{\text{box–text}} + S_{\text{owl}})/3 \tag{6}$$

We also compute a selection threshold T_S based on the mean and deviation of all $\tilde{S}_i$:

$$T_S = \frac{1}{n}\sum_{i=1}^{n}\tilde{S}_i \;+\; \alpha\sqrt{\frac{1}{n}\sum_{i=1}^{n}\Big(\tilde{S}_i - \frac{1}{n}\sum_{j=1}^{n}\tilde{S}_j\Big)^2} \tag{7}$$

We retain only those candidate boxes whose $\tilde{S}_i$ exceed T_S. This strategy effectively integrates both visual and textual modalities, ensuring that selected boxes accurately correspond to the target class.

Algorithm 1 *Semantic Mask Refinement (SMR)*

Input: selected box B_i, query image I^q, SAM predictor $\mathcal{S}$, CLIP vision encoder Φ, text feature f_t, threshold τ, refined mask set $\mathcal{M}_{\text{refine}}$
Output: final mask m^*

Step 1: Mask refinement
$\big((B_i^{m^1}, \text{Score}^1, \ell^1),(B_i^{m^2}, \text{Score}^2, \ell^2),(B_i^{m^3}, \text{Score}^3, \ell^3)\big) \leftarrow \mathcal{S}.\texttt{predict}(B_i)$
repeat
  $k_{\max} \leftarrow \arg\max_{k\in\{1,2,3\}} \text{Score}^k$
  $\big((B_i^{m^1}, \text{Score}^1, \ell^1),(B_i^{m^2}, \text{Score}^2, \ell^2),(B_i^{m^3}, \text{Score}^3, \ell^3)\big) \leftarrow \mathcal{S}.\texttt{predict}(B_i, \ell^{k_{\max}})$
until N *times*

Step 2: Mask filtering
$k_{\max} \leftarrow \arg\max_{k\in\{1,2,3\}} \text{Score}^k,\ M_i \leftarrow B_i^{m^{k_{\max}}}$
$f_{M_i} \leftarrow \text{LEVE}(M_i)\ \ M_{s,i} \leftarrow \langle f_{M_i}, f_t\rangle/(\|f_{M_i}\|\,\|f_t\|)$
if $M_{s,i} > \tau$ **then**
  $\mathcal{M}_{\text{refine}} \leftarrow \mathcal{M}_{\text{refine}} \cup \{M_i\}$
end
$m^* \leftarrow \bigcup_{M\in\mathcal{M}_{\text{refine}}} M$
return m^*

3.4. Semantic Mask Refinement

SAM [15] is highly sensitive to prompts due to its limited semantic understanding, making prompt quality critical for accurate segmentation [10]. To address the prompt sensitivity issues, we explicitly propose a *Semantic Mask Refinement (SMR)* (Figure 2 c)) that refines and specializes the

Table 1. Comparison of 1-shot and 5-shot Results on $COCO$-20^i [22]. * denotes training-free methods.

Method	Backbone	Params (MACs / FLOPs)	FPS	1-shot						5-shot					
				Fold0	Fold1	Fold2	Fold3	**mIoU**	**FB-IoU**	Fold0	Fold1	Fold2	Fold3	**mIoU**	**FB-IoU**
Mask FSS Method															
BAM [17]	ResNet50 [12]	52M (497.3G / 996.0G)	20.97	39.4	49.9	46.2	45.2	45.2	71.1	43.2	53.4	49.4	48.1	48.5	73.3
HDMNet [25]	ResNet50 [12]	50.9M (595.5G / 1.2T)	13.01	43.8	55.3	51.6	49.4	50.0	72.2	50.6	61.3	55.7	56.0	56.0	77.7
PerSAM* [38]	ViT-H [8]	0 (3.0T / 6.0T)	1.94	22.2	24.0	18.6	23.1	21.9	–	31.8	31.4	26.8	29.3	29.8	–
Matcher* [20]	DINOv2 [23] & ViT-H [8]	0 (11.5T / 23.0T)	0.24	52.7	53.5	52.6	52.1	52.7	73.6	52.7	53.5	52.6	52.1	52.7	77.0
Class-aware Mask FSS Method															
PI-CLIP* [33]	ViT-B/16 [8]	0 (595.5G / 959.9G)	4.35	49.3	65.7	55.8	56.3	56.8	–	56.4	66.2	55.9	58.0	59.1	–
PGMA-Net [6]	ResNet50 [12]	2.6M (67.85G / 135.7G)	9.74	49.9	56.7	55.8	54.7	54.3	75.8	49.5	61.7	59.1	57.9	57.1	76.7
PGMA-Net [6]	ResNet101 [12]	2.7M (126.46G / 252.9G)	7.95	55.2	62.7	60.3	59.4	59.4	78.5	55.9	65.9	63.4	61.9	61.8	79.4
Beyond-Mask [5]	ResNet50 [12]	7.6M (88.26G / 177.15G)	12.31	52.6	59.8	57.6	56.8	56.7	76.8	52.3	62.4	60.8	57.0	58.1	77.8
Beyond-Mask [5]	ViT-B/16 [8]	8.6M (26.53G / 53.16G)	5.30	51.2	61.8	58.0	55.6	56.7	77.0	53.1	62.4	59.2	56.8	57.9	77.8
Ours*	ViT-L/14 [8] & ViT-H [8]	0 (3.6T / 7.2T)	1.10	**68.2**	**71.4**	**70.8**	**71.6**	**70.5**	**83.6**	**68.1**	**72.0**	**70.9**	**72.6**	**70.9**	**83.9**

Table 2. Comparison of 1-shot and 5-shot Results on $Pascal$-5^i [29]. * denotes training-free methods.

Method	Backbone	Params (MACs / FLOPs)	FPS	1-shot						5-shot					
				Fold0	Fold1	Fold2	Fold3	mIoU	FB-IoU	Fold0	Fold1	Fold2	Fold3	mIoU	FB-IoU
Mask FSS Method															
BAM [17]	ResNet50 [12]	52M (272.7G / 546.2G)	21.07	69.0	73.6	67.6	61.1	67.8	79.7	70.6	75.1	70.8	67.2	70.9	82.2
HDMNet [25]	ResNet50 [12]	50.9M (334.0G / 671.3G)	13.73	71.0	75.4	68.9	62.1	69.4	–	71.3	76.2	71.3	68.5	71.8	–
PerSAM* [38]	ViT-H [8]	0 (4.5T / 9.0T)	1.95	45.4	51.1	43.8	38.5	44.7	–	52.6	57.7	51.9	46.5	52.2	–
Matcher* [20]	DINOv2 [23] & ViT-H [8]	0 (11.5T / 23.0T)	0.27	67.6	70.3	73.5	67.5	69.8	–	72.9	79.5	74.2	76.0	75.6	–
Class-aware Mask FSS Method															
PI-CLIP* [33]	ViT-B/16 [8]	0 (478.1G / 959.7G)	4.41	76.4	83.5	74.7	72.8	76.8	87.6	76.7	83.8	75.2	73.2	77.2	–
PGMA-Net [6]	ResNet50 [12]	2.6M (67.7G / 135.4G)	9.81	73.4	80.8	70.5	71.7	74.1	83.5	74.0	81.5	71.9	73.3	75.2	84.2
PGMA-Net [6]	ResNet101 [12]	2.7M (126.3G / 252.7G)	8.09	76.8	82.3	75.7	75.7	77.6	86.2	77.7	82.7	76.9	77.0	78.6	86.9
Beyond-Mask [5]	ResNet50 [8]	7.6M (88.3G / 177.2G)	12.42	76.2	80.4	68.0	76.9	75.4	83.9	76.3	80.8	68.9	77.8	76.0	84.2
Beyond-Mask [5]	ViT-B/16 [8]	8.6M (26.5G / 53.2G)	5.38	75.0	79.6	74.7	76.4	76.4	85.3	75.5	79.9	75.9	77.5	77.2	86.0
Ours*	ViT-L/14 [8] & ViT-H [8]	0 (3.0T / 6.0T)	1.15	**78.9**	**87.5**	**83.7**	**85.8**	**84.0**	**91.0**	**78.9**	**88.4**	**84.1**	**85.8**	**84.3**	**91.3**

prompts obtained from previous stages, precisely aligning them with SAM [15] to enable more accurate and semantically coherent segmentation masks.

The *SMR* step is summarized in Algorithm 1. First, for each candidate box produced by the previous *VLAS* stage, we feed the box prompt into SAM [15] to obtain an initial mask for that region. By default, SAM [15] returns up to three mask candidates per prompt, each with an associated predicted IoU score. We select the mask with the highest IoU score and use its predicted mask logits as a new prompt to SAM [15], obtaining a refined mask. This iterative refinement strategy is a common and effective way to quickly improve the initial mask's outline using SAM's own predictions [15, 38]. Next, we incorporate text information to filter the refined masks. We extract a mask feature f_{M_i} for each mask M_i using the *LEVE*, and we take the text feature f_t computed in the earlier *TCRP* stage for the target class. For each mask, we compute the cosine similarity $M_{s,i}$ between its mask feature and the text feature:

$$M_{s,i} = \langle f_{M_i},\, f_t \rangle / \|f_{M_i}\|\,\|f_t\| \tag{8}$$

where i indexes the i-th mask generated from a selected box prompt. We then filter the masks by selecting only those whose similarity exceeds a threshold τ:

$$\mathcal{M}_{\text{refine}} = \left\{ M_i \;\middle|\; i \in \{1,\ldots,n\},\; M_{s,i} > \tau \right\} \tag{9}$$

meaning that only masks sufficiently relevant to the text description (target class) are retained. Finally, we simply merge all the selected masks to produce the final output mask. This *SMR* refines the initially generated mask prompts to better align with SAM [15], ensuring that only the target object's mask is obtained.

4. Experiments

4.1. Implementation and Evaluation Details

Datasets. To verify the effectiveness and efficiency of the proposed approach, we evaluate ours and SOTA alternatives on publicly available FSS benchmarks, $Pascal$-5^i [29] and $COCO$-20^i [22]. Following previous studies [5, 6, 20, 38], we adopt the established splits into base and novel classes, conducting evaluations across all four folds for each dataset to ensure fair comparisons.

Evaluation metrics. In experiments, we report results using well-known FSS metrics including mean Intersection-over-Union (mIoU = $\frac{1}{C}\sum_{c=1}^{C}\frac{|P_c \cap G_c|}{|P_c \cup G_c|}$) [9], Foreground-Background IoU (FB-IoU = $\frac{\text{IoU}_{\text{Fg}}+\text{IoU}_{\text{Bg}}}{2}$) [28]. We evaluate mIoU and FB-IoU separately for each fold and their averages across all folds, while FPS is averaged over all test images.

We quantify efficiency using learnable parameters (Params), frames per second (FPS), multiply–accumulate

Table 3. Cross-domain evaluation on LVIS [11], WHU [14], Kvasir [26], and SBU [31]. * denotes training-free methods.

Method	Backbone	LVIS	WHU	Kvasir	SBU
		mIoU / FPS			
PerSAM* [38]	ViT-H	11.7 / 1.62	17.44 / 1.57	18.49 / 1.60	19.69 / 1.57
Matcher* [20]	DINOv2 & ViT-H	33.0 / 0.44	28.82 / 0.33	27.73 / 0.40	21.35 / 0.24
PI-CLIP* [33]	ViT-B/16	5.6 / 16.08	15.68 / 17.05	15.95 / 21.0	20.0 / 20.76
Ours*	ViT-L/14 & ViT-H	**46.0** / 0.80	**31.38** / 0.57	**40.15** / 0.85	**26.27** / 0.81

operations (MACs), and floating-point operations (FLOPs).

Implementation details. All experiments were conducted on a single NVIDIA 4090 GPU. For open-vocabulary detection component in our proposed approach, we use OWL-ViT [21] (ViT-L/14) pretrained on the Objects365 [30] and WebLI [7]. Visual and textual features were extracted using CLIP [27] (ViT-L/14) pretrained on web scale image–text pairs. Finally, masks were generated using SAM [15] (ViT-H) pretrained on SA-1B.

In our approach, we set the hyperparameter α in Eq. 7, used for computing the mean and standard deviation in the *VLAS* stage, to 0.3. Additionally, the threshold τ in Eq. 9 employed in the *SMR* stage was set to 0.2. In LEVE, we set the mid-layer feature extraction to the 8th layer, following SC-CLIP [2]. The experimental code can be found on our GitHub repository.

4.2. Performance Evaluation

Table 1 and Table 2 summarize our comparisons with SOTA FSS methods. We categorize these methods into two groups: Mask FSS methods, which utilize only mask annotations, and Class-aware FSS methods, which incorporate additional semantic (class) information.

Well-known mask FSS methods such as BAM [17] and HDMNet [25], trained solely on visual information from base classes, often suffer from base-class bias. Besides their quantitatively lower performance demonstrated in the tables, qualitative results (refer to Figure 6) for the 1-shot setting on the $Pascal\text{-}5^i$ [29] dataset also illustrate that BAM [17] and HDMNet [25] often fail to clearly distinguish novel target objects from adjacent base-class objects due to their inherent base-class bias.

Training-free SAM-based methods, such as PerSAM [38] and Matcher [20], rely purely on visual prompts derived from support images. PerSAM's limited prompt strategy results in lower accuracy, especially on $COCO\text{-}20^i$ [22]. Matcher [20] addresses this by generating richer prompts through bi-directional feature matching, achieving better results but still suffering from reliance on visual similarity alone. As illustrated in Figure 6 and the appendix available on our GitHub repository, Mask-only FSS methods visually exhibit relatively poor prediction performance, particularly when there are noticeable shifts in appearance or scene context between the query and support images.

Table 4. Ablation study on $COCO\text{-}20^i$ [22] and $Pascal\text{-}5^i$ [29]. FSS performance (mIoU, FB-IoU) and efficiency (FPS) results are presented in the format ($COCO\text{-}20^i$) / ($Pascal\text{-}5^i$). *M-S* denotes MobileSAM, a lightweight SAM variant for resource-constrained and real-time deployments.

Component	OWL	SAM	VLAS	LEVE	SMR	mIoU	FB-IoU	FPS
(a)	✓	✓				54.6 / 76.6	73.9 / 85.4	1.41 / 1.44
(b)	✓	✓			✓	59.6 / 79.4	77.0 / 87.6	1.08 / 1.25
(c)	✓	✓	✓			67.8 / 81.9	82.0 / 89.6	1.26 / 1.38
(d)	✓	✓	✓	✓		69.2 / 83.0	82.4 / 90.3	1.17 / 1.29
(e)	✓	✓	✓	✓	✓	70.5 / 84.0	83.6 / 91.0	1.10 / 1.15
(f)	✓	*M-S*	✓	✓	✓	66.0 / 76.2	80.8 / 85.1	1.18 / 1.35

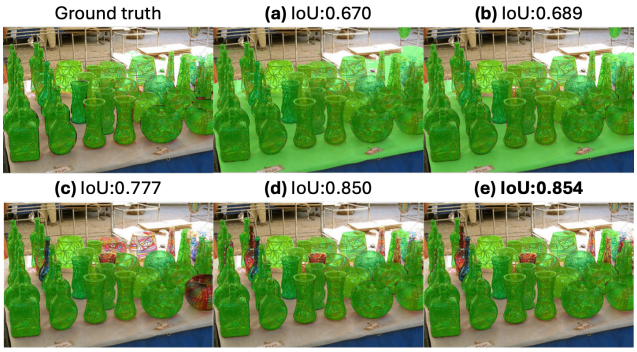

Figure 4. Qualitative visualization corresponding to the ablation study in Table 4.

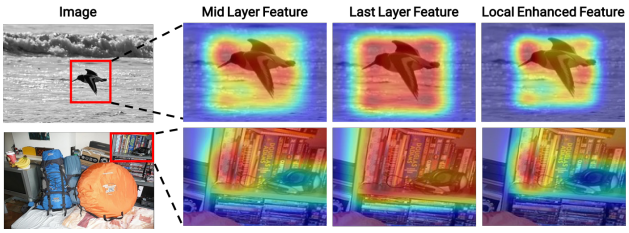

Figure 5. Visualization of attention maps from mid-layer feature, last layer feature, and the resulting local enhanced features in LEVE.

Class-aware FSS methods, including PI-CLIP [33] (training-free), PGMA-Net [6], and Beyond-Mask [5] (both requiring training), leverage semantic cues, typically achieving higher segmentation accuracy compared to mask-only FSS methods.

In comparison, our fully training-free approach achieves superior results across both datasets and settings by effectively integrating vision and text information, demonstrating robust and generalizable FSS performance.

To further evaluate the robustness of training-free FSS approach, we conduct cross-dataset experiments on four

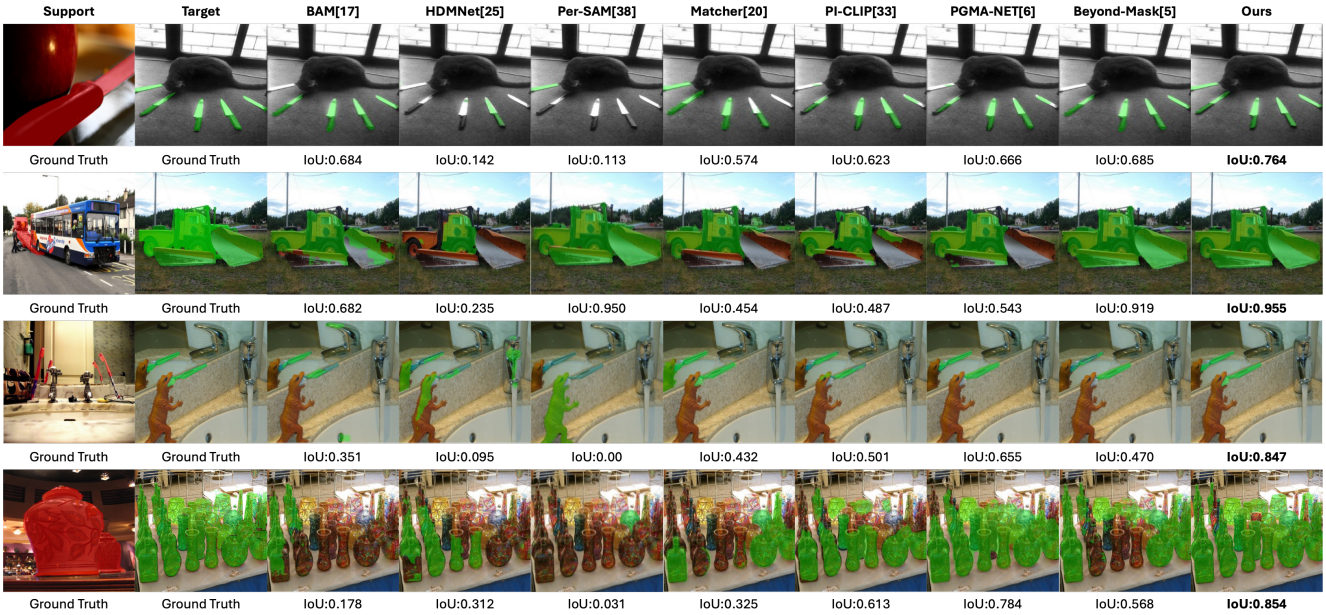

Figure 6. Qualitative comparison of ours and SOTA FSS methods under noticeable shifts in appearance or scene context between query and support images.

diverse datasets: LVIS [11] (large-scale general images), WHU [14] (satellite imagery), Kvasir [26] (medical images), and SBU [31] (shadow images). The results, summarized in Table 3, show that existing training-free FSS methods struggle on these unseen domains, exhibiting relatively low performance. In contrast, our method effectively leverages vision–language alignment to achieve the highest performance across all evaluated domains, demonstrating superior cross-domain generalization.

4.3. Ablation Study

We analyze the contribution of each proposed component through an ablation study, as summarized in Table 4. The baseline model, combining only OWL-ViT [21] and SAM [15], achieves relatively low performance, primarily due to numerous false-positive segmentations resulting from unfiltered object proposals. The *SMR* alone provides limited gains, as its effectiveness strongly depends on the initial bounding box quality. *VLAS* notably improves performance by semantically filtering OWL-ViT [21] proposals, substantially reducing incorrect masks. To examine the complexity–accuracy trade-offs in the training-free FSS via SAM, we conducted experiments with MobileSAM [37] as shown in (f). Qualitative results in Figure 4 visually confirm incremental mask refinement with each component, illustrating reduced false positives and improved precision.

Further incorporating the *LEVE*—which merges discriminative mid-layer local features and high-level global context—leads to additional accuracy improvements. Visualizations in Figure 5 highlight that the combination of mid-layer and high-layer features provided by *LEVE* effectively captures both detailed local structures and broad semantic contexts.

Overall, our ablation demonstrates the critical roles of *VLAS* and *LEVE*, collectively overcoming limitations of prior methods and achieving SOTA FSS performance without additional training.

5. Conclusion

We presented a simple training-free few-shot segmentation approach via vision-language guided prompting, achieving SOTA performance without any additional model updates. Our approach integrates three key stages: (i) *Text-Conditioned Region Proposal (TCRP)* for generating candidate bounding boxes and extracting effective region features; (ii) *Vision-Language Alignment Selection (VLAS)* for selecting semantically relevant boxes; and (iii) *Semantic Mask Refinement (SMR)* for precise mask prediction. Leveraging the complementary nature of vision and language information, our framework effectively guides SAM towards the intended object, robustly addressing appearance and contextual shifts between support and query images. Extensive experiments on standard benchmarks verify that our training-free, vision-language prompting approach surpasses existing methods, demonstrating its significant potential for few-shot segmentation.

6. Acknowledgement

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant (No.RS-2025-02219317; AI Star Fellowship(Kookmin University), No.RS-2025-02263754; Human-Centric Embodied AI Agents with Autonomous Decision-Making, IITP-2024-RS-2024-00397085; Leading Generative AI Human Resources, No.RS-2024-00357879; AI-based Biosignal Fusion and Generation Technology for Intelligent Personalized Chronic Disease Management and IITP-2024-RS-2024-00417958; Global Research Support Program in the Digital Field program) funded by the Korea government (MSIT).

References

[1] Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[2] Sule Bai, Yong Liu, Yifei Han, Haoji Zhang, and Yansong Tang. Self-calibrated clip for training-free open-vocabulary segmentation. *arXiv preprint arXiv:2411.15869*, 2024. 2, 3, 4, 7

[3] Hanbo Bi, Yingchao Feng, Wenhui Diao, Peijin Wang, Yongqiang Mao, Kun Fu, Hongqi Wang, and Xian Sun. Prompt-and-transfer: Dynamic class-aware enhancement for few-shot segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[4] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[5] Shijie Chang, Youwei Pang, Xiaoqi Zhao, Huchuan Lu, and Lihe Zhang. Beyond mask: Rethinking guidance types in few-shot segmentation. *Pattern Recognition*, 165:111635, 2025. 6, 7

[6] Shuai Chen, Fanman Meng, Runtong Zhang, Heqian Qiu, Hongliang Li, Qingbo Wu, and Linfeng Xu. Visual and textual prior guided mask assemble for few-shot segmentation and beyond. *IEEE Transactions on Multimedia*, 26:7197–7209, 2024. 6, 7

[7] Xi Chen, Xiao Wang, Soravit Changpinyo, Anthony J Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 7

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[9] 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*, 88:303–338, 2010. 6

[10] Qi Fan, Xin Tao, Lei Ke, Mingqiao Ye, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, Yu-Wing Tai, and Chi-Keung Tang. Stable segment anything model. *arXiv preprint arXiv:2311.15776*, 2023. 5

[11] 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*, pages 5356–5364, 2019. 7, 8

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 5

[14] Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on geoscience and remote sensing*, 57(1):574–586, 2018. 7, 8

[15] 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*, pages 4015–4026, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[16] Hyeokjun Kweon and Kuk-Jin Yoon. From sam to cams: Exploring segment anything model for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19499–19509, 2024. 2

[17] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Learning what not to segment: A new perspective on few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8057–8067, 2022. 2, 6, 7

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[19] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 2

[20] Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. *arXiv preprint arXiv:2305.13310*, 2023. 1, 2, 3, 6, 7

[21] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36:72983–73007, 2023. 2, 3, 4, 5, 7, 8

[22] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *Proceedings of*

the IEEE/CVF international conference on computer vision, pages 622–631, 2019. 2, 6, 7

[23] Maxime Oquab et al. Dinov2: Learning robust visual features without supervision. *arXiv*, 2023. 6

[24] Huihui Pan, Yuanduo Hong, Weichao Sun, and Yisong Jia. Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes. *IEEE Transactions on Intelligent Transportation Systems*, 24(3):3448–3460, 2022. 1

[25] Bohao Peng, Zhuotao Tian, Xiaoyang Wu, Chengyao Wang, Shu Liu, Jingyong Su, and Jiaya Jia. Hierarchical dense correlation distillation for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23641–23651, 2023. 2, 3, 6, 7

[26] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, et al. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 164–169, 2017. 7, 8

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 2, 3, 4, 7

[28] Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alyosha Efros, and Sergey Levine. Conditional networks for few-shot semantic segmentation. 2018. 6

[29] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017. 2, 6, 7

[30] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 7

[31] Tomás F Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *ECCV*, pages 816–832. Springer, 2016. 7, 8

[32] 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*, pages 3635–3647, 2024. 2

[33] Jin Wang, Bingfeng Zhang, Jian Pang, Honglong Chen, and Weifeng Liu. Rethinking prior information generation with clip for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3941–3951, 2024. 6, 7

[34] Defeng Xie, Ruichen Wang, Jian Ma, Chen Chen, Haonan Lu, Dong Yang, Fobo Shi, and Xiaodong Lin. Edit everything: A text-guided generative system for images editing. *arXiv preprint arXiv:2304.14006*, 2023. 2

[35] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023. 2

[36] Xiaohui Yuan, Jianfang Shi, and Lichuan Gu. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications*, 169: 114417, 2021. 1

[37] Chaoning Zhang et al. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv*, 2023. 8

[38] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Xianzheng Ma, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*, 2023. 1, 2, 3, 6, 7

[39] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in neural information processing systems*, 36:19769–19782, 2023. 2