The Power of Prior: Training-Free Open-Vocabulary Semantic Segmentation with LLaVA

Bingfeng Zhang[1] Siyue Yu[2*] Hui Li[3*] Jiahua Lin[4] Wenwu Wang[5] Jimin Xiao[2]
[1]China University of Petroleum (East China) [2]XJTLU [3]PolyU
[4]DIGITAL QINGDAO CONSTRUCTION CO.,LTD. [5]University of Surrey
bingfeng.zhang@upc.edu.cn, {siyue.yu02, jimin.xiao}@xjtlu.edu.cn, hui5li@polyu.edu.hk

Abstract

Multimodal Large Language Models (MLLMs) like LLaVA have demonstrated remarkable capabilities in multi-modal understanding and generation. This success motivates us to investigate whether the inherent prior knowledge embedded within such MLLMs contains sufficient spatial awareness for dense prediction tasks, without requiring any task-specific fine-tuning. Thus, in this paper, we explore the utilization of LLaVA for training-free open-vocabulary semantic segmentation. We discover that certain layers within the LLM part of LLaVA can generate localized features corresponding to given object classes. Building on this intrinsic capability, we design three modules: A question-answer pipeline to identify target classes in the image, a text-visual response module to extract initial reliable pixel-level activations for the target class, and a visual generation module to produce reliable refined prompts, which further serve as guidance for SAM to generate the predictions. Our LLaVA-based approach achieves new state-of-the-art performance on "Thing" category datasets, e.g.*, PASCAL VOC 2012 and COCO-object. Moreover, our method does not require explicit background class names, demonstrating its exceptional potential for handling open-world scenarios. Code is available at https://github.com/zbf1991/FSeg-LLaVA.*

1. Introduction

Open-vocabulary semantic segmentation (OVSS) [27, 46] aims to segment arbitrary objects in images based on textual queries, eliminating dependency on pre-defined training classes [31]. Based on training requirements, existing methods can be categorized into three paradigms: full supervision [8, 21, 50], weak supervision [43, 45], and training-free approaches [16, 22, 37, 47]. Among these, training-free methods have witnessed great attention due to their flexibility and zero-shot adaptability.

*Corresponding author.

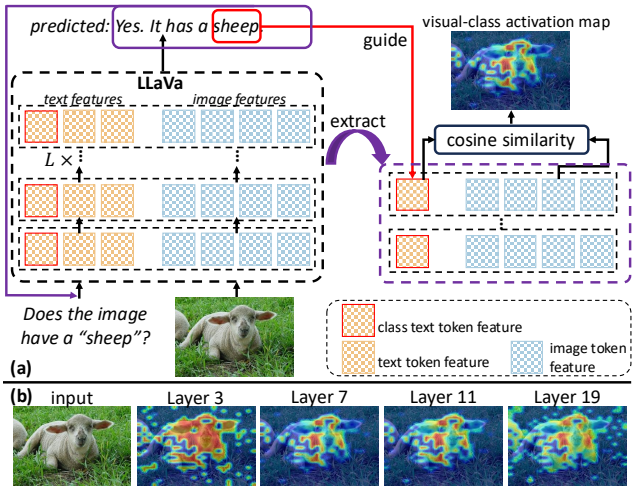

Figure 1. (a) Pipeline for visual-class activation map, which operates solely on foreground classes and motivates our FSeg-LLaVA. (b) Visual-class activation maps from different LLM layers in LLaVA. We compute cosine distances between visual token and class text token features as visual-class activation maps.

Most current training-free methods utilize a CLIP-based solution [18, 19, 36, 40], which produces segmentation by computing the embedding distance between image patches and class-related text sequences. To produce accurate predictions, previous methods [33, 40] focus on enhancing the attention maps of the vision encoder in CLIP to produce refined vision patch representations, which are then matched with the text embedding for mask prediction. Additionally, some approaches introduce extra vision foundation models [3, 30] to generate comprehensive semantic-aware attention maps [19].

Nevertheless, no matter which technique is employed, such CLIP-based solutions remain with several limitations: 1) To identify background and foreground, they must give explicit background subclasses definition, while such subclasses are inherently ambiguous to clarify or collect in

real-world scenarios. 2) They generate segmentation results solely based on the distance between each image patch and all predefined class name embeddings, possibly selecting similar but non-existent classes for certain patches. Such limitations essentially arise because CLIP adopts an encoder-only architecture, which can only perform segmentation in a discriminative, multi-class selection manner.

To overcome the above drawbacks, an intuitive solution is to adopt a non-discriminative architecture. We then consider whether Multimodal Large Language Models (MLLMs) [5, 23, 25, 41] can tackle this task, since most MLLMs use a decoder-only generative architecture with powerful knowledge generalization in their LLM part. Among them, LLaVA [23, 24] effectively bridges LLMs with visual perception through lightweight alignment, acquiring a strong multimodal understanding without the need for costly retraining. By combining a pre-trained vision encoder with an instruction-tuned LLM [7], LLaVA inherits the rich prior knowledge and instruction-following capabilities of large-scale language models. Therefore, we investigate whether LLaVA can be directly exploited to achieve segmentation in a non-discriminative manner.

However, LLaVA is a generative architecture [23], which is fundamentally different from the CLIP [32]. Applying it to segmentation will introduce new challenges: 1) How to produce a certain signal for LLaVA to indicate the presence of a specific class in the image; 2) How to leverage this signal to generate pixel-level segmentation masks. To solve these problems, as illustrated in Fig. 1 (a), we first attempt to input an image with a simple text instruction to get the text answer response. Then, we use the predicted answer to search for the corresponding class text token features and image patch token features in the LLM of LLaVA, and compute cosine distances between them as the visual-class activation map. As expected, we find that the visual-class activation map does highlight the target object tokens, as shown in the upper right of Fig. 1 (a). Moreover, we observe some interesting characteristics for LLaVA models that the features from the middle layers (*e.g.*, Fig. 1 (b) layer 7 & 11) exhibit precise class activation compared to shallow (layer 3) and deep (layer 19) layers.

Based on the above hypothesis and observations, we propose a new framework called FSeg-LLaVA for training-free open-vocabulary semantic segmentation. We design three main modules in FSeg-LLaVA: 1) A Question-answer pipeline module (QAP), which queries LLaVA with a specially designed question, to recognize the possible classes in the image based on the answer of LLaVA. 2) A text-visual response (TVR) module to compute visual-class activation maps from LLM layers in LLaVA. 3) A visual generation module (VGM), which involves a designed prototype block to further remove noisy regions in the selected activation maps from TVR, and then produces reliable prompts for SAM [17] to produce the final predictions.

Extensive experiments show that our approach achieves new state-of-the-art performances on the "*Thing*" category datasets with a clear margin. For example, our approach achieves 68.0% and 42.0% on the PASCAL VOC 2012 and COCO-Object, respectively. Meanwhile, we also achieve competitive performance on other datasets. Note that our approach does not require any background category as input, demonstrating its potential for handling complex cases.

Our contributions are summarized as follows:

- We propose a new LLaVA-centred framework (FSeg-LLaVA) for training-free open-vocabulary semantic segmentation, which verifies that the strong prior knowledge of MLLMs can tackle the dense prediction task.
- We carefully design three modules to utilize the prior knowledge of LLaVA to perform accurate segmentation when given arbitrary foreground classes.
- Our approach achieves state-of-the-art performance on "Thing" datasets and reveals characteristics of LLaVA, offering new insights into the application of MLLMs.

2. Related Work

2.1. Training-free Open-Vocabulary Semantic Segmentation

Training-free open-vocabulary semantic segmentation (OVSS) aims to leverage frozen vision–language models such as CLIP [32] for dense prediction without additional fine-tuning. Recent studies [26, 33] primarily reform inference process of CLIP to bridge its image-level supervision and pixel-level localization. CLIPtrase [33] recalibrates patch self-correlations to enhance local feature awareness suppressed by the global dominance of the [CLS] token, while CLIPSeg [26] exploits multi-layer feature coherence to integrate local semantics from middle layers. Feedback-driven methods [6] further enhance spatial consistency by propagating output-level correspondences back to intermediate attention. Other works [4, 13, 47] mitigate feature noise and semantic redundancy. SFP [13] suppresses propagated outliers to purify spatial features, and FreeCP [4] performs class purification to remove redundant and ambiguous categories, resulting in cleaner activation maps. From a data-centric perspective, ReME [47] demonstrates that constructing high-quality reference sets significantly boosts training-free segmentation performance. Beyond feature purification and inference redesign, recent works leverage external priors to enhance spatial reasoning. Trident [34] aggregates sub-image features with high-resolution priors from the Segment Anything Model (SAM) to refine CLIP predictions, while CorrCLIP [49] constrains patch correlations using SAM masks and self-supervised similarity. Despite these advances, they still struggle with the global bias and false positive activation

of CLIP, limiting their ability to mine discriminative and fine-grained segmentation cues. Thus, we try to explore new pipelines for training-free OVSS.

2.2. Multimodal Large Language Model

Multimodal large language models (MLLMs) have advanced rapidly in recent years, integrating visual encoders and large-scale language backbones to enable unified, cross-modal understanding and reasoning. Early frameworks, such as Flamingo [1], BLIP-2 [20], and LLaVA [23, 24], align visual and textual modalities through lightweight adapters or visual token projections, enabling models to follow instructions and perform open-ended visual reasoning. Subsequent systems, including InstructBLIP [9], Qwen-VL [41], and InternVL [5], further enhance fine-grained alignment and multimodal comprehension through large-scale instruction tuning. More recent developments, such as GPT-4V [11] and Gemini [38, 39], demonstrate that scaling both the visual encoder and the language backbone can yield impressive zero-shot capabilities across diverse visual tasks. These advances suggest that MLLMs inherently learn rich visual–semantic representations and emergent localization abilities, even without explicit dense supervision. Accordingly, we examine whether existing MLLMs, such as LLaVA, possess the ability to directly localize and segment accurate target masks in training-free OVSS.

3. Methodology

3.1. Overview

Fig. 2 shows the overall framework of our approach, which can be divided into the following steps:

- The target image and the pre-defined class set are input into our QAP for a fixed-format text query to indicate potential categories and the corresponding description, where LLaVA is deployed for text generation.
- Next, based on the generated class descriptions, our TVR is employed to extract features and attention maps of class and visual tokens from selected layers of LLaVA. Subsequently, the feature distances between the class and visual tokens are computed and integrated with attention maps to produce a reliable response map for the given image.
- Finally, the reliable response map is input to our VGM with extracted features of the visual token from selected LLaVA layers to further build refined masks for each class and then generate corresponding point prompt and box prompt for SAM to predict the final segmentation masks.

3.2. Question-Answer Pipeline

To enable LLaVA to perform training-free OVSS, an intuitive approach is to design question–answer templates that guide LLaVA to produce class-related responses. These responses can then be leveraged to reverse-trace the corresponding visual class-activated regions within LLaVA. Thus, we design a Question-Answer Pipeline (QAP) to generate structured class descriptions for the given pre-defined class set. Note here that all the classes in the set will be checked through our QAP. In the QAP, given a class name c_{name}, *e.g.*, bird, the following text is used as the query input T^c_{query} (c is the class name):

"Does the image have the class of a* { c_{name} }*? Answer yes or no to this question. If no, just say 'No.' If yes, the answer has to be two sentences, the first sentence is 'yes, it contains a* { c_{name} }*.' The second sentence is to describe* { c_{name} }*."

When such a text query is input to the LLaVA model with target image I, it will output a fixed-form text answer response. We can use the answer to recognize whether the class is in the image or not. Meanwhile, as illustrated in Fig. 1, the cosine distance between visual and class text token features in LLaVA can highlight the corresponding object, while also introducing some noisy regions. To produce accurate class activation maps, we extract the class token and visual features from the LLM part of LLaVA to generate the initial class activation map, and then utilize the attention map to filter the noisy pixels.

Formally, given an input image I and class text query T^c_{query}, the LLaVA is required to return the text response, all token features and attention maps as follows:

$$T^c_{\text{output}}, F_{\phi_m}, A_{\phi_m} = \phi_m(\phi_v(I), g(T^c_{\text{query}})), \tag{1}$$

where $\phi_m(\cdot)$ represents the LLM part of LLaVA. $\phi_v(\cdot)$ is the vision encoder of LLaVA, and $g(\cdot)$ is the text tokenizer for the text query. $F_{\phi_m} \in \mathbb{R}^{L\times(N_v+N_t)\times d}$ and $A_{\phi_m} \in \mathbb{R}^{L\times(N_v+N_t)\times(N_v+N_t)}$ are the visual token features and attention maps produced by L transformer blocks of ϕ_m, respectively. N_v is the number of visual tokens, N_t is the maximum number of output text tokens, and d is the channel size. T^c_{output} is the output text sequence, represents as following:

$$T^c_{\text{output}} = (S_0, S_1)\,, \tag{2}$$

where S_0 and S_1 are the pre-defined two sentences in the template, *i.e.*, $S_j = \left(t_0^{(j)}, t_1^{(j)}, \ldots, t_{N_j-1}^{(j)}\right)$, and $t_i^{(j)}$ denotes the i-th token in the j-th sentence, with N_j being the length of S_j (For simplicity, the tokenization process that splits words into sub-words is ignored).

Based on the response T^c_{output}, the word "yes/no" is set to recognize whether the image contains the pre-defined class, *i.e.*, when $t_0^{(0)}$ in S_0 is "yes" means that the class is in the image and LLaVA will also output S_1. Otherwise, the image does not contain the class, and there will be no S_1.

In this way, with our QAP, LLaVA is prompted to produce a fixed two-sentence response to indicate whether the queried class is present in the image or not. The first sentence explicitly includes the class name, while the second

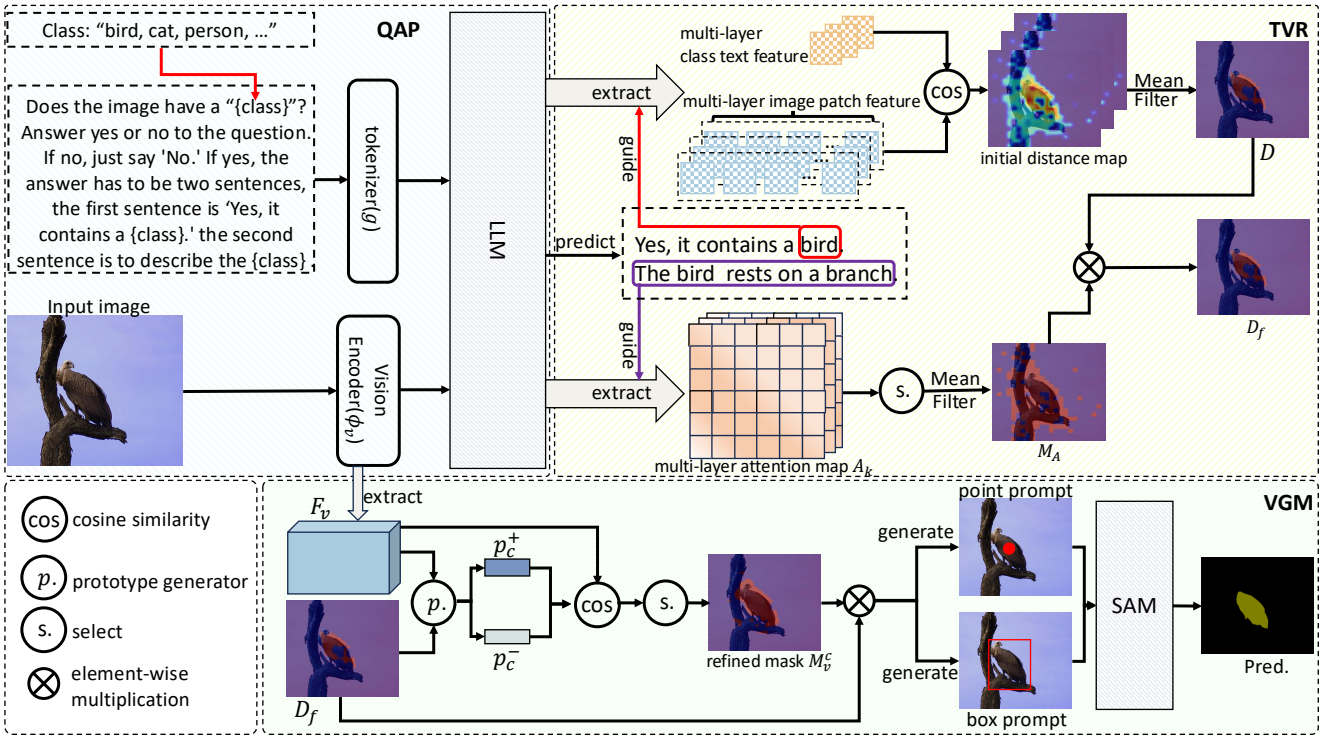

Figure 2. The Framework of the whole FSeg-LLaVA. Given an image with a predefined class set, we first use our QAP to generate the text response and retain the classes that LLaVA believes are present in the image. Then, our TVR extracts the features corresponding to the class text tokens and visual tokens, as well as the attention maps between class and visual tokens. All of these are used to produce the final reliable visual-class activation maps for each class. Finally, VGM uses the activation maps as input and generates the final predictions.

provides a detailed description. These two sentences are then utilized for subsequent processing to construct the final segmentation results.

3.3. Text-Visual Response Module

After generating the answer for a specific class using QAP, we design a Text-Visual Response Module (TVR) to produce visual-class activation responses by extracting the hidden state corresponding to the class token and the image from LLaVA. Specifically, for the given image I and class c, the current output of LLaVA is $T^c_{\text{output}} = (S_0, S_1)$ and $t_0^{(0)}$ in S_0 is “yes”, indicating that class c is in the image. Suppose the index of the class token c in S_0 is $t_c^{(0)}$, we can extract the corresponding feature map and compute the distance map as:

$$\tilde{D}(v_i, t_c^0) = \frac{1}{l_e - l_s + 1} \sum_{l=l_s}^{l_e} cos(F^l_{\phi_m}(v_i), F^l_{\phi_m}(t_c^0)), \quad (3)$$

where $\tilde{D} \in \mathbb{R}^{N_v \times 1}$ is the distance map, v_i is the index of the image token, and $v_i \in \{v_0, v_1, ..., v_{N_v-1}\}$. l_s and l_e are the indices of transformer layers in ϕ_m. $cos(\cdot)$ is used to compute the cosine distance. The simple distance map has shown the ability to localize the target class. We demonstrate some cases in Fig. 5. Surprisingly, it can also be observed that certain middle layers in LLaVA show a more precise semantic response. Specifically, layers 7 to 19 demonstrate more accurate semantic localization than shallower or deeper layers.

Accordingly, the challenge becomes how to distill accurate visual-class responses in the initial distance map $\tilde{D}$. In this way, we first propose a dynamic threshold τ to help filter low-confident regions:

$$\tau = \min\left(\alpha \cdot \max(\tilde{D}), \beta\right), \quad (4)$$

where α and β are hyperparameters to adjust the threshold.

After that, the high-confidence visual-class response map can be generated as:

$$D(v_i, t_c^0) = \begin{cases} \tilde{D}(v_i, t_c^0), & \text{if } \tilde{D}(v_i, t_c^0) > \tau \\ 0, & \text{else} \end{cases}. \quad (5)$$

Meanwhile, leveraging the detailed class descriptions in the second sentence S_1, we can generate an additional visual–class response by extracting cross-attention maps between text and visual tokens from each LLaVA layer.In de-

tail, for the j-th text token t_j^1 in S_1, its cross-attention map with visual token from the l-th LLaVA layer can be represented as $A_j^l \in \mathbb{R}^{N_v \times 1}$. Then, we can derive the visual-class responses for each text token in S_1 from all layers as $A \in \mathbb{R}^{L \times N_t \times N_v}$. However, such a visual-class response will definitely contain noise. Following [15], we extract the top K responses based on the spatial entropy, represented as $\left\{A_k^{\text{Top}}\right\} = topK(A)$, where $k \in [1, K]$. And the selected visual-class attention response map is computed as:

$$A_v = \frac{1}{K} \sum_{i=1}^{K} A_k^{\text{Top}}, \tag{6}$$

where $A_v \in \mathbb{R}^{N_v \times 1}$. Next, we can convert A_v to a binary mask M_A to suppress the low-confidence region:

$$M_A(i) = \begin{cases} 1, \text{ if } A_v(i) > \frac{1}{N_v} \sum_{j=1}^{N_v} A_v(j) \\ 0, \text{ else} \end{cases}. \tag{7}$$

Finally, the high-confidence class response map D and the binary visual-class attention response mask M_A are multiplicated to generate the reliable visual-class activation map D_f of the class c:

$$D_f^c = D \cdot M_A. \tag{8}$$

After reshaping, we can obtain $D_f^c \in \mathbb{R}^{1 \times h \times w}$, where h and w represent height and width and $h \times w = N_v$.

Repeating the above process (Eq. (5) - Eq. (8)) for all foreground classes present in the image, we can establish a set of visual-class activation maps $\{D_f^c\}_{c \in \mathcal{C}_\mathcal{I}}$, where $\mathcal{C}_\mathcal{I}$ denotes the set of detected object categories in the target image. To account for background influence, we further introduce a background response map $D_f^{\text{bg}} = \min(\alpha \cdot \max(\{D_f^c\}_{c \in \mathcal{C}_\mathcal{I}}), \beta)$. Then, D_f^{bg} is concatenated with $\{D_f^c\}_{c \in \mathcal{C}_\mathcal{I}}$ to form the final reliable visual-class activation maps $\hat{\mathbf{D}} = \left\{D_f^c\right\}_{c \in \mathcal{C}_\mathcal{I} \cup \{bg\}}$.

3.4. Visual Generation Module

In the above process, we are investigating the semantic association ability of LLM. However, since LLM is a generative model and is not directly designed for segmentation tasks. Our derived visual-class activation maps $\hat{\mathbf{D}}$ still lack fine-grained spatial awareness. We then explore how to enhance fine-grained semantic relationships. In LLaVA, which comprises a vision encoder, a text tokenizer, and an LLM, the vision encoder inherently captures rich spatial and semantic representations, as it is responsible for extracting visual features. Thus, we try to capitalize on the vision encoder and design a Visual Generation Module (VGM).

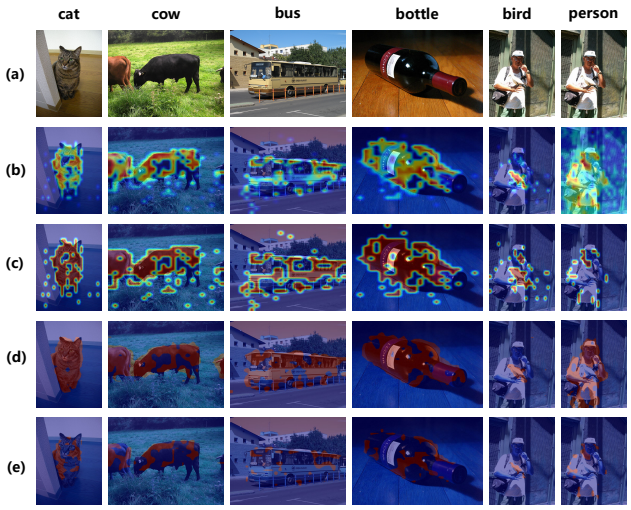

Figure 3. Visualization about the response maps of our method (FSeg-LLaVA1.5) on Pascal VOC 2012 dataset. (a) Original images. (b) The high-confidence visual-class response map D. (c) The selected visual-class attention response map A_v. (d) The prototype mask M_v. (e) The final refined binary mask M_r.

Specifically, we can use $\hat{\mathbf{D}}$ to generate the semantic guidance map $\hat{M}_f \in \mathbb{R}^{1 \times h \times w}$ as:

$$\hat{M}_f(i,j) = \underset{c \in \mathcal{C}_\mathcal{I} \cup \{bg\}}{\arg\max}\ D_f^c(i,j), \tag{9}$$

where (i, j) represents the spatial location.

Next, we extract the feature map $F_v \in \mathbb{R}^{d \times h \times w}$ from the vision encoder ϕ_v. Guided by $\hat{M}_f$, we can compute the foreground prototype $p_c^+ \in \mathbb{R}^d$ and background prototype $p_c^- \in \mathbb{R}^d$ for each class c as:

$$\begin{aligned} p_c^+ &= \frac{\sum_{i,j} \mathbb{1}[\hat{M}_f(i,j)=c] \cdot F_v(i,j)}{\sum_{i,j} \mathbb{1}[\hat{M}_f(i,j)=c] + \epsilon}, \\ p_c^- &= \frac{\sum_{i,j} (1 - \mathbb{1}[\hat{M}_f(i,j)=c]) \cdot F_v(i,j)}{\sum_{i,j} (1 - \mathbb{1}[\hat{M}_f(i,j)=c]) + \epsilon}, \end{aligned} \tag{10}$$

where $F_v(i,j) \in \mathbb{R}^d$ is the visual feature at the spatial location (i, j), $\mathbb{1}[\cdot]$ is the indicator function, outputting 1 if the condition is true and 0 otherwise. ϵ is a small constant for numerical stability.

After we get the prototypes, we compute the cosine similarity between F_v and the two prototypes for a prototype mask $M_v^c \in \mathbb{R}^{h \times w}$ as:

$$M_v^c(i,j) = \mathbb{1}\left[\cos(F_v(i,j), p_c^+) > \cos(F_v(i,j), p_c^-)\right], \tag{11}$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity.

Following that, we obtain the refined binary mask M_r^c by element-wise multiplication of M_v^c and $\hat{M}_f^c$:

$$M_r^c(i,j) = M_v^c(i,j) \cdot \hat{M}_f^c(i,j). \tag{12}$$

Fig. 3 shows different response maps of our method in each step of the inference time. As different semantic activation maps are used for filtering, we eventually obtain visual tokens with very high confidence, whose regions are often discontinuous and fail to fully cover the entire target object. Nevertheless, such activated regions can serve as strong prompts for SAM [17] to segment objects.

Thereby, we extract box prompts and point prompts based on M_r^c for SAM [17]. For the point prompts S_p, we perform morphological denoising on M_r^c, and extract valid connected components. The centroid of each region is then taken as a positive prompt point. Meanwhile, the box prompt S_b is obtained as the minimal bounding rectangle of the binary mask M_r^c. With these prompts, SAM produces the final segmentation mask for the class c as:

$$M_{\text{sam}}^c = \phi_{\text{sam}}(I|S_p, S_b), \tag{13}$$

where ϕ_{sam} denotes the SAM predictor.

4. Experiments

4.1. Dataset and Evaluation Metrics

We conduct a comprehensive evaluation on several well-established datasets following [19, 40]. These datasets are divided into two main categories: (i) With background class: PASCAL VOC (VOC 21) [10], PASCAL Context (Context60) [29], and COCO Object (Object) [2]; (ii) Without background class: COCO Stuff (Stuff) [2], and ADE20K (ADE) [52].

The mean Intersection over Union (mIoU) metric is used to evaluate semantic segmentation performance.

4.2. Implementation details

To validate the effectiveness of our approach, we conduct experiments on LLaVA-1.5 [23] and LLaVA-1.6 [24], covering three model variants: LLaVA-1.5-7B, LLaVA-1.6-7B, and LLaVA-1.6-13B. Specifically, LLaVA-1.5 is built upon the Vicuna-v1.5 [7, 51] backbone, while LLaVA-1.6 utilizes three different language backbones: Vicuna-7B [7, 51], Mistral-7B [12], and Vicuna-13B [7, 51].

To mitigate potential localization errors, we revise the image preprocessing pipeline by replacing padding with direct resizing to the target input size. Additionally, we disable the multi-resolution inference strategy introduced in LLaVA-1.6 and instead use only the original image resolution during evaluation. In Eq. (3), l_s is set to 7 for all backbones, while l_e is set as 13, 13, and 23 when the LLM backbone is Vicuna-7B, Mistral-7B, and Vicuna-13B, respectively. In Eq. (4), α is set as 0.75 for all backbones, while β is set as 0.19, 0.16, and 0.19 when the LLM backbone is Vicuna-7B, Mistral-7B, and Vicuna-13B, respectively. All other settings are kept at their default configurations. We use ViT-H SAM model [17] for final predictions.

4.3. Comparison with state-of-the-art

In Tab. 1, we compare our approach with other methods. It can be seen that our FSeg-LLaVA1.5 (vicuna-7B) generates the new state-of-the-art performance on two "*Thing*" categories datasets, *i.e.*, VOC21 and COCO-object datasets, outperforming other approaches by a clear margin. While our approach only generates competitive performance on the datasets that contain some "*Stuff*" categories, since the large-area or inherently complex "stuff" categories are difficult to fully cover with the point and box prompts. Moreover, it can also be found that compared to FSeg-LLaVA1.5, our FSeg-LLaVA1.6 (vicuna-7B) brings 1.0% mIoU increase on the average performance, and using different LLM, *e.g.*, FSeg-LLaVA1.6 (Mistral-7B), can also achieve competitive results, which illustrates the effectiveness of our approach. Finally, using a larger model (FSeg-LLaVA1.6-Vicuna-13B) does not achieve higher performance, showing that a larger-scale model does not provide more accurate localization ability in LLaVA.

Results visualization is demonstrated in Fig. 4. It can be observed that our FSeg-LLaVA performs well in single-class scenarios (the first and second rows). Moreover, as the model size increases, the predicted segmentation masks become more precise, even capturing clear and accurate object boundaries. In more complex multi-class scenarios (the third row), FSeg-LLaVA is also capable of accurately segmenting multiple categories within the same scene. It effectively distinguishes adjacent objects and produces segmentation masks with well-defined boundaries, demonstrating its strong capability in handling challenging scenarios.

4.4. Ablation Study

We use FSeg-LLaVA1.5 (Vicuna-7B) and FSeg-LLaVA1.6 (Vicuna-7B) to conduct an ablation study on VOC21 unless explicitly stated.

In Tab. 2, we compare the performance with different input text queries, using "*Does the image contain a* { } ..." leads to a noticeable performance drop compared to the other two formulations, while other two input text queries yield consistently stronger and comparable results, suggesting subtle differences in phrasing to the final segmentation.

In Tab. 3, we conduct an ablation study about using visual responses from different layers. Using 7-13 layers produces higher performance than other choices, demonstrating a notable localization ability. As the layers go deeper, the performance drops rapidly. This degradation can be attributed to the increasing semantic abstraction and loss of fine-grained spatial details in deeper layers of VLM.

Besides, in Fig. 5, we present qualitative results illustrating the visual activation responses of the "class" token across different layers of the LLM component in FSeg-LLaVA. These responses are computed using Eq. (5). It can be seen that shallow layers exhibit diffuse activation

Table 1. Open-vocabulary semantic segmentation comparison with other state-of-the-art methods. Note that unlike all other methods, our FSeg-LLaVA does not require additional background definition.

Method	Backbone	Pub.	Train	VOC21	Context60	Object	Stuff	ADE	Avg.
CoDe [42]	CLIP	CVPR'24	✓	57.5	30.5	32.3	23.9	17.7	32.4
DINOiser [44]	CLIP+DINO	ECCV'24	✓	62.1	32.4	34.8	24.6	20.0	34.8
CaR [36]	CLIP	CVPR'24	✗	48.5	30.5	36.6	-	17.7	-
PnP-OVSS [28]	BLIP	CVPR'24	✗	-	-	36.2	17.9	14.2	-
ClearCLIP [18]	CLIP	ECCV'24	✗	51.8	32.6	33.0	23.9	16.7	31.6
SCLIP [40]	CLIP	ECCV'24	✗	61.7	31.5	32.1	23.9	17.8	33.4
ProxyCLIP [19]	CLIP+DINO	ECCV'24	✗	61.3	35.3	37.5	26.5	20.2	36.2
LaVG [14]	CLIP+DINO	ECCV'24	✗	62.1	31.6	34.2	23.2	15.8	33.4
LPOSS [35]	CLIP+DINO	CVPR'25	✗	62.9	34.1	35.1	**26.5**	**22.2**	36.2
ResCLIP [48]	CLIP	CVPR'25	✗	61.1	33.5	35.0	24.7	18.0	34.5
FreeCP [4]	CLIP	ICCV'25	✗	65.8	**35.3**	37.2	24.9	18.4	36.3
FSeg-LLaVA1.5	Vicuna-7B	-	✗	**68.0**	30.6	**42.0**	21.2	16.9	35.7
FSeg-LLaVA1.6	Mistral-7B	-	✗	65.1	32.4	40.0	23.2	17.9	35.7
FSeg-LLaVA1.6	Vicuna-7B	-	✗	65.9	33.4	41.6	23.1	20.0	**36.8**
FSeg-LLaVA1.6	Vicuna-13B	-	✗	64.1	32.9	39.1	23.3	18.7	35.6

Table 2. Ablation study about using different input text prompts for LLaVA.

T_{query}	Fseg-LLaVA1.5	Fseg-LLaVA1.6
Is a photo of a { }? Answer yes or no for this question...	67.9	65.7
Does the image contain a { }? Answer yes or no for this question...	67.1	62.4
Does the image have the class of { }? Answer yes or no for this question...	**68.0**	**65.9**

Table 3. Ablation study about using different layers in Eq. (3).

l_s	l_e	FSeg-LLaVA1.5	FSeg-LLaVA1.6
5	9	63.7	64.5
7	9	65.8	64.6
7	11	66.7	65.8
7	13	**68.0**	**65.9**
7	17	66.8	65.2
10	13	67.6	64.4
15	22	49.1	55.8
22	30	15.2	19.7

patterns that often include regions unrelated to the target class. In contrast, deeper layers, *e.g.*, layer 19, tend to activate numerous small, spurious regions; layer 30 responds to the whole image, and neither of them can build an accurate object localization response. Notably, intermediate layers (*e.g.*, layers 7 and 11), yield more focused activation maps that better align with the true class region, illustrating that middle layers have a stronger segmentation ability.

In Tab. 4, we evaluate the influence of the prompts from M_r for final mask generation. By observation, point prompts produce higher performance than box prompts, indicating that they have a more accurate localization ability. Moreover, using both box and points performs better than using either alone across both FSeg-LLaVA1.5 and FSeg-LLaVA1.6 variants. This demonstrates that complementary spatial cues from multiple prompts are useful in FSeg-LLaVA for generating multi-class segmentation masks.

Table 4. Ablation study about different prompts for SAM.

Point	Box	FSeg-LLaVA1.5	FSeg-LLaVA1.6
✓		61.4	59.5
	✓	57.8	56.6
✓	✓	**68.0**	**65.9**

Table 5. Ablation study about different strategies to produce the reliable response map D_f in Eq. (8).

Strategy	FSeg-LLaVA1.5	FSeg-LLaVA1.6
$D_f = D$	67.3	65.6
$D_f = M_A$	27.3	34.7
$D_f = D \cdot M_A$	**68.0**	**65.9**

Tab. 5 shows the influence of different reliable response maps D_f. When $D_f = D$, only the distance between the

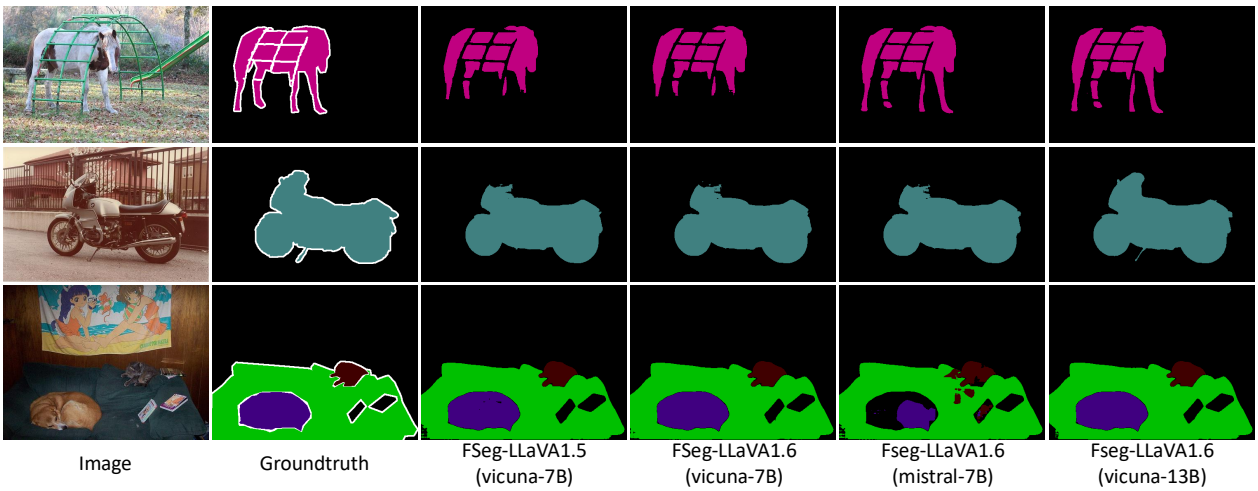

Figure 4. Qualitative visualization of our FSeg-LLaVA. It can be observed that ours can clearly localize and segment different classes.

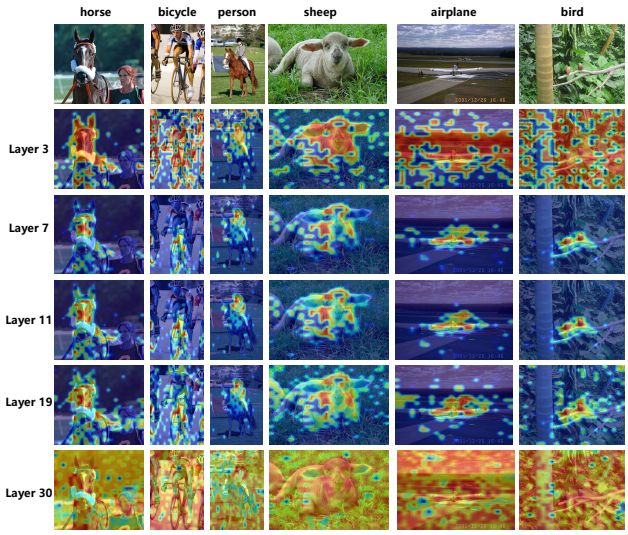

Figure 5. Initial distance map (Eq. (3)) among different layers in FSeg-LLaVA1.5. The shallow layers (*e.g.*, layer 3) tend to activate large regions; the deeper layers (*e.g.*, layer 19) introduce small noisy activations; the deepest layer (*e.g.*, layer 30) may even be dispersed. Instead, the middle layers (*e.g.*, layers 7 and 11) generate more accurate and semantically aligned visual activations.

text class token and the visual class token is used. The performance still remains at a high level, which demonstrates that the features in the LLM part of LLaVA contain rich semantic information for segmentation. When $D_f = M_A$, which means only the attention map is selected, the performance drops significantly, indicating that relying solely on the attention map lacks accurate localization between the text and visual modalities. In contrast, when $D_f = D \cdot M_A$, the model achieves the best performance, showing that M_A is useful to filter noisy regions in D.

Table 6. Ablation study about different strategies to produce the refined binary mask M_r in Eq. (12).

Strategy	FSeg-LLaVA1.5	FSeg-LLaVA1.6
$M_r = M_v$	48.4	60.0
$M_r = \hat{M}_f$	67.3	65.1
$M_r = M_v \cdot \hat{M}_f$	**68.0**	**65.9**

In Tab. 6, we comparing different strategies for generating the final refined binary mask, When $M_r = \hat{M}_f$, *i.e.*, only the mask generated using visual-text token distance and attention maps, the performance is close to $M_r = M_v \cdot \hat{M}_f$, which demonstrates that visual-text token distance is the key to the final segmentation. Besides, when $M_r = M_v$, *i.e.*, only the prototype mask is used, FSeg-LLaVA1.6 performs more stably than FSeg-LLaVA1.5, showing strong visual-text semantic alignment ability in LLaVA1.6.

5. Conclusion

In this paper, we propose a new method FSeg-LLaVA, utilizing the advanced multimodal large language model LLaVA to tackle training-free open-vocabulary semantic segmentation. We design three main modules: 1) a question-answer pipeline to classify the classes in the image, 2) a text-visual response module that extracts the text and visual features with their cross attention maps to produce a reliable visual-class activation map, and 3) a visual generation module to construct prototypes as guidance to further remove the noise and generate precise prompts for SAM to generate the final semantic segmentation maps. Extensive experiments demonstrate that FSeg-LLaVA achieves the best performance to date, indicating that current MLLMs can effectively tackle dense prediction tasks without requiring additional fine-tuning.

Acknowledge: This work was supported by the National Natural Science Foundation of China (No. 62301613, 62301451, 62471405), Shandong Natural Science Foundation (No. ZR2023QF046), Project 25-3-1-3-zyyd-jch supported by Qingdao Natural Science Foundation, and University of Surrey IAS Fellowship.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, pages 23716–23736, 2022. 3

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, pages 1209–1218, 2018. 6

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 1

[4] Qi Chen, Lingxiao Yang, Yun Chen, Nailong Zhao, Jianhuang Lai, Jie Shao, and Xiaohua Xie. Training-free class purification for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2508.00557*, 2025. 2, 7

[5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024. 2, 3

[6] Zhixiang Chi, Yanan Wu, Li Gu, Huan Liu, Ziqiang Wang, Yang Zhang, Yang Wang, and Konstantinos Plataniotis. Plug-in feedback self-adaptive attention in clip for training-free open-vocabulary segmentation. In *ICCV*, pages 22815–22825, 2025. 2

[7] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. 2, 6

[8] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *CVPR*, pages 4113–4123, 2024. 1

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, pages 49250–49267, 2023. 3

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 6

[11] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 3

[12] AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, D de Las Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. Mistral 7b. corr, abs/2310.06825, 2023. doi: 10.48550. *arXiv preprint ARXIV.2310.06825*, 10, 2023. 6

[13] Shuo Jin, Siyue Yu, Bingfeng Zhang, Mingjie Sun, Yi Dong, and Jimin Xiao. Feature purification matters: suppressing outlier propagation for training-free open-vocabulary semantic segmentation. In *ICCV*, pages 20291–20300, 2025. 2

[14] Dahyun Kang and Minsu Cho. In defense of lazy visual grounding for open-vocabulary semantic segmentation. In *ECCV*, pages 143–164, 2024. 7

[15] Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *CVPR*, pages 9339–9350, 2025. 5

[16] Chanyoung Kim, Dayun Ju, Woojung Han, Ming-Hsuan Yang, and Seong Jae Hwang. Distilling spectral graph for object-context aware open-vocabulary semantic segmentation. In *CVPR*, pages 15033–15042, 2025. 1

[17] 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 *ICCV*, pages 4015–4026, 2023. 2, 6

[18] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *ECCV*, pages 143–160, 2024. 1, 7

[19] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *ECCV*, pages 70–88, 2024. 1, 6, 7

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. pages 19730–19742, 2023. 3

[21] Yongkang Li, Tianheng Cheng, Bin Feng, Wenyu Liu, and Xinggang Wang. Mask-adapter: The devil is in the masks for open-vocabulary segmentation. In *CVPR*, pages 14998–15008, 2025. 1

[22] Yuqi Lin, Minghao Chen, Kaipeng Zhang, Hengjia Li, Mingming Li, Zheng Yang, Dongqin Lv, Binbin Lin, Haifeng Liu, and Deng Cai. Tagclip: A local-to-global framework to enhance open-vocabulary multi-label classification of clip without training. In *AAAI*, pages 3513–3521, 2024. 1

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, pages 34892–34916, 2023. 2, 3, 6

[24] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024. 2, 3, 6

[25] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024. 2

[26] Yajie Liu, Guodong Wang, Jinjin Zhang, Qingjie Liu, and Di Huang. Unveiling the knowledge of clip for training-free open-vocabulary semantic segmentation. In *AAAI*, pages 5649–5657, 2025. 2

[27] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. pages 23033–23044, 2023. 1

[28] Jiayun Luo, Siddhesh Khandelwal, Leonid Sigal, and Boyang Li. Emergent open-vocabulary semantic segmentation from off-the-shelf vision-language models. In *CVPR*, pages 4029–4040, 2024. 7

[29] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, pages 891–898, 2014. 6

[30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[31] Jie Qin, Jie Wu, Pengxiang Yan, Ming Li, Ren Yuxi, Xuefeng Xiao, Yitong Wang, Rui Wang, Shilei Wen, Xin Pan, et al. Freeseg: Unified, universal and open-vocabulary image segmentation. In *CVPR*, pages 19446–19455, 2023. 1

[32] 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. pages 8748–8763, 2021. 2

[33] Tong Shao, Zhuotao Tian, Hang Zhao, and Jingyong Su. Explore the potential of clip for training-free open vocabulary semantic segmentation. In *ECCV*, pages 139–156, 2024. 1, 2

[34] Yuheng Shi, Minjing Dong, and Chang Xu. Harnessing vision foundation models for high-performance, training-free open vocabulary segmentation. In *ICCV*, pages 23487–23497, 2025. 2

[35] Vladan Stojnić, Yannis Kalantidis, Jiří Matas, and Giorgos Tolias. Lposs: Label propagation over patches and pixels for open-vocabulary semantic segmentation. In *CVPR*, pages 9794–9803, 2025. 7

[36] Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. Clip as rnn: Segment countless visual concepts without training endeavor. In *CVPR*, pages 13171–13182, 2024. 1, 7

[37] Lv Tang, Peng-Tao Jiang, Haoke Xiao, and Bo Li. Towards training-free open-world segmentation via image prompt foundation models. *IJCV*, 133(1):1–15, 2025. 1

[38] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 3

[39] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 3

[40] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *ECCV*, pages 315–332, 2024. 1, 6, 7

[41] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 3

[42] Ji-Jia Wu, Andy Chia-Hao Chang, Chieh-Yu Chuang, Chun-Pei Chen, Yu-Lun Liu, Min-Hung Chen, Hou-Ning Hu, Yung-Yu Chuang, and Yen-Yu Lin. Image-text co-decomposition for text-supervised semantic segmentation. In *CVPR*, pages 26794–26803, 2024. 7

[43] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *arXiv preprint arXiv:2310.01403*, 2023. 1

[44] Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation. In *ECCV*, pages 320–337, 2024. 7

[45] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Yi Wang, Yu Qiao, and Weidi Xie. Learning open-vocabulary semantic segmentation models from natural language supervision. In *CVPR*, pages 2935–2944, 2023. 1

[46] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *CVPR*, pages 2945–2954, 2023. 1

[47] Xiwei Xuan, Ziquan Deng, and Kwan-Liu Ma. Reme: A data-centric framework for training-free open-vocabulary segmentation. In *ICCV*, pages 20954–20965, 2025. 1, 2

[48] Yuhang Yang, Jinhong Deng, Wen Li, and Lixin Duan. Resclip: Residual attention for training-free dense vision-language inference. In *CVPR*, pages 29968–29978, 2025. 7

[49] Dengke Zhang, Fagui Liu, and Quan Tang. Corrclip: Reconstructing patch correlations in clip for open-vocabulary semantic segmentation. In *ICCV*, pages 24677–24687, 2025. 2

[50] Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. Dreamlip: Language-image pre-training with long captions. In *ECCV*, pages 73–90, 2024. 1

[51] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 6

[52] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017. 6