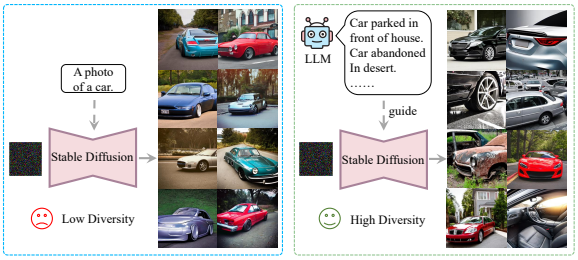
Training-free Open-Vocabulary Semantic Segmentation via Diverse Prototype Construction and Sub-region Matching

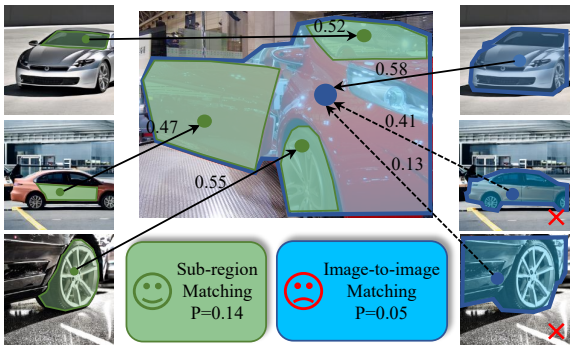
Xuanpu Zhao[1,2,3], Dianmo Sheng[1,2,3], Zhentao Tan[1,2,3], Zhiwei Zhao[1,2,3]
Tao Gong[1,2,3], Qi Chu[1,2,3*], Bin Liu[1,2,3], Nenghai Yu[1,2,3]
[1]School of Cyber Science and Technology, University of Science and Technology of China
[2]Anhui Province Key Laboratory of Digital Security
[3]the CCCD Key Lab of Ministry of Culture and Tourism
{zhaoxuanpu,dmsheng,tzt,zwzhao98}@mail.ustc.edu.cn, {tgong,qchu,fowice,ynh}@ustc.edu.cn

Abstract

Open-vocabulary semantic segmentation (OVSS) aims to segment images of arbitrary categories specified by class labels. While previous approaches relied on extensive image-text pairs or dense semantic annotations, recent training-free methods attempted to overcome these limitations by constructing semantic prototypes in the construction stage and image-to-image matching (i.e., prototype matching) during testing. However, these methods often struggle to effectively capture the visual characteristics of categories and fail to utilize local features during prototype matching. To deal with these problems, we propose a novel training-free framework for OVSS that constructs diverse prototypes and performs fine-grained sub-region matching. Specifically, our method leverages Large Language Models (LLMs) to guide support image generation by descriptions of different attributes of categories and employs coarse-fine clustering to obtain diverse and robust part-level prototypes in the construction stage. During testing, we propose a sub-region matching method, which assigns part-level prototypes to sub-regions utilizing optimal transport, to fully utilize local image features among part-level prototypes. Extensive experiments demonstrate the effectiveness of our method and show that our method achieves state-of-the-art performance, outperforming previous methods across five datasets.



(a) Diverse Prototype Construction.



(b) Sub-region Matching

Figure 1: Motivation of this work. (a) We construct diverse prototypes with the guidance of LLM. (b) We assign different part-level prototypes to sub-regions for fine-grained matching and utilization of complementary information.

Introduction

Open-vocabulary semantic segmentation (OVSS) (Ghiasi et al. 2022) classifies pixels in an image into a set of arbitrary categories that are specified by textual input. It is challenging since it requires assigning pixels to the correct semantic label from a large vocabulary. To solve this challenging task, some works (Xu et al. 2022b; Liang et al. 2023; Xu et al. 2023b; Han et al. 2023) generate class-agnostic mask proposals firstly and then classify the proposals using vision language models, *i.e.*, CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021). However, These methods require dense semantic annotation and face issues of transferring the image classification capabilities of vision-language models to region classification(Zhou et al. 2023; Liu et al. 2024; Jiao et al. 2023). Other works perform contrast learning through weak supervision from image-text pairs (Xu et al. 2022a; Luo et al. 2023; Xu et al. 2023a) to overcome this limitation. These methods require a large number of image-text pairs but still struggle to output high-quality mask proposals due to a lack of fine-grained supervision.

Recently, some training free manners (Karazija et al. 2023; Wang et al. 2024; Barsellotti et al. 2024) achieve promising performance following a two-stage framework, *i.e.*, semantic prototypes construction in the construction stage and similarity comparison during the test stage. To be specific, OVDiff first proposes to construct semantic prototypes via stable diffusion model (Rombach et al. 2022) and compare the similarity between the test image and con-

*Corresponding author.

structed prototypes. RIM (Wang et al. 2024) introduces a ranking-based matching process into the framework and FreeDA (Barsellotti et al. 2024) proposes to collect prototypes from a large set of captions.

We argue that there are two key points to achieve good performance for OVSS, 1) diverse prototypes to capture the visual characteristics, and 2) appropriate image-to-image matching method to better utilize the prototypes. However, the above methods (Karazija et al. 2023; Wang et al. 2024; Barsellotti et al. 2024) have failed to meet these two key points. They either depend on labeled captions of the segmentation dataset or use naive prompts like 'A good photo of a cat.' for generating images. The former requires a great quantity of manual labeling and the constructed prototypes can't well represent a category not included in these captions. The latter leads to a lack of diversity in the generated prototypes and therefore does not encapsulate the characteristics of real data, which is usually variable. Besides, the category of the region is determined simply by the whole-level prototype (i.e., representing the global feature within the foreground area of the generated image) that is most similar to the regional features. This loses a lot of local information and does not utilize the information among different prototypes. To be specific, for a certain region of a car to be classified (as shown in Figure 1), different sub-regions of the region have a high similarity with parts of different support images belonging to a car (*e.g.*, the door in the region is similar to the door in a support image, while the wheel in the region is similar to the wheel in another support image). However, the similarity between the whole region and the whole foreground area of the support image may not be high due to average pooling. Additionally, more than one support image may have a high similarity with the region but only the highest similarity is used and others are discarded, leading to an inadequate utilization of prototypes.

To tackle the first issue mentioned above, we expand the naive prompt with various descriptions about different attributes generated by LLMs (Touvron et al. 2023a,b; Meta 2024) to inject diversity into image synthesis. To retain local features of the synthesized images, we map them to part-level prototypes. Further, we conduct prototypes fusing with our coarse-fine clustering method to get more representative part-level prototypes and filter out noise caused by some low-quality parts. Through the analysis of the second question in the previous paragraph, we believe that it's more suitable to compare a proposal to be classified with parts from all support images than those from a single support image. Therefore, we propose to assign all the part-level prototypes to sub-region features of a proposal and a dustbin via optimal transport (OT) (Rubner, Tomasi, and Guibas 2000). This process activates the part-level prototypes from different support images that have a high similarity to the sub-regions. Our approach enables a fine-grained similarity comparison and avoids the underutilization existing in previous methods. To sum up, the contributions of this paper are as follows:

- We leverage LLM-guided image synthesis and a coarse-fine prototype fusing method to construct diverse and robust prototypes.
- We propose a sub-region matching method that utilizes OT to realize a fine-grained similarity comparison between test images and support images and full utilization of the information among all prototypes.
- Experiments show that our method achieves state-of-the-art performance on six datasets, without requiring training and extra datasets.

Related Work

Open-Vocabulary Semantic Segmentation

SimSeg (Xu et al. 2022b) proposed a simple two-stage framework using CLIP for open-vocabulary classification of mask proposals. The following works (Xu et al. 2023b; Liang et al. 2023; Xu et al. 2023b) attempt to align the proposal generator and CLIP. However, these methods are hindered by their reliance on costly dense annotations, posing a challenge in cases where such annotations are difficult to obtain. Other works have attempted contrastive learning from large-scale image-text pairs (Xu et al. 2022a; Luo et al. 2023; Xu et al. 2023a). Although not using dense annotation, these works often encounter spatial confusion in dense prediction tasks (Yi et al. 2023).

Training-Free OVSS

Some recent works (Karazija et al. 2023; Wang et al. 2024; Barsellotti et al. 2024) solve OVSS via support images synthesis and similarity comparison. Among them, OVDiff first proposes to construct semantic prototypes via stable diffusion model (Rombach et al. 2022) and compare the similarity between test image and prototypes. RIM (Wang et al. 2024) introduces a ranking-based matching process into the framework, and FreeDA (Barsellotti et al. 2024) proposes to collect prototypes from a large set of captions. Although achieve promising results in training-free OVSS, they only use naive prompts or require a great quantity of manually labeled captions for support images synthesis. Additionally, they failed to perform fine-grained matching between images and fully utilize the information among prototypes.

Optimal Transport

Optimal transport (OT) is a classical optimization problem that can be visualized as moving a pile of earth to match another pile with the least effort (Rubner, Tomasi, and Guibas 2000). It is adopted in many computer vision applications. In domain adaptation, (Flamary et al. 2016) proposed using OT for aligning source and target distributions. For local feature matching, (Ni et al. 2023) developed PATS, which employs OT to match image patches. Semantic correspondence has also benefited from OT formulations. (Liu et al. 2020) cast semantic correspondence as an OT problem, allowing for more robust matching between images with significant appearance variations. In panoptic segmentation, (Li et al. 2023) proposed Point2Mask to formulate point-supervised segmentation as an OT problem and enable the generation of high-quality segmentation masks from sparse point annotations. To our knowledge, we are the first to adopt OT in the field of Open-vocabulary Semantic Segmentation.

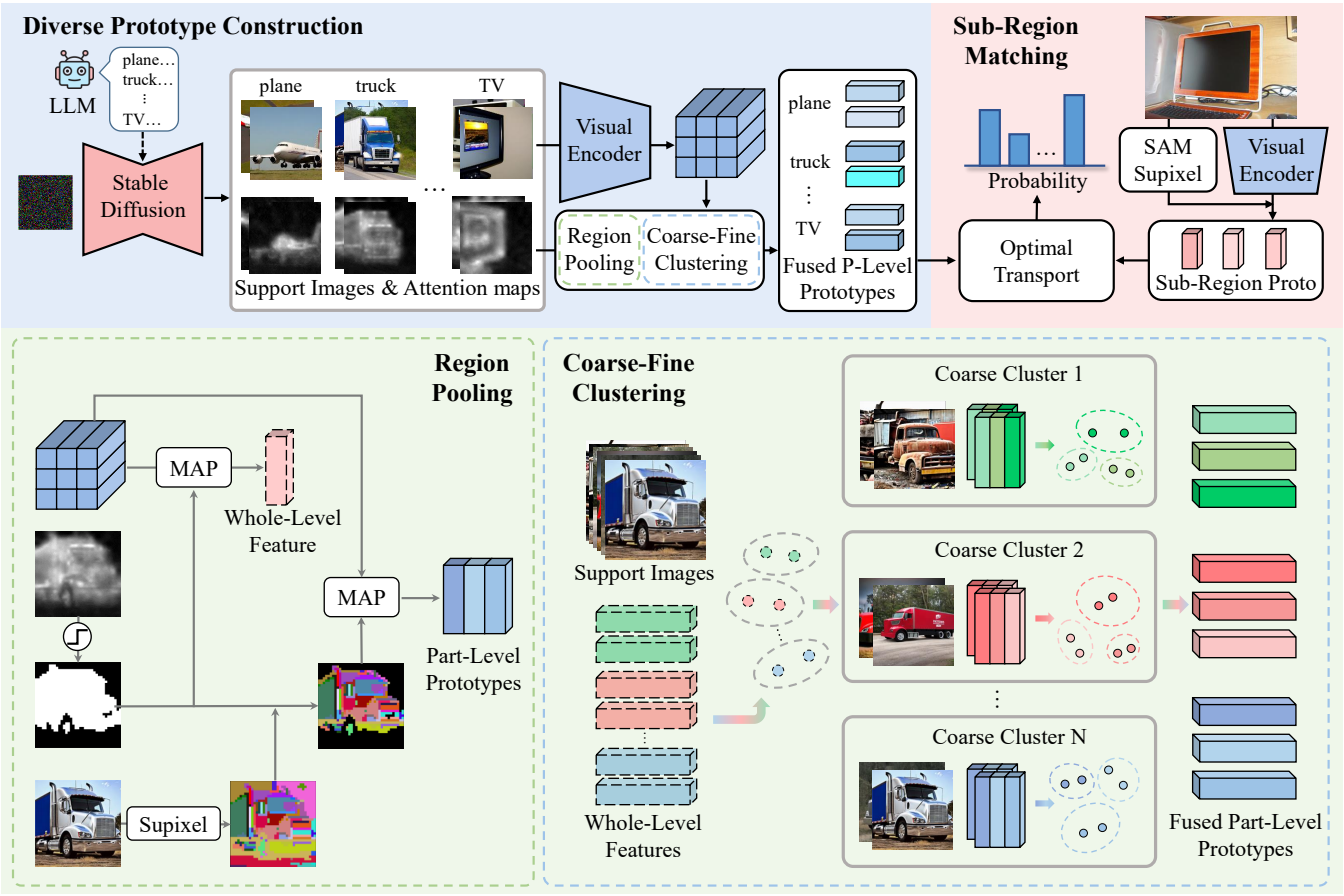


Figure 2: The first row shows the overview of our method. In **Diverse Prototype Construction**, we adopt LLM and coarse-fine clustering to get diverse and robust semantic prototypes. In **Sub-Region Matching**, similarity comparison is carried out between sub-region features and fused part-level prototypes during testing. The second row shows the details of Region Pooling and Coarse-Fine Clustering.

Method

Overview

We show the framework of our method in the first line of Figure 2. In diverse prototype construction, we use LLM (Touvron et al. 2023a,b; Meta 2024) to guide the diffusion model (Rombach et al. 2022) in generating support images, accompanied by attention maps that locate the foreground area (Hertz et al. 2022; Tang et al. 2022). The feature of these support images extracted by feature extractor and attention maps are mapped to fused part-level prototypes (denoted by 'Fused P-level Prototypes' in Figure 2) via region pooling and coarse-fine clustering. During the test stage, SAM and superpixel (denoted by 'Supixel' in Figure 2) are adopted to generate mask proposals for the test image, and the sub-region features of each proposal are extracted via the same feature extractor. Sub-region matching between fused part-level prototypes and sub-region features to determine each proposal's class. The details are as follows.

Diverse Prototype Construction

LLM-Guided Image Synthesis To inject diversity into images synthesized by stable diffusion model, we resort to large language model to generate K various descriptions for all C candidate classes about each of the following four attributes: state, feature, co-occurrence object and location. These descriptions can be denoted by $T_c = \{t^1_{c,a}, t^2_{c,a}, \cdots, t^K_{c,a} | a \in (state, feature, object, location), c \in C\}$. For example, when generating descriptions about the '**location**' attribute for car, we ask LLM by '**List different locations in which a car may appear in an image. Generate K short descriptions about car in these locations. Each description should start as A photo of car**'. We generate a support image with each description and get a set of support images of a category $S_c = \{S^1_c, S^2_c, \cdots, S^N_c | c \in C\}, N = 4K$.

Next, we locate the foreground area of category c on support images S_c and split the area into parts to construct both whole- and part-level prototypes. Specifically, for a support image $S^i_c, i \in \{1, \cdots, N\}$ of category c, we compute a score map by averaging the cross-attention maps (Hertz et al.

2022; Tang et al. 2022)

$$A(S_c^i, c) = \frac{1}{TLH} \sum_{t,l,h} A(S_c^i, c)_{t,l,h}, \quad (1)$$

where t, l and h index diffusion time steps, denoising layers and cross-attention heads respectively. We then normalize the score map to $[0, 1]$ and threshold it to γ and get a binary mask $M_{fg}(S_c^i, c) \in \{0, 1\}^{H \times W}$ for foreground area. Finally, the foreground area of S_c^i is mapped to whole-level feature as

$$\omega_c^i = MAP(\varphi_1(S_c^i), \zeta_1(M_{fg})), \quad (2)$$

where φ_1 is the UNet of stable diffusion, $\zeta_1(\cdot)$ denotes the bilinear interpolation which resizes binary mask to the size of feature map extracted by φ_1 and MAP denotes mask average pooling. Further, we exploit a superpixel algorithm (Felzenszwalb and Huttenlocher 2004) to partition S_c^i by grouping pixels into non-overlapping parts, which can be denoted by $R_1, R_2, \cdots, R_F \in \{0, 1\}^{H \times W}$ and F is the number of superpixels. These parts of S_c^i are mapped to part-level prototypes

$$p_c^{i,f} = MAP(\varphi_2(S_c^i), \zeta_2(M_{fg} \odot R_f)), f \in F, \quad (3)$$

where φ_2 is the feature extractor of DINOv2, $\zeta_2(\cdot)$ denotes the bilinear interpolation which resizes binary mask to the size of DINOv2's feature map. We denote the whole-level feature and part-level prototypes of all categories as $\Omega = \{\omega_c^i | i = 1, 2, \cdots, N\}_c^C$ and $P = \{p_c^{i;1,\cdots,F} | i = 1, 2, \cdots, N\}_c^C$ respectively.

Coarse-Fine Clustering To make the constructed prototypes more diverse and robust, we increase the scale of prototypes by κ times and fuse these prototypes back to its original scale with our coarse-fine clustering method. To be specific, there are support images $S = \{S_c^i | i = 1, 2, \cdots, \kappa N\}_c^C$, whole-level features $\Omega = \{\omega_c^i | i = 1, 2, \cdots, \kappa N\}_c^C$ and part-level prototypes $P = \{p_c^{i;1,\cdots,F} | i = 1, 2, \cdots, \kappa N\}_c^C$ after scaling up. Take category c for an example, we first group the support images $\{S_c^i | i = 1, 2, \cdots, \kappa N\}$ into N coarse clusters $\{Clu_c^1, Clu_c^2, \cdots, Clu_c^N\}$ with their corresponding whole-level features using the K-means algorithm. Inside a coarse cluster (i.e., Clu_c^i), we once again use the K-means algorithm to classify the part-level prototypes $\{p_c^{j;1,\cdots,F} | S_c^j \in Clu_c^i\}$ into fine cluster and take the clusters center as the fused part-level prototypes $\hat{p}_c^{i;1,\cdots,F}$, where i denote which coarse cluster these part-level prototypes belong to and F is the number of fused parts or fine clusters in the coarse cluster Clu_c^i. Finally, we denote all the fused part-level prototypes as $\hat{P} = \{\hat{p}_c^{i,1,\cdots,F} | i = 1, 2, \cdots, N\}_c^C$.

Sub-Region Matching

Revisit to Optimal Transport Suppose a set of suppliers $\mathcal{S} = \{s_i | i = 1, 2, \cdots, m\}$ are required to transport goods to another set of consumers $\mathcal{D} = \{d_i | i = 1, 2, \cdots, n\}$, where s_i denotes the supply units of supplier i and d_j represents the demand of demander j. Meanwhile, the total supply units should equal total demand units, $\sum_i s_i = \sum_j d_j$. A cost function c_{ij} specifies the cost of transporting one unit of goods from supplier i to consumer j. The goal of OT (Rachev 1985) is to find a transportation plan $X = \{x_{ij} | i = 1, 2, \cdots, m, j = 1, 2, \cdots, n\}$ that minimizes the total transportation cost:

$$\begin{aligned} \min_X \quad & \sum_{i=1}^m \sum_{j=1}^n c_{ij} x_{ij} \\ s.t. \quad & \sum_{j=1}^n x_{ij} = s_i, \ i = 1, \cdots, m \\ & \sum_{i=1}^m x_{ij} = d_j, \ j = 1, \cdots, n \\ & x_{ij} \geq 0, \ i = 1, \cdots, m, j = 1, \cdots, n. \end{aligned} \quad (4)$$

This optimization problem can be efficiently tackled with Sinkhorn Iteration (Cuturi 2013).

Optimal Transport for Sub-Region Matching For clarity, we first explain the case when $\kappa = 1$ (no coarse-fine clustering is conducted in this case).

For a test image, we generate class-agnostic mask proposals and sub-regions with SAM and superpixel algorithm respectively. Similar to part-prototypes construction in equation 3, we get the features of sub-regions in a mask proposal

$$q^f = MAP(\varphi_2(I_{test}), \zeta_2(M_p \odot R_f)), f \in F. \quad (5)$$

In equation 5, M_p, I_{test} and F represent mask proposal, the test image and the number of sub-regions respectively while the meaning of other symbols is the same as in equation 3.

Based on our analysis before, we propose to activate and assign some of the part-level prototypes to $Q = \{q^f, f = 1, ..., F\}$ while restraining and discarding others according to the similarity between Q and the previously constructed P. Once this matching process is completed, we can calculate the quantity of prototypes assigned to Q in different categories and decide category of the proposal with the quantity. To realize this design, we define all part-level prototypes in P as the m suppliers, features of sub-regions in Q as the n demanders and one minus cosine similarity

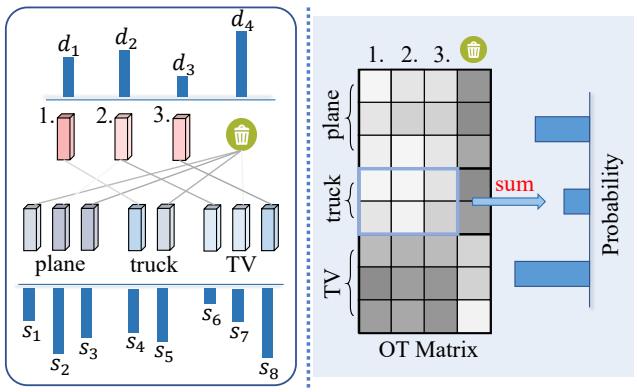


Figure 3: Illustration for sub-region matching. For clarity, we show an example with only three categories.

between them as the cost

$$c_{ij} = 1 - \frac{\mathcal{S}_i^T \cdot \mathcal{D}_j}{\|\mathcal{S}_i\|_2 \cdot \|\mathcal{D}_j\|_2}. \tag{6}$$

Intuitively, a supplier (demander) will play a more important role in the optimal transport problem if it has more supply units. Therefore, if the area corresponding to a part-level prototype (sub-region) occupies a larger proportion in the foreground (mask proposal), we allocate more supply (demand) units to it. Because we believe that area can, to some extent, reflect the importance of a certain part within the whole foreground. From this, we derive the supply units of the i^{th} supplier (corresponding to $p_c^{n;f}$ in P) and the demand units of the j^{th} demander (corresponding to q^f in Q)

$$s_i = \frac{1}{N} \cdot \frac{SUM(M_{fg} \odot R_f)}{\sum_{f \in F} SUM(M_{fg} \odot R_f)} \tag{7}$$

$$d_j = \frac{SUM(M_p \odot R_f)}{\sum_{f \in F} SUM(M_p \odot R_f)}, \tag{8}$$

where SUM represents the summation over all elements in matrix, R_f in equation 7 and equation 8 represents superpixels of the support images S_c^n and a test image I_{test} respectively. With these supply units, demand units and cost, we formulate our optimization problem as follows

$$\begin{aligned} \min_X \quad & \sum_{i=1}^m \sum_{j=1}^n c_{ij} x_{ij} \\ s.t. \quad & \sum_{j=1}^n x_{ij} \leq s_i,\ i = 1, ..., m \\ & \sum_{i=1}^m x_{ij} = d_j,\ j = 1, ..., n \\ & x_{ij} \geq 0,\ i = 1, ..., m, j = 1, ..., n. \end{aligned} \tag{9}$$

Note that there is an imbalance between the supply and demand ($\sum_i s_i = C \sum_j d_j = C$), which means only part of the supply units are delivered to demanders. This imbalance is necessary because suppliers can choose to transmit only the goods with small transmission cost (part-level prototypes that have high similarity with sub-region features) rather than all the goods. In the latter case, we cannot obtain any useful information because all categories choose to transmit one unit of goods. To solve problem 9, we add a dustbin as the $(n+1)^{th}$ demander demanding $(C-1)$ units of goods and set the cost from suppliers to the dustbin demander as zero, which balances supply and demand thus derives our final problem with the same form as problem 4:

$$\begin{aligned} \min_X \quad & \sum_{i=1}^m \sum_{j=1}^n c_{ij} x_{ij} + \sum_{i=1}^m 0 \cdot x_{i(n+1)} \\ s.t. \quad & \sum_{j=1}^n x_{ij} + x_{i(n+1)} = s_i,\ i = 1, \cdots, m \\ & \sum_{i=1}^m x_{ij} = d_j,\ j = 1, \cdots, n \\ & \sum_{i=1}^m x_{i(n+1)} = C - 1 \\ & x_{ij} \geq 0,\ i = 1, \cdots, m, j = 1, \cdots, n, n+1. \end{aligned} \tag{10}$$

Problem 10 is equivalent to Problem 9 and is a standard optimal transport problem that can be solved efficiently using the Sinkhorn iteration. We illustrate the Problem 10 in the left part of Figure 3 with an example when there are three categories. As shown in right part of Figure 3, we obtain the probability distribution of the proposal's category from the transportation plan X:

$$P(c|X) = \sum_{i \in \Gamma_c} \sum_{j=1}^n x_{ij} \tag{11}$$

where $\Gamma_c = \{i | \mathcal{S}_i \in P_c\}$.

There are only a few changes when it comes to the case where $\kappa > 1$ (coarse-fine clustering is conducted in this case). The part-level prototypes P should be replaced by fused part-level prototypes $\hat{P}$ and the supply units of the i^{th} supplier (corresponding to $\hat{p}_c^{i';f}$) should be replaced by following summation

$$\hat{s}_i = \frac{1}{\kappa} \sum_{j \in \Psi_c^{i';f}} s_j, \tag{12}$$

where $\Psi_c^{i';f}$ denotes the set of index of part-level prototypes used to fuse $\hat{q}_c^{i';f}$.

Experiments

Dataset and Evaluation Metric

We evaluate our method on the validation splits of traditional semantic segmentation benchmarks, namely Pascal VOC 2012 (Everingham et al. 2010), Pascal Context (Mottaghi et al. 2014), ADE20K (Zhou et al. 2017, 2019) and COCO Stuff (Caesar, Uijlings, and Ferrari 2018). We also validate our method when considering an additional 'background' class. For this case, we evaluate our method on the validation splits of Pascal Context (denoted by Context-60) and COCO Object (Lin et al. 2014) with an additional 'background' class. We use the mean of class-wise intersection over union (mIoU) to measure the performance.

Implementation Details

We use the Llama3 (Meta 2024) as our LLM to generate diverse descriptions for each class. For the support image generation, we employ Stable Diffusion v2-base (Rombach et al. 2022) with 50 diffusion steps. We employ DINOv2 (Oquab et al. 2023) with a ViT-B (Dosovitskiy et al. 2020) with an input image size of 518×518. This leads to dense features with a size corresponding to 37×37. SAM (Kirillov et al. 2023) with ViT-B is adopted as our mask proposal generator for the test image. We collect 32x32 prompt points in a grid manner to generate mask proposals for the test images. We also use the Felzenszwalb's algorithm (Felzenszwalb and Huttenlocher 2004) for extracting superpixels. To get diverse support images, we set the K and κ to be 16 and 8 respectively. For the foreground mask, we set the threshold $\gamma = 0.4$ which can filter out most of the background area.

Comparison With the State-of-the-Art Methods

We first compare our method with recent state-of-the-art methods for training-free open-vocabulary semantic segmentation. Specifically, we include DiffSeg (Wang et al.

Method	Training Dataset	Supervision	Training-free	mIoU					
				VOC-20	Context-59	Stuff	ADE	Context-60	Object
OpenSeg (Ghiasi et al. 2022)	COCO Stuff	SM	✗	60.0	36.9	15.3	-	-	-
SimSeg (Xu et al. 2022b)	COCO Stuff	SM	✗	88.4	47.7	20.5	-	-	-
OvSeg (Liang et al. 2023)	COCO Stuff	SM	✗	92.6	53.3	24.8	-	-	-
ReCo (Shin, Xie, and Albanie 2022)	ImageNet1k⋆	-	✓	57.7	22.3	16.3	11.2	19.9	15.7
MaskCLIP (Zhou, Loy, and Dai 2022)	LAION	IT	✗	74.9	26.4	16.4	9.8	23.6	20.6
GroupViT (Xu et al. 2022a)	CC12M	IT	✗	81.5	23.8	15.4	9.2	18.7	27.9
TCL (Cha, Mun, and Roh 2023)	CC3M+CC12M	IT	✗	83.2	33.9	22.4	17.1	30.4	30.4
DiffSeg (Wang et al. 2023)	-	-	✓	-	-	-	-	27.5	37.9
SCLIP (Wang, Mei, and Yuille 2023)	-	-	✓	80.4	33.0	22.4	14.6	30.4	30.5
GEM (Bousselham et al. 2024)	-	-	✓	-	34.5	-	17.1	-	-
ClearCLIP (Lan et al. 2024a)	-	-	✓	80.9	35.9	23.9	16.7	32.6	33.0
ProxyCLIP (Lan et al. 2024b)	-	-	✓	80.3	39.1	26.5	20.2	35.3	37.5
LaVG (Kang and Cho 2024)	-	-	✓	82.5	34.7	23.2	15.8	31.6	34.2
OVDiff (Karazija et al. 2023)	-	-	✓	81.7	33.7	-	14.9	30.1	34.8
RIM (Wang et al. 2024)	-	-	✓	77.8	34.3	-	17.0	-	-
FreeDA (Barsellotti et al. 2024)	COCO Captions⋆	Cap	✓	85.6	43.1	27.8	22.4	38.3	37.4
Ours	-	-	✓	**86.5**	**43.9**	**28.5**	**24.2**	**39.4**	**38.4**

Table 1: Comparison with state-of-the-art unsupervised open-vocabulary semantic segmentation models. The mark ⋆ refers to datasets used for support only. For the supervision type, 'SM' denotes segmentation mask, 'IT' denotes image-text pairs, and 'Cap' denotes captions only.

Proto Construction		Sub-Region Matching	mIoU	
LLM	Clustering		Context-59	ADE
✗	✗	✗	37.46	21.87
✓	✗	✗	40.66	22.33
✓	✓	✗	41.16	22.40
✗	✗	✓	37.30	21.22
✓	✗	✓	42.76	23.28
✓	✓	✓	**43.88**	**24.18**

Table 2: Ablation study on different components.

Proto Construction		Matching Method				mIoU
LLM	Clustering	single	img-to-cls	all	img-to-img	
✓	✗				✓	40.66
✓	✗	✓				41.25
✓	✗		✓			41.72
✓	✗			✓		42.76
✓	✓	✓				41.05
✓	✓		✓			42.56
✓	✓			✓		**43.88**

Table 3: Ablation study on different matching method.

Clustering Method				mIoU
no clustering	no scaling up	plain	corse-fine	
✓				42.23
	✓			42.76
		✓		43.34
			✓	**43.88**

Table 4: Ablation study on coarse-fine clustering

2023), SCLIP (Wang, Mei, and Yuille 2023), GEM (Bousselham et al. 2024), ClearCIIP (Lan et al. 2024a), ProxyCLIP (Lan et al. 2025) and LaVG (Kang and Cho 2025) that utilize image-text similarity and ReCo (Shin, Xie, and Albanie 2022), OVDiff (Karazija et al. 2023), RIM (Wang et al. 2024), and FreeDA (Barsellotti et al. 2024) which also exploit the arbitrary input categories to obtain a set of visual references. We also compare with MaskCLIP (Zhou, Loy, and Dai 2022), which introduces some modifications to the CLIP architecture to exploit its multimodal embedding space, and GroupViT (Xu et al. 2022a) and TCL (Cha, Mun, and Roh 2023) which rely on extensive contrastive training on large-scale image-text pairs to learn a textual-visual alignment. Further, we list the performance of OpenSeg (Ghiasi et al. 2022), SimSeg (Xu et al. 2022b) and OvSeg (Liang et al. 2023) for reference, which trains a proposal network and finetunes a VLM by supervision of densely labeled segmentation masks.

We report the results on four benchmarks in Table 1. As shown, our solution achieve the best results on all datasets, surpassing all the competitors by a consistent margin. Specifically, we achieve an average improvement of 5.6 mIoU points with respect to OVDiff, which adopts the same setting as ours. We also surpass FreeDA which uses an extra dataset for support image synthesis. It's a surprise that we even achieved performance comparable to SimSeg, which indicates that the method of generating visual reference features based on diffusion models has a promising future.

Ablation Studies

Ablation Study on Different Components We first conduct a series of ablation studies to thoroughly investigate the impact of each component of the proposed method on Context-59 and ADE datasets. As shown in Table 2, the 1^{st} to 3^{rd} rows are ablation studies without sub-region matching and the 4^{th} to 6^{th} rows are those with sub-region matching. In the 1^{st} and 2^{nd} rows, we conduct image matching via

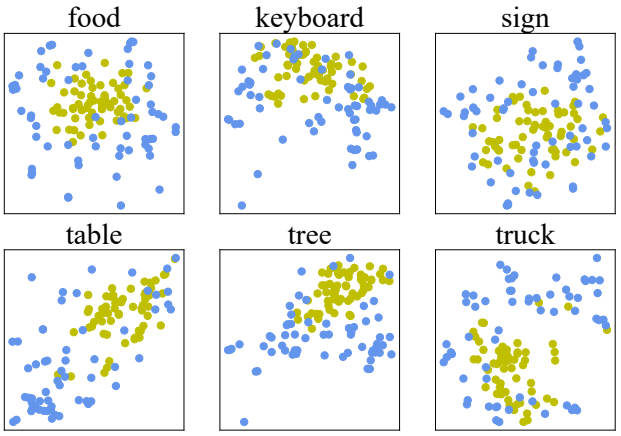


Figure 4: T-SNE visualization of whole-level prototypes constructed from naive prompt (yellow dots) and those constructed from LLM's descriptions (blue dots).

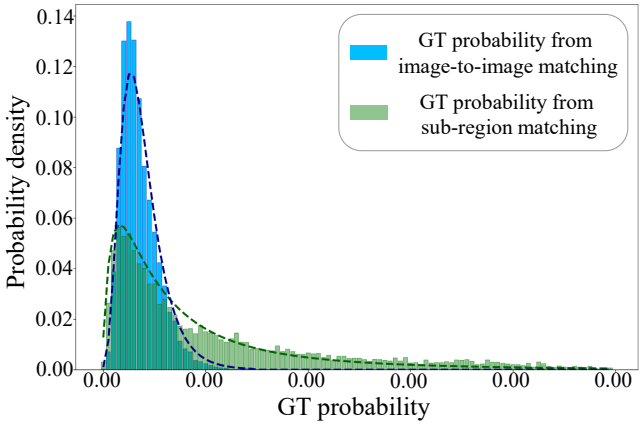


Figure 5: Distribution of the predicted probability of the ground truth class.

cosine similarity with the whole-level prototypes (different from whole-level feature used only for guiding coarse-fine clustering) following previous works. The whole-level prototype is constructed similarly as whole-level feature except that the feature map comes from DINOV2. In the 3^{rd} row, we use a weighted combination of the cosine similarity between the sub-region features of a proposal and the part-level prototypes from a support image. It's shown that the LLM's descriptions, coarse-fine clustering and sub-region matching are all usefull. Interestingly, in the 3^{rd} line, the sub-region matching algorithm causes a slight performance degradation when used without LLM and we attribute this to the lack of diversity of part-level prototypes.

In Figure 4, we show the T-SNE visualization of whole-level prototypes constructed from naive prompts (yellow dots) and those constructed from LLM's descriptions (each sub-figure represents a category). It can be observed that the prototypes generated from the descriptions of the LLM cover a wider area in the feature space. The predicted probability of the ground truth category is counted while using image-to-image matching and sub-region matching respectively and we show their distributions in Figure 5. When using image-to-image matching, the predicted probability of the GT category is mainly centered on 0.01 to 0.13, which illustrates the confusion of model. However, after using sub-region matching, the probability is spread out to larger values, indicating that the model became more deterministic in classification.

Ablation Study on the Matching Strategy We also compare the performance of different matching methods on Context-59. In the 1^{st} row of Table 3, we compute the cosine similarity between the feature of a mask proposal and all the whole-level prototypes. We categorize a proposal as category c if the most similar prototype belongs to category c. This is the simple image-to-image matching strategy adopted by previous work (corresponding to the 2^{nd} row in Table 2). Based on this strategy, we replace the cosine similarity between a mask proposal and a prototype with the EMD distance (Rubner, Tomasi, and Guibas 2000) between sub-region features of the proposal and the part-level prototypes a support image (2^{nd} and 5^{th} row), which enables finer-grained matching. At this point, we question whether it is better to match different sub-regions of a proposal to prototypes from different support images, rather than prototypes from just one. To verify this, we calculate the EMD distance between the sub-region features of a proposal and all part-level prototypes of each category and took the nearest category as the predicted category (3^{rd} and 6^{th} row). Based on this, we propose our final sub-region matching method (4^{th} and 7^{th} row).

As shown in table 3, replacement of cosine distance to EMD distance can bring some performance improvement. After matching the sub-region features with all the part-level prototypes, the performance further improves , demonstrating that the complementary information is very useful in the matching.

Ablation Study on the Coarse-Fine Clustering In this section, we first compare our coarse-fine clustering with the plain clustering method, which directly uses the K-means algorithm to fuse the part-level prototypes (the 3^{rd} and 4^{th} row in Table 4). It's shown that our coarse-fine clustering is better than the plain clustering. Next we discard the clustering and use the prototypes directly after scaling up (the 1^{st} row in Table 4). It can be observed that there is a significant performance degradation without clustering. We attribute this to the noise introduced to the part-level prototypes when the number of generated support images scales up. This indicates that clustering is necessary to ensure the quality of part-level prototypes while the diversity increases.

Conclusions

In this work, we presented a training-free framework to tackle OVSS. We construct diverse prototypes with the guidance of LLM and coarse-fine clustering. Then we conduct sub-region matching between sub-region features and constructed prototypes. Experimentally, we achieve state-of-the-art results on six datasets.

Acknowledgments

This work was supported by Anhui Provincial Science and Technology Major Project (No. 2023z020006), the National Natural Science Foundation of China (No. U20B2047) and Fundamental Research Funds for the Central Universities.

References

Barsellotti, L.; Amoroso, R.; Cornia, M.; Baraldi, L.; and Cucchiara, R. 2024. Training-Free Open-Vocabulary Segmentation with Offline Diffusion-Augmented Prototype Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3689–3698.

Bousselham, W.; Petersen, F.; Ferrari, V.; and Kuehne, H. 2024. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3828–3837.

Caesar, H.; Uijlings, J.; and Ferrari, V. 2018. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1209–1218.

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

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.

Felzenszwalb, P. F.; and Huttenlocher, D. P. 2004. Efficient graph-based image segmentation. *International journal of computer vision*, 59: 167–181.

Flamary, R.; Courty, N.; Tuia, D.; and Rakotomamonjy, A. 2016. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell*, 1(1-40): 2.

Ghiasi, G.; Gu, X.; Cui, Y.; and Lin, T.-Y. 2022. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, 540–557. Springer.

Han, K.; Liu, Y.; Liew, J. H.; Ding, H.; Liu, J.; Wang, Y.; Tang, Y.; Yang, Y.; Feng, J.; Zhao, Y.; et al. 2023. Global knowledge calibration for fast open-vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 797–807.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Jiao, S.; Wei, Y.; Wang, Y.; Zhao, Y.; and Shi, H. 2023. Learning mask-aware clip representations for zero-shot segmentation. *Advances in Neural Information Processing Systems*, 36: 35631–35653.

Kang, D.; and Cho, M. 2024. In Defense of Lazy Visual Grounding for Open-Vocabulary Semantic Segmentation. *arXiv preprint arXiv:2408.04961*.

Kang, D.; and Cho, M. 2025. In defense of lazy visual grounding for open-vocabulary semantic segmentation. In *European Conference on Computer Vision*, 143–164. Springer.

Karazija, L.; Laina, I.; Vedaldi, A.; and Rupprecht, C. 2023. Diffusion models for zero-shot open-vocabulary segmentation. *arXiv preprint arXiv:2306.09316*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Lan, M.; Chen, C.; Ke, Y.; Wang, X.; Feng, L.; and Zhang, W. 2024a. ClearCLIP: Decomposing CLIP Representations for Dense Vision-Language Inference. *arXiv preprint arXiv:2407.12442*.

Lan, M.; Chen, C.; Ke, Y.; Wang, X.; Feng, L.; and Zhang, W. 2024b. ProxyCLIP: Proxy Attention Improves CLIP for Open-Vocabulary Segmentation. *arXiv preprint arXiv:2408.04883*.

Lan, M.; Chen, C.; Ke, Y.; Wang, X.; Feng, L.; and Zhang, W. 2025. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *European Conference on Computer Vision*, 70–88. Springer.

Li, W.; Yuan, Y.; Wang, S.; Zhu, J.; Li, J.; Liu, J.; and Zhang, L. 2023. Point2mask: Point-supervised panoptic segmentation via optimal transport. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 572–581.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7061–7070.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, Y.; Bai, S.; Li, G.; Wang, Y.; and Tang, Y. 2024. Open-vocabulary segmentation with semantic-assisted calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3491–3500.

Liu, Y.; Zhu, L.; Yamada, M.; and Yang, Y. 2020. Semantic correspondence as an optimal transport problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4463–4472.

Luo, H.; Bao, J.; Wu, Y.; He, X.; and Li, T. 2023. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, 23033–23044. PMLR.

Meta, A. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.-G.; Lee, S.-W.; Fidler, S.; Urtasun, R.; and Yuille, A. 2014. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 891–898.

Ni, J.; Li, Y.; Huang, Z.; Li, H.; Bao, H.; Cui, Z.; and Zhang, G. 2023. Pats: Patch area transportation with subdivision for local feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17776–17786.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Rachev, S. T. 1985. The Monge–Kantorovich mass transference problem and its stochastic applications. *Theory of Probability & Its Applications*, 29(4): 647–676.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Rubner, Y.; Tomasi, C.; and Guibas, L. J. 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40: 99–121.

Shin, G.; Xie, W.; and Albanie, S. 2022. Reco: Retrieve and co-segment for zero-shot transfer. *Advances in Neural Information Processing Systems*, 35: 33754–33767.

Tang, R.; Liu, L.; Pandey, A.; Jiang, Z.; Yang, G.; Kumar, K.; Stenetorp, P.; Lin, J.; and Ture, F. 2022. What the daam: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, F.; Mei, J.; and Yuille, A. 2023. Sclip: Rethinking self-attention for dense vision-language inference. *arXiv preprint arXiv:2312.01597*.

Wang, J.; Li, X.; Zhang, J.; Xu, Q.; Zhou, Q.; Yu, Q.; Sheng, L.; and Xu, D. 2023. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *arXiv preprint arXiv:2309.02773*.

Wang, Y.; Sun, R.; Luo, N.; Pan, Y.; and Zhang, T. 2024. Image-to-Image Matching via Foundation Models: A New Perspective for Open-Vocabulary Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3952–3963.

Xu, J.; De Mello, S.; Liu, S.; Byeon, W.; Breuel, T.; Kautz, J.; and Wang, X. 2022a. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18134–18144.

Xu, J.; Hou, J.; Zhang, Y.; Feng, R.; Wang, Y.; Qiao, Y.; and Xie, W. 2023a. Learning open-vocabulary semantic segmentation models from natural language supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2935–2944.

Xu, M.; Zhang, Z.; Wei, F.; Hu, H.; and Bai, X. 2023b. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2945–2954.

Xu, M.; Zhang, Z.; Wei, F.; Lin, Y.; Cao, Y.; Hu, H.; and Bai, X. 2022b. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, 736–753. Springer.

Yi, M.; Cui, Q.; Wu, H.; Yang, C.; Yoshie, O.; and Lu, H. 2023. A simple framework for text-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7071–7080.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 633–641.

Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127: 302–321.

Zhou, C.; Loy, C. C.; and Dai, B. 2022. Extract free dense labels from clip. In *European Conference on Computer Vision*, 696–712. Springer.

Zhou, Z.; Lei, Y.; Zhang, B.; Liu, L.; and Liu, Y. 2023. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11175–11185.