International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02185-6

Towards Training-Free Open-World Segmentation via Image Prompt Foundation Models

Lv Tang[1] · Peng-Tao Jiang[1] · Haoke Xiao[1] · Bo Li[1]

Received: 16 December 2023 / Accepted: 17 June 2024

Abstract
The realm of computer vision has witnessed a paradigm shift with the advent of foundational models, mirroring the transformative influence of large language models in the domain of natural language processing. This paper delves into the exploration of open-world segmentation, presenting a novel approach called Image Prompt Segmentation (IPSeg) that harnesses the power of vision foundational models. IPSeg lies the principle of a training-free paradigm, which capitalizes on image prompt techniques. Specifically, IPSeg utilizes a single image containing a subjective visual concept as a flexible prompt to query vision foundation models like DINOv2 and Stable Diffusion. Our approach extracts robust features for the prompt image and input image, then matches the input representations to the prompt representations via a novel feature interaction module to generate point prompts highlighting target objects in the input image. The generated point prompts are further utilized to guide the Segment Anything Model to segment the target object in the input image. The proposed method stands out by eliminating the need for exhaustive training sessions, thereby offering a more efficient and scalable solution. Experiments on COCO, PASCAL VOC, and other datasets demonstrate IPSeg's efficacy for flexible open-world segmentation using intuitive image prompts. This work pioneers tapping foundation models for open-world understanding through visual concepts conveyed in images.

Keywords Open-world Segmentation · Vision Foundations models · Image Prompt

1 Introduction

In recent years, large language models (LLMs) (Chowdhery et al, 2023; Touvron et al, 2023; Zhang et al, 2022) have sparked a revolution in natural language processing (NLP). These foundational models exhibit remarkable transfer capabilities, extending far beyond their initial training objectives. LLMs showcase robust generalization abilities and excel in a multitude of open-world language tasks, including language comprehension, generation, interaction, and reasoning. Inspired by the success of LLMs, vision foundational models such as CLIP (Radford et al, 2021), DINOv2 (Oquab et al, 2023), BLIP (Li et al, 2022), and SAM (Kirillov et al, 2023) have also emerged. These models, once trained, can seamlessly apply their knowledge to various downstream tasks. Such a trend has further motivated researchers to explore ways of open-world visual understanding.

Pioneering works (Liu et al, 2023a; Dai et al, 2023; Zhu et al, 2023a) have mainly focused on how to understand images as a whole in the open world. Herein, we project our viewpoint to open-world understanding at the object level, specifically for the task of open-world segmentation (Qi et al, 2022). When approaching open-world segmentation tasks, there are three primary strategies for leveraging foundational models. The most widely studied approach (Liang et al, 2023; Oin et al, 2023; Ghiasi et al, 2022) is to utilize a vision foundation model like CLIP or DINOv2 and cooperate it with a specific segmentation header or adapter to complete the

Communicated by Hong Liu.

L. Tang, P. Jiang and H. Xiao have contributed equally to this work.

✉ Bo Li
libra@vivo.com

Lv Tang
lvtang@vivo.com

Peng-Tao Jiang
pt.jiang@vivo.com

Haoke Xiao
xiaohaoke@vivo.com

[1] vivo Mobile Communication Co., Ltd, Shanghai, China

Published online: 16 July 2024

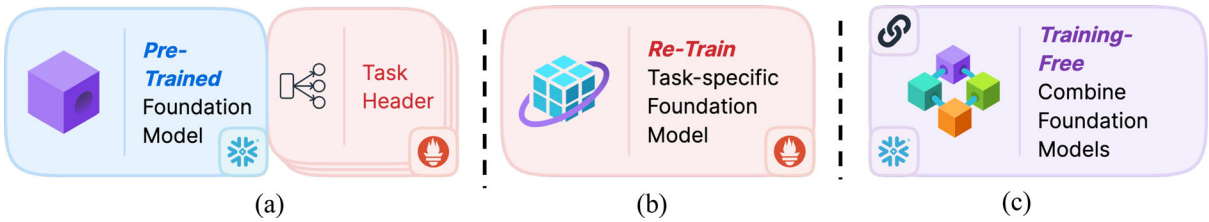

Fig. 1 Comparison of different open-world segmentation frameworks based on foundation models. From left to right, they are foundation model adaptions, task-specific foundation models training from scratch, and training-free foundation models

open-world segmentation task. Such methods (Fig. 1a) often require fine-tuning or training the segmentation header or adapter. In addition to the above methods combining foundation model with adapter, some researchers have tried to draw on the successful experience in NLP and directly train a foundation model for generic dense-prediction vision problems, as demonstrated in works like Painter (Wang et al, 2023a). Such models (Fig. 1b) can complete open-world segmentation simply with a task-specific prompt. Lately, the Segment Anything Model (SAM) (Kirillov et al, 2023) has attained remarkable zero-shot segmentation results. It presents researchers with the prospect of devising an alternative way to accomplish open-world segmentation without the need for training (Fig. 1c). For example, PerSAM (Zhang et al, 2023b) effectively transfers SAM to open-world object segmentation tasks in a training-free manner through the design of the cross-attention layer in SAM's decoder, thereby tapping into the potential of vision foundational models to a significant extent. While these approaches have achieved excellent performance, incorporating more vision foundation models to improve the generalization capability and segmentation quality for open-world segmentation remains an avenue for further inspection.

In addition to the architectural design of the foundation model for open-world segmentation tasks, another critical aspect is the development of flexible and user-friendly prompts. This ensures that the model accurately grasps the visual concepts users desire. As shown in Fig. 2a, b existing works typically rely on predefined textual descriptions or high-quality annotations for a given image as the segmentation prompt, which lacks flexibility. Yet, in the context of open-world scenarios, we not only expect the network to perform well on various open-set datasets but also need it to handle object segmentation tasks with more versatile prompt information. Therefore, a fundamental question emerges: Could we prompt the foundational models, such as SAM, to segment specific objects based on the prompt of the user-given image that contains objects with a clear subjective concept?

Motivated by this question, we present a novel open-world segmentation framework, which utilizes image prompts to instruct the training-free vision foundational models to segment open-world objects. The proposed Image Prompt Segmentation (IPSeg) network is a straightforward yet highly effective framework, comprising three main components, i.e., feature extraction, feature interaction, and segmentation. For the feature extraction, we design two branches, including the **prompt** and the **input** branches. The **prompt** branch is dedicated to capturing general representations of subjective objects belonging to a specific category from the prompt image, and the extracted representations are employed to identify the objects in the input image.

The **input** branch is designed to capture the feature representation of the input image to be segmented, following the same architecture proposed in the **prompt** branch. For the feature interaction, we've devised a feature interaction module to facilitate interaction between the input image features and the given image prompt features, thereby accentuating the pixel points of the target objects. Finally, the generated pixel points serve as the prompt information for SAM (Kirillov et al, 2023), guiding SAM in predicting the final segmentation map.

In summary, the key contributions are listed as follows:

- We propose a training-free open-world object segmentation framework based on foundational models. We take the pioneering step of utilizing image prompts with clear target objects to query generic object representations from foundational models. Such a framework can potentially inspire researchers to address open-world segmentation from a fresh perspective.
- We introduce a simple but effective framework, coined as IPSeg, which contains three effective components. They are utilized to extract discriminative features of target objects identified in the given image prompt and generate accurate points to prompt SAM models to generate object masks.
- We validate the proposed IPSeg framework on widely used segmentation datasets, including COCO-20^i (Nguyen and Todorovic, 2019), FSS-1000 (Li et al, 2020) and PerSeg (Zhang et al, 2023b). Compared to methods PerSAM and Painter, our proposed method can achieve a 30.6% and 42.8% improvement in the mIoU metric with flexible prompts under a training-free mechanism.

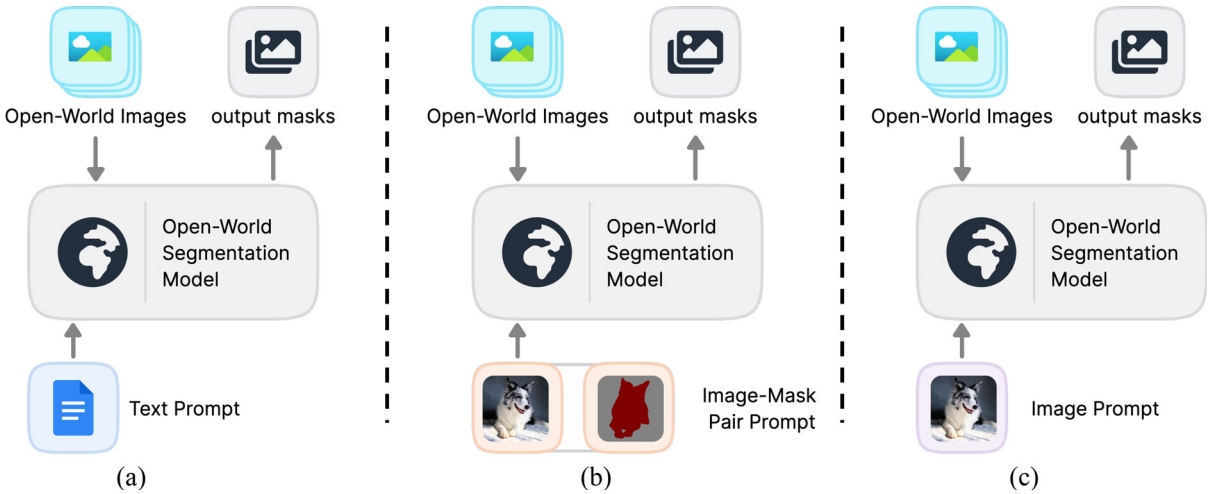

Fig. 2 Different prompt forms in existing open-world segmentation methods. The left is the prompt of predefined textual descriptions or categories. The middle is the prompt form used in existing one-shot object segmentation works (Liu et al, 2023b; Zhang et al, 2023b). The right is the prompt form used in this paper, which only uses one image containing a salient object with specific visual concepts

2 Related Works

2.1 Large Vision Models (LVMs)

Prompted by the powerful generalized ability of large language models (Devlin et al, 2018; Lu et al, 2019; Brown et al, 2020; Radford et al, 2018, 2019; Zhang et al, 2023a) in nature language processing, large vision models (Oquab et al, 2023; Kirillov et al, 2023; Radford et al, 2021) have emerged. Among these large vision models, CLIP (Radford et al, 2021) align the image and text feature spaces through contrastive learning on the huge number of image-text pairs, whose models show powerful zero-shot generalization ability on various downstream vision tasks (Xu et al, 2023), such as open-world segmentation (Qi et al, 2022; Cen et al, 2021). SAM (Kirillov et al, 2023) train a prompt-based large segmentation model on 1 billion masks. The prompt-based segmentation model can accurately segment objects in images from different domains. Such ability has facilitated different applications, such as object tracking (Yang et al, 2023; Cheng et al, 2023; Zhu et al, 2023b), image segmentation (Zhang and Liu, 2023; Chen et al, 2023; Tang et al, 2023; Jiang and Yang, 2023), 3D reconstruction (Cen et al, 2023; Shen et al, 2023) etc. Besides, DINOv2 (Oquab et al, 2023) learn powerful object-level representations in an unsupervised manner. Such powerful representations facilitate downstream dense scene parsing tasks, such as semantic segmentation (Chen et al, 2017; Long et al, 2015), and depth estimation (Ranftl et al, 2021).

2.2 Open-World Segmentation

Open-world segmentation aims to extend traditional close-set segmentation models (Long et al, 2015; Chen et al, 2017) to enable open-set pixel classification, making them more versatile and capable of generalization. The models of open-world segmentation (Cui et al, 2020; Cen et al, 2021; Qi et al, 2022) need to be able to handle unknown classes. There exist several kinds of open-world segmentation methods. The first line of works attempts (Xia et al, 2020; Cen et al, 2021; Angus et al, 2019; Hammam et al, 2023) to classify the pixels of objects out of the training set's distribution to 'anomaly'. They do not distinguish different novel classes in "anomaly", in detail. The second line of works (Xian et al, 2019; Bucher et al, 2019) usually trains segmentation models on datasets with a fixed number of seen classes and utilizes the models to segment images with unseen classes. They strive to improve the generalization of segmentation embedding to unseen classes.

Recently, owing to LVMs, such as CLIP (Radford et al, 2021) have shown significant zero-shot classification ability, researchers attempt to transfer their image-level classification ability to region-level classification. These methods (Luo et al, 2023; Xu et al, 2023; Ma et al, 2022; Liang et al, 2023; Xu et al, 2022b; Zhou et al, 2023c; Liu et al, 2022) adapt CLIP models to the open-world segmentation models by training on the datasets with seen classes to align the predicted region features and text features. Among the methods using LVMs, some works (Zhou et al, 2022; Liu et al, 2023b; Zhang et al, 2023b) also attempt to utilize training-free LVMs and design prompts to conduct open-world segmentation. Without fine-tuning LVMs, they directly extract object seg-

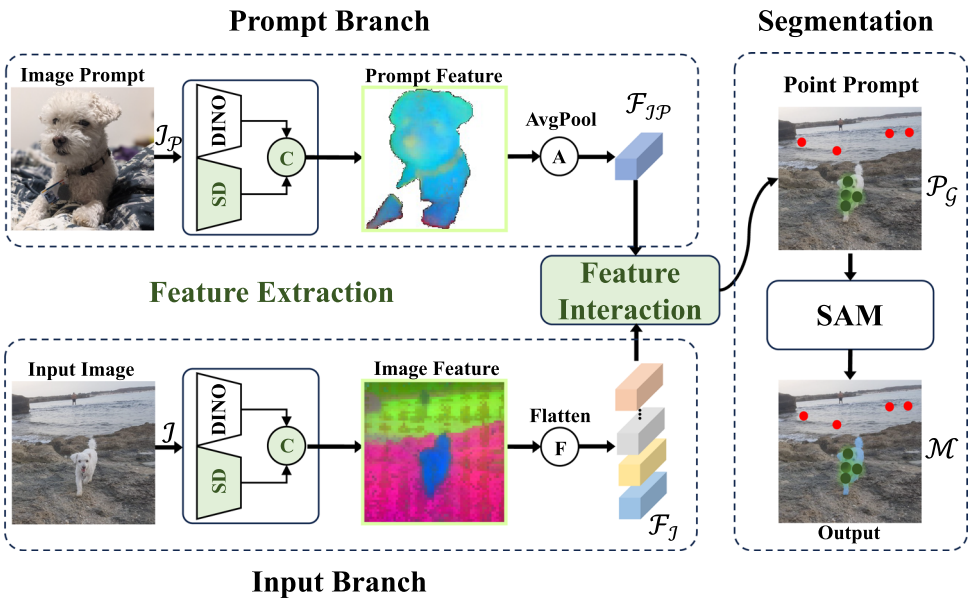

Fig. 3 The framework of our proposed IPSeg framework. Importantly, all parameters in the network remain frozen, eliminating the need for additional training. The green point in $\mathcal{P}_\mathcal{G}$ represents the positive point prompts sent to SAM, while the red point represents the negative point prompts sent to SAM (Color figure online)

mentation masks from them. Zhou et al. Zhou et al (2022) conduct minimal modification of the CLIP model to extract segmentation masks of open-world categories. Liu et al. Liu et al (2023b) and Zhang et al. Zhang et al (2023b) utilize an image with an object mask to extract prompts. Then, the prompts are used to instruct the SAM model to segment objects of the target category indicated in the provided image.

Our proposed method also falls into the training-free LVMs categories. Different from previous works using image-mask pairs, we only utilize an image containing the objects of the target concept as prompts to conduct open-world segmentation. Image prompts are more flexible than image-mask pairs, as humans do not need to annotate the objects of the target class. Besides, we also utilize off-of-the-shelf LVMs, such as DINOv2, to extract discriminative feature representations of image prompts. Then, discriminative feature representations are used to prompt LVMs to segment target objects in test images.

3 Method

We first introduce the preliminaries about the Segmentation Anything Models (SAM) (Kirillov et al, 2023), used in this paper. Then, we introduce the proposed IPSeg framework, which is shown in Fig. 3. Given an image prompt with a clear concept, IPSeg is capable of segmenting any semantically identical object under the open-world setting.

3.1 Preliminaries

SAM consists of three components: a prompt encoder Enc$\mathcal{P}$, an image encoder Enc$\mathcal{I}$, and a lightweight mask decoder Dec$\mathcal{M}$. As a prompt-based framework, SAM takes as input an image $\mathcal{I}$, and prompts $\mathcal{P}$ (like specific points). Specifically, SAM initially utilizes Enc$\mathcal{I}$ to extract features from the input image and employs Enc$\mathcal{P}$ to encode the provided prompts into prompt tokens:

$$F_\mathcal{I} = Enc\mathcal{I}(\mathcal{I}), \quad T_\mathcal{P} = Enc\mathcal{P}(\mathcal{P}). \tag{1}$$

Afterwards, the encoded image $F_\mathcal{I}$ and prompts $T_\mathcal{P}$ are input into the decoder Dec$\mathcal{M}$ for feature interaction. It's worth noting that SAM constructs the decoder's input by concatenating several learnable mask tokens $T_\mathcal{M}$ as prefixes to the prompt tokens $T_\mathcal{P}$. These mask tokens are responsible for generating the mask output, formulated as:

$$\mathcal{M} = Dec\mathcal{M}(F_\mathcal{I}, Concat(T_\mathcal{M}, T_\mathcal{P})), \tag{2}$$

where $\mathcal{M}$ denotes the final segmentation masks predicted by SAM.

As discussed above, SAM can segment objects in an image based on the given prompt. Therefore, the core of this paper lies in how to find semantically matching points in the image $\mathcal{I}$ to be segmented when given an image prompt $\mathcal{I}_\mathcal{P}$ that contains clear visual concepts. This, in turn, guides SAM in generating segmentation results. Note we focus on constructing an image-prompt open-world framework. Exploring prompts, like bounding boxes, is out of the scope of this paper.

3.2 Overview

The pipeline of our method is shown in Fig. 3. The proposed IPSeg framework comprises three components: feature

extraction, feature interaction and SAM. The feature extraction module is used in the **prompt** branch and **input** branch, which can extract the discriminative feature representations of both input image $\mathcal{I}$ and image prompt $\mathcal{I}_{\mathcal{P}}$. Then, the prompt feature $\mathcal{F}_{\mathcal{IP}}$ interacts with the input image feature $\mathcal{F}_{\mathcal{I}}$ in the feature interaction module, to generate specialized prompts $\mathcal{P}_{\mathcal{G}}$ such as points in the input image, which contains the same semantic information with the prompt image. Finally, the generated prompt $\mathcal{P}_{\mathcal{G}}$ and the input image $\mathcal{I}$ are sent to SAM, generating the final prediction $\mathcal{M}$. We will provide detailed explanations of the first two components in the subsequent subsections.

3.3 Feature Extraction

Extracting a robust feature representation from both the prompt image $\mathcal{I}_{\mathcal{P}}$ and the input image $\mathcal{I}$, which effectively captures the visual semantic information in both sets of images, also ensures that the network can find a consistent semantic object between these two sets of images. Generally, the feature representation of an image can be divided into high-level feature representation and low-level feature representation. In this paper, we explore how to extract a feature representation of an image from both of these aspects.

In the following, we first introduce the feature extraction process. Then, we introduce how we utilize the feature extraction to constitute the **prompt** and **input** branch of the IPSeg framework.

3.3.1 Feature Extraction

High-level Feature Extraction

Previous study (Oquab et al, 2023) has established that features from Vision Transformers, particularly those from DINOv2, are rich in explicit information pertinent to semantic segmentation and are highly effective when used as K-Nearest Neighbors classifiers. DINOv2, in essence, excels at extracting semantic content with high accuracy from each image. Consequently, we have chosen to utilize the features extracted by the foundational model DINOv2 to represent the semantic information of each image, denoted as $\mathcal{F}_{\mathcal{D}}$.

Low-Level Feature Extraction

DINOv2 is proficient in capturing significant high-level semantic information, yet it has limitations in providing intricate low-level detail information. As illustrated in the second column of Fig. 4, the visual features generated exclusively through DINOv2 might miss out on fine-grained low-level details. Notably, there is a discernible research gap in augmenting features extracted by DINOv2 with low-level detail information without necessitating additional training.

In our proposed IPSeg, integrating a pre-trained model that specializes in capturing low-level detail information becomes vital. Such a model is capable of effectively compensating for the detailed information that might be overlooked by DINOv2. Notably, Stable Diffusion (SD) (Rombach et al, 2022) has recently been recognized for its exceptional prowess in generating high-quality images, underscoring its ability to robustly represent images with comprehensive content and detailed information. Consequently, our primary focus is to explore the potential benefits of combining SD features with DINOv2 in enhancing the overall quality of feature representations.

The architecture of SD consists of three key components: an encoder $\mathcal{E}_{nc}$, a decoder $\mathcal{D}_{ec}$, and a denoising U-Net $\mathcal{U}_{net}$ operating within the latent space. We initiate the process by projecting an input image I_0 into the latent space using the encoder $\mathcal{E}_{nc}$, resulting in a latent code $x_0 = \mathcal{E}_{nc}(I_0)$. Subsequently, we introduce Gaussian noise ϵ to the latent code, following a predefined time step t. Finally, utilizing the latent code x_t at time step t, we extract the SD features $\mathcal{F}_{\mathcal{S}}$ through the denoising U-Net:

$$\mathcal{F}_{\mathcal{S}} = \mathcal{U}_{net}(x_t, t),\ x_t = \sqrt{\bar{a}_t}x_0 + \sqrt{1-\bar{a}_t}\epsilon. \tag{3}$$

$\bar{a}_t$ is utilized to determine the noise schedule (Ho et al, 2020).

Feature Fusion

Building upon the discussions mentioned earlier, we present a straightforward yet notably effective fusion strategy. This strategy is designed to capitalize on the strengths of both SD and DINOv2 features:

$$\mathcal{F}_{\mathcal{F}} = Cat(\mathcal{F}_{\mathcal{S}}, \mathcal{F}_{\mathcal{D}}), \tag{4}$$

where $Cat(,)$ denotes feature concatenation along the channel dimension. In the third and sixth columns of Fig. 4, the fused feature aids in generating a smoother and more resilient visual feature, which helps for feature matching. Specifically, the addition of SD enhances the internal features of foreground objects, making them smoother and more consistent, thereby assisting the network in extracting target objects from segmented images.

3.3.2 Input and Prompt Branches

After introducing the pipeline of the feature extraction, we utilize the visual encoder to extract features for the input image (**input** branch) and image prompt (**prompt** branch), respectively.

Input Branch

For the input image $\mathcal{I}$, we use the above process to extract the feature $\mathcal{F}_{\mathcal{I}} \in \mathbb{R}^{H\times W\times C}$, where H, W mean the spatial

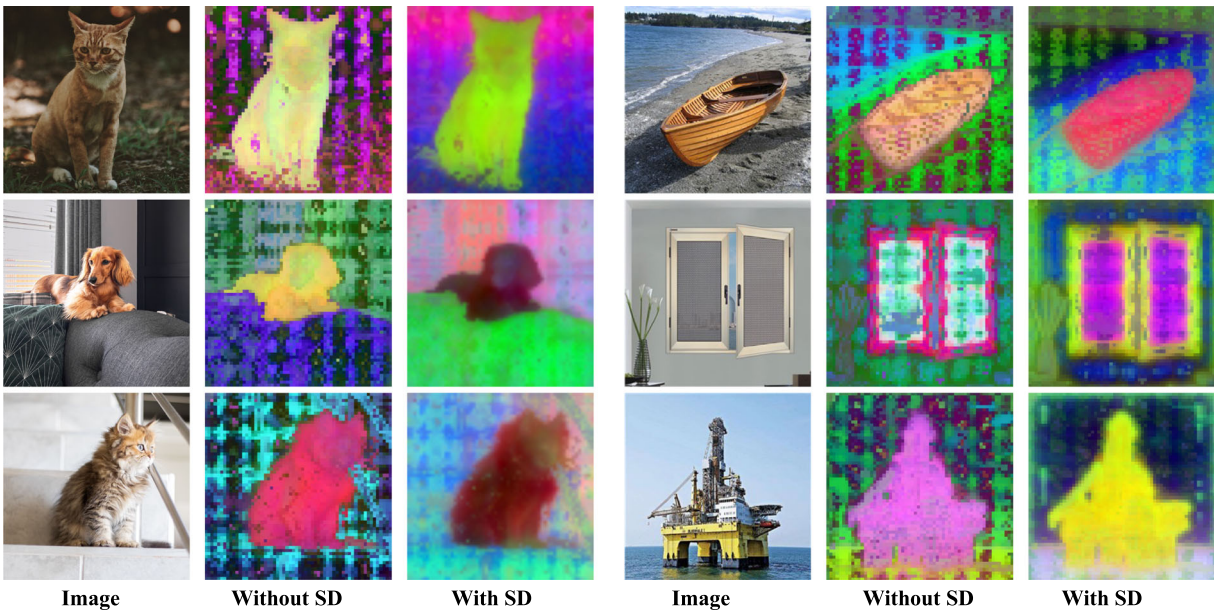

Fig. 4 Visualization results of features extracted from different models. The second and fifth columns indicate the use of only the DINOv2 model for feature extraction, while the third and sixth columns denote the use of both DINOv2 and SD models for this purpose

resolution of the feature and C means the channel number. Then, we reshape the $\mathcal{F}_{\mathcal{I}}$ to $\mathbb{R}^{HW\times C}$, where HW means the total number in the feature and the representation of each pixel is $\mathbb{R}^{C}$.

Prompt Branch

For the image prompt $\mathcal{I}_{\mathcal{P}}$, we also extract its feature $\mathcal{F}_{\mathcal{IP}} \in \mathbb{R}^{H\times W\times C}$ through the above process. Since we do not care about the background information of this feature, we use an unsupervised salient object detection method TSDN (Zhou et al, 2023b) to filter these pixels belong to the background, then use the average pooling (*Avgpool*) operation to generate the prompt embedding:

$$\mathcal{F}_{IP} = Avgpool(\mathcal{F}_{IP} \odot \mathcal{MS}), \tag{5}$$

where $\odot$ denotes pixel-wise multiplication. The object map $\mathcal{MS}$ is directly obtained by the unsupervised method TSDN.

3.4 Feature Interaction and Segment

After generating input image feature $\mathcal{F}_{\mathcal{I}}$ and input prompt feature vector $\mathcal{F}_{\mathcal{IP}}$, we can obtain specific point prompt for the input image $\mathcal{I}$ by performing interaction between $\mathcal{F}_{\mathcal{I}}$ and $\mathcal{F}_{\mathcal{IP}}$.

Concretely, for input image feature which contains HW pixels, the feature representation of each pixel is denoted as $\mathcal{F}_{\mathcal{I}}^{l}$, where $l \in [1, HW]$. Firstly, we calculate the correlation score between $\mathcal{F}_{\mathcal{IP}}$ and $\mathcal{F}_{\mathcal{I}}^{l}$ through cosine similarity. Secondly, we utilize a TopK algorithm to select the points in the input image that are most semantically similar to the prompt image, which are at position P_{coord}:

$$S = \mathcal{F}_{IP} \otimes \mathcal{F}_{I},\ P_{coord} = \mathrm{TopK}(S) \in \mathbb{R}^{K}, \tag{6}$$

where $\otimes$ means matrix multiplication. As shown in Fig. 5, The foreground object in the prompt image and the object to be segmented in the input image maintain good semantic consistency, ensuring the effectiveness of our TopK algorithm.

Finally, we further refine the P_{coord} into c clustering centers as the positive point prompts for SAM. In addition, using the same pipeline, we also selected K points that are the least similar to the prompt image feature and clustered them into c cluster centers as negative point prompts for SAM. We set $K = 32$ and $c = 4$ in this paper. The generated positive/negative point prompts and the input image $\mathcal{F}_{\mathcal{I}}$ are sent to SAM to predict final segmentation results $\mathcal{M}$.

4 Experiment

4.1 Experimental Setup

We employ the Stable Diffusion v1.5 and DINOv2 models as our feature extractors. The DDIM timestep in the denoising process is set to 50 by default. All experiments are conducted on a single RTX A6000 GPU with only 13 G GPU memory. This means that our proposed training-free framework can run on cheaper graphics cards such as RTX3090, providing a good perspective for researchers with limited computing power to explore foundational models.

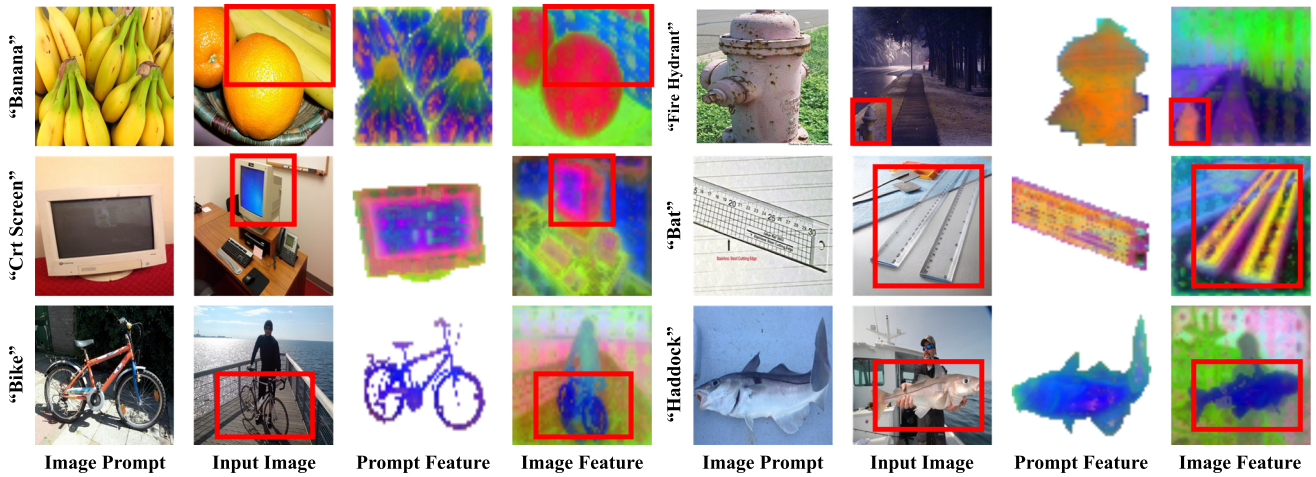

Fig. 5 Visualizing the features of foreground objects in the prompt image and all objects in input prompt

Table 1 Comparison of the few-shot semantic segmentation performance between our proposed method and five typical generalist models. Painter (Wang et al, 2023a), SegGPT (Wang et al, 2023b) and DeLVM (Guo et al, 2024) are three methods which require the extra training process

		COCO-20^i						
Methods	Pub & Year	Fold0	Fold1	Fold2	Fold3	Mean	FSS	PerSeg
Painter	CVPR 2023	31.2	35.3	33.5	32.4	33.1	61.7	56.4
SegGPT	ICCV 2023	56.3	57.4	58.9	51.7	56.1	85.6	95.5
DeLVM	ArXiv 2024	12.6	13.6	10.1	10.5	11.7	36.9	9.8
PerSAM	ICLR 2024	23.1	23.6	22.0	23.4	23.0	81.6	89.5
Matcher	ICLR 2024	52.7	53.5	52.6	52.1	52.7	87.0	94.9
Matcher-Z	ICLR 2024	22.9	23.2	22.2	22.8	22.8	81.2	88.3
Ours	Year 2024	**40.9**	**44.9**	**40.1**	**46.2**	**43.0**	**82.7**	**92.7**

PerSAM (Zhang et al, 2023b) and Matcher (Zhao et al, 2023) are two training-free few-shot works. Matcher-Z means the performance of Matcher under zero-shot setting. **Note that, for IPSeg, we do not utilize the ground truth corresponding to the prompt image to select its foreground object.** We report mIoU (%) in this table

Bold values indicate the performance of our model

4.2 Evaluation Datasets

Following PerSAM (Zhang et al, 2023b), we conduct few-shot experiments on three datasets, including COCO-20^i (Nguyen and Todorovic, 2019), FSS-1000 (Li et al, 2020) and PerSeg (Zhang et al, 2023b) to evaluate the performance of our proposed IPSeg network in the open-world scene. Note that PerSeg is a new dataset collected by PerSAM, which comprises a total of 40 objects from various categories, including daily necessities, animals, and buildings. For each object, there are 5 to 7 images and masks, representing different poses or scenes. We use the same setting in the paper PerSAM to perform experiments. Unlike previous few-shot works utilizing the image-mask pair as input, our method only needs a randomly sampled image as the image prompt.

Moreover, inspired by the work ViL-Seg (Liu et al, 2022), we employ three datasets, including COCO-Stuff (Caesar et al, 2018), PASCAL-VOC (Everingham et al, 2010) and PASCAL-Context (Mottaghi et al, 2014) to evaluate the performance of our IPSeg network in the zero-shot setting. We use the same experimental setting of ViL-Seg to perform the experiments. For the above datasets, **15** classes (frisbee, skateboard, cardboard, carrot, scissors, suitcase, giraffe, cow, road, wall concrete, tree, grass, river, clouds, playingfield) is out of the 183 object categories in COCO-Stuff; **5** classes (potted plant, sheep, sofa, train, tv-monitor) is out of the 20 object categories in PASCAL-VOC; **4** classes (cow, motorbike, sofa, cat) is out of the 59 object categories in PASCAL-Context.

4.3 Quantitative Evaluation

Herein, we do not utilize the ground-truth mask corresponding to the prompt image to select its foreground object for IPSeg.

4.3.1 Compared to Generalist Models

We select five representative open-world object segmentation methods that employ foundational models in distinct ways: Painter (Wang et al, 2023a), SegGPT (Wang et al,

Table 2 Comparison of the zero-shot segmentation performance between our proposed methods and seven typical specialist models

Methods	Pub &Year	NT	COCO-Stuff	PASCAL-VOC	PASCAL-Context
SPNet	CVPR 2019	✓	8.7	15.6	4.0
ZS3	NeurIPS 2019	✓	9.5	17.7	7.7
CaGNet	MM 2020	✓	13.9	29.9	15.0
SIGN	ICCV 2021	✓	15.5	28.9	14.9
ViL-Seg	ECCV 2022	✓	16.4	34.4	16.3
GroupVit	CVPR 2022	✓	16.1	79.0	49.2
TCL	CVPR 2023	✓	27.6	84.5	62.0
Ours	Year 2024	✗	**32.7**	**57.9**	**67.7**

We report mIoU (%) in this table
NT Need Training
Bold values indicate the performance of our model

2023b), DeLVM (Guo et al, 2024), PerSAM (Zhang et al, 2023b) and Matcher (Zhao et al, 2023). Painter, SegGPT and DeLVM are based on a generalized foundation model that is directly trained for various tasks, allowing the use of image-mask pairs for open-world object segmentation. In contrast, PerSAM and Matcher efficiently adapt SAM for open-world object segmentation tasks without the need for additional training. The comparative results are in Table. 1.

As indicated in Table. 1, our proposed method consistently outperforms Painter, DeLVM and PerSAM. This demonstrates the efficacy of our IPSeg network. Specifically, our approach shows significant mIoU performance improvements over PerSAM on the COCO-20^i, FSS, and PerSeg datasets, with improvements of 87.0%, 1.3%, and 3.6%, respectively. A noteworthy point is that Painter, DeLVM and PerSAM rely on image-mask pair inputs, which are more stringent and less flexible approaches compared to our method. This observation suggests that the use of a single image as a prompt, as proposed in our method, is a promising avenue for further research. This approach could serve as an alternative or supplement to the traditional image-mask pair prompts, potentially broadening the scope of research in open-world segmentation tasks.

Note that, our proposed IPSeg is designed for the zero-shot open-world segmentation. Therefore, for fair comparison, we also evaluate the performance of Matcher under zero-shot setting. The zero-shot setting means using the unsupervised salient object detection method TSDN (Zhou et al, 2023b) to filter the background of image prompts instead of their corresponding ground truth. As shown in Table. 1, IPSeg can surpass Matcher's performance in the zero-shot setting (Matcher-Z) by a large margin, which further illustrates the validity of our IPSeg.

4.3.2 Compared to Specialist Models

We have conducted a comparison of our proposed IPSeg network with several well-known specialist zero-shot segmentation methods, including SPNet (Xian et al, 2019), ZS3 (Bucher et al, 2019), CaGNet (Gu et al, 2020), SIGN (Cheng et al, 2021), ViL-Seg (Liu et al, 2022), GroupVit (Xu et al, 2022a) and TCL (Cha et al, 2023). It is important to note that these specialist methods are designed with specific segmentation models, each trained on particular datasets. The comparative results are displayed in Table. 2. Our IPSeg network demonstrates superior performance compared to these specialist models. Notably, it outperforms the CLIP-based ViL-Seg method on the COCO-Stuff, Pascal-VOC, and Pascal-Context datasets, with mIoU performance improvements of 99.4%, 68.3%, and a remarkable 231%, respectively. It is worth mentioning that the Pascal-Context dataset, primarily comprising four common classes, represents relatively simpler scenarios. This aspect may have contributed to the substantial superiority of IPSeg over ViL-Seg in this dataset. Compared to TCL, our method has also achieved competitive performance.

In conclusion, our training-free IPSeg network consistently surpasses specialist open-world object segmentation methods. This success underscores the potential of exploring open-world object segmentation from a novel angle, combining foundational models in a training-free approach. Such an endeavor could significantly enhance the efficiency and applicability of segmentation tasks in diverse real-world scenarios.

4.4 Qualitative Evaluation

In Fig. 6, we showcase the visualization results from our IPSeg network. These visualizations highlight the network's capability in effectively segmenting objects within a variety of complex scenes. This serves as a testament to the effectiveness of our approach from a visual standpoint. Particularly noteworthy is the network's performance in intricate scenarios involving multiple objects, such as scenes labelled 'Dogs' and 'Elephants.' In these cases, our IPSeg network accurately segments the target objects, underscoring its proficiency in

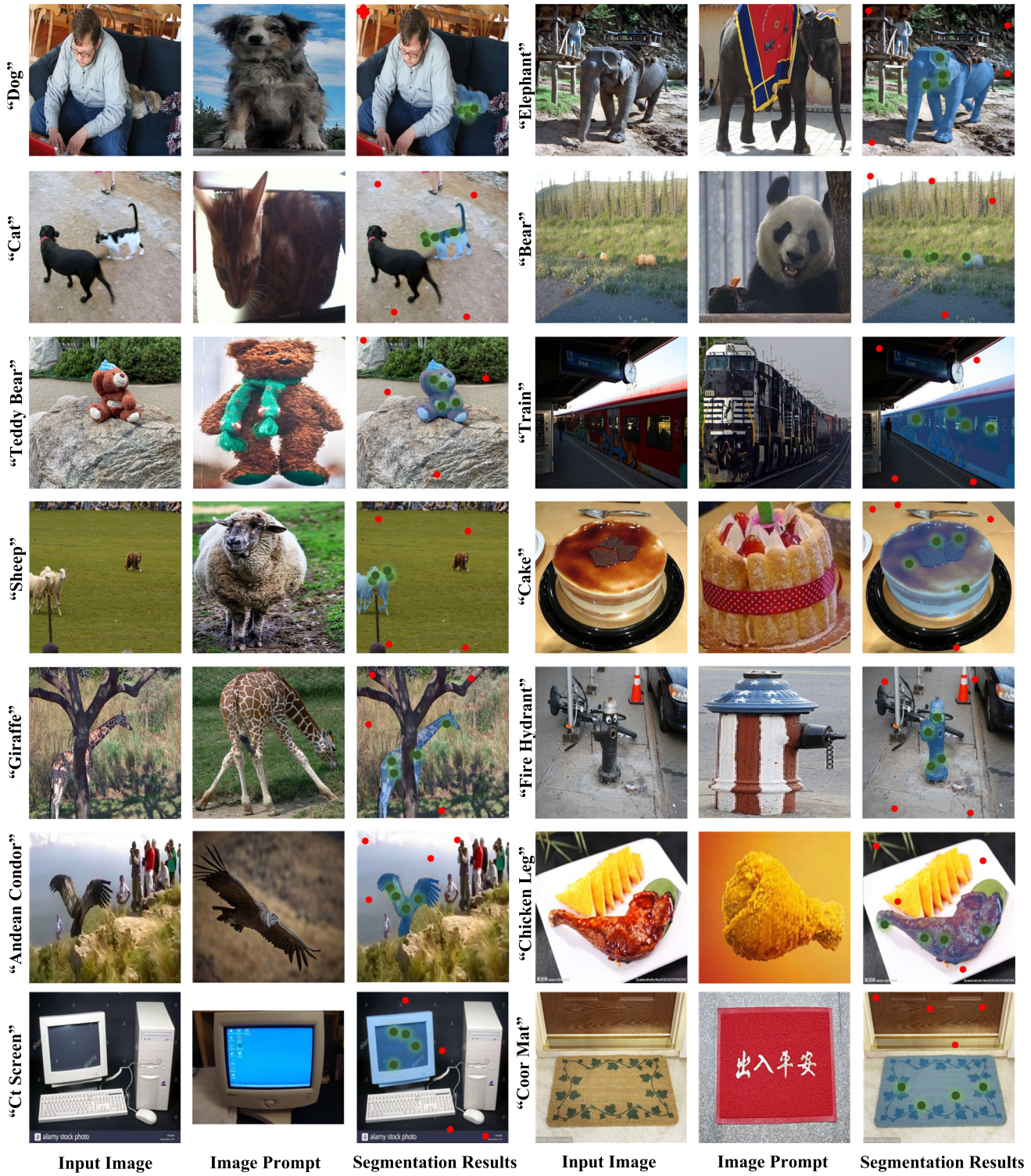

Fig. 6 Qualitative segmentation results of the proposed IPSeg framework. It can be seen that the proposed method can effectively segment the objects contained in the prompt image in the input images from different scenarios. The green point represents positive point prompts sent to SAM, while the red point represents negative point prompts sent to SAM (Color figure online)

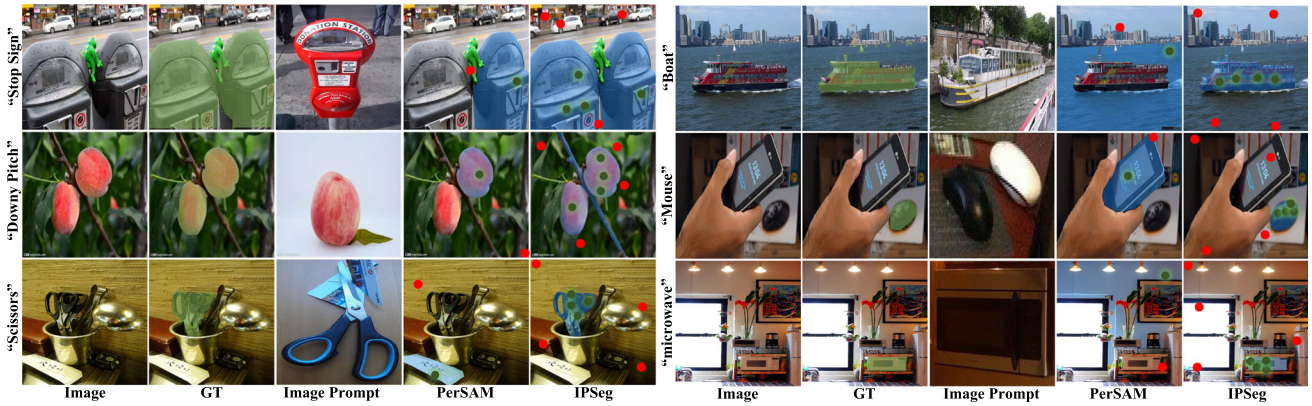

Fig. 7 Qualitative segmentation results of the proposed IPSeg and PerSAM using same image prompts. The green point represents positive point prompts sent to SAM, while the red point represents negative point prompts sent to SAM (Color figure online)

Table 3 Ablation studies of the combination of SD and DINOv2 in this paper

Methods	COCO-20^i					FSS	PerSeg
	Fold0	Fold1	Fold2	Fold3	Mean		
Ours (w/o DINOv2)	21.6	22.0	21.8	22.2	21.9	63.5	72.9
Ours (w/o SD)	39.7	43.5	39.4	44.0	41.7	80.2	90.1
Ours (DINOv2 + SD)	**40.9**	**44.9**	**40.1**	**46.2**	**43.0**	**82.7**	**92.7**

We report mIoU (%) in this table
w/o means without
Bold values indicate the performance of our model

Fig. 8 Further analysis about why adding SD can help improve the performance. The second and fifth columns indicate the use of only the DINOv2 model for feature extraction, while the third and sixth columns denote the use of both DINOv2 and SD models for this purpose

Table 4 Hyperparameters setting in the feature interaction module

Methods	COCO-20^i					FSS	PerSeg
	Fold0	Fold1	Fold2	Fold3	Mean		
Ours(K=32,c=32)	26.1	27.6	26.7	32.6	28.3	50.6	57.8
Ours(K=32,c=16)	40.6	44.2	40.4	44.0	42.3	81.8	90.5
Ours(K=32,c=8)	40.2	45.0	40.3	45.3	42.7	81.2	90.7
Ours(K=32,c=4)	**40.9**	**44.9**	**40.1**	**46.2**	**43.0**	**82.7**	**92.7**
Ours(K=32,c=2)	37.2	40.5	37.8	40.9	39.1	74.7	91.3
Ours(K=4,c=4)	37.5	40.7	37.3	44.7	40.1	70.0	82.2
Ours(K=8,c=4)	39.0	43.5	39.4	46.2	42.0	76.1	84.1
Ours(K=16,c=4)	39.9	44.7	39.9	46.3	42.7	78.2	88.3
Ours(K=32,c=4)	**40.9**	**44.9**	**40.1**	**46.2**	**43.0**	**82.7**	**92.7**
Ours(K=64,c=4)	37.8	42.3	38.2	42.0	40.1	81.6	90.9

We report mIoU (%) in this table
Bold values indicate the performance of our model

correctly identifying objects in the input image that have semantic correspondence with those in the image prompt. This ability showcases the robustness and adaptability of the IPSeg network in dealing with diverse and challenging segmentation tasks. To further illustrate the validity of our method, we conduct some visual comparisons with PerSAM in Fig. 7. It can be seen that in different complex scenes, the performance of our IPSeg is better than that of PerSAM under the same image prompts.

4.5 Ablation Studies

For ablation studies, similar to the experimental setting above, we do not utilize the ground truth corresponding to the prompt image to select its foreground object.

4.5.1 Combination of SD and DINOv2

In the process of feature extraction, our IPSeg network considers both high-level and low-level details from the input and prompt images. Recognizing the limitations of the DINOv2 model in capturing low-level features, we integrate the SD model to address this gap. As shown in Table. 3, incorporating SD significantly boosts the performance of our IPSeg network. This improvement is further evidenced by the visual results in Fig. 4, where the inclusion of SD is observed to result in smoother feature representations. Moreover, as shown in Table. 3, the performance of using solely SD as the feature extractor is clearly inferior to that of using a combination of DINOv2 and SD. One primary reason is that the features extracted by SD lack high-level semantic information. As illustrated in Fig. 8, incorporating features extracted by the SD model allows IPSeg to more distinctly differentiate between foreground and background. This enhancement significantly boosts the performance of IPSeg.

4.5.2 Hyperparameters in Feature Interaction

In feature interaction, we introduce a simple yet effective approach for generating point prompts to guide SAM in generating the corresponding segmentation results. In this module, we compute the similarity between each pixel in the prompt image and the input image using cosine similarity. We use the TopK algorithm to select the TopK most/least similar points, followed by the clustering algorithm to group these points into c cluster centers. In Table. 4, we investigate the impact of different values of K and c on performance. We observe that using the TopK algorithm alone helps the model achieve initial performance (Ours($K = 32, c = 32$)), and further application of the clustering algorithm improves performance even more.

4.5.3 Image Prompt Robustness

In this paper, we introduce a more flexible approach to using image prompts. To further validate the robustness of our model with different image prompt combinations, we randomly selected three different image prompt combinations. Specifically, we prepare appropriate prompt images based on their categories. For all prompt images, we firstly manually choose different prompt images with clear visual representations in certain classes. Then, we randomly compose prompt set-1 to set-3 from these prompt images. Note that, all prompt images are chosen from the used benchmark based on categories, such as COCO and FSS, and we make sure that the selected prompt image and evaluation datasets do not have the same image. From Table. 5, it can be observed that our method maintains good robustness across different prompt inputs. As shown in Fig. 9, when given the same input image with different image prompts, our proposed IPSeg network can consistently generate satisfactory results. This experiment further indicates that in future improvement of this

Table 5 Image prompt robustness of this paper

Methods	COCO-20^i						
	Fold0	Fold1	Fold2	Fold3	Mean	FSS	PerSeg
Ours(Prompt Set-1)	40.8	42.3	39.4	45.1	41.9	82.6	92.1
Ours(Prompt Set-2)	38.9	45.9	40.3	44.8	42.5	82.5	91.9
Ours(Prompt Set-3)	40.9	44.9	40.1	46.2	43.0	82.7	92.7

We report mIoU (%) in this table

Fig. 9 Qualitative results of the proposed IPSeg framework when using different image prompts. When given the same input image with different image prompts, our proposed IPSeg network can consistently generate satisfactory results. This also indicates the robustness of our method. The green point represents positive point prompts sent to SAM, while the red point represents negative point prompts sent to SAM (Color figure online)

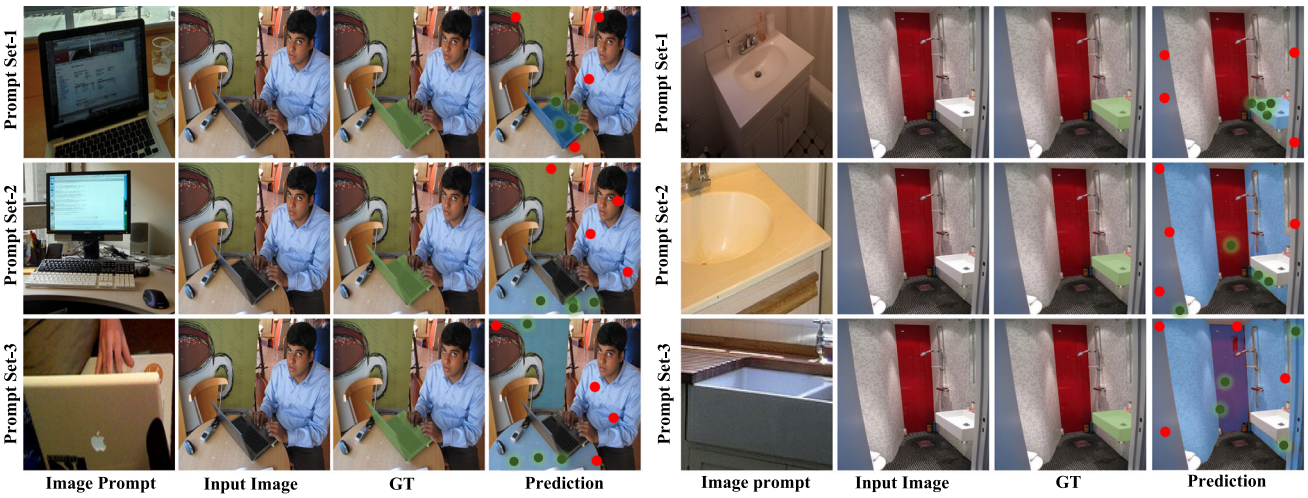

Fig. 10 Some failure prediction results of our IPSeg under different image prompts. The green point represents positive point prompts sent to SAM, while the red point represents negative point prompts sent to SAM (Color figure online)

framework, researchers can have a more flexible choice of prompts, reaffirming the potential of our IPSeg.

Moreover, in Fig. 10, we share some failed cases. Specifically, we showcase objects that are correctly segmented in prompt set-1 but failed in prompt set-2 and set-3. This group of examples indicates that choosing image prompts with only a single, complete target object can significantly aid IPSeg in achieving accurate segmentation results. Hence, when preparing image prompts, we strive to adhere to these two principles for collecting image prompts. However, while aiming for optimal performance, we do not want our framework to be constrained by the reference images. Consequently, the three prompt sets designed in Table. 5 are not deliberately combined. This design approach ensures that the results

Table 6 The impact of background noise on IPSeg

Methods	COCO-20^i					FSS	PerSeg
	Fold0	Fold1	Fold2	Fold3	Mean		
Ours (w/o TSDN)	8.9	10.2	9.1	9.4	9.4	32.8	50.2
Ours (with TSDN)	40.9	44.9	40.1	46.2	43.0	82.7	92.7
Ours (with A2S)	40.1	44.3	39.3	45.9	42.4	82.5	92.0
Ours (with GT)	44.7	48.8	46.9	48.9	47.3	84.7	93.6

We report mIoU (%) in this table
w/o means without operation

Table 7 The performance of IPSeg using different feature extractors

Methods	COCO-20^i					FSS	PerSeg
	Fold0	Fold1	Fold2	Fold3	Mean		
Ours (MAE)	9.9	12.2	9.9	11.6	10.7	36.7	53.7
Ours (CLIP)	20.3	26.8	21.6	29.9	24.7	65.0	89.7
Ours (DINOv2)	40.9	44.9	40.1	46.2	43.0	82.7	92.7

We report mIoU (%) in this table

Table 8 Comparison between IPSeg and PerSAM on other four datasets, containing DAVIS2017 (Pont-Tuset et al, 2017), Pascal-Part (Morabia et al, 2020), PACO-Part (Ramanathan et al, 2023) and LVIS-92^i (Gupta et al, 2019)

Methods	Pub & Year	VOS-DAVIS2017		Part Segmentation		Semantic Segmentation
		J	F	Pascal-Part	PACO-Part	LVIS-92^i
PerSAM	ICLR 2024	71.3	75.1	32.5	22.5	15.6
IPSeg	Year 2024	**75.3**	**77.3**	**34.3**	**29.0**	**20.3**

For video object segmentation (VOS), we report **J** and **F** scores. For part segmentation and semantic segmentation, we report mIoU (%)
Bold values indicate the performance of our model

obtained by IPSeg are reliable and credible and indirectly shows that our framework does not rely on carefully selected image prompts that require extensive time investment.

4.5.4 Impact of Background Noise

To investigate the impact of background noise on our method, we conduct experiments under the following settings: without using the unsupervised salient object detection (USOD) method TSDN (Zhou et al, 2023b) to filter the background, using TSDN to filter the background, and using the ground truth corresponding to the referring image to filter the background. The results are shown in Table. 6. Initially, it is evident that not filtering the background significantly affects our experimental performance. Further improvements are observed upon utilizing TSDN. Finally, by utilizing the ground truth to filter the background noise in the referring image, we can achieve best performance. Moreover, if we use another USOD method A2S (Zhou et al, 2023a), IPSeg's performance does not fluctuate dramatically. This experiment shows that IPSeg requires the USOD method to provide a relatively less noisy image prompt, but is not dependent on a particular USOD method.

4.5.5 Different Feature Extractors

To demonstrate the impact of different feature extractors, inspired by Matcher (Zhao et al, 2023), we use MAE (He et al, 2022) and CLIP (Radford et al, 2021) as feature extractors, and the performance is shown in Table. 7. Using DINOv2 as feature extractor achieves the best performance on all datasets. Additionally, this experiment demonstrates that IPSeg, as a training-free framework, facilitates the integration of various feature extractors.

4.5.6 Transferability of IPSeg

We conduct experiments on several datasets to further demonstrate the effectiveness and transferability of our IPSeg, as shown in Table. 8. These datasets contain video object segmentation benchmark DAVIS2017 (Pont-Tuset et al, 2017), semantic segmentation benchmark LVIS-92^i (Gupta et al, 2019), and part segmentation benchmarks Pascal-Part (Morabia et al, 2020) and PACO-Part (Ramanathan et al, 2023). As shown in Table. 8, the performance of our method can outperform PerSAM for all datasets. This point once again illustrates the validity of our IPSeg.

5 Conclusion

In this paper, we introduce the IPSeg framework for open-world segmentation using visual concepts from a single image. IPSeg is a simple yet highly effective approach designed to inspire researchers to approach open-world segmentation from two pivotal perspectives: efficient utilization of foundational models and a flexible setup for prompt information. Through our exploration of how to optimally combine diverse foundational models, our method attains outstanding performance on six widely utilized datasets. Furthermore, our research underscores the importance of adaptability in foundational models, emphasizing their potential to revolutionize the way we approach complex computer vision challenges. We believe that our contributions will pave the way for future research endeavors, pushing the boundaries of what's possible in open-world segmentation and setting new standards for efficiency and versatility in the field.

Data Availability Statement All the data used in this study are available from third-party institutions. Researchers can access the data through the instructions presented in the original works of the corresponding datasets. However, researchers should follow specific regulations stated by these datasets and use them for only academic purposes.

Code Availability The code of this work is released at https://github.com/luckybird1994/IPSeg.

References

Angus, M., Czarnecki, K., & Salay, R. (2019). Efficacy of pixel-level ood detection for semantic segmentation. arXiv preprint arXiv:1911.02897

Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Bucher, M., Vu, T. H., Cord, M., et al. (2019). Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1906.00817

Caesar, H., Uijlings, JRR. & Ferrari, V. (2018). Coco-stuff: Thing and stuff classes in context. *In: Proceedings of the IEEE conference on computer vision and pattern recognition, computer vision foundation / IEEE computer society*, pp 1209–1218

Cen, J., Yun, P., Cai, J., et al. (2021). Deep metric learning for open world semantic segmentation. In *International conference on computer vision* , pp 15,333–15,342

Cen, J., Zhou, Z., Fang, J., et al. (2023). Segment anything in 3d with nerfs. arXiv preprint arXiv:2304.12308

Cha, J., Mun, J., & Roh, B. (2023). Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Conference on computer vision and pattern recognition IEEE*, pp 11,165–11,174

Chen, L. C., Papandreou, G., Kokkinos, I., et al. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(4), 834–848.

Chen, T., Mai, Z., Li, R., et al. (2023). Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation. arXiv preprint arXiv:2305.05803

Cheng, J., Nandi, S., Natarajan, P., et al. (2021). SIGN: Spatial-information incorporated generative network for generalized zero-shot semantic segmentation. In *International conference on computer vision IEEE*, pp 9536–9546

Cheng, Y., Li, L., Xu, Y., et al. (2023). Segment and track anything. arXiv preprint arXiv:2305.06558

Chowdhery, A., Narang, S., Devlin, J., et al. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research, 24*(240), 1–113.

Cui, Z., Longshi, W., & Wang, R. (2020). Open set semantic segmentation with statistical test and adaptive threshold. In *2020 IEEE International conference on multimedia and expo (ICME), IEEE*, pp 1–6

Dai, W., Li, J., Li, D., et al. (2023). Instructblip: Towards general-purpose vision-language models with instruction tuning. CoRR abs/2305.06500

Devlin, J., Chang, MW., Lee, K., et al. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Everingham, M., Gool, L. V., Williams, C. K. I., et al. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision, 88*(2), 303–338.

Ghiasi, G., Gu, X., Cui, Y., et al. (2022). Scaling open-vocabulary image segmentation with image-level labels. In *European conference on computer vision*, vol 13696. Springer, pp 540–557

Gu, Z., Zhou, S., Niu, L., et al. (2020). Context-aware feature generation for zero-shot semantic segmentation. In *Proceedings of the 28th ACM international conference on multimedia*, pp 1921–1929

Guo, J., Hao, Z., Wang, C., et al. (2024). Data-efficient large vision models through sequential autoregression. arXiv preprint arXiv:2402.04841

Gupta, A., Dollár, P., & Girshick, RB. (2019). LVIS: A dataset for large vocabulary instance segmentation. In *CVPR computer vision foundation / IEEE*, pp 5356–5364

Hammam, A., Bonarens, F., Ghobadi, SE., et al. (2023). Identifying out-of-domain objects with dirichlet deep neural networks. In *International conference on computer vision*, pp 4560–4569

He, K., Chen, X., Xie, S., et al. (2022). Masked autoencoders are scalable vision learners. In *Conference on computer vision and pattern recognition*, pp 16,000–16,009

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems, 33*, 6840–6851.

Jiang, PT., & Yang, Y. (2023). Segment anything is a good pseudo-label generator for weakly supervised semantic segmentation. arXiv preprint arXiv:2305.01275

Kirillov, A., Mintun, E., Ravi, N., et al. (2023). Segment anything. arXiv preprint arXiv:2304.02643

Li. J., Li, D., Xiong, C., et al. (2022). BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning, Proceedings of Machine Learning Research, PMLR*, vol 162, pp 12,888–12,900

Li, X., Wei, T., Chen, YP., et al. (2020). FSS-1000: A 1000-class dataset for few-shot segmentation. In *Conference on computer vision and pattern recognition, Computer vision foundation / IEEE*, pp 2866–2875

Liang, F., Wu, B., Dai, X., et al. (2023). Open-vocabulary semantic segmentation with mask-adapted CLIP. In *Conference on computer vision and pattern recognition. IEEE*, pp 7061–7070

Liu, H., Li, C., Wu, Q., et al. (2023a). Visual instruction tuning. CoRR abs/2304.08485

Liu, Q., Wen, Y., Han, J., et al. (2022). Open-world semantic segmentation via contrasting and clustering vision-language embedding. *European conference on computer vision*, Springer, pp. 275–292

Liu, Y., Zhu, M., Li, H., et al. (2023b). Matcher: Segment anything with one shot using all-purpose feature matching. arXiv preprint arXiv:2305.13310

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Conference on computer vision and pattern recognition*, pp 3431–3440

Lu, J., Batra, D., Parikh, D., et al. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32

Luo, H., Bao, J., Wu, Y., et al. (2023). Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. *International Conference on Machine Learning, PMLR*, pp. 23033–23044

Ma, C., Yang, Y., Wang, Y., et al. (2022). Open-vocabulary semantic segmentation with frozen vision-language models. arXiv preprint arXiv:2210.15138

Morabia, K., Arora, J., & Vijaykumar, T. (2020). Attention-based joint detection of object and semantic part. CoRR abs/2007.02419

Mottaghi, R., Chen, X., Liu, X., et al. (2014). The role of context for object detection and semantic segmentation in the wild. In *Conference on computer vision and pattern recognition. IEEE Computer Society*, pp 891–898

Nguyen, K., & Todorovic, S. (2019). Feature weighting and boosting for few-shot segmentation. In *International conference on computer vision, IEEE*, pp 622–631

Oin, J., Wu, J., Yan, P., et al. (2023). Freeseg: Unified, universal and open-vocabulary image segmentation. In *Conference on computer vision and pattern recognition. IEEE*, pp 19,446–19,455

Oquab, M., Darcet, T., & Théo Moutakanni, ea. (2023). Dinov2: Learning robust visual features without supervision. CoRR

Pont-Tuset, J., Perazzi, F., Caelles, S., et al. (2017). The 2017 DAVIS challenge on video object segmentation. CoRR abs/1704.00675

Qi, L., Kuen, J., Wang, Y., et al. (2022). Open world entity segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*, 8743–8756.

Radford, A., Narasimhan, K., Salimans, T., et al. (2018). Improving language understanding by generative pre-training. OpenAI

Radford, A., Wu, J., Child, R., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*(8), 9.

Radford, A., Kim, JW., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on machine learning, Proceedings of Machine Learning Research, PMLR*, vol 139, pp 8748–8763

Ramanathan, V., Kalia, A., Petrovic, V., et al. (2023). PACO: parts and attributes of common objects. In *CVPR, IEEE*, pp 7141–7151

Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision transformers for dense prediction. In *Conference on computer vision and pattern Recognition*, pp 12,179–12,188

Rombach, R., Blattmann, A., Lorenz, D., et al. (2022). High-resolution image synthesis with latent diffusion models. In *Conference on computer vision and pattern recognition*

Shen, Q., Yang, X., & Wang, X. (2023). Anything-3d: Towards single-view anything reconstruction in the wild. arXiv preprint arXiv:2304.10261

Tang, L., Xiao, H., & Li, B. (2023). Can sam segment anything? when sam meets camouflaged object detection. arXiv preprint arXiv:2304.04709

Touvron, H., Lavril, T., Izacard, G., et al. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971

Wang, X., Wang, W., Cao, Y., et al. (2023a). Images speak in images: A generalist painter for in-context visual learning. In *Conference on computer vision and pattern recognition. IEEE*, pp 6830–6839

Wang, X., Zhang, X., Cao, Y., et al. (2023b). Seggpt: Towards segmenting everything in context. In *International conference on computer vision. IEEE*, pp 1130–1140

Xia, Y., Zhang, Y., Liu, F., et al. (2020). Synthesize then compare: Detecting failures and anomalies for semantic segmentation. *European conference on computer vision*, Springer, pp. 145–161

Xian, Y., Choudhury, S., He, Y., et al. (2019). Semantic projection network for zero-and few-label semantic segmentation. In *Conference on computer vision and pattern recognition*, pp 8256–8265

Xu, J., Mello, SD., Liu, S., et al. (2022a). Groupvit: Semantic segmentation emerges from text supervision. In *Conference on computer vision and pattern recognition. IEEE*, pp 18,113–18,123

Xu, M., Zhang, Z., Wei, F., et al. (2022). A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. *European conference on computer vision*, Springer, pp. 736–753

Xu, M., Zhang, Z., Wei, F., et al. (2023). Side adapter network for open-vocabulary semantic segmentation. In *Conference on Computer Vision and Pattern Recognition*, pp 2945–2954

Yang, J., Gao, M., Li, Z., et al. (2023). Track anything: Segment anything meets videos. arXiv preprint arXiv:2304.11968

Zhang, K., & Liu, D. (2023). Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785

Zhang, R., Han, J., Zhou, A., et al. (2023a). Llama-adapter: Efficient fine-tuning of language models with zero-init attention. arXiv preprint arXiv:2303.16199

Zhang, R., Jiang, Z., Guo, Z., et al. (2023b). Personalize segment anything model with one shot. arXiv preprint arXiv:2305.03048

Zhang, S., Roller, S., Goyal, N., et al. (2022). Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068

Zhao, W., Rao, Y., Liu, Z., et al. (2023). Unleashing text-to-image diffusion models for visual perception. CoRR

Zhou, C., Loy, C. C., & Dai, B. (2022). Extract free dense labels from clip. *European conference on computer vision*, Springer, pp. 696–712

Zhou, H., Chen, P., Yang, L., et al. (2023). Activation to saliency: Forming high-quality labels for unsupervised salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology, 33*(2), 743–755.

Zhou, H., Qiao, B., Yang, L., et al. (2023b). Texture-guided saliency distilling for unsupervised salient object detection. In *Conference on computer vision and pattern recognition*

Zhou, Z., Lei, Y., Zhang, B., et al. (2023c). Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *IEEE conference on computer vision and pattern recognition*, pp 11,175–11,185

Zhu, D., Chen, J., Shen, X., et al. (2023a). Minigpt-4: Enhancing vision-language understanding with advanced large language models. CoRR abs/2304.10592

Zhu, J., Chen, Z., Hao, Z., et al. (2023b). Tracking anything in high quality. arXiv preprint arXiv:2307.13974