International Journal of Computer Vision (2026) 134:296
https://doi.org/10.1007/s11263-026-02886-0

CMPF: Harmonizing Cross-Model Prior Fusion for Open-Vocabulary Segmentation

Sicheng Zhao[1] · Xi Chen[2] · Hongxun Yao[2] · Haosen Yang[3] · Yanhao Zhang[4] · Sheng Jin[5] · Xiatian Zhu[3] · Haonan Lu[4] · Kui Jiang[2] · Guiguang Ding[6]

Received: 12 December 2025 / Accepted: 8 May 2026

Abstract

Open-vocabulary segmentation poses significant challenges, as it requires segmenting and recognizing objects across an open set of categories in unconstrained environments. Building on the success of powerful vision-language (ViL) foundation models, such as CLIP, recent efforts sought to harness their zero-shot capabilities to recognize unseen categories. Despite notable performance improvements, these models still encounter the critical issue of *generating and recognizing precise mask proposals for unseen categories and scenarios*, resulting in inferior segmentation perfaormance eventually. To address this challenge, we introduce a novel **C**ross-**M**odel **P**rior **F**usion (CMPF) framework, an innovative framework that fuses visual knowledge from a localization foundation model (e.g., SAM) and text knowledge from a ViL model (e.g., CLIP), leveraging their complementary knowledge priors to overcome inherent limitations in mask proposal generation. Taking the ViL model's visual encoder as the feature backbone, we propose Query Injector and Feature Injector to inject the visual localization feature into the learnable queries and CLIP features respectively, within a transformer decoder. In addition, an OpenSeg Ensemble strategy is designed to further improve mask quality by incorporating SAM's universal segmentation masks during inference. To fully exploit pre-trained knowledge while minimizing training overhead, we freeze both foundation models, focusing optimization efforts solely on a lightweight transformer decoder for mask proposal generation – the performance bottleneck. Extensive experiments demonstrate that CMPF advances state-of-the-art results across various segmentation benchmarks, trained exclusively on COCO panoptic data, and tested in a zero-shot manner. Code is available at https://github.com/chenxi52/CMPF.

Keywords Open-Vocabulary Segmentation · Panoptic Segmentation · Vision-Language Model · SAM

Communicated by Hong Liu.

Sicheng Zhao and Xi Chen have contributed equally to this work.

✉ Sicheng Zhao
schzhao@tsinghua.edu.cn

Xi Chen
xichen98cn@gmail.com

Hongxun Yao
h.yao@hit.edu.cn

Haosen Yang
haosen.yang.6@gmail.com

Yanhao Zhang
zhangyanhao@oppo.com

Sheng Jin
Jsh.hit.doc@gmail.com

Xiatian Zhu
eddy.zhuxt@gmail.com

Haonan Lu
luhaonan@oppo.com

Kui Jiang
jiangkui@hit.edu.cn

Guiguang Ding
dinggg@tsinghua.edu.cn

[1] Department of Psychological and Cognitive Sciences, Tsinghua University, Beijing 100084, China

[2] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China

[3] Surrey Institute for People-Centred Artificial Intelligence, University of Surrey, Guildford GU2 7XH, U.K.

[4] OPPO AI Center, Beijing 100006, China

Published online: 29 May 2026

1 Introduction

Image segmentation is a fundamental task in computer vision, enabling a wide range of applications such as object recognition (Cheng et al., 2021), scene understanding (Zhao et al., 2024), medical understanding (Ma et al., 2024), and image manipulation. Traditional methods, however, are often tailored to specific datasets and segmentation tasks, limiting their generalization and adaptability. This results in a significant gap when compared to human visual intelligence, which excels at perceiving and interpreting diverse visual concepts in open-world settings. While many domain adaptation methods (Pan et al., 2020; Zhao et al., 2021, 2024) seek to adapt models for open-domain scenarios, they are fundamentally constrained in their ability to classify unseen or open classes. To bridge this disparity, the concept of open-vocabulary segmentation (Liang et al., 2023; Ding et al., 2022; Xu et al., 2022; Ghiasi et al., 2022; Chen et al., 2023) has emerged as a promising paradigm. Open-vocabulary segmentation leverages large pre-trained Vision-Language (ViL) models (Radford et al., 2021; Jia et al., 2021; Li et al., 2022), which align text and image embeddings in a shared representation space. By mapping text embeddings to visual features, these models enable recognition and segmentation of arbitrary categories, mimicking the human ability to generalize across a vast and dynamic vocabulary of visual concepts.

Existing approaches to open-vocabulary segmentation can be broadly categorized into two paradigms: two-stage and one-stage learning. Two-stage methods (Liu et al., 2024; Xu et al., 2022, 2023a; Liang et al., 2023; Chen et al., 2023), including MaskCLIP (Ding et al., 2023), combine ViL models with mask proposal classifiers. These methods typically rely on class-agnostic mask generation (Cheng et al., 2021), which can lead to inefficiencies in processing and suboptimal classification of segmented regions, particularly for complex or unseen categories. The separation of proposal generation and classification leads to misalignment between the features used for segmentation and those used for semantic understanding. In contrast, one-stage methods, such as FC-CLIP (Yu et al., 2023), adopt CLIP's visual encoder (Radford et al., 2021) as the image backbone to directly generate mask proposals. While it opts for a CNN-based CLIP over a ViT-based one to handle larger input sizes, it frequently overfits to training categories, restricting the model's ability to generalize its mask proposals to unseen classes.

The core limitation of these methods stems from the foundational ViL models they employ, such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021). While these models exhibit strong semantic generalization across unseen categories due to their rich language-based textual representations, the absence of pixel-level annotations restricts their ability to perform precise dense image-text alignment, a critical requirement for high-quality segmentation. This limitation is evident in Fig. 1(left), where the segmentation performance of the open-vocabulary state-of-the-art (SOTA) method (Yu et al., 2023) demonstrates significantly poorer performance on unseen classes compared to seen classes, reducing its practical utility. Further analysis, shown in Fig. 1(right), reveals that existing CLIP-based approaches (Ding et al., 2023; Yu et al., 2023) underperform in both average recall (AR) and mean intersection-over-union (mIoU) metrics for unseen categories, underscoring their reduced recognition ability for such classes. Although recent works have sought to address these challenges through techniques like class-agnostic mask learning (Chen et al., 2023; Ding et al., 2023; Xu et al., 2022; Ghiasi et al., 2022; Xu et al., 2023b) and teacher-student self-training (Xu et al., 2023a), they have yet to resolve the fundamental issue of optimizing mask proposals for unseen classes using ViL's features.

To overcome these limitations, a new perspective from integration with visual knowledge prior of the localization foundation model – Segment Anything Model (SAM) (Kirillov et al., 2023), is pioneering in this study. We propose **CMPF**, a framework that harnesses the spatial capabilities of frozen SAM image encoder to synergistically and efficiently enhance the coarse semantic strength of the frozen CLIP model. By harmonizing spatial localization and semantic reasoning, our approach improves the robustness and generalization of open-vocabulary segmentation in open-world scenarios. This direction aligns well with the goals of the Special Issue on *Ensuring Trustworthiness in Open-World Visual Recognition*, particularly in terms of robustness enhancement, dynamic adaptability to novel categories, and the integration of large vision foundation models for open-world visual recognition. Our method incorporates three key modules: (1) Query Injector, which aggregates local space-aware features from SAM to serve as the spatial queries for corresponding mask regions, enhancing the learnability of queries in a transformer decoder (Cheng et al., 2021). (2) Feature Injector, designed to enrich each pixel's CLIP feature by incorporating comprehensive global spatial information from SAM, improving semantic understanding. (3) OpenSeg Ensemble Module, designed to further boost the quality of mask predictions based on the spatial information injection of SAM during training by ensembling with zero-shot mask proposals from SAM in the inference stage. By synergizing these two powerful foundation models, CMPF establishes a unified approach that enhances spatial precision and general-

[5] S-Lab, Nanyang Technological University, Singapore 637335, Singapore

[6] BNRist, School of Software, Tsinghua University, Beijing 100084, China

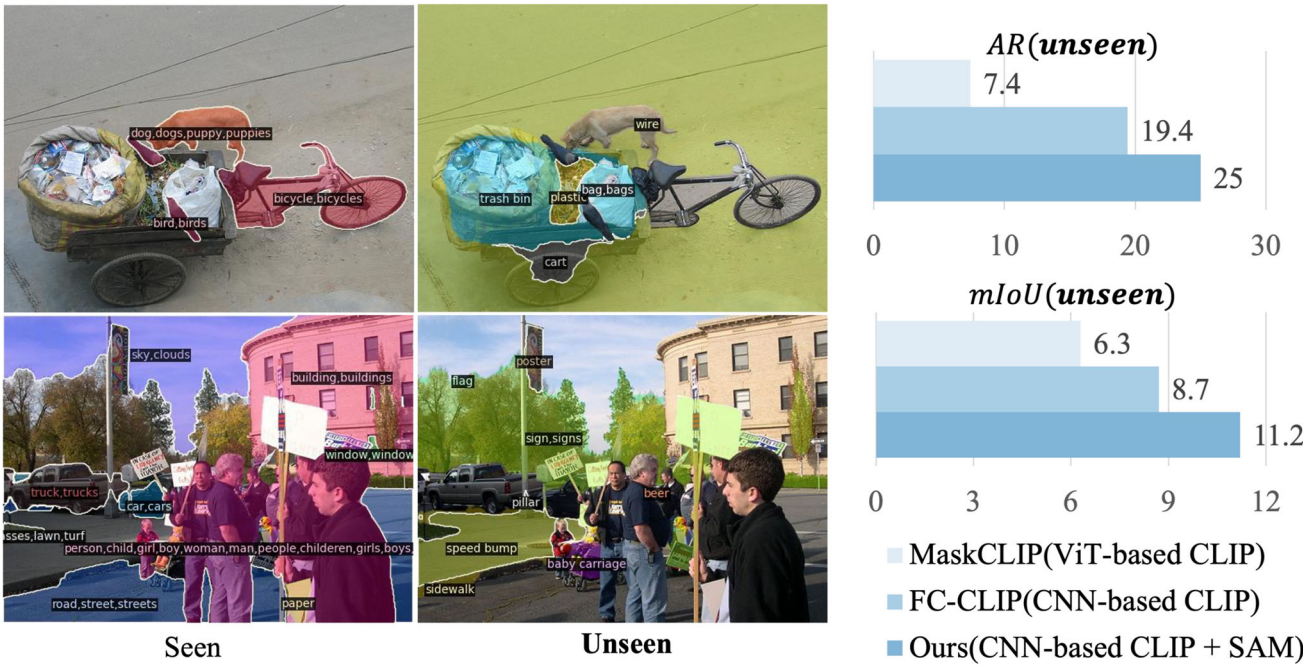

Fig. 1 (Left): Segmentation results of the existing SOTA (Yu et al., 2023) on seen and unseen classes in the PC-459 (Mottaghi et al., 2014) dataset, highlighting significantly poorer performance on unseen classes compared to seen classes. (Right): AR and mIoU scores for unseen classes in PC-459, comparing existing methods, MaskCLIP (Ding et al., 2023) and FC-CLIP (Yu et al., 2023), with our approach. Our method integrates SAM (Kirillov et al., 2023) and CNN-based CLIP (Radford et al., 2021) to enhance performance in open-vocabulary segmentation task

ization, enabling robust segmentation of unseen categories in open-vocabulary scenarios. As demonstrated in Fig. 1(right), CMPF achieves significant improvements in average recall (AR) and mIoU metrics of unseen categories on the challenging PC-459 dataset, validating its effectiveness in addressing the limitations of existing methods and advancing the state of open-vocabulary segmentation for open-word application.

Our contributions can be summarized as follows:

- Addressing an acknowledged limitation in mask proposal quality, we introduce CMPF, a novel framework that fuses knowledge-based priors of foundational models to tackle the open-vocabulary segmentation task effectively.
- We propose three critical components – Query Injector, Feature Injector, and the OpenSeg Ensemble Module – to integrate SAM's visual features into the transformer decoder module, thereby enhancing the final mask generation and generalization capability.
- Extensive experiments on various segmentation tasks demonstrate the superiority of our CMPF in producing high-quality mask proposals and achieving enhanced final performance.

The remainder of this paper is organized as follows. Section 2 reviews related work in open-vocabulary segmentation and foundation models for vision-language tasks. Section 3 formulates the open-vocabulary segmentation problem and introduces our proposed framework. Section 4 details the experimental setup, presents main results across panoptic, semantic, and instance segmentation benchmarks, evaluates generalization to unseen classes, and provides comprehensive analytical studies. Finally, Section 5 concludes the paper.

2 Related Work

2.1 Open-vocabulary Segmentation

Open-vocabulary segmentation aims to segment objects even without seeing those classes during training. Previous approaches (Liu et al., 2024; Ding et al., 2023; Xu et al., 2022, 2023a; Liang et al., 2023; Chen et al., 2023) typically adopt a two-stage process: first using an additional segmentation model to generate class-agnostic mask proposals (Cheng et al., 2021), then establishing interaction between ViL features and these multiple mask proposals. For instance, OPSNet (Chen et al., 2023) combines mask query embeddings with the final-layer CLIP embeddings for out-of-domain learning and applies an IoU branch to filter less informative proposals. Similarly, MaskCLIP (Ding

et al., 2023) integrates class-agnostic masks with learnable mask tokens to interact with CLIP embeddings. While these approaches demonstrate progress, critical challenges remain in tightly coupling segmentation architectures with ViL models, particularly in improving cross-modal alignment and reducing proposal noise for unseen categories.

Alternatively, one-stage approaches (Yu et al., 2023; Wu et al., 2024; Liu et al., 2025) extend ViL models as segmenter backbone, eliminating dedicated segmentation models and addressing proposal overfitting through an end-to-end format. Research such as FC-CLIP (Yu et al., 2023) and CLIPSelf (Wu et al., 2024) indicates that convolutional CLIP models generally exhibit superior generalization capabilities compared to ViT-based (Dosovitskiy et al., 2021) counterparts, primarily due to their capability to handle larger input resolutions effectively–a crucial advantage for segmentation tasks. However, adapting CLIP's image-text pretraining to segmentation requires overcoming its inherent region-level biases.

Consequently, a fundamental issue persists: generating accurate mask proposals for unseen categories and scenarios. This challenge stems from relying on frozen ViL models that lack the capacity for fine-grained pixel-level understanding during mask generation.

2.2 Foundation Models for Vision-Language Tasks

Recent advances in large-scale foundation models, pretrained on extensive datasets, have showcased exceptional zero-shot generalization capabilities. Notably, multi-modal foundation models (Radford et al., 2021; Jia et al., 2021) exhibit strong transfer performance across diverse downstream tasks through efficient strategies. Building upon these developments, researchers have successfully adapted CLIP for dense prediction tasks (Rao et al., 2022) by enhancing multimodal dense-level alignment through prompt learning (Zhou et al., 2022). While this approach effectively bridges image-text alignment to pixel-text correspondence, it currently remains constrained to closed-set class recognition scenarios.

In the realm of the segmentation foundation models, recent research has made significant strides in developing unified foundation models for diverse segmentation tasks, including instance, semantic, panoptic, referring, and part segmentation. Several notable approaches have emerged in this domain: SEEM (Zou et al., 2023a) unifies spatial prompts (points, boxes, scribbles, and masks) and text prompts to a joint image-text representation space, enabling generalization across multiple segmentation tasks. Similarly, X-Decoder (Zou et al., 2023b) achieves task unification through the use of different query types. The SAM (Kirillov et al., 2023) represents a major breakthrough, demonstrating remarkable zero-shot generalization capabilities through large-scale pretraining on the SA-1B dataset. SAM's adaptability has been further enhanced through various prompting strategies, *e.g.*, low-level prompts (Li et al., 2025), and weak supervision (Chen et al., 2024). Recent works (Yang et al., 2024; Yuan et al., 2024) have explored using bounding boxes generated through open-vocabulary detection methods as prompts, combining SAM and ViL model to exploit their complementary strengths in open-vocabulary segmentation. While these advancements are impressive, key challenges remain in achieving fully automatic open-vocabulary segmentation tasks and transitioning from instance segmentation to semantic and panoptic segmentation. In contrast to these approaches, our method introduces an end-to-end framework that simultaneously generates and optimizes location proposals, enabling truly universal segmentation capabilities.

3 Method

In this section, we first lay the groundwork by clearly defining the problem setup, then introduce our CMPF framework that effectively fuses frozen foundation models for open-vocabulary segmentation. The fusion is achieved through two key components: the Query Injector and the Feature Injector, as illustrated in Fig. 2. Finally, we detail our inference strategy, the OpenSeg Ensemble Module, in Fig. 3.

3.1 Problem Definition

Open-vocabulary segmentation involves training with ground-truth masks corresponding to a predefined set of class labels, $\mathbf{C}_{train}$. During testing, the model is evaluated on a distinct set of class labels, $\mathbf{C}_{test}$, which includes novel classes not seen during training. This task requires segmenting images in an open-world context, where the model must categorize pixels into semantic classes for semantic segmentation, identify individual instances for instance segmentation, and address both tasks simultaneously in panoptic segmentation. The notation $\mathbf{C}$ represents either $\mathbf{C}_{train}$ or $\mathbf{C}_{test}$, depending on the training or testing phase.

3.2 Our Approach CMPF

3.2.1 Overall Architecture

Building upon Mask2Former (Cheng et al., 2021), our framework extends the architecture for open-vocabulary segmentation following (Xu et al., 2023b; Yu et al., 2023), as illustrated in Fig. 2. The core transformer decoder is comprised of L layers. The image encoders of both CLIP and SAM, as well as the CLIP text encoder, are frozen to maintain their prior knowledge.

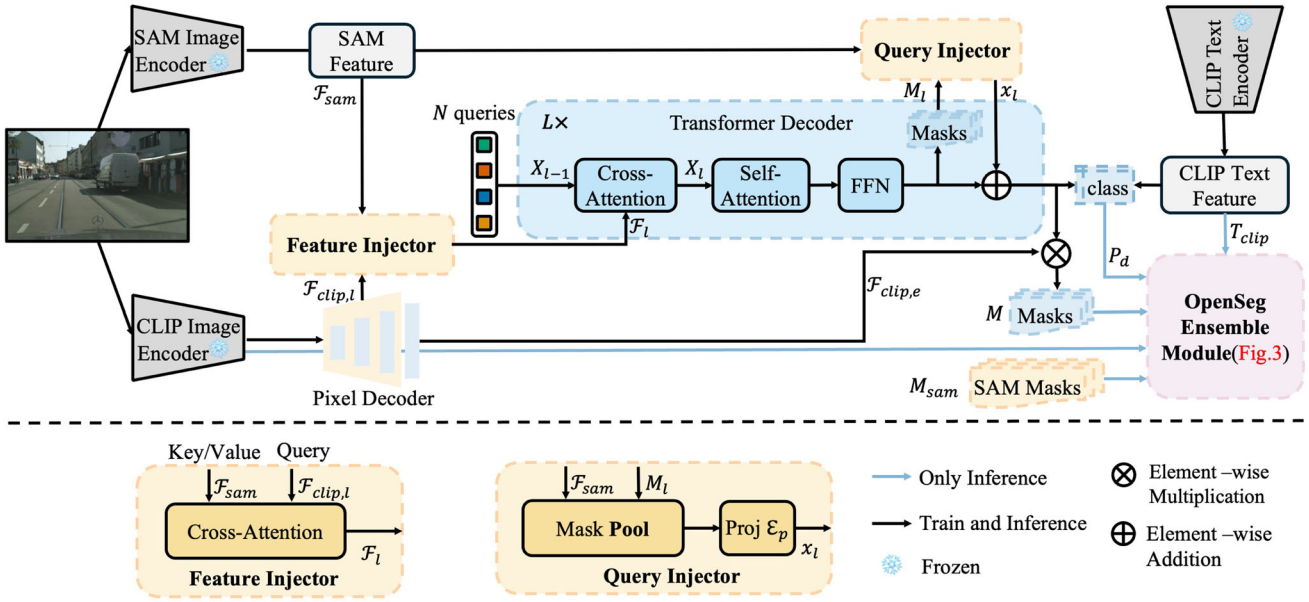

Fig. 2 Overview of our CMPF (Top): We introduce three key components: the Query Injector, Feature Injector, and OpenSeg Ensemble Module (detailed in Fig. 3) to enhance open-vocabulary dense-level understanding. The framework processes parallel feature streams from SAM and CLIP. Given N queries, spatial information from SAM is injected into these queries within intermediate layers of the transformer decoder (Cheng et al., 2021), leading to N class predictions and N corresponding mask predictions–both conditioned on the mask feature representation from CLIP. The OpenSeg Ensemble Module then integrates these predictions with zero-shot SAM masks to generate the final results. (Bottom) Detailed design of the two injectors

Feature Input: The system processes visual semantic information from CLIP through parallel feature extraction pathways. The first pathway extracts multi-scale features through its encoder and pixel decoder, yielding $\mathcal{F}_{clip,l}\}_{l=1}^{L}$ where $\mathcal{F}_{clip,l} \in \mathbb{R}^{C\times H_l\times W_l}$, and mask feature $\mathcal{F}_{clip,e} \in \mathbb{R}^{C\times H_e\times W_e}$, with C being the channel dimension. Concurrently, a separate pathway processes final-layer SAM features to produce $\mathcal{F}_{sam} \in \mathbb{R}^{C_s\times H_s\times W_s}$.

Transformer Decoder: The transformer decoder operates on a set of N learnable queries representing all entities in the input image. Each decoder layer dynamically interacts with the corresponding hierarchical image feature $\mathcal{F}_l$ through three core operations: cross-attention with feature maps, self-attention among queries, and feed-forward network transformation, where cross-attention implements masked attention:

$$X_l = \text{SoftMax}\left(\frac{Q_{l-1}K_l^\top}{\sqrt{d_k}} + \mathcal{M}_{l-1}\right)V_l + X_{l-1},$$
$$\text{s.t.} \quad Q_{l-1} = f_Q(X_{l-1}), \quad K_l/V_l = f_{K/V}(\mathcal{F}_l), \tag{1}$$

where $X_l \in \mathbb{R}^{N\times D}$ represents the evolving query features, where N is the number of queries and D is the embedding dimension. $f_{Q/K/V}$ are linear transformation functions, $\sqrt{d_K}$ is scaling factor, and attention mask $\mathcal{M}_{l-1}$ at feature location [x,y] is derived from the previous prediction to focus computation on predicted foreground regions:

$$\mathcal{M}_{l-1,[x,y]} = \begin{cases} 0, & \text{if } M_{l-1,[x,y]} > 0, \\ -\infty, & \text{otherwise,} \end{cases} \tag{2}$$

where M_{l-1} refers to the continuous mask prediction (Eq. (3)). Note that the initial mask prediction M_0 is derived from X_0 before decoder processing.

Predictions: The predictions utilize three projection functions to enable task-specific transformations: $\mathcal{E}_m$: projects query features for mask prediction; $\mathcal{E}_p$: transforms pooled mask features; $\mathcal{E}_c$: prepares query features for CLIP text alignment. At each decoder layer l, the mask prediction is generated through a linear projection of query features, followed by multiplication with CLIP mask features:

$$M_l = \mathcal{E}_m(X_l) \cdot \mathcal{F}_{clip,e}, \tag{3}$$

with final-layer output results being denoted as $M = \{m_i\}_{i=1}^{N}$. For open-vocabulary predictions, the detector replaces conventional classifiers with CLIP text feature T_{clip} alignment, yielding class predictions $P_d = \{p_{i,d}\}_{i=1}^{N}$:

$$P_d = \text{SoftMax}\left(\mathcal{E}_c(\cdot) \cdot T_{clip}^\top/\tau\right),$$
$$\text{s.t.} \quad \mathcal{E}_c(\cdot) = \mathcal{E}_c\left(X_l + \mathcal{E}_p(\mathbf{pool}_M\mathcal{F}_{clip,e})\right), \tag{4}$$

where τ denotes a learnable temperature parameter for logit calibration, and $\mathbf{pool}_M$ represents the mask pooling process (Ghiasi et al., 2022), which aggregates the features of the region with the proposal masks $M > 0$ through dot-product.

To address CLIP's inherent feature coarseness, we introduce two specialized modules: the Query Injector integrates SAM's precise spatial information directly into mask queries, while the Feature Injector enhances CLIP features with SAM's local detail. Unlike ViT-Adapter (Chen et al., 2023), which globally incorporates spatial information into vision transformers, our approach specifically targets query-level enhancement. The OpenSeg Ensemble Module (detailed in Section 3.2.4) further improves segmentation performance during inference through complementary mask prediction fusion by integrating with SAM's geometrically precise zero-shot masks.

3.2.2 Query Injector

While freezing both foundation model backbones preserves their pretrained knowledge, it also restricts the adaptability of spatial representations during training. As a result, the Mask2Former transformer decoder must rely on a limited number of learnable queries to discover object regions without strong spatial guidance, which becomes particularly challenging in open-vocabulary scenarios.

In the decoder, learnable queries are responsible for discovering object regions through iterative cross-attention. However, these queries are randomly initialized and gradually optimized using training data. Without explicit spatial guidance, they may overfit to the spatial patterns of seen categories, leading to suboptimal mask proposals for unseen objects.

To address this limitation, we introduce the Query Injector to explicitly inject spatial localization cues from SAM into the query refinement process. Specifically, at each decoder layer l, the predicted masks M_l are used to **pool** and extract region-aware features from SAM. These features serve as spatial anchors that guide the queries toward more accurate and spatially coherent object regions:

$$x_l = \mathcal{E}_p(\mathbf{pool}_{M_l}\mathcal{F}_{sam}) \in \mathbb{R}^{N\times D}. \tag{5}$$

Subsequently, the spatial query x_l is combined with the queries X_l through element-wise addition and forwarded to the subsequent decoder layer. This design progressively improves the spatial alignment between queries and image regions across layers, enabling the model to generate more precise mask proposals.

3.2.3 Feature Injector

While the Query Injector enhances the spatial grounding of object queries, the CLIP visual features used for mask generation still suffer from coarse spatial resolution. This limitation originates from the image-text contrastive pretraining of CLIP, which focuses on global semantics rather than fine-grained spatial correspondence. To address this issue, we introduce the Feature Injector to refine CLIP features by integrating pixel-level spatial awareness from SAM.

The Feature Injector can be interpreted as a spatial refinement process. Instead of directly using CLIP features for mask prediction, we enhance them with spatial cues from SAM through the multi-head cross-attention (MHCA) (Vaswani et al., 2017) . This allows each pixel feature in CLIP to selectively aggregate relevant spatial context from SAM, leading to more accurate and coherent mask boundaries.

Specifically, before the cross-attention between learnable queries and CLIP image features, the Feature Injector integrates semantic content from CLIP. The formulation for this feature integration is given by:

$$\mathcal{F}_l = \text{SoftMax}\left(\frac{f_q(\mathcal{F}_{clip,l})\, f_k(\mathcal{F}_{sam})^\intercal}{\sqrt{d_k}}\right) f_v(\mathcal{F}_{sam}) \in \mathbb{R}^{C\times H_l\times W_l}. \tag{6}$$

Here, $f_{q/k/v}$ are linear projection functions in MHCA, while $\sqrt{d_k}$ denotes the scaling factor.

Mechanism When segmenting an unseen category, CLIP may recognize its semantic similarity but fail to localize it precisely. The Query Injector enables the corresponding query to focus on spatially consistent regions identified by SAM, while the Feature Injector further refines pixel-level features to capture object boundaries. As a result, the model can produce more accurate mask proposals for novel categories.

3.2.4 Inference Strategy

Previous studies (Liu et al., 2024; Yu et al., 2023) have established that region-level features cropped from CLIP retain open-vocabulary recognition capabilities. Consequently, these works leverage a detector-ViL ensemble (referred to as class ensemble) by combining CLIP's feature-based predictions with detector scores. However, even with the enhanced mask proposals produced by our CMPF during training, the final predictions can still benefit from complementary priors at inference time.

To this end, we propose an OpenSeg Ensemble Module with a two-stage fusion design, where the two stages are

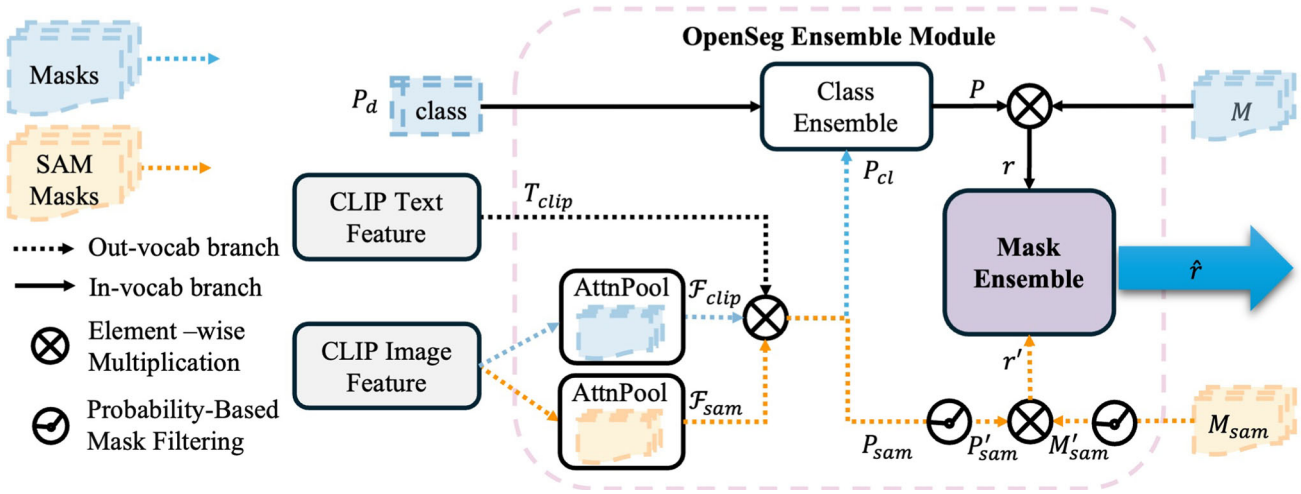

Fig. 3 Architecture of the OpenSeg Ensemble Module. The framework first generates class-agnostic masks via SAM using uniformly sampled point prompts, then performs hierarchical fusion through: (1) *class ensemble* that combines CLIP-aligned semantic scores with detector outputs via weighted fusion, and (2) *mask ensemble* that injects SAM's spatial predictions into novel class mask proposals predictions. This dual-stage approach enhances open-vocabulary generalization by preserving both semantic accuracy (via CLIP) and spatial coherence (via SAM)

functionally complementary: Stage 1 (Class Ensemble) performs semantic score fusion between the detector head and CLIP alignment to calibrate class distributions, while Stage 2 (Mask Ensemble) performs spatial mask fusion by incorporating SAM's zero-shot masks to refine the pixel-wise prediction, especially for novel categories. This two-stage design (illustrated in Fig. 3) explicitly couples semantic calibration with spatial regularization in a structured manner.

Stage 1: Class Ensemble. Follow previous works (Yu et al., 2023), we first integrate predictions from both the detector head ($p_{i,d}$) and CLIP itself ($p_{i,cl}$) via a weighted geometric mean, producing a calibrated class distribution p_i ($P = \{p_i\}_{i=1}^{N}$) for each query:

$$p_i(j) = \begin{cases} (p_{i,d}(j))^{1-\alpha} \cdot (p_{i,cl}(j))^{\alpha}, & \text{if } j \in \mathbf{C}_{train}, \\ (p_{i,d}(j))^{1-\beta} \cdot (p_{i,cl}(j))^{\beta}, & \text{else.} \end{cases} \tag{7}$$

Here, $\alpha, \beta \in [0, 1]$ are weighting parameters that balance the contributions of the two predictions, and $p_i(j)$ denotes the combined probability for class j in proposal i. Specifically, $P_{cl} = \{p_{i,cl}\}_{i=1}^{N}$ is derived by aligning the mask-pooled frozen CLIP image features with text features. The mask pool here is achieved by a multi-head attention layer (AttnPool) at the end of CLIP's image encoder, where the globally average-pooled feature works as the query $\bar{q}$, while the spatial features generate key-value pair (k, v):

$$\mathcal{F}_{clip} = \text{AttnPool}(\bar{q}, k, v; \mathcal{M}), \tag{8}$$

$$P_{cl} = \text{SoftMax}(\mathcal{F}_{clip} \cdot T_{clip}^{\top}/\tau), \tag{9}$$

where $\mathcal{M}$ denotes the mask attention mechanism (see Eq. (2)).

The mask predictions r for N queries are then generated by aggregating the products of these probability-mask pairs: $r = \sum_{i=1}^{N} p_i(c) \cdot m_{i,[x,y]}$, $r \in \mathbb{R}^{C \times HW}$. H, W denote the spatial size of input image.

Stage 2: Mask Ensemble. Given the calibrated class distribution from Stage 1, we further integrate SAM's zero-shot masks M_{sam} to provide a complementary spacial prior. This stage is designed to inject geometry-consistent mask proposals from SAM to improve spatial completeness and boundary coherence for novel categories.

Concretely, the SAM masks are generated by uniformly sampling point prompts across the image. These masks are used to pool CLIP image features and derive classification scores P_{sam} by aligning with CLIP text features:

$$\mathcal{F}_{sam} = \text{AttnPool}(\bar{q}, k, v; \mathcal{M}_{sam}), \tag{10}$$

$$P_{sam} = \text{SoftMax}(\mathcal{F}_{sam} \cdot T_{clip}^{\top}/\tau). \tag{11}$$

A threshold $\xi = 0.5$ is applied to filter these low-quality masks based on the maximum probability, resulting in the selected probability-mask pairs $(M'_{sam}, P'_{sam}) = \{(m'_i, p'_i) \mid \text{argmax}_c\, p_i > \xi, p_i \in P_{sam}\}_{i=1}^{N'}$.

In the context of semantic segmentation, the SAM mask predictions, denoted as $\hat{r}$, are computed similarly as follows: $r' = \sum_{i=1}^{N'} p'_i(c) \cdot m'_{i,[x,y]}$. The final mask prediction, $\hat{r}$, is obtained by integrating the predictions r and r' through a mask ensemble approach:

$$\hat{r}_{[x,y]}(j) = \begin{cases} r_{[x,y]}(j), & \text{if } j \in \mathbf{C}_{train}, \\ (1-\epsilon) r_{[x,y]}(j) + \epsilon r'_{[x,y]}(j), & \text{else}, \end{cases} \tag{12}$$

where $\hat{r}_{[x,y]}(j)$ denotes the combined probability for class j in pixel $[x, y]$. Subsequently, the final semantic segmentation result for each query with index i is determined by assigning each pixel $[x, y]$ a class based on $\text{argmax}_{c\in\{1,\ldots,|\mathbf{C}|\}} \hat{r}_{[x,y]}$.

In panoptic segmentation, the effectiveness of the results is largely dependent on the performance of individual queries. This reliance diminishes the effectiveness of integrating the more-noisy SAM masks. Consequently, the final segmentation results of are determined by assigning each pixel $[x, y]$ to one of the N queries' predictions. The assignment is given by the following expression: $\text{argmax}_{i\in 1,\ldots,N}\{p_i(c_i)\cdot m_{i,[x,y]}\}$, where c_i represents the most likely class label, calculated as $c_i = \text{argmax}_{c\in\{1,\ldots,|\mathbf{C}|,\emptyset\}} p_i$ ($\emptyset$ represents void regions with confidence threshold).

4 Experiments

4.1 Datasets and Evaluation Protocol

We use the COCO panoptic (Lin et al., 2014) as the training dataset, which comprises 133 distinct classes. Our evaluation encompasses a range of open-vocabulary segmentation tasks, including panoptic, semantic, and instance segmentation, all conducted in a zero-shot way. In the domain of panoptic segmentation, we evaluate performance on three datasets: ADE20K (Zhou et al., 2017), Cityscapes (Cordts et al., 2016), and BDD100K (Yu et al., 2020). For semantic segmentation, we assess models on ADE20K, which provides a dual challenge: a semantic task subset with 150 classes (A-150) and a comprehensive version with 847 classes (A-847), as well as PASCAL-Context (Mottaghi et al., 2014), an extension of PASCAL VOC (Everingham et al., 2010), with the full-class version referred to as PC-459 and the 59 most frequent classes as PC-59. For instance segmentation, we evaluate on LVIS v1.0 (Gupta et al., 2019), which features 337 rare categories. The evaluation metrics include panoptic quality (PQ), average precision (AP), and mean intersection-over-union (mIoU) for panoptic segmentation, mIoU for semantic segmentation, and AP for instance segmentation. As FC-CLIP (Yu et al., 2023) uses a checkpoint selection strategy based on ADE20K validation set, which is unsuitable for open-vocabulary testing scenarios, we report experimental results using the official code to ensure fair comparisons.

4.2 Implement Details

We use $N = 250$ queries for both training and testing, with CLIP as the backbone for open-vocabulary text-image alignment. Specifically, we employ the RN50×64 and ConvNext-Large versions of the CLIP image encoder. Unless otherwise stated, experiments are conducted using the ConvNeXt-Large CLIP model. Additionally, we validate our approach using the ViT-Base model from SAM, with the selection rationale detailed in the ablation studies. To obtain multi-level semantic features, we apply feature pyramid networks (FPN) after CLIP. SAM processes input images $\mathbf{I}$ of 1024×1024. As demonstrated by PlainViT (Li, 2022), the deepest feature of ViT contains sufficient information for multi-scale object recognition, and given that SAM is frozen, we do not use FPN for SAM. Instead, we utilize a single convolution layer to project the features to the necessary resolution and then feed them into a single-scale deformable attention transformer (Zhu et al., 2021). Our transformer decoder comprises $L = 9$ layers. Feature maps with resolutions of 1/8, 1/16, and 1/32 are processed by successive decoder layers in a round-robin fashion. During training, we adhere to the strategy and loss functions outlined in (Cheng et al., 2021). The model from the final iteration is selected for reporting the primary results. Training is conducted on 4 Tesla A100 GPUs with a batch size of 16.

4.3 Inference Details

During inference, images are resized such that the shortest side is 800 pixels for general datasets and 1024 for the Cityscapes. We employ a 32×32 grid of points research to generate masks from the SAM ViT-Huge model. The default parameters are set as follows: $\alpha = 0.4$ and $\beta = 0.8$ in Eq. (7), and a mask ensemble parameter $\epsilon = 0.2$ in Eq. (12).

4.4 Evaluation on Open-vocabulary Segmentation

4.4.1 Open-vocabulary Panoptic Segmentation

Table 1 presents a comprehensive comparison of CMPF with leading methods in zero-shot open-vocabulary panoptic segmentation. Our approach, CMPF with RN50×64, notably surpasses other works and the FC-CLIP baseline, achieving improvements of +1.8 PQ, +0.3 AP, and +2.0 mIoU on ADE20K; +2.6 PQ, +1.6 AP, and +0.9 mIoU on Cityscapes; and +4.8 mIoU on BDD100K. When equipped with the ConvNeXt-L backbone, CMPF further extends its performance on open-set recognition tasks. The model delivers consistent improvements across all metrics, including +0.8 PQ and +1.6 mIoU on ADE20K, +1.5 PQ, +0.5 AP, and +0.8 mIoU on Cityscapes, and +1.4 PQ and +2.9 mIoU on BDD10K. These results demonstrate the robustness and

Table 1 Performance of open-vocabulary panoptic segmentation. We present results obtained using both CLIP RN50x64 and ConvNext-L. 'SD' denotes the Stable Diffusion model (Rombach et al., 2022) pre-trained on a subset of the LAION dataset. **Bold** represents best, underline indicates second best.

Method	ViL Model	ADE20K			Cityscapes			BDD100K	
		PQ	AP	mIoU	PQ	AP	mIoU	PQ	mIoU
OPSNet (Chen et al. 2023) (ICCV2023)	RN50	19.0	-	25.4	41.5	-	-	-	-
MaskCLIP (Ding et al. 2023) (ICML2023)	ViT-L/14	15.1	6.0	23.7	-	-	-	-	-
MasQCLIP (Xu et al. 2023a) (ICCV2023)	ViT-L/14	23.3	-	30.4	-	-	-	-	-
ODISE (Xu et al. 2023b) (CVPR2023)	ViT-L/14+SD	22.6	14.4	29.9	-	-	-	-	-
ODISE(caption) (Xu et al. 2023b) (CVPR2023)	ViT-L/14+SD	23.4	13.9	28.7	-	-	-	-	-
ODISE+CLIPSelf (Wu et al. 2024) (ICLR2024)	ViT-L/14+SD	23.7	13.6	30.1	-	-	-	-	-
FC-CLIP (Yu et al. 2023) (NeurIPS2023)	RN50×64	21.3	13.2	28.7	42.6	27.3	55.1	13.8	41.4
CMPF(ours)	RN50×64	23.1	13.5	30.7	45.2	**28.9**	56.0	12.9	46.2
FC-CLIP (Yu et al. 2023) (NeurIPS2023)	ConvNeXt-L	25.1	16.4	32.8	44.3	27.9	56.0	17.9	49.4
EOV-Seg (Niu et al. 2025) (AAAI2025)	ConvNeXt-L	24.5	13.7	32.1	-	-	-	-	
CMPF(ours)	ConvNeXt-L	**25.9**	**16.5**	**34.4**	**45.8**	28.4	**56.8**	**19.3**	**52.3**

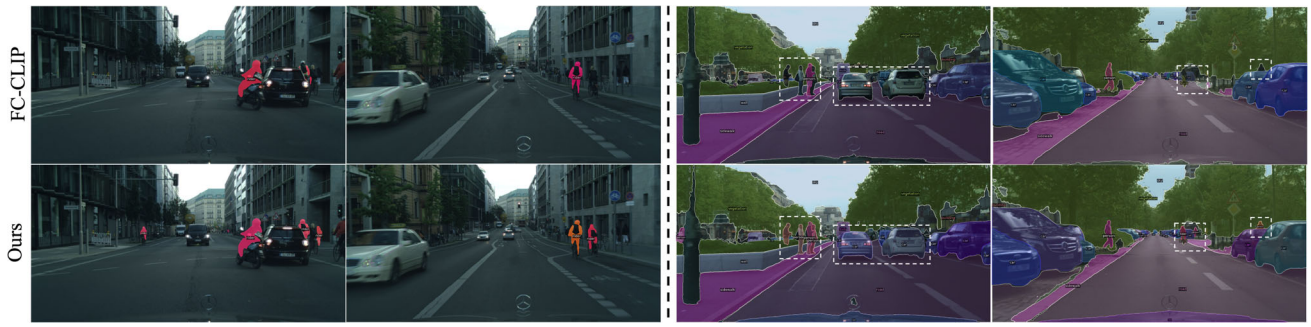

Fig. 4 Qualitative comparison of panoptic segmentation results on the Cityscapes dataset. The left panel shows panoptic segmentation results for unseen classes, while the right panel displays overall panoptic segmentation results. Our method demonstrates superior generalization to the unseen class 'rider'. White boxes in overall segmentation maps highlight areas with significant differences, where CMPF exhibits improved performance in detecting small instances and recognizing individual entities

Fig. 5 Qualitative comparison of panoptic segmentation results on the BDD100K dataset. The left panel shows panoptic segmentation results for unseen classes, while the right panel displays overall panoptic segmentation results. Our method demonstrates superior generalization to the unseen class 'vegetation' and 'street light'. White boxes in overall segmentation maps highlight areas with significant differences, where CMPF exhibits improved performance in detecting small instances and recognizing individual entities

scalability of our approach across different backbone architectures. Qualitative results of panoptic segmentation on two street datasets, shown in Figs. 4 and 5, showcase improvements in segmentation, particularly for novel classes such as

Table 2 Performance of cross-dataset open-vocabulary semantic segmentation. 'SD' is the Stable Diffusion model, 'IN' refers to the ImageNet (50K) dataset, 'Panop.+Cap.' signifies the combined use of COCO panoptic and COCO caption datasets, and 'LN' denotes the Localized Narrative dataset. See corresponding works for dataset details. **Bold** represents best, underline indicates second best

Method	ViL Model	Training Dataset	mIoU			
			PC-459	PC-59	A-847	A-150
ZegFormer (Ding et al. 2022) (CVPR2022)	ViT-B/16	COCO Stuff	-	-	-	16.4
SimBase (Xu et al. 2022)(ECCV2022)	ViT-B/16	COCO Stuff	-	47.7	7.0	20.5
OpenSeg (Ghiasi et al. 2022) (ECCV2022)	ALIGN	COCO Panoptic+LN	11.2	45.9	6.8	24.8
OPSNet (Chen et al. 2023) (ICCV2023)	RN50	COCO Panoptic+IN	-	54.3	-	25.4
Ovseg (Liang et al. 2023) (ICCV2023)	ViT-L/14	COCO Stuff	12.4	55.7	9.0	29.6
MaskCLIP (Ding et al. 2023) (ICML2023)	ViT-L/14	COCO Panoptic	10.0	45.9	8.2	23.7
MasQCLIP (Xu et al. 2023a) (ICCV2023)	ViT-L/14	COCO Panoptic	18.2	57.8	10.7	30.4
ODISE(caption) (Xu et al. 2023b) (CVPR2023)	ViT-L/14+SD	COCO Panop.+Cap.	13.8	55.3	11.0	28.7
ODISE (Xu et al. 2023b) (CVPR2023)	ViT-L/14+SD	COCO Panoptic	14.5	57.3	11.1	29.9
ODISE+CLIPSelf (Wu et al. 2024) (ICLR2024)	ViT-L/14+SD	+COCO Stuff	-	-	-	30.1
SCAN (Liu et al. 2024) (CVPR2024)	ViT-L/14	COCO Stuff	16.7	**59.3**	14.0	33.5
FC-CLIP (Yu et al. 2023) (NeurIPS2023)	RN50×64	COCO Panoptic	15.6	54.8	10.8	28.7
CMPF(ours)	RN50×64	COCO Panoptic	18.7	57.5	11.8	30.7
FC-CLIP (Yu et al. 2023) (NeurIPS2023)	ConvNeXt-L	COCO Panoptic	17.3	56.6	14.0	32.8
EOV-Seg (Niu et al. 2025) (AAAI2025)	ConvNeXt-L	COCO Panoptic	16.8	56.9	12.8	32.1
CMPF(ours)	ConvNeXt-L	COCO Panoptic	**19.7**	58.4	**14.8**	**34.4**

'rider', 'vegetation', and 'streetlight'. They also highlight our enhanced detection of small objects and improved individual entity recognition.

4.4.2 Open-vocabulary Semantic Segmentation

Table 2 presents a comparative analysis of CMPF in open-vocabulary semantic segmentation. Using the RN50×64 backbone, CMPF significantly outperforms the baselines. Compared to FC-CLIP, CMPF achieves gains of +3.1 mIoU on PC-459, +2.7 mIoU on PC-59, +1.0 mIoU on A-847, and +2.0 mIoU on A-150. These improvements are also reflected in the ConvNeXt-L configuration. Overall, CMPF establishes a new performance benchmark across the PC-459, A-847, and A-150 datasets. Notably, the COCO Stuff dataset used by SCAN (Liu et al., 2024) includes 171 classes, more than the 133 classes in COCO Panoptic dataset. However, SCAN is limited to semantic segmentation tasks, whereas our method, trained on fewer categories, is capable of both semantic and panoptic segmentation. Considering these factors, our improvements on PC-59 are particularly noteworthy. For qualitative insights, Fig. 6 and Fig. 7 present qualitative comparisons on the PC-459 and A-847 datasets and show that CMPF achieves more accurate segmentations compared to CLIP-based methods (Ding et al., 2023; Yu et al., 2023), particularly in preserving structural completeness and contextual coherence. This confirms CMPF's effectiveness in handling open-word tasks.

4.4.3 Open-vocabulary Instance Segmentation

Table 3 presents the results for rare categories in the LVIS (Gupta et al., 2019) dataset, comparing CMPF with methods that integrate SAM and CLIP for open-vocabulary segmentation tasks, specifically RegionSpot and Open-Vocabulary SAM. These methods rely on external proposals as box prompts and are trained on expansive datasets beyond COCO panoptic, such as V3Det, which includes up to 13,029 categories, and SAM-1B, containing 1.1 billion high-quality segmentation masks. In contrast, CMPF adopts an end-to-end framework to generate proposals, significantly reducing the reliance on additional training resources and external proposals. Moreover, it is not constrained to box-instance segmentation tasks. Notably, CMPF outperforms Open-Vocabulary SAM by 1.6 AP_r. Furthermore, the synergistic optimization of SAM and CLIP in our framework addresses their individual limitations, achieving an additional +0.6 AP_r improvement over (Yu et al., 2023).

4.4.4 Generalization Ability to Unseen Classes

To rigorously evaluate our CMPF framework's generalization capability on previously unseen categories, we compare its performance with other methods across the challenging PC-459, A-847, and A-150 datasets, which consist of 459, 847, and 150 classes, respectively. As shown in Table 4, CMPF consistently outperforms other methods, not only

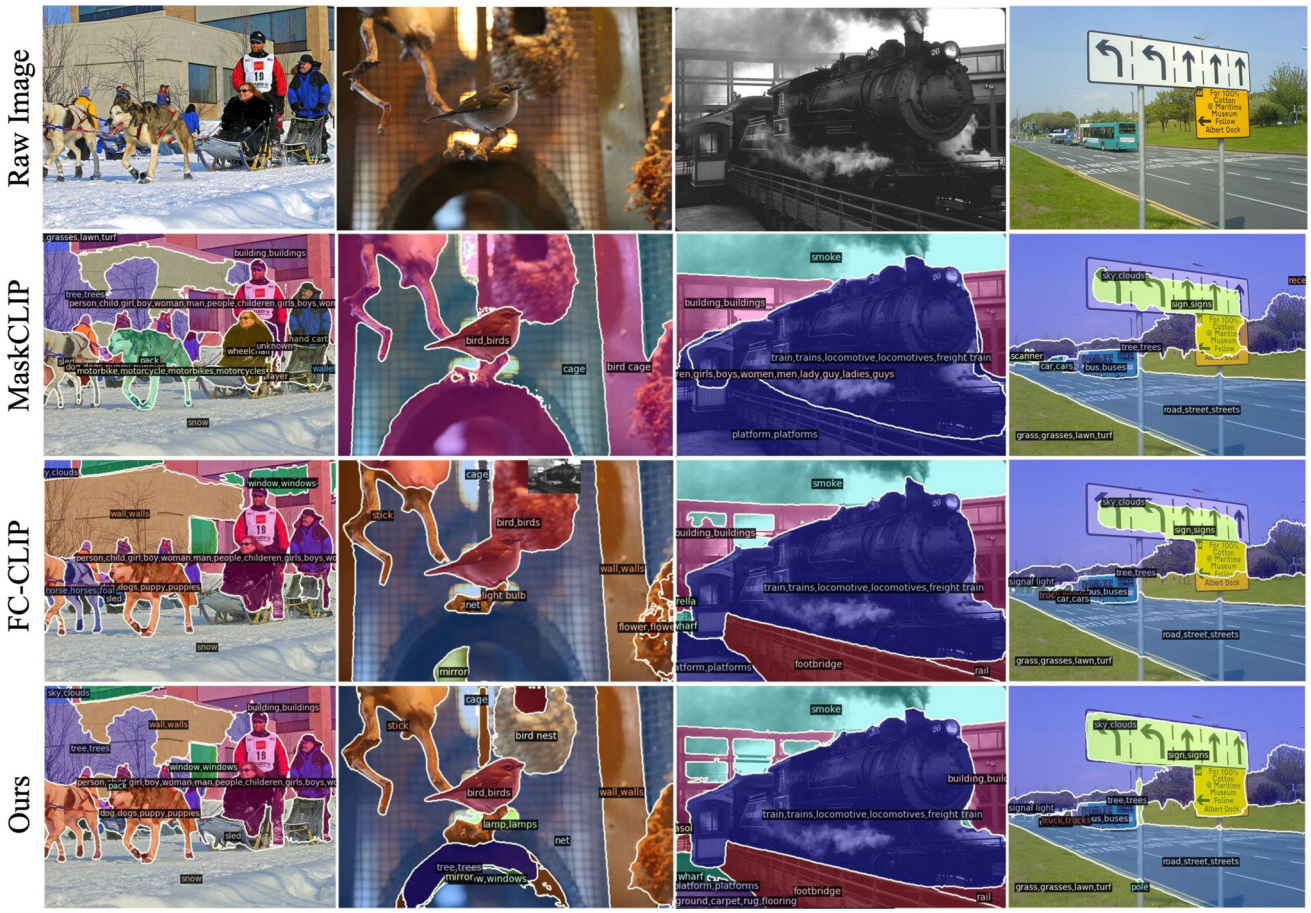

Fig. 6 Qualitative comparison of semantic segmentation results on PC-459

Table 3 Performance (AP_r) of open-vocabulary instance segmentation on 'rare' categories in LVIS v1.0 dataset. Object365 (O365), OpenImages (OI), and V3Det (V3D) datasets are used in RegionSpot. LVIS‡represents a dataset variant containing only the 'normal' and 'frequent' classes from the LVIS training set. See corresponding works for dataset and proposal details

Method	ViL Model	Training Dataset	External Proposals	AP_r
RegionSpot (Yang et al. 2024) (NeurIPS2024)	ViT-B/16	O365	GLIP-T(B)	12.7
	ViT-L/14	O365, OI, V3D	SAM-B	14.3
	ViT-B/16	O365, OI, V3D	GLIP-T	20.0
Open-Vocabulary SAM (Yuan et al. 2024) (ECCV2024)	RN50×16	SAM-1B+LVIS‡	Detic	24.0
FC-CLIP (Yu et al. 2023) (NeurIPS2023)	ConvNeXt-L	COCO Panoptic	-	25.0
CMPF(ours)	ConvNeXt-L	COCO Panoptic	-	**25.6**

Table 4 Performance of open-vocabulary semantic segmentation for seen and unseen classes

Method	PC-459		A-847		A-150	
	$mIoU_{seen}$	$mIoU_{unseen}$	$mIoU_{seen}$	$mIoU_{unseen}$	$mIoU_{seen}$	$mIoU_{unseen}$
MaskCLIP (Ding et al. 2023) (ICML2023)	30.3	6.3	16.7	6.0	32.7	16.5
FC-CLIP (Yu et al. 2023)(NeurIPS2023)	46.0	8.7	34.0	11.3	47.3	22.0
CMPF(ours)	**48.1**	**11.2**	**34.7**	**12.1**	**49.0**	**23.6**

Fig. 7 Qualitative comparison of semantic segmentation results on A-847 dataset

Table 5 Evaluating open-vocabulary recall across datasets with semantic segmentation annotations under three IoU thresholds, providing insights into seen (S) and unseen (U) classes

	PC-459 (S/U)			A-847 (S/U)			A-150 (S/U)		
Method	0.5	0.75	0.9	0.5	0.75	0.9	0.5	0.75	0.9
FC-CLIP	91.4/77.7	73.6/56.2	43.9/28.4	**84.6**/66.6	**60.0**/41.3	**27.6**/14.3	83.7/72.0	58.9/45.5	**27.7**/17.4
CMPF(ours)	**92.6/81.0**	**76.0/59.9**	**47.0/31.4**	78.9/**82.9**	52.3/**59.4**	23.6/**24.6**	**84.7/73.2**	**60.1/47.2**	27.6/**18.0**

enhancing the performance on seen cases but also improving results on unseen classes, achieving gains of +2.5 on PC-459, +0.8 on A-847 and +1.6 on A-150 in terms of mIoU for unseen classes. Additional qualitative comparisons on unseen classes are presented in Fig. 8. Compared to state-of-the-art methods, our approach demonstrates superior performance in segmenting and classifying novel classes.

4.4.5 Comparative Recall Across Datasets

Table 5 presents the recall rates of our method and FC-CLIP across three datasets, evaluating the quality of proposal generation within an end-to-end framework using a ViL backbone. The comparison is based on the predicted mask proposals and class-agnostic semantic ground truth, with recall rates detailed at IoU thresholds of 0.5, 0.75, and 0.9. The results indicate that CMPF, with SAM assistance, consistently outperforms in generating mask proposals for unseen classes.

4.5 Ablation Studies

4.5.1 The Effectiveness of Each Component

We conduct ablation studies to assess the effectiveness of each component of our method. Table 6 presents the results of these ablations on three challenging out-of-vocabulary datasets, highlighting the contribution of each component to overall performance. Specifically, Model 1 represents the scenario where only proposals from SAM are utilized. In this setup, SAM masks are used to pool CLIP features, providing basic semantic understanding without explicit semantic guid-

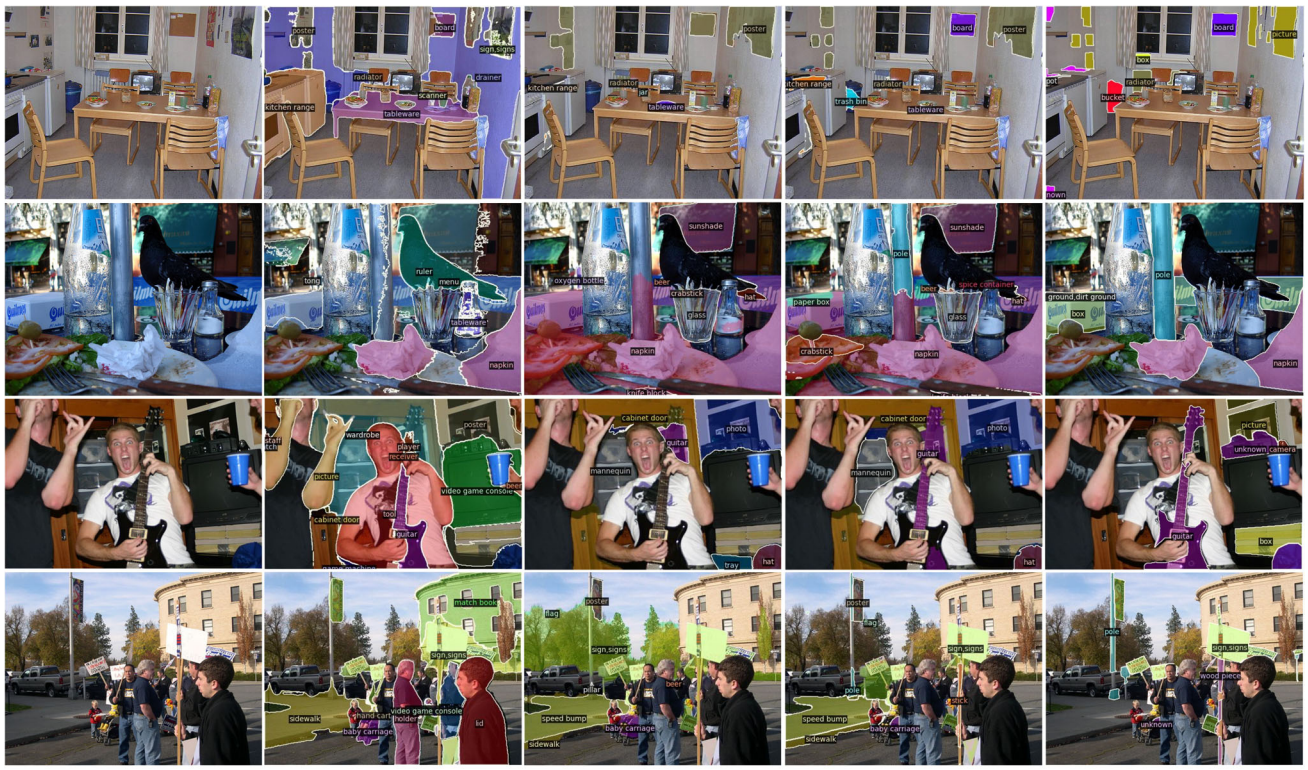

Fig. 8 Zero-shot semantic segmentation results on unseen classes in PC-459 dataset. Compared to existing SOTA approaches, MaskCLIP (Ding et al., 2023) and FC-CLIP (Yu et al., 2023), our method successfully segments novel classes, such as 'poster', 'board' (first row), and 'pole' (second row), showcasing improved performance

Table 6 Ablations of the proposed modules: results after complete training iterations. Model 1 relies solely on SAM-generated proposals, whereas Models 2 to 5 incrementally integrate modules within the unified framework of CMPF.

				PC-459	A-847	ADE20K		
#	Query Injector	Feature Injector	OpenSeg Ensemble	mIoU	mIoU	PQ	AP	mIoU
Model 1	×	×	×	6.6	6.5	-	-	25.4
Model 2	×	×	×	17.3	14.0	25.1	16.4	32.8
Model 3	×	×	✓	17.6	14.5	-	-	33.5
Model 4	✓	×	×	18.1	14.2	25.7	16.5	33.6
Model 5	✓	✓	×	18.5	14.4	**25.9**	**16.5**	33.8
CMPF	✓	✓	✓	**19.7**	**14.8**	-	-	**34.4**

ance. This configuration achieves approximately 6.5 mIoU on PC-459 and A-847, and 25.4 mIoU on ADE20K (A-150), demonstrating the fundamental generalization capability of SAM masks. Therefore, we introduce the OpenSeg Ensemble module during inference to address the limitation of unseen mask proposals. This enhancement is evident in the comparison between Model 2 and Model 3, as well as between Model 5 and CMPF. Notably, without the two injectors, results are suboptimal when only OpenSeg Ensemble is used.

Table 7 Impact of selected insertion layer in transformer decoder on Query Injector performance: results after 55K iterations. The 'Size' column is the relative interacted image feature size of multi-level feature maps

		COCO(seen)			Cityscapes		
Size	Layers	PQ	AP	mIoU	PQ	AP	mIoU
1/32	1, 4, 7	52.6	42.7	62.6	**40.4**	20.8	53.0
1/16	2, 5, 8	52.5	42.5	62.4	40.1	20.7	53.3
1/8	3, 6, 9	**52.7**	**42.8**	**62.7**	40.0	**21.6**	**54.0**

Table 8 Open-vocabulary performance with SAMs on Cityscapes, PC-459, A-847, and PC-59. **Bold** highlights optimal results

SAM	Cityscapes			PC-459	A-847	PC-59
	PQ	AP	mIoU	mIoU	mIoU	mIoU
ViT-T Zhang et al. (2023)	43.0	25.9	55.0	17.6	14.0	56.8
ViT-B (Kirillov et al. 2023)(ICCV2023)	**45.8**	**28.4**	**56.8**	**18.5**	**14.4**	**58.1**
ViT-L (Kirillov et al. 2023)(ICCV2023)	44.2	27.1	56.2	17.4	14.2	56.7
ViT-H (Kirillov et al. 2023)(ICCV2023)	44.2	28.0	56.2	17.3	13.9	56.6

Table 9 Comparative analysis of FPS performance and Trainable vs. Frozen parameter counts using a single A100. All results are obtained from the average time on the validation set and exclude pre- and post-processing time

Method	ADE20K		A-847		Parmas[M]	
	PQ↑	FPS↑	PQ↑	FPS↑	Frozen↓	Trainable↓
FC-CLIP (Yu et al. 2023)(NeurIPS2023)	25.1	2.7	14.0	3.1	200.0	21.0
SCAN (Liu et al. 2024)(CVPR2024)	-	-	14.0	0.9	731.6	158.7
EOV-Seg (Niu et al. 2025)(AAAI2025)	24.5	11.6	12.8	11.8	203.7	21.8
CMPF	25.9	2.2	14.4	2.9	293.5	26.5

4.5.2 Where to Inject

Table 7 presents an ablation examining the impact of layer insertion for the Query Injector within the transformer decoder, which consists of a total of 9 layers. Since SAM Vision Transformers provide the final layer features as the most relevant feature maps, we investigate the optimal layer for query injection based on their interaction with corresponding CLIP feature maps. Results show that injecting SAM query features at layers $l = 3, 6, 9$ yields the most significant improvement, with subsequent layer $l + 1$ refining the newly introduced queries for enhanced performance. For the Feature Injector, to mitigate the exponential increase in computational complexity associated with cross-attention as feature size grows, we limit its application to 1/32-sized features, specifically at layers $l = 1, 4, 7$.

4.5.3 Comparison with Different SAMs

In Table 8, we provide detailed results of CMPF (w/o. Mask Ensemble), using ConvNeXt-L CLIP (Radford et al., 2021) alongside different size of co-trained SAM (Kirillov et al., 2023): ViT-T (Tiny) (Zhang et al., 2023), ViT-B (Base), ViT-L (Large) and ViT-H (Huge). Across the board, the ViT-B configuration stands out, delivering better performance in our evaluations. We also provide visualizations of the k-means clustering results for feature embeddings from the SAM image encoders in Fig 9. The visualization demonstrates that ViT-B balances segmentation accuracy and connectivity, offering precise segmentation with good instance connectivity. ViT-T provides coarse boundaries, while ViT-L and ViT-H, though more precise, have reduced instance connectivity and may be less effective for panoptic segmentation with CLIP. Thus, ViT-B's balanced performance makes it a robust choice.

4.5.4 Speed and Model Size

Table 9 reports the inference speed (FPS) and parameter statistics of different methods on a single NVIDIA A100 GPU. We compare both the total frozen parameters and the number of trainable parameters to better assess practical efficiency. From an overall efficiency perspective, CMPF achieves 2.2 FPS on ADE20K and 2.9 FPS on A-847, remaining comparable to the same CLIP-based methods FC-CLIP and SCAN. While EVO-Seg attains higher FPS, it adopts a lighter architecture without integrating large foundation models and achieves lower segmentation accuracy. SCAN, in contrast, shows lower speed on A-847. These results indicate that CMPF maintains competitive runtime despite incorporating both SAM and CLIP.

4.6 Visualization Analysis

To further analyze our model prediction, we first visualize SAM's feature representation in Fig. 10. As shown, SAM exhibits robust generalization capabilities: when provided with box prompts, its predicted mask probability maps exhibit heightened attention toward target objects while suppressing peripheral edges, thereby confirming its proficiency in boundary delineation.

We then present instance-level attention maps under different ablation settings in Fig. 11. Compared with the CLIP-only baseline, introducing the Query Injector enables the model to activate previously unrecognized object regions, especially for novel categories. The Feature Injector, on the other hand, refines the spatial distribution of attention, producing more complete and coherent activation patterns. These results indicate that the two injectors play complementary roles in enhancing object localization and spatial consistency.

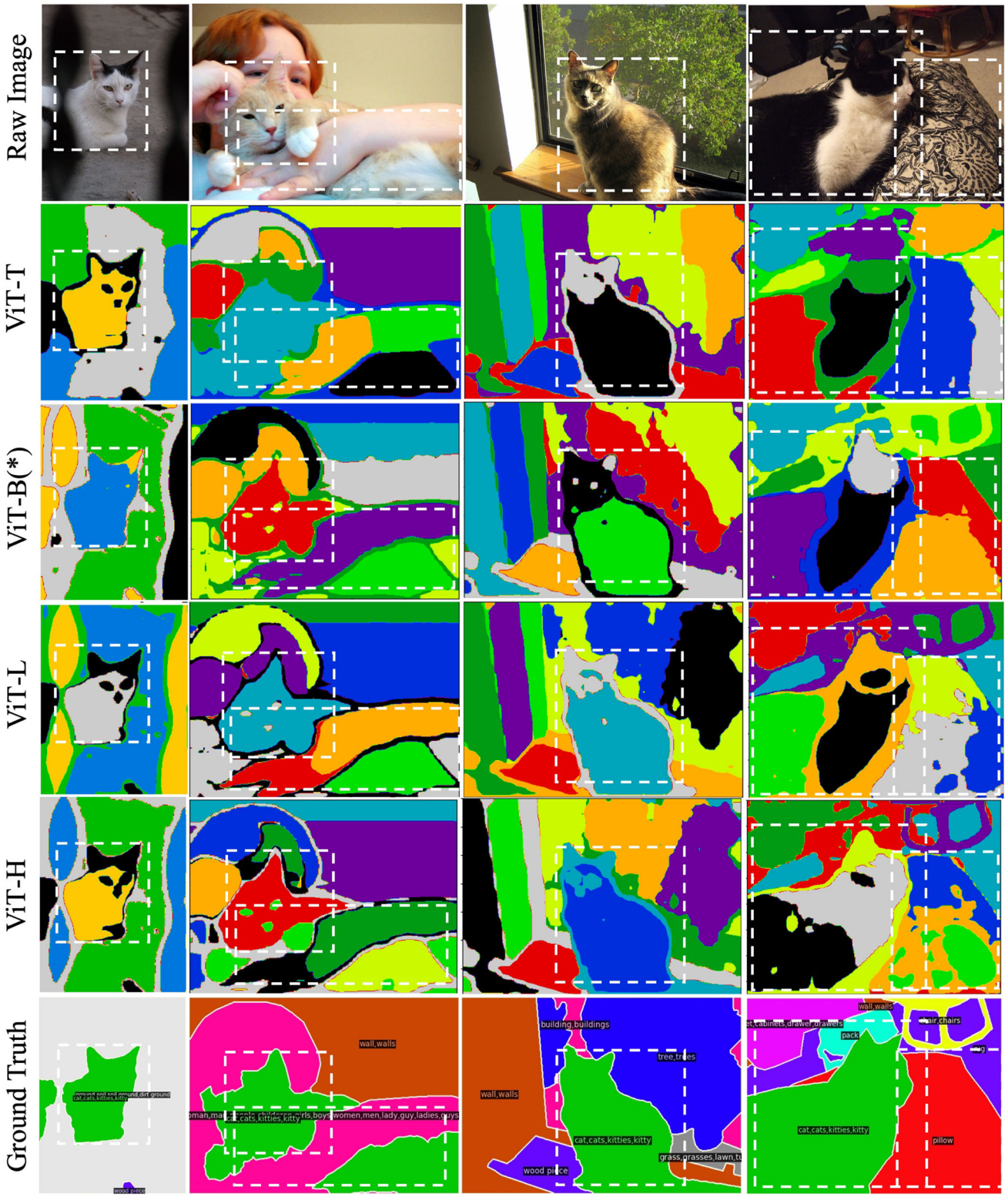

Fig. 9 K-means clustering visualization of feature embeddings from various SAM image encoders on the PC-459 dataset. White boxes in the clustering maps highlight areas with significant differences. The ViT-B encoder exhibits better instance-level connectivity understanding. Note that the cluster colors are randomly assigned.

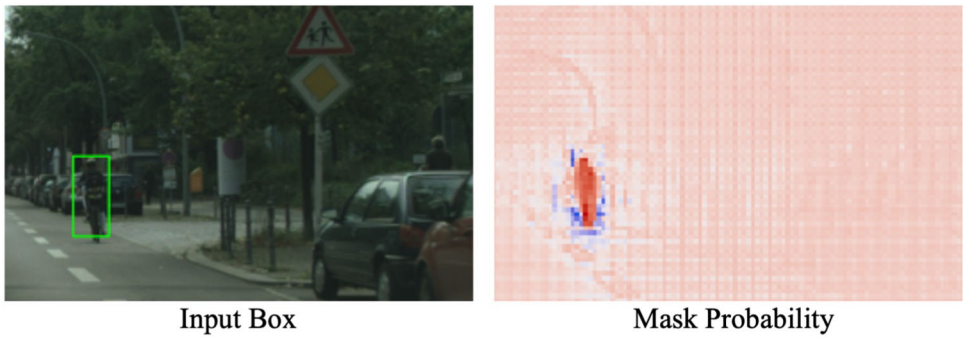

Fig. 10 Probability distribution heatmap of SAM-generated segmentation masks under box prompting

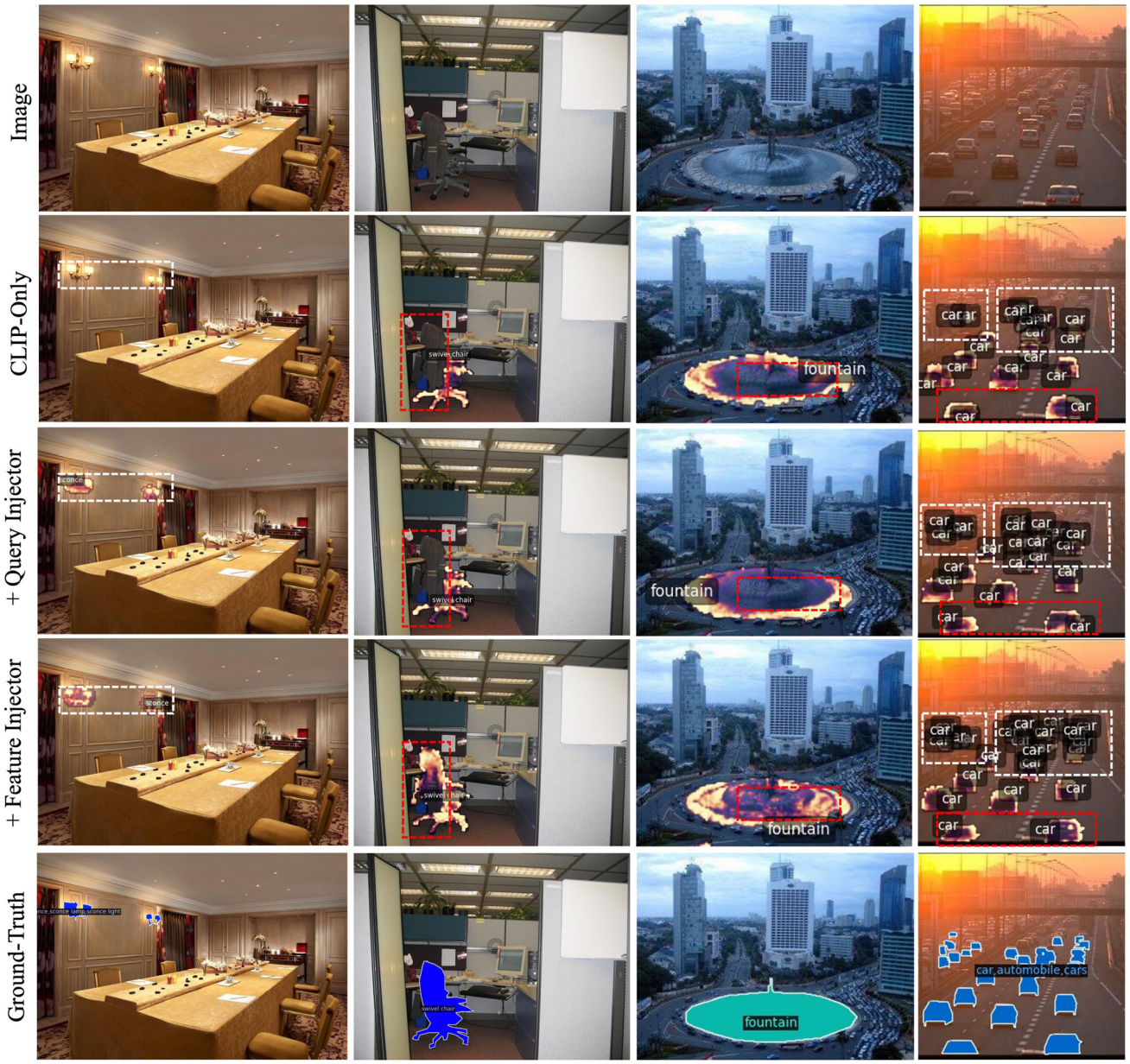

Fig. 11 Instance-level attention map ablation comparison on the ADE20K dataset. The first three columns correspond to novel categories, while the last column shows a base-class example. White boxes highlight newly recognized regions, and red boxes indicate refined attention areas, demonstrating the complementary roles of the two injectors

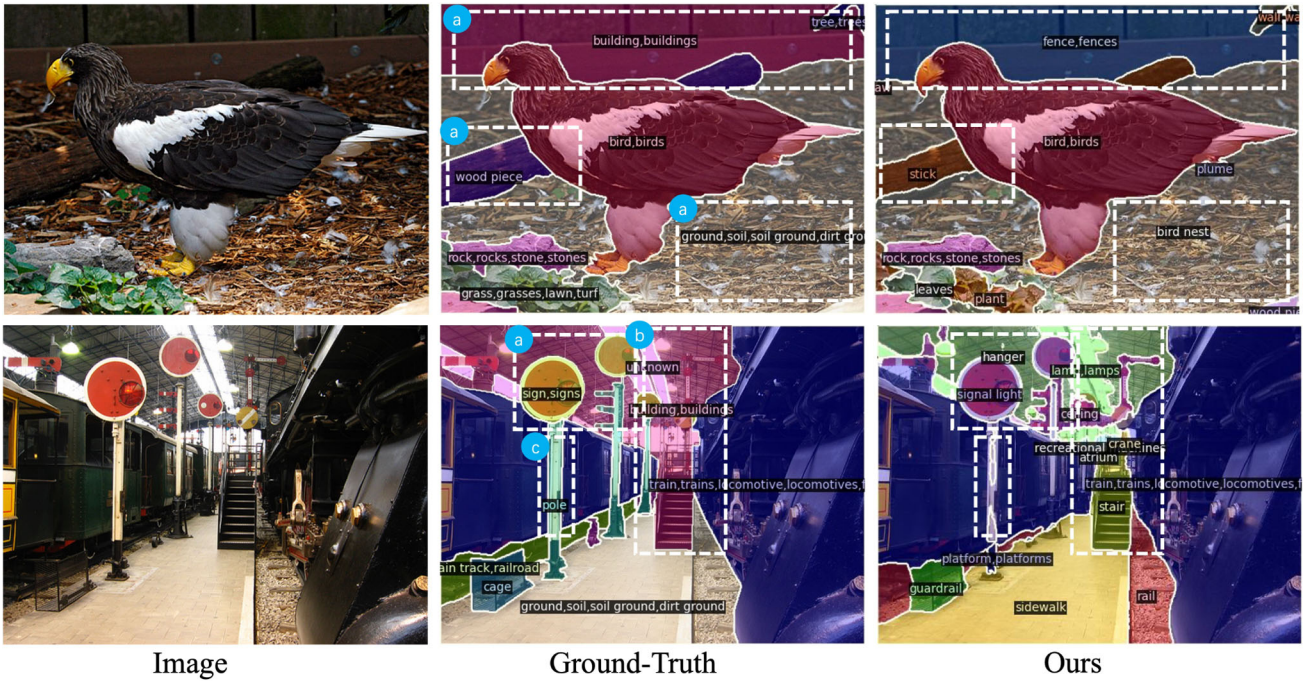

Fig. 12 Failure case examples on the PC-459 dataset: (a) semantic ambiguity; (b) granularity mismatch; (c) ambiguous boundaries

Finally, we illustrate representative failure cases in Fig. 12. Most errors arise from semantic ambiguity between visually similar categories, such as "*wood-piece*" vs. *"stick"* or *"sign"* vs. *"signal light"*, as well as granularity mismatch between predictions and annotations. For example, small structures (e.g., "*lamp*" or "*stair*") may be labeled as part of "*building*" in the ground truth but predicted separately by our model. In addition, thin objects such as "*pole*" often exhibit ambiguous boundaries. These cases reflect the inherent challenges of open-vocabulary segmentation.

5 Conclusion

In this study, we propose CMPF, a novel framework that significantly enhances mask proposal quality for unseen categories by cross-modal prior fusion, setting a new state-of-the-art in open-vocabulary segmentation. By integrating SAM's visual dense-prediction capabilities with CLIP's textual semantic understanding, CMPF leverages the Query Injector and Feature Injector modules to fuse SAM's visual features with learned queries and CLIP's features, refining mask proposals through multiple transformer decoder layers. Additionally, the OpenSeg Ensemble Module is introduced during inference to aggregate zero-shot SAM masks, further improving predictions for out-of-vocabulary categories. Experimental results highlight the effectiveness and adaptability of CMPF in open-vocabulary scenarios.

Limitation and Future Work. While CMPF leverages the SAM image encoder to inject spatial information, the prompt encoder and mask decoder of SAM are not utilized in our current framework. These components play an important role in SAM's segmentation capability. Integrating them into dense prediction tasks may further improve spatial reasoning and mask quality, but would inevitably introduce additional computational overhead. In addition, as discussed in the failure case analysis, CMPF may still encounter challenges in scenarios with strong semantic ambiguity or ambiguous object boundaries. For future work, we plan to explore more effective integration of SAM's promptable segmentation mechanism and stronger semantic reasoning modules. We will also investigate whether scaling such joint integration with larger vision-language and segmentation foundation models can further enhance open-vocabulary segmentation performance.

Funding This work was supported by Beijing Natural Science Foundation (No. L252009), the National Natural Science Foundation of China (Nos. 62571294, 62476069), and CCF-DiDi GAIA Collaborative Research Funds.

Data Availability The data in our paper is openly available in a public repository. The data that support the findings of this study are openly available as follows:

- COCO: https://cocodataset.org/#download
- Cityscapes: https://www.cityscapes-dataset.com/downloads/
- ADE20K (A150): http://sceneparsing.csail.mit.edu/
- BDD100K: http://bdd-data.berkeley.edu/
- A847: https://groups.csail.mit.edu/vision/datasets/ADE20K/

- PC-459 and PC-59: https://cs.stanford.edu/ roozbeh/pascal-context/
- LVIS: https://www.lvisdataset.org/

Declarations

Conflicts of Interest The authors have no relevant financial or non-financial interests to disclose.

References

Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., & Qiao, Y. (2023). Vision transformer adapter for dense predictions. International Conference on Learning Representations.

Chen, X., Li, S., Lim, S.-N., Torralba, A. & Zhao, H. (2023). Open-vocabulary panoptic segmentation with embedding modulation. In: IEEE/CVF International Conference on Computer Vision, pp. 1141–1150 .

Chen, Q., Chen, Y., Huang, Y., Xie, X., & Yang, L. (2024). Region-based online selective examination for weakly supervised semantic segmentation. *Information Fusion, 107*, Article 102311.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R. (2021). Masked-attention mask transformer for universal image segmentation. *Advances in Neural Information Processing Systems* (pp. 1290–1299).

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13213–13223).

Ding, Z., Wang, J., & Tu, Z. (2023). Open-vocabulary universal image segmentation with MaskCLIP. *International Conference on Machine Learning* (Vol. 202, pp. 8090–8102).

Ding, J., Xue, N., Xia, G.-S., & Dai, D. (2022). Decoupling zero-shot semantic segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11573–11582).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision, 88*(2), 303–338.

Ghiasi, G., Gu, X., Cui, Y., & Lin, T. Y. (2022). Scaling open-vocabulary image segmentation with image-level labels. *European Conference on Computer Vision* (pp. 540–557).

Gupta, A., Dollár, P., & Girshick, R. (2019). Lvis: A dataset for large vocabulary instance segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5351–5359).

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., & Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. *International Conference on Machine Learning* (Vol. 139, pp. 4904–4916).

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollar, P., & Girshick, R. (2023). Segment anything. *IEEE/CVF International Conference on Computer Vision* (pp. 4015–4026).

Li, Y. (2022). Exploring plain vision transformer backbones for object detection. *European Conference on Computer Vision* (pp. 280–296).

Li, Y., Fan, J., Pan, Y., Yao, T., Lin, W. & Mei, T. (2022). Uni-eden: Universal encoder-decoder network by multi-granular vision-language pre-training. ACM Transactions on Multimedia Computing, Communications, and Applications **18**(2) .

Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P. & Marculescu, D. (2023). Open-vocabulary semantic segmentation with mask-adapted clip. In: IEEE/CVF International Conference on Computer Vision, pp. 7061–7070 .

Li, J., Chen, T., Wang, X., Zhong, Y., & Xiao, X. (2025). Adapting the segment anything model for multi-modal retinal anomaly detection and localization. *Information Fusion, 113*, Article 102631.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. (2014). Microsoft COCO: Common objects in context. *European Conference on Computer Vision* (pp. 740–755).

Liu, Y., Bai, S., Li, G., Wang, Y., & Tang, Y. (2024). Open-vocabulary segmentation with semantic-assisted calibration. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3491–3500).

Liu, Y., Ge, P., Wang, G., Liu, Q., & Huang, D. (2025). Multi-grained contrastive learning for text-supervised open-vocabulary semantic segmentation. ACM Transactions on Multimedia Computing, Communications, and Applications **21**(3) .

Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. (2024). Segment anything in medical images. *Nature Communications, 15*, 1–9.

Mottaghi, R., Chen, X., Liu, X., Cho, N.-G., Lee, S.-W., Fidler, S., Urtasun, R., & Yuille, A. (2014). The role of context for object detection and semantic segmentation in the wild. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 891–898).

Niu, H., Hu, J., Lin, J., Jiang, G., & Zhang, S. (2025). Eov-seg: efficient open-vocabulary panoptic segmentation. AAAI Conference on Artificial Intelligence.

Pan, F., Shin, I., Rameau, F., Lee, S., & Kweon, I. S. (2020). Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3763–3772).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning* (Vol. 139, pp. 8748–8763).

Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., & Lu, J. (2022). DenseCLIP: Language-guided dense prediction with context-aware prompting. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18061–18070).

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10684–10695).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems (Vol. 30).

Wu, S., Zhang, W., Xu, L., Jin, S., Li, X., Liu, W., & Loy, C.C. (2024). CLIPSelf: Vision transformer distills itself for open-vocabulary dense prediction. In: International Conference on Learning Representations .

Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., & De Mello, S. (2023). Open-vocabulary panoptic segmentation with text-to-image diffusion models. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2955–2966).

Xu, X., Xiong, T., Ding, Z., & Tu, Z. (2023). Masqclip for open-vocabulary universal image segmentation. *IEEE/CVF International Conference on Computer Vision* (pp. 887–898).

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

Yang, H., Ma, C., Wen, B., Jiang, Y., Yuan, Z., & Zhu, X. (2024). Recognize any regions. *Advances in Neural Information Processing Systems* (pp. 51312–51332).

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., & Darrell, T. (2020). BDD100k: A diverse driving dataset for heterogeneous multitask learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2636–2645 .

Yu, Q., He, J., Deng, X., Shen, X. & Chen, L.-C. (2023). Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional CLIP. *Advances in Neural Information Processing Systems* (Vol. 36, pp. 32215–32234).

Yuan, H., Li, X., Zhou, C., Li, Y., Chen, K., & Loy, C.C. (2024). Open-vocabulary SAM: Segment and recognize twenty-thousand classes interactively. In: European Conference on Computer Vision, pp. 419–437 .

Zhang, C., Han, D., Qiao, Y., Kim, J. U., Bae, S.-H., Lee, S., & Hong, C. S. (2023). Faster segment anything: Towards lightweight sam for mobile applications arXiv preprint. arXiv:2306.14289

Zhao, X., Feng, W., Zhang, Z., Lv, J., Zhu, X., Lin, Z., Hu, J., & Shao, J. (2024). Cbnet: A plug-and-play network for segmentation-based scene text detection. *International Journal of Computer Vision, 132*(8), 3119–3138.

Zhao, S., Li, B., Xu, P., Yue, X., Ding, G., & Keutzer, K. (2021). Madan: Multi-source adversarial domain aggregation network for domain adaptation. *International Journal of Computer Vision, 129*(8), 2399–2424.

Zhao, S., Yao, H., Lin, C., Gao, Y., & Ding, G. (2024). Multi-source-free domain adaptive object detection. *International Journal of Computer Vision,132*(12), 5950–5982.

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., & Torralba, A. (2017). Semantic understanding of scenes through the ADE20k dataset. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 302–321).

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to prompt for vision-language models. *International Journal of Computer Vision, 130*(9), 2337–2348.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2021). Deformable detr: Deformable transformers for end-to-end object detection. International Conference on Learning Representations.

Zou*, X., Dou*, Z.-Y., Yang*, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., Peng, N., Wang, L., Lee*, Y.J., & Gao*, J.: Generalized decoding for pixel, image and language. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15116–15127 (2023).

Zou*, X., Yang*, J., Zhang*, H., Li*, F., Li, L., Wang, J., Wang, L., Gao*, J., & Lee*, Y.J. (2023). Segment everything everywhere all at once. In: Advances in Neural Information Processing Systems, pp. 19769–19782 .