Aligning Vision-Language Model with Fine-grained Semantics for Open-Vocabulary Segmentation

Yong Xien Chng*, Xuchong Qiu*, Yizeng Han, Kai Ding, Wan Ding, and Gao Huang✉, *Member, IEEE,*

***Abstract*—Pre-trained Vision-Language Models (VLMs) are often used to tackle the challenging task of Open-vocabulary Segmentation (OVS). To preserve the valuable pre-trained knowledge of VLM-based mask classifiers, most existing approaches freeze their parameters during training. However, our comprehensive analysis identifies a previously overlooked limitation: the performance of OVS is primarily constrained by mask classification. Specifically, VLMs pre-trained using globally pooled image-text representations often fail to capture localized, region-specific semantics necessary for accurate segmentation. This discovery motivates us to improve the fine-grained alignment between word-level text features and pixel-level image features extracted by VLMs. To this end, we propose the Fine-grained Semantic Reconstruction (FiSeR), a novel auxiliary task designed to enrich the spatial semantic detail of visual features. FiSeR trains the model to predict a randomly masked target class label using the image features and the remaining unmasked text. This encourages the model to link the specific words to the corresponding image regions, improving its ability to recognize and segment objects at the region level. FiSeR is broadly applicable and can be incorporated into various VLM-based segmentation models to improve their performance. Additionally, we introduce the Text-guided Visual Aligner (TeVA), a lightweight network module that injects relevant fine-grained semantics from the text information early in the visual encoding process. This enables the model to condition its visual processing on the target text categories from the beginning, improving its ability to associate text with the correct spatial regions. Collectively, these innovations culminate in our proposed framework FOV-Seg. Notably, FOV-Seg achieves new state-of-the-art results across multiple representative OVS benchmarks, improving performance consistently and reducing training costs by nearly 5× compared to previous best methods. Our code and data will be released.**

***Index Terms*—open-vocabulary, segmentation, fine-grained, alignment, vision-language model**

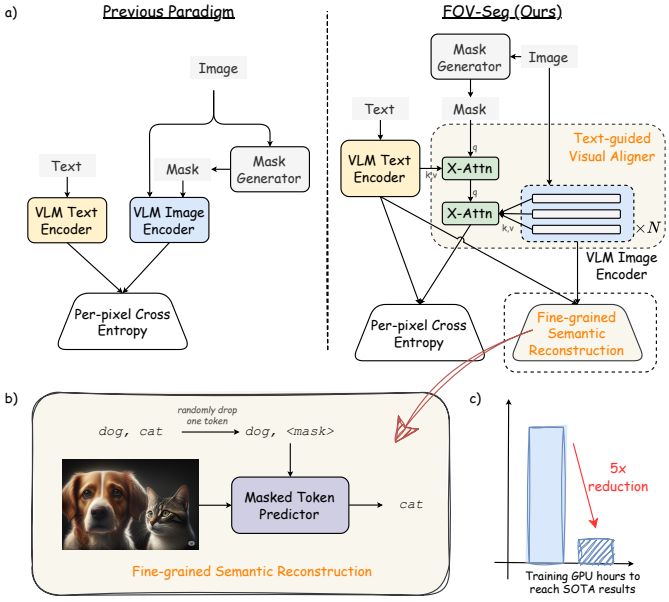


Fig. 1: a) Comparison between our proposed FOV-Seg and previous open-vocabulary segmentation (OVS) paradigm. FOV-Seg introduces two key innovations, namely Fine-grained Semantic Reconstruction (FiSeR) and Text-guided Visual Aligner (TeVA) to improve OVS performance b) FiSeR strengthens the fine-grained alignment between text and image features by explicitly teaching VLM to reconstruct the masked semantics from visual cues and remaining context. c) FOV-Seg trains 5× faster than previous SOTA model.

I. Introduction

Open-vocabulary segmentation [1, 2] is an important task that combines semantic segmentation [3] of unseen background elements with instance segmentation [4] of unseen foreground objects. Its application has profound implications for enhancing scene comprehension in domains like autonomous driving [5, 6] and robotics [7, 8], leading to widespread research interest. Despite considerable progress, existing methods still show limited real-world performance and require substantial computational resources for training [1, 2], hindering their widespread adoption.

Current state-of-the-art open-vocabulary segmentation methods [9, 10, 11] rely heavily on Vision-Language Models (VLMs) [12, 13, 14] for their robust zero-shot capabilities [15]. These methods extract visual features from frozen VLMs and propose various techniques to utilize these features. They typically focus on training the mask generators and keep the VLMs frozen. The VLMs are kept frozen during training to preserve their valuable pre-trained knowledge and to minimize the high computational cost of adapting the large VLMs. However, since VLMs are generally not trained to process individual image regions, they may require some adaptation to perform optimally for dense segmentation tasks that require precise categorization of image parts. To verify this hypothesis, we conduct several analytical experiments using the highly modular and efficient MaskCLIP model [9]. As discussed in Sec. III, our analysis reveals a previously overlooked limitation: *mask classification is the primary*

Y. X. Chng, Y. Han and G. Huang are with the Department of Automation, BNRist, Tsinghua University, Beijing, China. X. Qiu, K. Ding and W. Ding are with the Bosch Corporate Research, Shanghai, China.
Email: chngyx10@mails.tsinghua.edu.cn, gaohuang@tsinghua.edu.cn.
* Equal contribution.
✉Corresponding author.

performance bottleneck for open-vocabulary segmentation. In light of these observations, we decide to explore an alternative approach for open-vocabulary segmentation in this work. Instead of freezing the VLM, we focus on optimizing the VLM-based mask classifier. To guide our strategy for improving mask classification, we conduct further investigation, which indicates that one of the main bottlenecks for mask classification stems from the *lack of fine-grained image-text alignment in the visual features extracted by VLMs*. This suggests that enhancing the semantic awareness of these features could be a promising approach for improving mask classification performance.

Based on insights gained from our preliminary analysis, we propose FOV-Seg, a novel framework illustrated in Fig. 1, that seamlessly integrates two innovative modules to enhance fine-grained alignment between word-level text and pixel-level image features. First, we introduce the Fine-grained Semantic Reconstruction (FiSeR), a novel auxiliary reconstruction task that trains the model to predict the identity of a randomly masked class label by leveraging both the input image and the remaining unmasked text. This approach encourages the model to associate specific words with corresponding image regions, thereby enhancing its ability to recognize and segment objects at the region level. FiSeR can be easily applied to a variety of VLM-based mask classifiers to achieve improved performance. Second, we propose the Text-guided Visual Aligner (TeVA), a lightweight network module that modifies CLIP's attention mechanisms to inject relevant fine-grained semantic information from the text early in the visual encoding process. TeVA operates in two stages. First, it cross-attends mask tokens with target class tokens, conditioning them on the corresponding semantic context. Then, these enriched mask tokens cross-attend with image tokens, enabling the model to extract more meaningful visual features and ultimately improving mask classification performance. Additionally, to enhance training efficiency and preserve the VLM's pretrained knowledge when adapting to new distribution, we selectively fine-tune only a small portion of the VLM's parameters.

Comprehensive experiments and ablations confirm the superiority of our method. Compared to the previous best method, MAFT+, FOV-Seg achieves new state-of-the-art results with gains of up to 1.3 PQ and 3.1 mIoU on key benchmarks. Notably, it accomplishes this while reducing training costs by nearly 5×. Our main contributions are summarized as follows:

1) We carefully analyze existing open-vocabulary segmentation methodology, revealing that *mask classification is the main performance bottleneck for this task, and its weak performance mainly arises from a lack of fine-grained image-text alignment in the extracted visual features*.
2) We propose FOV-Seg, a novel framework incorporating two key innovations: Fine-grained Semantic Reconstruction, which enriches visual features extracted by the VLM classifier with important localized semantic information, and Text-guided Visual Aligner, which strengthens fine-grained alignment between textual and visual features.
3) We ablate FOV-Seg extensively, showing that it is 5× more efficient and outperforms the previous state-of-the-art method across multiple representative datasets.

II. Related Work

Open-vocabulary segmentation [16, 17] combines both semantic and instance segmentation of unseen classes. Current methods primarily adopt Vision-Language Models (VLMs) such as CLIP [12, 18, 14] that can perform zero-shot classification. Given the complexity of this task, research in this field begins with the exploration of methods focusing exclusively on open-vocabulary semantic segmentation [19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29]. LSeg [30] directly fine-tunes a CLIP model to learn dense image features. While OpenSeg [31], ZSseg [32], and ZegFormer [33] all share a common approach of generating region proposals before applying CLIP classification, each implements this strategy differently. OVSeg [34] collects mask-image pairs to improve CLIP's performance on masked images. SAN [35] employs a side adapter network that leverages outputs from a frozen CLIP model to perform mask prediction and classification. CAT-Seg [36] introduces a novel cost-aggregation method to refine CLIP's dense predictions. SED [37] further enhances CAT-Seg by using a hierarchical CLIP model to generate hierarchical dense predictions. As open-vocabulary semantic segmentation techniques mature and VLMs become increasingly sophisticated, attention shifts to the more challenging task of full open-vocabulary segmentation. Most methods for open-vocabulary segmentation initially adopt a two-stage approach for its simplicity and training efficiency. The pioneering MaskCLIP [9] introduces a novel Relative Mask Attention mechanism to extract regional mask information. MasQCLIP [38] enhances MaskCLIP by using progressive distillation to improve mask generation and adding a query adapter to enhance model adaptation. Since the two-stage approach generally lacks synergy between mask classification and generation, recent methods shift towards a one-stage approach to enhance performance. ODISE [39], DiffSegmenter [40] and Texture Diffusion [41] explore using frozen internal representations of Stable Diffusion [42] for open-vocabulary segmentation, while FC-CLIP [10] investigates using a CNN-based CLIP model that efficiently provides feature maps with much higher resolution. Recent works have started to pay attention to the quality of extracted visual feature representation. For instance, MaskAdapter [43] additionally extract semantic activation maps using a mask consistency loss, while MAFT+ [11] jointly optimizes CLIP's vision and text representations. However, they both neglect the importance of fine-grained image-text alignment.

The most similar approaches to our own are MasQCLIP and MAFT+. Our method differs by employing a novel text reconstruction task to efficiently fine-tune a VLM-based mask classifier for better fine-grained alignment between textual labels and the semantic information within visual features. The resulting, more robust feature representation greatly improves open-vocabulary segmentation performance.

Efficient Adaptation Methods for VLMs can significantly reduce the computational demands required for training these models. Among these approaches, adapter-based methods [44, 45, 46] introduce minimal trainable parameters at strategic locations within the model, whereas prompt tuning [47, 48]

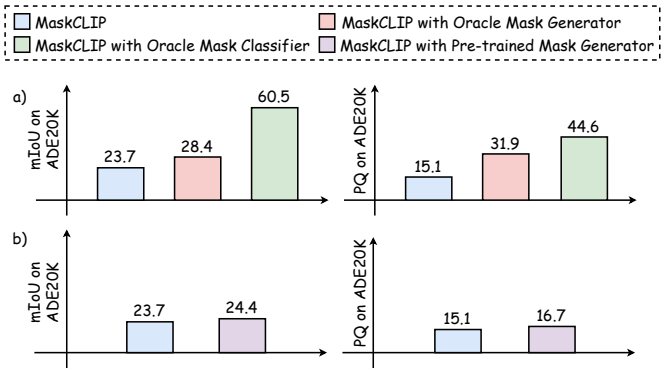


Fig. 2: a) MaskCLIP shows a much greater performance gain with a perfect "oracle" mask classifier than with a perfect "oracle" mask generator, highlighting *mask classification as the main performance bottleneck for open-vocabulary segmentation*. b) Using a pre-trained mask generator performs as well as one re-trained from scratch, indicating that the mask generator can be frozen to enhance training efficiency without performance loss.

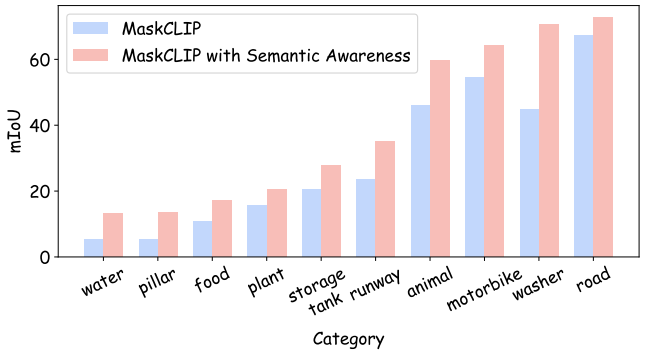


Fig. 3: The incorporation of fine-grained semantic awareness significantly improves MaskCLIP's performance across many out-of-domain classes in ADE20K. Compared to the baseline MaskCLIP model trained on COCO, this approach substantially improves performance, with gains of up to 13.7 points in mIoU. These results highlight the lack of fine-grained alignment as a key factor influencing performance in open-vocabulary segmentation.

injects these parameters into the input space. Low-rank Adaptation (LoRA) and its variants [49, 50] avoid additional parameters by low-rank adapting only the linear layers. Alternatively, using more efficient backbones [51, 52, 53] or adapting the normalization layers [54] or the network biases [55] are also very effective in minimizing learnable parameters. In contrast to previous open-vocabulary segmentation methods that entirely freeze their VLM-based mask classifiers, we explore fine-tuning a minimal subset of the VLM's parameters to improve its performance for open-vocabulary segmentation.

Masked Language Modeling (MLM) is a widely used technique in natural language processing that involves training a model to recover omitted or corrupted tokens within a text. First introduced in BERT [56], MLM has become a foundational pre-training strategy for both language models [57, 58] and vision-language models [59, 60, 61]. Recent research underscores its scalability [18, 62] and effectiveness in adapting to a wide range of downstream tasks [62, 63]. Building on this, MaskedVLM [64] extends MLM to a multimodal setting [65, 66] by reconstructing masked elements in one modality (e.g., vision) using contextual information from another (e.g., language), thereby enabling joint vision-language modeling. Closest to our work is Mask Grounding [67], which incorporates mask-guided text reconstruction as an auxiliary task to encourage fine-grained alignment for referring image segmentation. However, Mask Grounding is limited to tasks where a single text label corresponds to a single ground-truth mask. As such, it is not suitable for open-vocabulary segmentation, which involves multiple labels and multiple masks. In this work, we propose a novel form of text reconstruction as an auxiliary task tailored for open-vocabulary segmentation. Our method enables the model to learn fine-grained semantic alignments in a multi-label, multi-mask context, ultimately leading to improved segmentation performance.

III. Preliminary Analysis

In this section, we carefully analyze the seminal MaskCLIP [9] method to identify key components affecting performance in open-vocabulary segmentation. This process yields critical insights that shape our approach in developing a method that performs effectively and trains efficiently. The results of our analysis are as follows:

❶ **Between mask generation and mask classification, which step is the main performance bottleneck for open-vocabulary segmentation?** To answer this question, we conduct an experiment comparing the effects of a perfect mask generator and a perfect mask classifier on performance. We either replace the mask generator with an "oracle" one that provides ground-truth masks, or replace the mask classifier with an "oracle" one that assigns ground-truth labels on the predicted masks. Fig. 2(a) shows that MaskCLIP with the "oracle" classifier greatly outperforms MaskCLIP with the "oracle" mask generator, achieving an mIoU of 60.5 and a PQ of 44.6 on the ADE150 dataset. This huge improvement of 32.1 points in mIoU and 12.7 points in PQ demonstrates that *mask classification is the main performance bottleneck for open-vocabulary segmentation*.

❷ **Can we improve training efficiency while maintaining model performance by freezing the pre-trained mask generator?** To explore this possibility, we replace the mask generator with a COCO pre-trained version from Mask2Former's model zoo [68], keeping it frozen during training. As depicted in Fig. 2(b), the performance of the pre-trained mask generator matches that of a newly trained one. This suggests that *it is possible to freeze the mask generator, allowing us to enhance training efficiency without degrading performance.*

❸ **What leads to the limited classification performance in existing open-vocabulary segmentation networks?** Upon examining existing networks, we observe that they primarily extract unimodal visual features while neglecting the critical

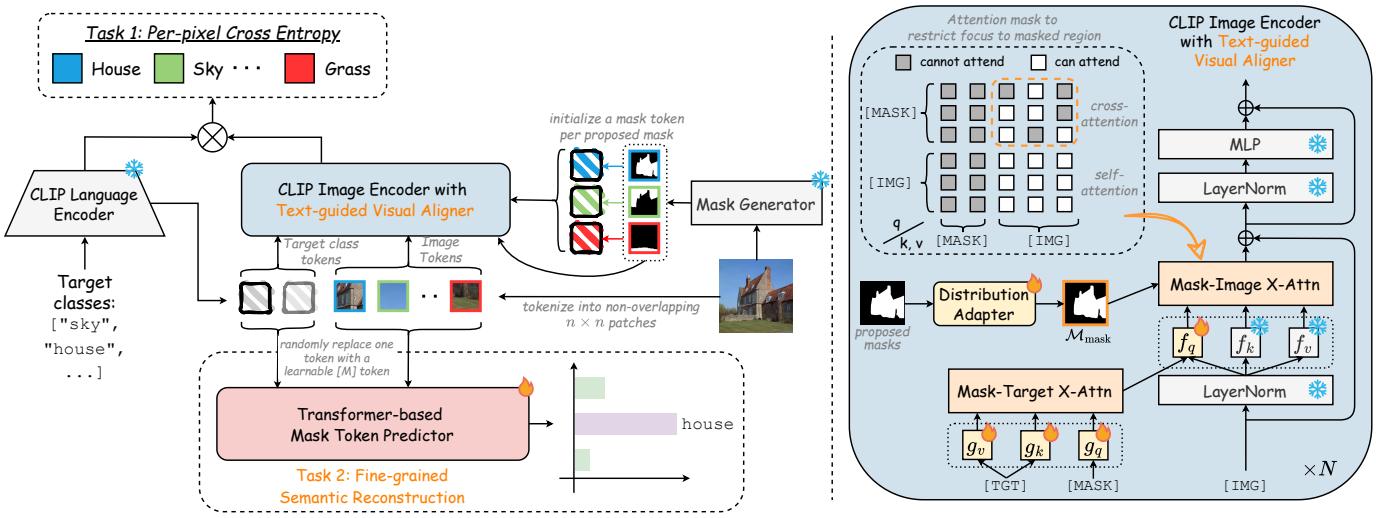


Fig. 4: Overview of FOV-Seg. Guided by the insight that *mask classification is the main performance bottleneck and its weak performance mainly arises from the lack of fine-grained image-text alignment*, FOV-Seg introduces two key innovations for this task. First, it employs the Fine-grained Semantic Reconstruction (FiSeR), an auxiliary semantic reconstruction task that enriches the extracted image features with localized, region-specific semantic information. Second, it introduces the Text-guided Visual Aligner (TeVA), a dedicated network module that injects relevant text-label semantic awareness into the visual encoding process in order to improve the model's ability to better align visual and textual modalities.

semantic information available in text labels [9]. This can be verified by observing the attention maps in Fig. 6, which clearly show the failure of CLIP's text tokens to focus on the relevant regions in the image. This omission is problematic because text information is crucial for aligning pixel-level visual features with word-level semantic information. To validate this hypothesis, we compare MaskCLIP with our proposed variant that explicitly incorporates fine-grained semantics early into the feature extraction process. Specifically, we modify MaskCLIP to perform cross-attention with target-domain class labels before extracting visual features. As shown in Fig. 3, this simple modification substantially improves out-of-domain performance in ADE20K, highlighting *lack of fine-grained alignment as a crucial factor affecting network performance*.

Summary: Our analysis reveals a critical need to enhance mask classification performance in the development of open-vocabulary segmentation networks. Through examination of existing networks, we uncover a key limitation: the lack of fine-grained alignment between extracted visual features and word-level semantic details. This limitation arises from insufficient interaction between visual features and text labels during the feature extraction process and results in visual representations that often fail to capture the fine-grained semantic information essential for accurate segmentation. Building upon these insights, we now present our proposed solution.

IV. Method

In this section, we first describe the architecture of FOV-Seg. Following that, we explain in detail the core components of FOV-Seg, namely 1) Fine-grained Semantic Reconstruction (FiSeR) and 2) Text-guided Visual Aligner (TeVA). Finally, we present the overall training loss function of our method.

Architecture Overview. FOV-Seg is a simple yet effective framework that significantly enhances the performance of existing open-vocabulary segmentation networks. As shown in Fig. 4, following the common design [9, 38, 11], our method operates through the sequential generation and classification of mask proposals. The process begins with a mask generator, which can be any conventional pre-trained segmentation network that produces a set of candidate mask proposals. These proposals are then classified using a vision-language model capable of zero-shot classification. Following previous work, we use CLIP [12] for this purpose. CLIP consists of an image encoder and a language encoder. The image encoder extracts features from image tokens, while the language encoder processes text labels. Zero-shot classification is performed by computing the cosine similarity between mask and label embeddings and assigning each mask to the label with the highest similarity score. To enable regional classification within our task, we follow prior work by using a mask token for each mask proposal. Each mask token attends only to image tokens within its corresponding masked region. These mask tokens are analogous to CLIP's [CLS] token but instantiated per region. Building on the insights from Section III, FOV-Seg improves mask classification by freezing the mask generator and introducing two novel components that facilitate better fine-grained image-text alignment.

1) **Fine-grained Semantic Reconstruction (FiSeR).** An auxiliary task that enriches the semantic representation of image features by requiring the model to predict a randomly masked class label based on the input image and the remaining ground-truth labels. This prediction task encourages the model to capture more detailed and discriminative semantic cues.
2) **Text-guided Visual Aligner (TeVA).** A network mod-

ule that enhances the visual representations of vision-language models by integrating relevant text-label semantics early in the visual encoding process. By conditioning visual processing on target text labels from the outset, the model's ability to associate text with the correct spatial regions is improved. This is achieved by applying cross-attention between image features and class-level label embeddings during visual encoding, which enables more semantically informed feature learning.

Fine-grained Semantic Reconstruction (FiSeR). Inspired by recent advances in masked input modeling [69, 56, 70] and grounded language understanding [67], we propose a novel auxiliary task, FiSeR, to enhance the fine-grained alignment between visual features and their semantic labels. For a given image and its set of class labels, we randomly replace one target class token with a learnable [M] token and task the model with reconstructing the masked label using only the visual content and the remaining unmasked tokens. This masked reconstruction objective encourages the model to learn contextual and discriminative semantic cues grounded in the visual input, thereby promoting more robust feature learning. Although our default configuration masks only a single label per image per training step for simplicity, the method can be readily extended to multiple masked labels.

To implement this, we tokenize the image into non-overlapping $n \times n$ patches [71] and the labels using a Byte Pair Encoding (BPE) tokenizer [12]. These tokens, together with the learnable [M] token replacing one target class token, are fed into a Transformer-based "Predictor." Importantly, the Predictor receives the same number of tokens as in the unmasked case. We do not employ a separate classifier head; instead, the hidden representation of the [M] token is passed through a linear classifier to predict its true identity. Logits are computed as cosine similarities between this predicted embedding and the pre-computed text embeddings of the dataset's target categories. The model is trained with a cross-entropy loss over this dynamic candidate set. This auxiliary task is used only during training and introduces no additional computational overhead during inference. Owing to its versatility, FiSeR can be seamlessly integrated into various VLM-based classifiers to improve performance. To the best of our knowledge, this is the first work to apply masked input modeling to open-vocabulary segmentation tasks.

A brief mathematical formulation for FiSeR is as follows. Given an image $\mathbf{I}$ and its corresponding set of K text labels $\mathbf{T} = \{c_1, \ldots, c_K\}$, the process is defined by:

$$\mathbf{P} = \text{CLIPImageEncoder}(\text{Tokenize}(\mathbf{I})) \tag{1}$$

$$\mathbf{O} = \text{CLIPLanguageEncoder}(\text{Mask}(\text{Tokenize}(\mathbf{T}))) \tag{2}$$

$$\mathbf{z}_{\text{pred}} = \text{Predictor}(\text{Concat}([\mathbf{O}, \mathbf{P}])) \tag{3}$$

$$\mathbf{e}_i = \text{CLIPLanguageEncoder}(\text{Tokenize}(c_i)) \tag{4}$$

$$l_i = \text{sim}(\mathbf{z}_{\text{pred}}, \mathbf{e}_i) \tag{5}$$

$$\mathcal{L}_{\text{FiSeR}} = \mathcal{L}_{\text{CE}}(\text{softmax}([l_1, \ldots, l_K]), \mathbf{y}_{\text{gt}}) \tag{6}$$

where $\mathbf{P}$ and $\mathbf{O}$ are the visual and unmasked text features, respectively. The "Predictor" outputs an embedding $\mathbf{z}_{\text{pred}}$ for the masked position. For each candidate class label $c_i \in \mathbf{T}$, a corresponding text embedding $\mathbf{e}_i$ is used. The logits l_i are computed via cosine similarity (sim) between $\mathbf{z}_{\text{pred}}$ and $\mathbf{e}_i$. Finally, the cross-entropy loss $\mathcal{L}_{\text{CE}}$ is calculated against the one-hot ground-truth label $\mathbf{y}_{\text{gt}}$.

Text-guided Visual Aligner (TeVA). Current open-vocabulary segmentation methods typically rely on mask attention [9] for extracting regional information. However, this approach fails to leverage the semantic richness contained in text labels during visual feature extraction. This omission is problematic because textual information plays an essential role in aligning visual features with semantic content. To address this limitation, we propose TeVA, an innovative multimodal attention mechanism. Crucially, the module is integrated into every block of the CLIP image encoder. It replaces the original mask-token self-attention, enabling iterative, text-guided refinement of mask features during the visual encoding process.

Our approach involves two complementary steps. First, in the *Mask-Target Cross-Attention* step, the mask tokens cross-attend with target class tokens generated by CLIP's language encoder to infuse semantic understanding into the mask tokens, enabling them to capture contextually relevant information. Second, in the *Mask-Image Cross-Attention* step, these semantically-aware mask tokens cross-attend with tokens within the masked image regions to extract all task-specific and contextually relevant information.

Before applying the cross-attention mechanism, we introduce a lightweight convolutional Distribution Adapter to mitigate the substantial distributional gap between the mask proposals and CLIP's expected input—a challenge arising from the mismatch between CLIP's training data and the nature of mask proposals, as noted in prior work [34]. Specifically, CLIP's visual encoder is pre-trained on natural images rich in texture and semantic detail, whereas the generated mask proposals are typically binary or low-entropy spatial maps that differ markedly in structure. Feeding these masks directly into the encoder can therefore lead to suboptimal feature representations. The proposed adapter serves as a learned translator, transforming the raw masks into feature distributions that are more compatible with CLIP's visual encoder.

Mathematically, TeVA is computed as follows: Given m mask tokens $[\texttt{MASK}] \in \mathbb{R}^{m \times C}$, n image tokens $[\texttt{IMG}] \in \mathbb{R}^{n \times C}$, t target class tokens $[\texttt{TGT}] \in \mathbb{R}^{t \times C}$, CLIP's query, key, value projections f_q, f_k, f_v, randomly initialized query, key, value projections g_q, g_k, g_v for *Mask-Target Cross Attention* and Softmax operator σ,

$$\text{TeVA}([\texttt{MASK}], [\texttt{IMG}], [\texttt{TGT}]) = \sigma(\hat{\mathbf{q}}_{\text{mask}} \mathbf{k}_{\text{img}}^T + \mathcal{M}_{\text{mask}}) \cdot \mathbf{v}_{\text{img}}, \tag{7}$$

$$\hat{\mathbf{q}}_{\text{mask}}, \mathbf{k}_{\text{img}}, \mathbf{v}_{\text{img}} = f_q(\hat{[\texttt{MASK}]}), f_k([\texttt{IMG}]), f_v([\texttt{IMG}]), \tag{8}$$

$$\hat{[\texttt{MASK}]} = \sigma(\mathbf{q}_{\text{mask}} \mathbf{k}_{\text{tgt}}^T) \cdot \mathbf{v}_{\text{tgt}}, \tag{9}$$

$$\mathbf{q}_{\text{mask}}, \mathbf{k}_{\text{tgt}}, \mathbf{v}_{\text{tgt}} = g_q([\texttt{MASK}]), g_k([\texttt{TGT}]), g_v([\texttt{TGT}]), \tag{10}$$

where $\mathcal{M}_{\text{mask}} \in \mathbb{R}^{m \times n}$ is a binary attention mask that restricts each mask token to attend only to the image patches it covers.

Formally, its elements are defined as:

$$\mathcal{M}_{\text{mask}}(i,j) = \begin{cases} 0, & \text{if mask}_i \text{ contains any pixel from patch}_j, \\ -\infty, & \text{otherwise.} \end{cases} \tag{11}$$

In this formula, mask_i refers to the binary attention mask corresponding to the i-th mask token and patch_j refers to the j-th image patch from the input image. Self-attention for image tokens is omitted here for brevity, as it remains unchanged from the original CLIP model.

Synergistic Interaction between FiSeR and TeVA. While FiSeR and TeVA both leverage textual information to improve feature representation, they serve distinct yet synergistic roles within our framework. TeVA is a core architectural enhancement to the visual encoder. By integrating a multimodal attention mechanism into every transformer block, it fundamentally changes how mask features are processed during the forward pass. Its purpose is to create semantically-rich visual representations by iteratively guiding the feature extraction process with text. As an architectural component, TeVA is active during both training and inference. In contrast, FiSeR is an auxiliary training objective designed to improve the quality of the learned representations. It is not part of the final model architecture, but rather introduces a supplementary pretext task during training where the model must reconstruct masked text labels from visual and semantic context. This task acts as a powerful regularizer, compelling the model to learn a more robust and fine-grained alignment between the visual features (which are being refined by TeVA) and their corresponding text labels. In essence, TeVA provides the architectural *pathway* for text to guide vision, while FiSeR provides the learning *discipline* to ensure that the resulting representations are deeply and contextually aligned. The "Predictor" module for FiSeR is discarded after training, leaving no computational overhead at inference time. This synergistic combination allows the TeVA-enhanced encoder to not only process textual guidance but also to develop a more sophisticated understanding of the visual-semantic space, leading to superior segmentation performance.

Efficient Knowledge-Preserving Training. Although foundation models like CLIP possess the open-vocabulary knowledge, they often require fine-tuning to adapt to new distributions. To guide our adaptation method, we follow best practices from previous works [54, 38], strategically fine-tuning only two parameter sets: those introduced for TeVA and the query projection layers within CLIP. All other layers in the mask classifier remain frozen. As shown in Table IV(a), our careful optimization approach is critical for enabling efficient and effective adaptation to new domains, significantly outperforming both fully fine-tuned and frozen CLIP models.

Loss Function. The total loss function $\mathcal{L}$ is a weighted combination of the auxiliary reconstruction loss $\mathcal{L}_{\text{FiSeR}}$ from our FiSeR task and the primary cross-entropy loss $\mathcal{L}_{\text{CE}}$ for mask classification:

$$\mathcal{L} = \lambda_{\text{FiSeR}}\mathcal{L}_{\text{FiSeR}} + \lambda_{\text{CE}}\mathcal{L}_{\text{CE}}. \tag{12}$$

In this work, we set the weights $\lambda_{\text{FiSeR}} = 1$ and $\lambda_{\text{CE}} = 2$.

V. Experiments

In this section, we first describe the datasets (Sec. V-A) and evaluation metrics (Sec. V-B) used. Next, we describe our implementation details (Sec. V-C). Then, we quantitatively and qualitatively compare our method with leading open-vocabulary segmentation methods (Sec. V-D). Finally, we carefully ablate our proposed method (Sec. V-E).

A. Training and Evaluation Datasets

We train our method on the COCO-Panoptic *train* dataset [73] and evaluate its performance using the COCO, Mapillary Vistas [74], ADE20K [75] and PASCAL Context [76] *val* datasets. Splitting into *train* and *val* datasets with distinct labels is the standard practice in open-vocab. segmentation. Note that ADE20K has two subsets, ADE150 and ADE847, containing 150 and 847 classes, respectively. Similarly, PASCAL Context has two subsets, PC59 and PC459, containing 59 and 459 classes, respectively.

B. Evaluation Metrics

We evaluate our method using two main metrics: Panoptic Quality (PQ) for panoptic segmentation and mean intersection-over-union (mIoU) for semantic segmentation. mIoU measures the average overlap between the predicted mask and the ground truth across all classes, while PQ measures the overall quality of a panoptic segmentation by combining semantic and instance segmentation accuracy.

C. Implementation Details

Our method is implemented within the Detectron2 framework, built upon a Mask R-CNN baseline trained on the COCO-Panoptic dataset. The architecture integrates a pre-trained ViT-L/16-336 CLIP model as the *mask classifier* and a Swin-B Mask2Former as the *mask generator*. The generator provides class-agnostic mask proposals, while the classifier performs semantic assignment.

Architectural Choice and Design Rationale. Our model's design follows standard practices to allow for a fair and transparent evaluation against other approaches.

Mask Classifier. For the classification branch, we adopt the ViT-L/16-336 CLIP model as the visual-text encoder. This choice aligns with seminal two-stage vision-language segmentation methods such as MaskCLIP [9] and MasQCLIP [38], enabling direct and fair comparisons. While recent works [11, 10] employ stronger backbones such as ConvNeXt-Large, our competitive performance demonstrates that the improvements stem primarily from the proposed FiSeR and TeVA modules rather than from reliance on a more powerful backbone.

Mask Generator. For mask proposal generation, we employ a Swin-B Mask2Former model, using only its generated mask proposals without class predictions. As shown in Table IV(e), when using a weaker ResNet-50 backbone comparable to prior works, our model already surpasses the baselines. The stronger Swin-B backbone is ultimately used in our main experiments to showcase the full potential of our framework while maintaining fairness in baseline comparisons.

TABLE I: Comparison with leading open-vocabulary panoptic segmentation methods. * indicates close-vocabulary evaluation. **Bold** indicates best.

Method	COCO*		ADE150		Mapillary		ADE847	PC59	PC459
	PQ	mIoU	PQ	mIoU	PQ	mIoU	mIoU	mIoU	mIoU
MaskCLIP [9]	-	-	15.1	23.7	-	-	8.2	45.9	10.0
FreeSeg [72]	-	-	16.3	24.6	-	-	-	-	-
ODISE [39]	55.4	65.2	22.6	29.9	14.2	-	11.1	57.3	14.5
MasQCLIP [38]	48.5	62.0	23.3	30.4	-	-	10.7	57.8	18.2
FC-CLIP [10]	54.4	63.7	26.8	34.1	18.2	27.9	14.8	58.4	18.2
MAFT+ [11]	-	-	27.1	36.1	-	-	15.1	59.4	21.6
FOV-Seg (Ours)	**56.6**	**67.4**	**28.4**	**37.1**	**19.1**	**30.1**	**16.5**	**62.5**	**23.9**

TABLE II: Efficiency and performance of leading open-vocabulary segmentation methods. Inference is benchmarked on both ADE150 and ADE847, while performance is evaluated on the standard ADE150 benchmark.

Method	Training		Trainable	ADE847 Inference		ADE150 Inference		ADE150	
	GPU Hrs↓	Iters(K)↓	Params(M)↓	FPS↑	Mem (GB)↓	FPS↑	Mem (GB)↓	PQ↑	mIoU↑
MaskCLIP	17	6	63	2.83	10.8	3.04	10.8	15.1	23.7
ODISE	4760	369	28	0.39	23.2	0.41	23.2	22.6	29.9
MasQCLIP	17	5	25	3.06	9.9	3.48	9.9	23.3	30.4
FC-CLIP	424	369	21	2.54	17.1	2.71	17.1	26.8	34.1
MAFT+	224	60	220	2.75	13.7	2.94	13.7	27.1	36.1
FOV-Seg (Ours)	45	25	25	2.42	10.2	2.63	10.2	28.4	37.1

Distribution Adapter. Complementing these components, the Distribution Adapter bridges the raw mask proposals and the CLIP-based semantic space. It is a lightweight, two-layer convolutional module composed of sequential 3×3 convolutional layers (stride = 1). The first layer projects low-dimensional mask features into a richer latent space, followed by a non-linear activation to enhance representational capacity. The second layer refines these features and condenses them into a single-channel representation. Despite its simplicity, the adapter effectively harmonizes spatial mask structures with CLIP's feature distribution, improving mask-text correspondence without introducing significant computational overhead.

Data Preparation and Augmentation. We follow standard Mask R-CNN [4] conventions for data preprocessing and augmentation. Input images are resized to 896×896, and CLIP's positional embeddings are adjusted via bilinear interpolation. The text inputs consist of the category names for each dataset. During training, we apply Large-Scale Jittering augmentation by sampling random scales between 0.1 and 2.0, followed by a fixed crop of 1024×1024. During inference, images are resized with the shorter side set to 800 and the longer side capped at 1333 pixels, following standard Mask R-CNN practice.

Training Configuration. The model is optimized using AdamW with a learning rate of 1×10^{-4}, a weight decay of 0.05, and a learning rate multiplier of 0.1 for the backbone. Training is conducted for 25,000 iterations with a batch size of 16. Unless otherwise noted, all hyperparameters follow those used in MaskCLIP to ensure consistent experimental settings.

Handling Variable Class Numbers. Both the FiSeR and TeVA modules must handle varying numbers of target classes per image. To enable batch-level parallelization, we adopt a padding and masking strategy. For each batch, we identify the maximum number of classes $K_{\max}$, pad the text token sequences of other images with a special [PAD] token, and apply an attention mask to ignore these tokens during attention computations. During loss computation, logits corresponding to padded tokens are masked out so that the FiSeR loss is computed only over valid classes. This approach preserves computational efficiency while maintaining the semantic integrity of class-specific predictions.

D. Main Results

In this subsection, we quantitatively and qualitatively compare our method against other leading approaches using the COCO [73], ADE20K [75] PASCAL [76] and Mapillary Vistas [74] datasets.

Open-Vocabulary Panoptic Segmentation. Tab. I shows that our method, FOV-Seg, outperforms both two-stage and one-stage approaches across various panoptic and semantic segmentation datasets. Compared to two-stage methods like MaskCLIP and MasQCLIP, FOV-Seg achieves a PQ improvement of up to 13.3 points and mIoU improvement of up to 6.7 on ADE150. Compared to one-stage methods like ODISE, FC-CLIP and MAFT+, FOV-Seg attains a PQ improvement of up to 5.8 points and mIoU improvement of up to 7.2 on both the indoor ADE150 and outdoor Mapillary Vistas datasets, establishing itself as the new state-of-the-art in this domain.

Open-Vocabulary Semantic Segmentation. Many leading open-vocabulary semantic segmentation methods are trained on COCO-Stuff [77], which offers richer annotations for se-

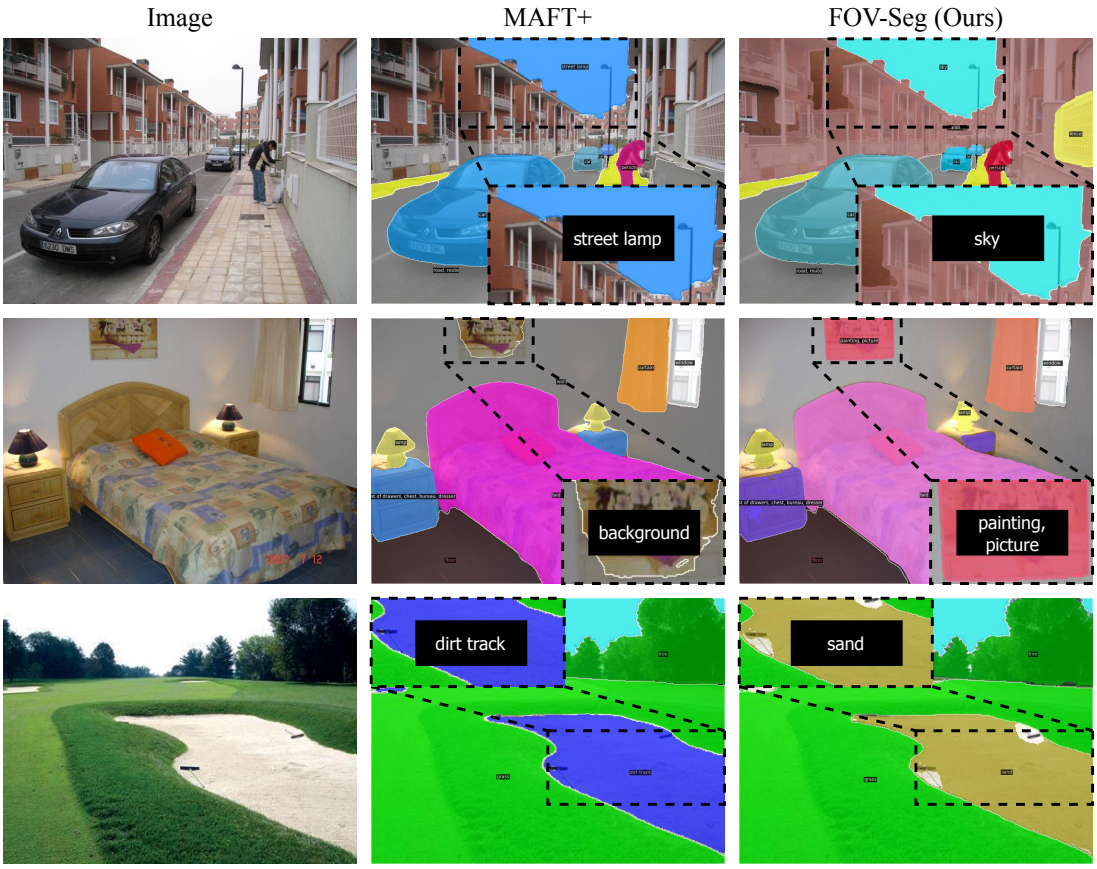


Fig. 5: Qualitative comparison on open-vocab. panoptic segmentation. Unlike MAFT+, which often misses the predictions of certain objects, our method produces more masks and achieves higher class prediction accuracy. Zoom-in for better view.

TABLE III: Comparison with leading open-vocabulary semantic segmentation methods. **Bold** indicates best.

Method	ADE150 mIoU	ADE847 mIoU	PC59 mIoU	PC459 mIoU
OVSeg [34]	29.6	9.0	57.7	15.7
SAN [35]	33.3	13.7	60.2	17.1
SED [37]	35.2	13.9	60.6	22.6
CAT-Seg [36]	37.9	16.0	63.3	23.8
FOV-Seg (Ours)	**38.0**	**16.3**	**63.8**	**24.5**

mantic tasks. To ensure a fair comparison, we train our model using the same COCO-Stuff dataset. As shown in Tab. III, our method outperforms all previously leading approaches under this setting. Notably, our model achieves this despite being designed for the more challenging task of panoptic segmentation. Compared to CAT-Seg, the previous best-performing method, FOV-Seg achieves improvements of 0.1, 0.3, 0.5, and 0.7 mIoU on ADE150, ADE847, PC59, and PC459, respectively.

Efficiency Analysis. As shown in Tab. II, FOV-Seg is designed for efficient training and inference. It requires only 45 GPU hours for training, a 5× reduction compared with MAFT+, and maintains a compact 10.2 GB memory footprint. The model also demonstrates stable inference performance. It achieves 2.63 FPS on the ADE150 dataset, with only a minor decrease to 2.42 FPS on the much larger ADE847 dataset. This slight drop in speed is a common trend also observed in competing methods. Our approach maintains this stability, even with a large number of classes, because we precompute and cache the text embeddings for each benchmark's fixed vocabulary. This one-time process eliminates the need to run the language encoder for every image, making our computational pipeline largely independent of vocabulary size. The few vocabulary-dependent operations—specifically, the TeVA module's cross-attention and the final classification layer—are computationally trivial, constituting only approximately 0.09% of total FLOPs. Consequently, the impact on performance is negligible. The minor drop in speed stems only from interacting with a larger set of cached embeddings. Similarly, the total inference memory remains constant because additional memory required to store embeddings for 847 classes versus 150 is negligible compared to the memory footprint of the main vision encoder. Furthermore, our model is far more memory-efficient than competitors like FC-CLIP and MAFT+. These methods rely on heavy inference pipelines with additional refinement modules and geometric ensembling, which drive up memory costs. In contrast, our streamlined approach omits these high-overhead components. This design substantially reduces resource usage while preserving accuracy, enabling our method to deliver state-of-the-art performance with sig-

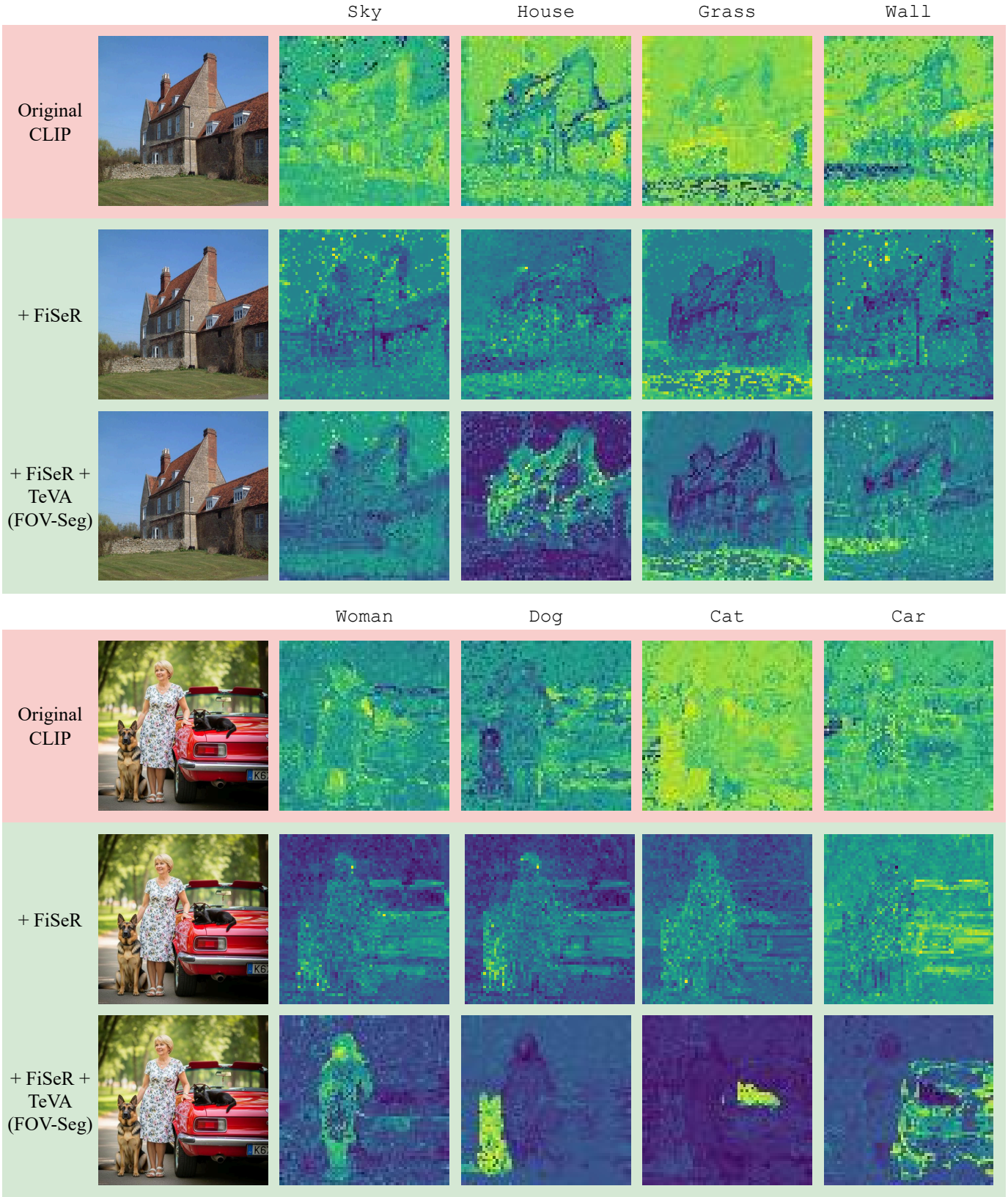


Fig. 6: Visualization of cross-attention maps between text and image tokens extracted by the original CLIP and our meticulously fine-tuned CLIP in FOV-Seg. As demonstrated, our proposed fine-grained text-image alignment methods are crucial for redirecting network's attention to relevant image regions, enabling FOV-Seg to significantly outperform existing open-vocabulary segmentation networks.

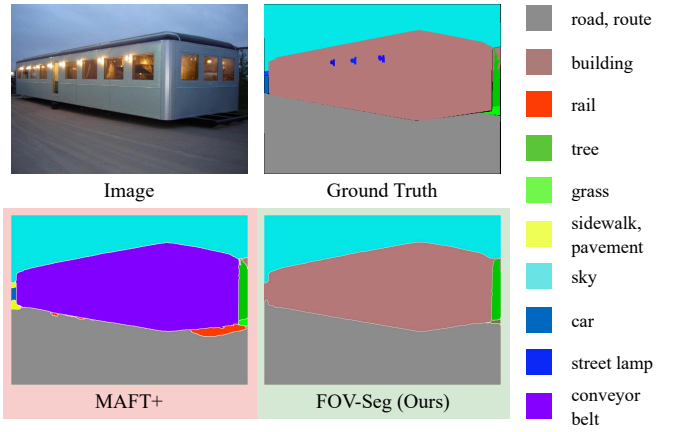


Fig. 7: Qualitative comparison on open-vocab. semantic segmentation. Unlike MAFT+, our method accurately identifies buildings with uncommon shapes and textures while avoiding false predictions, such as misclassifying objects as *rail*.

Fig. 8: Robustness of FOV-Seg towards ambiguous and noisy input.

nificantly lower computational demands.

Qualitative Results. In Fig. 5 and Fig. 7, we present some mask predictions of FOV-Seg on ADE150 for both panoptic segmentation and semantic segmentation. Compared to MAFT+, the previous best method for open-vocabulary segmentation, FOV-Seg generates more masks and predicts mask classes more accurately.

E. Ablations

For a fair comparison, all experiments here are run with a training schedule of 10000 iterations, causing the results to be different from Tab. I.

Importance of FOV-Seg for better fine-grained image-text alignment. To demonstrate the advantages of our approach, Fig. 6 provides a comparative visualization of cross-attention maps between text and image tokens. These maps are generated from both the original CLIP model and our proposed FOV-Seg, which incorporates a carefully fine-tuned version of CLIP. The visualization highlights a key limitation of the original CLIP architecture: its inability to capture fine-grained details, as it was pre-trained on image-level objectives rather than tasks requiring precise localization. As a result, it struggles to accurately identify relevant entities and performs suboptimally in open-vocabulary segmentation tasks. In contrast, our method effectively addresses this shortcoming by enhancing attention to relevant image regions, leading to significantly improved segmentation performance.

Robustness of FOV-Seg towards ambiguous and noisy input. To validate its performance in real-world scenarios, we tested FOV-Seg on ambiguous and noisy inputs. As shown in Figure 8, our model successfully resolves semantic ambiguity by correctly segmenting fine-grained categories, such as different dog breeds that share similar visual characteristics. It also demonstrates strong robustness to data degradation, maintaining decent segmentation accuracy even in challenging low-light and high-blur conditions.

Importance of Fine-grained Semantic Reconstruction (FiSeR) and Text-guided Visual Aligner (TeVA). In Tab. IV(a), we incrementally integrate our proposed modules into the baseline model, which consists of a frozen pre-trained mask generator and a CLIP model with only its query projection layer tuned. This is a simple but strong baseline as it already outperforms some recent methods like MaskCLIP [9] and MasQCLIP [38]. However, our proposed FiSeR module still significantly outperforms this baseline, achieving a large improvement of +2.3 PQ. Additionally, when combined with TeVA, the performance further improves by +3.2 PQ, resulting in a final model that reaches 27.8 PQ, achieving a new state-of-the-art for this task.

Effect of tuning components other than the CLIP mask classifier. To efficiently preserve CLIP's pre-trained knowledge while giving it the flexibility to adapt to new distributions, we optimize only a small subset of CLIP's parameters. In Tab. IV(b), we ablate the effect of tuning additional pre-trained modules, namely the mask generator and language encoder. The results reveal a sharp performance decline after these adjustments. This outcome validates the benefits of our optimization strategy and underscores the importance of careful parameter adjustment for cross-domain generalization.

Effect of tuning different parameters within the CLIP mask classifier. In Tab. IV(c), we examine the impact of adjusting various parameters in the mask classifier. As indicated in the table, tuning all parameters results in inferior performance, whereas tuning fewer parameters generally leads to better outcomes. In comparison to adjusting the query projection layer f_q, tuning all parameters or MLP parameters results in a performance decline of 10.6 PQ and 8.0 PQ, respectively. Among the different projection layer adjustments, our method of tuning f_q performs the best.

Compatibility with different mask classifiers. In Tab. IV(d), we investigate the compatibility of our method with respect to various VLM backbones. As shown, FOV-Seg is able to work well with different VLM-based classifiers, allowing it to benefit easily from future advancements in VLM backbones.

TABLE IV: **Ablation experiments**. All experiments here are run with a shorter training schedule of 10000 iterations, causing the results to be different from Tab. I. The entries marked in gray are the same, which is our proposed FOV-Seg.

(a) Importance of Fine-grained Semantic Reconstruction (FiSeR) and Semantic Augmenter (TeVA) in FOV-Seg. Both greatly improves perf.

Method	PQ
Baseline	24.6
+ FiSeR	26.9 (+2.3)
+ TeVA	25.6 (+1.0)
+ FiSeR + TeVA	**27.8** (+3.2)

(b) Effect of tuning components other than the CLIP mask classifier. Tuning language encoder and mask generator do not improve performance.

Parameters Tuned	PQ
FOV-Seg	**27.8**
+ tune language encoder	25.1 (-2.7)
+ tune mask generator	22.5 (-5.3)

(c) Effect of tuning different parameters within the CLIP mask classifier. Tuning only the CLIP's query projection layer yields the best performance.

CLIP Parameters Tuned	PQ
f_q	**27.8**
f_k	27.1 (-0.7)
f_v	23.1 (-4.7)
f_q, f_k	26.1 (-1.7)
f_q, f_v	24.7 (-3.1)
f_q, f_k, f_v	23.5 (-4.3)
MLP	19.8 (-8.0)
All	17.2 (-10.6)

(d) Compatibility with different mask classifiers. FOV-Seg is compatible with various VLM mask classifiers.

Mask Classifier Backbone	PQ	mIoU
ViT-B-16	25.8	34.1
ViT-L-14-336	**27.8**	**36.4**
EVA01-g-14-plus	27.1	36.3

(e) Compatibility with different mask generators. FOV-Seg is compatible with different mask generators.

Mask Generator Backbone	PQ	mIoU
ResNet-50	26.3	36.0
Swin-T	26.5	36.7
Swin-B	**27.8**	**36.4**

(f) Effect of using different number of layers in Distribution Adapter. Two layers give the best performance.

Number of Layers	PQ
0 (w/o Dist. Adapter)	27.2
1	27.4 (+0.2)
2	**27.8** (+0.6)
3	27.5 (+0.3)

(g) Comparison with the leading parameter-efficient fine-tuning method, LoRA [49]. Our query projection tuning strategy consistently outperforms LoRA across all ranks.

Method	Rank	PQ
Tune f_q	-	**27.8**
LoRA	256	25.6 (-2.2)
LoRA	128	25.9 (-1.9)
LoRA	64	26.5 (-1.3)

(h) Preservation of CLIP's Pretrained Knowledge after fine-tuning. The referring segmentation performance of CLIP model in FOV-Seg does not degrade after fine-tuning.

Case	RefCOCO	RefCOCO+
before tuning	23.9	25.0
after tuning	**24.6** (+0.7)	**25.9** (+0.9)

(i) Importance of Mask-Image and Mask-Target X-Att.

Method	PQ
Baseline	24.6
+ Mask-Image X-Att	24.9
+ Mask-Target X-Att	25.2
+ TeVA (Mask-Target X-Att + Mask-Image X-Att)	**25.6**

(j) FOV-Seg Perf. with Oracle Mask Generator (OMG).

Method	PQ	mIoU
MaskCLIP	15.1	23.7
MaskCLIP with OMG	28.4	31.9
FOV-Seg	27.8	36.4
FOV-Seg with OMG	**34.0**	**41.7**

Compatibility with different mask generators. To demonstrate our method's compatibility with various mask generators, we conduct ablation studies using different pre-trained mask generators. As shown in Tab. IV(e), all tested mask generators produce meaningful results. While stronger mask generators show slight improvements, the overall performance gains are minimal. This observation further confirms that mask classification is the main bottleneck for this task, validating our approach of focusing on this aspect.

Effect of using different number of layers in Distribution Adapter. In Tab. IV(f), we evaluate the sensitivity of Distribution Adapter in TeVA to different number of convolutional layers. We observe that using two layers achieves the best performance and adhere to this design choice.

Comparison with Other Efficient Fine-Tuning Methods. Low-Rank Adaptation (LoRA) [49] is widely used for efficiently fine-tuning pre-trained networks in transfer learning. In Tab. IV(g), we compare our query projection tuning strategy with LoRA. Following common practice, we apply LoRA to the attention projection layers of the CLIP model [12]. As shown in the table, our method significantly outperforms LoRA across different ranks, demonstrating its effectiveness in fine-tuning CLIP for open-vocabulary segmentation.

Preservation of CLIP's Pre-trained Knowledge after fine-tuning. To show that CLIP's internal knowledge is preserved after fine-tuning, we compare the original CLIP backbone's performance with our adapted version on referring image segmentation [67]. As shown in Tab. IV(h), performance remains unchanged after adaptation. This is possible because of FOV-Seg's minimal adaptation approach, which restricts weight updates to a select few strategically chosen parameters and involves only a small number of tuning iterations.

Importance of Mask-Image and Mask-Target Cross Attention in TeVA. We conduct an ablation study to analyze the individual contributions of our two proposed cross-attention modules in TeVA. As presented in Tab. IV(i), incorporating either the Mask-Image or the Mask-Target cross-attention module individually yields performance gains over the base-

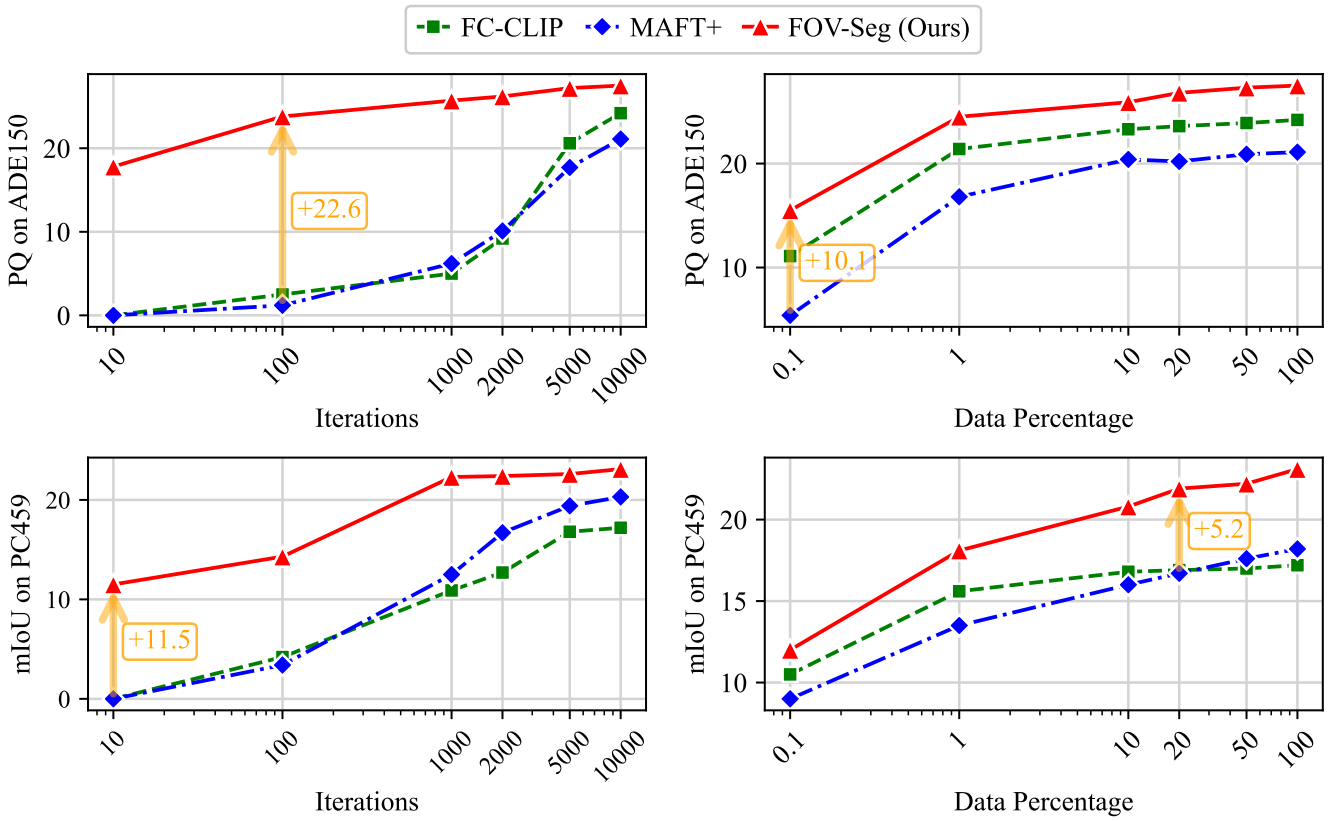


Fig. 9: Effect of training length and data size on model performance. Our method consistently outperforms other leading methods across all training schedules and data sizes.

line. Combining both modules results in the best performance, reaching a PQ of 25.6. This confirms that both components are complementary and essential for achieving the optimal performance of our model, effectively aligning textual descriptions with both the mask regions and the broader image context.

FOV-Seg Performance with Oracle Mask Generator. To isolate the performance of the mask classification component, we evaluate our method with an Oracle Mask Generator that uses ground-truth masks. As shown in Tab. IV(j), substituting the standard mask generator with the oracle generator leads to a significant performance surge for both FOV-Seg and MaskCLIP. In particular, FOV-Seg's performance improves from 27.8 to 34.0 in PQ and from 36.4 to 41.7 in mIoU.

Robustness to compute- and data-limited scenarios. Fig. 9 demonstrates FOV-Seg's robustness under limited training iterations and data sizes, with consistent advantages over leading methods. In compute-constrained settings, FOV-Seg achieves a +22.6 improvement on ADE150 over MAFT+ with just 100 training iterations. This lead is maintained even at 10 iterations, with our method securing an +11.5 mIoU advantage over MAFT+ on PC459. FOV-Seg also excels in data-scarce scenarios. When trained on a mere 0.1% sample of the COCO-Panoptic dataset, it surpasses MAFT+ by +10.1 PQ on ADE150. Similarly, with 20% of the data, our method outperforms MAFT+ by +5.2 on PC459. This highlights FOV-Seg's efficiency as an open-vocab. segmentation framework.

VI. Conclusion

In this paper, we rethink the prevailing practice of freezing Vision Language Model (VLM) parameters for Open Vocabulary Segmentation (OVS). Our work stems from the critical insight that mask classification is the primary performance bottleneck. This issue arises because VLMs, pretrained on global representations, fail to capture the localized, region-specific semantics essential for this task. To address this challenge, we propose FOV-Seg, a novel and efficient framework enhancing fine-grained alignment between textual words and visual pixels. FOV-Seg introduces two synergistic innovations tackling the problem from different angles. First, the Text-guided Visual Aligner (TeVA) injects textual guidance early within the visual encoder, conditioning feature extraction on target categories and ensuring the model attends to relevant semantics. Complementing this, the Fine-grained Semantic Reconstruction (FiSeR) auxiliary task enriches visual features by compelling the model to associate specific word-level semantics with corresponding image regions. By pairing TeVA's early guidance with FiSeR's semantic enrichment, our framework adapts the VLM to produce spatially precise, semantically rich features that overcome the mask classification bottleneck. As a result, FOV-Seg achieves new state-of-the-art results on multiple OVS benchmarks while reducing training costs by nearly $5\times$ compared to prior methods.

ACKNOWLEDGMENTS

This work is supported in part by the National Key R&D Program of China under Grant 2021ZD0140407, the National Natural Science Foundation of China under Grants 62321005 and 62276150, and the THU-Bosch JCML.

REFERENCES

[1] J. Wu, X. Li, S. Xu, H. Yuan, H. Ding, Y. Yang, X. Li, J. Zhang, Y. Tong, X. Jiang *et al.*, "Towards open vocabulary learning: A survey," *IEEE TPAMI*, 2024.

[2] C. Zhu and L. Chen, "A survey on open-vocabulary detection and segmentation: Past, present, and future," *IEEE TPAMI*, 2023.

[3] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE TPAMI*, 2017.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017.

[5] F. Codevilla, E. Santana, A. M. López, and A. Gaidon, "Exploring the limitations of behavior cloning for autonomous driving," in *ICCV*, 2019.

[6] M. Toromanoff, E. Wirbel, and F. Moutarde, "End-to-end model-free reinforcement learning for urban driving using implicit affordances," in *CVPR*, 2020.

[7] S. Pate, W. Xu, Z. Yang, M. Love, S. Ganguri, and L. L. Wong, "Natural language for human-robot collaboration: Problems beyond language grounding," *arXiv:2110.04441*, 2021.

[8] H. Ahn, S. Choi, N. Kim, G. Cha, and S. Oh, "Interactive text2pickup networks for natural language-based human–robot collaboration," *IEEE Robotics and Automation Letters*, 2018.

[9] Z. Ding, J. Wang, and Z. Tu, "Open-vocabulary universal image segmentation with maskclip," in *ICML*, 2023.

[10] Q. Yu, J. He, X. Deng, X. Shen, and L.-C. Chen, "Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip," in *NeurIPS*, 2023.

[11] S. Jiao, H. Zhu, J. Huang, Y. Zhao, Y. Wei, and S. Humphrey, "Collaborative vision-text representation optimizing for open-vocabulary segmentation," in *ECCV*, 2024.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[13] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, "Reproducible scaling laws for contrastive language-image learning," in *CVPR*, 2023.

[14] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "Eva-clip: Improved training techniques for clip at scale," *arXiv:2303.15389*, 2023.

[15] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," *IEEE TPAMI*, 2018.

[16] Y. Pan, R. Sun, Y. Wang, W. Yang, T. Zhang, and Y. Zhang, "Purify then guide: A bi-directional bridge network for open-vocabulary semantic segmentation," *IEEE TCSVT*, 2025.

[17] W. Xu, C. Wang, X. Feng, R. Xu, L. Huang, Z. Zhang, L. Guo, and S. Xu, "Generalization boosted adapter for open-vocabulary segmentation," *IEEE TCSVT*, 2025.

[18] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, 2021.

[19] Y. Zheng, H. Yang, and D. Huang, "Deep common feature mining for efficient video semantic segmentation," *IEEE TCSVT*, 2024.

[20] F. Ding and J. Li, "Multi-level collaborative learning for multi-target domain adaptive semantic segmentation," *IEEE TCSVT*, 2024.

[21] Y. Huang, D. Kang, L. Chen, W. Jia, X. He, L. Duan, X. Zhe, and L. Bao, "Card: Semantic segmentation with efficient class-aware regularized decoder," *IEEE TCSVT*, 2024.

[22] Y. Huang, D. Kang, S. Gao, W. Li, and L. Duan, "High-level feature guided decoding for semantic segmentation," *IEEE TCSVT*, 2024.

[23] Q. Ren, S. Lu, Q. Mao, and M. Dong, "Exploring prototype-anchor contrast for semantic segmentation," *IEEE TCSVT*, 2024.

[24] Z. Yang, H. Yu, W. Sun, L. Cheng, and A. Mian, "Domain-invariant prototypes for semantic segmentation," *IEEE TCSVT*, 2024.

[25] G. Zhu, R. Wang, Y. Liu, Z. Zhu, C. Gao, L. Liu, and N. Sang, "An adaptive post-processing network with the global-local aggregation for semantic segmentation," *IEEE TCSVT*, 2024.

[26] L. Chen, H. Dai, and Y. Zheng, "Rafnet: Reparameterizable across-resolution fusion network for real-time image semantic segmentation," *IEEE TCSVT*, 2024.

[27] Y. Zhang, S. Tian, M. Liao, Z. Zhang, W. Zou, and C. Xu, "Fine-grained self-supervision for generalizable semantic segmentation," *IEEE TCSVT*, 2024.

[28] C. You, L. Jiao, L. Li, X. Liu, F. Liu, W. Ma, and S. Yang, "Contour knowledge-aware perception learning for semantic segmentation," *IEEE TCSVT*, 2025.

[29] Z. Guo, L. Bian, H. Wei, J. Li, H. Ni, and X. Huang, "Dsnet: A novel way to use atrous convolutions in semantic segmentation," *IEEE TCSVT*, 2025.

[30] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," in *ICLR*, 2022.

[31] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, "Scaling open-vocabulary image segmentation with image-level labels," in *ECCV*, 2022.

[32] M. Xu, Z. Zhang, F. Wei, Y. Lin, Y. Cao, H. Hu, and X. Bai, "A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model," in *ECCV*, 2022.

[33] J. Ding, N. Xue, G.-S. Xia, and D. Dai, "Decoupling zero-shot semantic segmentation," in *CVPR*, 2022.

[34] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu, "Open-vocabulary semantic segmentation with mask-adapted clip," in *CVPR*, 2023.

[35] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, "Side adapter network for open-vocabulary semantic segmentation," in *CVPR*, 2023.

[36] S. Cho, H. Shin, S. Hong, S. An, S. Lee, A. Arnab, P. H. Seo, and S. Kim, "Cat-seg: Cost aggregation for open-vocabulary semantic segmentation," in *CVPR*, 2024.

[37] B. Xie, J. Cao, J. Xie, F. S. Khan, and Y. Pang, "Sed: A simple encoder-decoder for open-vocabulary semantic segmentation," in *CVPR*, 2024.

[38] X. Xu, T. Xiong, Z. Ding, and Z. Tu, "Masqclip for open-vocabulary universal image segmentation," in *ICCV*, 2023.

[39] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. D. Mello, "Open-vocabulary panoptic segmentation with text-to-image diffusion models," in *CVPR*, 2023.

[40] J. Wang, X. Li, J. Zhang, Q. Xu, Q. Zhou, Q. Yu, L. Sheng, and D. Xu, "Diffusion model is secretly a training-free open vocabulary semantic segmenter," *IEEE TIP*, 2025.

[41] W. Sun, Y. Li, Q. Ye, J. Jiao, and Y. Zhou, "Depth-guided texture diffusion for image semantic segmentation," *IEEE TCSVT*, 2025.

[42] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[43] Y. Li, T. Cheng, W. Liu, and X. Wang, "Mask-adapter: The devil is in the masks for open-vocabulary segmentation," in *CVPR*, 2025.

[44] R. Zhang, J. Han, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, P. Gao, and Y. Qiao, "Llama-adapter: Efficient fine-tuning of language models with zero-init attention," in *ICLR*, 2024.

[45] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2024.

[46] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *IJCV*, 2024.

[47] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *ECCV*, 2022.

[48] X. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *ACL*, 2021.

[49] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *ICLR*, 2022.

[50] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," in *NeurIPS*, 2024.

[51] J.-Y. He, S.-H. Liang, X. Wu, B. Zhao, and L. Zhang, "Mgseg: Multiple granularity-based real-time semantic segmentation network," *IEEE TIP*, 2021.

[52] T. Wu, S. Tang, R. Zhang, J. Cao, and Y. Zhang, "Cgnet: A light-weight context guided network for semantic segmentation," *IEEE TIP*, 2021.

[53] Q. Huang, Z. Jie, L. Ma, L. Shen, and S. Lai, "A pyramid fusion mlp for dense prediction," *IEEE TIP*, 2025.

[54] B. Zhao, H. Tu, C. Wei, J. Mei, and C. Xie, "Tuning layernorm in attention: Towards efficient multi-modal llm finetuning," in *ICLR*, 2024.

[55] H. Cai, C. Gan, L. Zhu, and S. Han, "Tinytl: Reduce activations, not trainable parameters for efficient on-device learning," in *NeurIPS*, 2020.

[56] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NACL: HLT*, 2019.

[57] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language models," *JMLR*, 2024.

[58] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin *et al.*, "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.

[59] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *ICML*, 2023.

[60] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *ICML*, 2021.

[61] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," in *ICLR*, 2020.

[62] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *NeurIPS*, 2020.

[63] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv:2302.13971*, 2023.

[64] G. Kwon, Z. Cai, A. Ravichandran, E. Bas, R. Bhotika, and S. Soatto, "Masked vision and language modeling for multi-modal representation learning," in *ICLR*, 2023.

[65] J. Gao, M. Chen, and C. Xu, "Learning probabilistic presence-absence evidence for weakly-supervised audio-visual event perception," *IEEE TPAMI*, 2025.

[66] Z. Wu, J. Gao, and C. Xu, "Open-vocabulary video scene graph generation via union-aware semantic alignment," *ACM MM*, 2024.

[67] Y. X. Chng, H. Zheng, X. Q. Y. Han, and G. Huang, "Mask grounding for referring image segmentation," in *CVPR*, 2024.

[68] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *CVPR*, 2022.

[69] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, and F. Wei, "Image as a foreign language: Beit pretraining for

all vision and vision-language tasks," in *CVPR*, 2023.
[70] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022.
[71] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.
[72] J. Qin, J. Wu, P. Yan, M. Li, R. Yuxi, X. Xiao, Y. Wang, R. Wang, S. Wen, X. Pan *et al.*, "Freeseg: Unified, universal and open-vocabulary image segmentation," in *CVPR*, 2023.
[73] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.
[74] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *ICCV*, 2017.
[75] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *IJCV*, 2019.
[76] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *CVPR*, 2014.
[77] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in *CVPR*, 2018.

VII. Biography Section

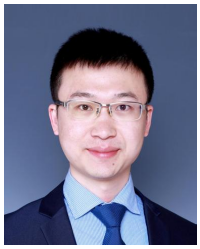
Yong Xien Chng received the BS degree from the School of Computing, National University of Singapore, Singapore. He is currently working toward the PhD degree with the Department of Automation, Tsinghua University, Beijing, China. His research interests include computer vision and deep learning, especially in multimodal foundation model.

Xuchong Qiu received the BS degree in Physics from Wuhan University, in 2014, the MS degree in Robotics from Sorbonne University, in 2016, and the PhD degree in Computer Science from Polytechnic Institute of Paris, in 2021. His research interests include Machine Learning, Computer Vision and Robotics.

Yizeng Han received the BS and PhD degrees from the Department of Automation, Tsinghua University, Beijing, China, in 2018. His current research interests include computer vision and deep learning, especially in dynamic neural networks.

Kai Ding received the MS and PhD degrees in Electrical and Computer Engineering from Technische Universität Dresden in 2016 and 2019, respectively. His research interests include robotics and deep learning.

Wan Ding received the BS degree in Mechatronics from Hubei Polytechnic University, in 2009, the MS and PhD degree in Mechanism and Design Theory from Beijing Jiaotong University, in 2015, the Postdoc research in IGMR, RWTH Aachen University, in 2017. His research interests include Robotics, Mechanism Theory, and Computer Vision.



Gao Huang (Member, IEEE) received the BS degree from the School of Automation Science and Electrical Engineering, Beihang University, in 2009, and the PhD degree from the Department of Automation, Tsinghua University, in 2015. He was was a postdoctoral researcher with the Department of Computer Science, Cornell University, Ithaca, from 2015 to 2018. He is currently an Associate Professor with the Department of Automation, Tsinghua University. His research interests include machine learning and computer vision.