Unified Open-World Segmentation with Multi-Modal Prompts

Yang Liu[1*] Yufei Yin[2*] Chenchen Jing[3] Muzhi Zhu[1] Hao Chen[1†] Yuling Xi[1]
Bo Feng[4] Hao Wang[4] Shiyu Li[4] Chunhua Shen[1]

[1] Zhejiang University [2] Hangzhou Dianzi University [3] Zhejiang University of Technology [4] Apple

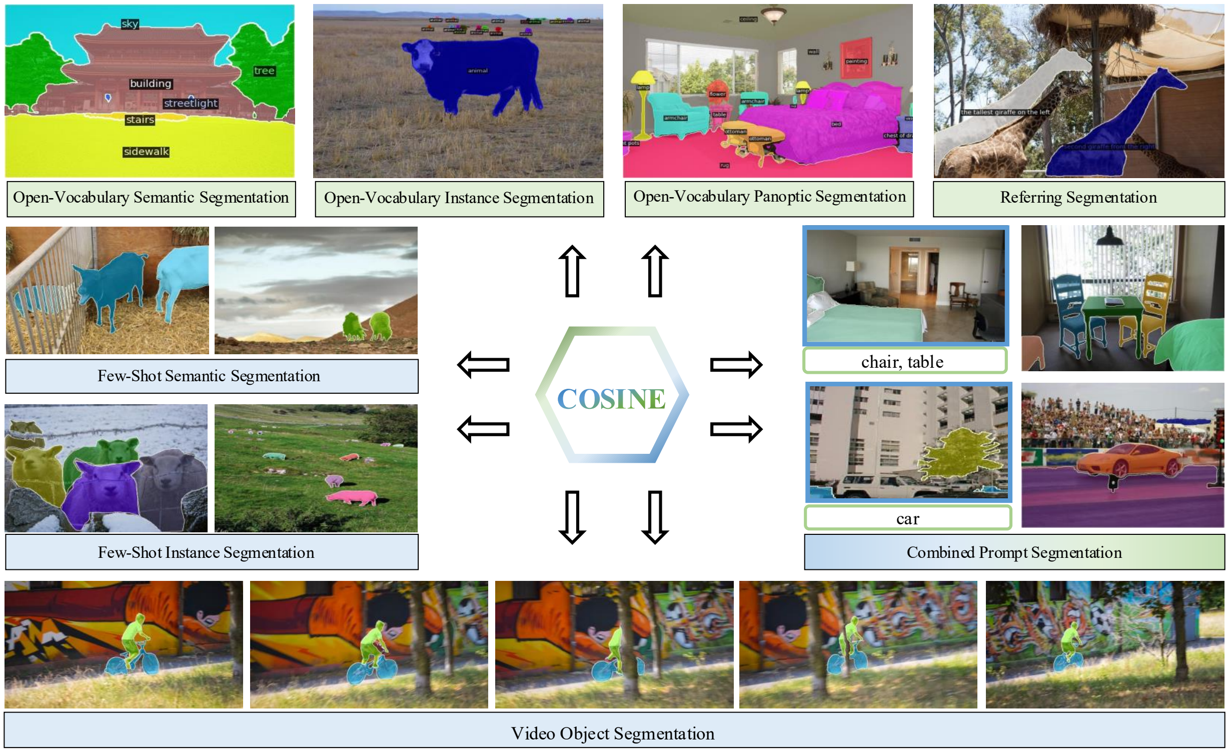

sky
tree
building
streetlight
stairs
sidewalk
Open-Vocabulary Semantic Segmentation
Open-Vocabulary Instance Segmentation
Open-Vocabulary Panoptic Segmentation
Referring Segmentation
Few-Shot Semantic Segmentation
Few-Shot Instance Segmentation
COSINE
chair, table
car
Combined Prompt Segmentation
Video Object Segmentation

Figure 1. COSINE is a unified open-world segmentation model that consolidates open-vocabulary and in-context segmentation. COSINE can simultaneously support text prompts (green boxes) and image prompts (blue boxes) as inputs to perform various segmentation tasks, including semantic segmentation, instance segmentation, panoptic segmentation, referring segmentation, and video object segmentation. In addition, COSINE can collaboratively use different types of prompts to perform various segmentation tasks.

Abstract

*In this work, we present COSINE, a unified open-world segmentation model that **C**onsolidates **O**pen-vocabulary **S**egmentation and **IN**-context s**E**gmentation with multi-modal prompts (e.g., text and image). COSINE exploits foundation models to extract representations for an input image and corresponding multi-modal prompts, and a SegDecoder to align these representations, model their interaction, and obtain masks specified by input prompts across different granularities. In this way, COSINE overcomes architectural discrepancies, divergent learning objectives, and distinct representation learning strategies of previous pipelines for open-vocabulary segmentation and in-context segmentation. Comprehensive experiments demonstrate that COSINE has significant performance improvements in both open-vocabulary and in-context segmentation tasks. Our exploratory analyses highlight that the synergistic collaboration between using visual and textual prompts leads to significantly improved generalization over single-modality approaches. Our code is released at* `https://github.com/aim-uofa/COSINE`.

1. Introduction

Image segmentation [17, 24, 61] aims to provide accurate conceptual localization, enabling precise understanding and

[*] Equal contribution. [†] HC is the corresponding author.

content analysis of images at the pixel level. Recently, image segmentation has gradually evolved from closed-world scenarios [3, 8, 13, 29, 41, 59] to open-world settings [18, 20, 26, 44, 48, 55, 62]. Unlike traditional closed-world segmentation models, which are limited to recognizing a fixed set of categories encountered during training, open-world segmentation models can localize arbitrary relevant objects in the wild based on user-provided prompts. Such models can enhance adaptability and robustness in unpredictable dynamic environments, such as autonomous driving and interactive robotics.

The current landscape of open-world segmentation research primarily revolves around two distinct paradigms: (1) Open-vocabulary segmentation [11, 23, 48, 49, 55, 56], which replaces learnable classifiers with textual embeddings derived from category descriptors, thereby extending conventional closed-set segmentation frameworks to recognize novel categories through natural language alignment; and (2) In-context segmentation [20, 26, 27, 32, 44, 57], which leverages contextual cues from example images to facilitate adaptive object segmentation in query images. Despite significant progress in both directions, current methods predominantly address these tasks in isolation, failing to holistically tackle the complexities of open-world segmentation. Relying solely on text may lead to insufficient fine-grained semantic abstraction, whereas image-based exemplars often lack explicit category boundaries and semantic alignment. This raises a critical question: *Can we unify open-vocabulary segmentation and in-context segmentation within a single framework, leveraging the complementary strengths of textual and visual modalities to enhance open-world segmentation capabilities?*

We compare these two paradigms and identify the following problems in unifying them: 1) Architectural Discrepancies: Existing methods exhibit substantial structural differences. For instance, SegGPT [44] employs a ViT-like [10] encoder-only architecture for in-context segmentation, whereas ODISE [48] adopts a Mask2Former-like [3] encoder-decoder structure for open-vocabulary segmentation. 2) Divergent Learning Objectives: Open-vocabulary segmentation primarily focuses on image-text semantic alignment, aiming to learn the association between images and class descriptions for recognizing novel categories in an open-world setting. In contrast, in-context segmentation emphasizes reference-query relationship modeling, leveraging contextual cues from example images to achieve adaptive target segmentation. 3) Distinct Representation Learning Strategies: Open-vocabulary segmentation typically relies on multimodal models (e.g., CLIP [38], Stable Diffusion [39]) to leverage text embeddings for category matching, whereas in-context segmentation predominantly utilizes vision foundation models (e.g., MAE [14], DINOv2 [36]), performing target localization on visual features. These fundamental differences pose challenges in designing a unified framework that effectively integrates both paradigms while preserving their advantages.

To address these challenges, we present COSINE, a unified open-world segmentation model that **C**onsolidates **O**pen-vocabulary **S**egmentation and **IN**-context s**E**gmentation. Specifically, we first deploy a frozen Model Pool consisting of multiple foundation models to extract representations for the target image and different modality prompts (e.g., text and image), and convert them into token sequences. This standardization of the input format facilitates structural unification across both tasks, thereby enabling a single decoder-only segmentation model, named SegDecoder, to jointly process them. The SegDecoder contains an Image-Prompt Aligner module and a Multi-Modality Decoder. The Image-Prompt Aligner module aligns the image with various prompts, reducing the modality gap and learning a unified multi-modal representation space. The Multi-Modality Decoder is used to model the interaction between object queries, the image, and different modality prompts, enabling the effective generation of masks at different granularities (e.g., semantic, instance). COSINE optimizes the lightweight SegDecoder only, effectively unleashing the potential of the foundation models for open-world segmentation. As illustrated in Fig. 1, COSINE can concurrently address open-vocabulary and in-context segmentation tasks at multiple granularities, including semantic, instance, and panoptic segmentation. Comprehensive experiments demonstrate that COSINE effectively unifies both settings within a single model, achieving state-of-the-art performance. In particular, our exploratory analyses highlight the synergistic collaboration between the visual and textual branches, leading to significantly improved generalization over single-modality approaches. We believe our findings offer valuable insights to the research community.

Our main contributions are as follows: (1) To our knowledge, our method is the first to unify in-context segmentation and open-vocabulary segmentation. We present a simple but effective framework, COSINE, which unleashes the potential of frozen foundation models across various segmentation tasks. (2) Our comprehensive experiments demonstrate that COSINE achieves significant performance improvements in both open-vocabulary and in-context segmentation tasks. (3) We further observe that the synergistic collaboration between different modality branches enhances generalization in open-world segmentation, providing valuable insights for the research community.

2. Related Work

Open-World Segmentation. Image segmentation [17, 24, 61], which aims to localize and organize meaningful concepts at the pixel level, has evolved from closed-world scenarios [3, 8, 13, 29, 41, 59] to open-world settings [18, 20, 26, 44, 48, 55, 62]. Open-world segmentation can be broadly categorized into two paradigms: Open-vocabulary segmentation [11, 23, 48, 49, 55, 56] focuses on recognizing and segmenting objects from an open set of categories. ODISE [48] leverages text-to-image diffusion models to learn rich semantic representations, enabling the generation of open-vocabulary panoptic masks. FC-CLIP [55] develops a segmentation model built on a shared frozen convolutional CLIP backbone, utilizing two-way open-vocabulary classification to improve performance on novel classes. OpenSeeD [56] unifies open-vocabulary segmentation and detection by jointly learning from segmentation and detection datasets, addressing task and data discrepancies through decoupled decoding and conditioned mask decoding. In-context segmentation [20, 26, 27, 32, 44, 57] enables segmentation guided by example images. SegGPT [44] introduces a random coloring scheme within in-context learning to enhance model generalization. PerSAM [57] employs one-shot data to generate positive-negative location priors and infuses high-level semantics from SAM [18] to guide personalized segmentation. Matcher [26] proposes a training-free framework that utilizes vision foundation models for few-shot segmentation tasks. DINOv [20] integrates visual in-context prompting to unify referring and generic segmentation tasks, introducing a specialized encoder-decoder architecture. SINE [27] resolves ambiguity in in-context segmentation by unifying tasks across various granularities. Unlike these methods, COSINE unifies open-vocabulary and in-context segmentation, fostering synergistic collaboration between these two branches, and significantly enhancing the model's open-world generalization performance.

Vision Foundation Models. Leveraging contrastive and generative training, vision foundation models, when combined with large-scale single-/multi-modal datasets, exhibit remarkable generalization capabilities. Pre-trained by masked image modeling [1, 10, 28, 47], MAE [14] demonstrating strong transfer performance across downstream tasks. DINOv2 [36], through image and patch-level discriminative self-supervised learning, learns versatile visual features applicable to a wide range of downstream tasks. CLIP[38] learns multi-modal representations by image-text contrastive learning and demonstrates outstanding zero-shot image classification performance. The Segment Anything Model (SAM)[18] achieves impressive zero-shot, class-agnostic segmentation performance by training on a large-scale segmentation dataset. This work aims to train an open-world segmentation model with strong generalization capabilities, leveraging the potential of existing foundation models, even under constraints in data and computational resources.

3. Method

We present COSINE, a unified open-world segmentation model that **C**onsolidates **O**pen-vocabulary **S**egmentation and **IN**-context s**E**gmentation. COSINE supports diverse modalities of input, such as images and text, by framing them as promptable segmentation tasks, offering powerful open-world perception capabilities. Our goal is not to achieve state-of-the-art performance across all open-vocabulary or in-context segmentation tasks, but to validate that diverse modal information can synergistically collaborate and extend knowledge across tasks, thereby enhancing the model's modeling capabilities and generalization performance.

3.1. Overview

As shown in Fig. 2(a), COSINE follows a simple design philosophy, consisting of a Model Pool for extracting features from a target image and different modality prompts (e.g., text and example image), and a decoder-only segmentation model, named SegDecoder, which takes image and prompt features as inputs and outputs results for various segmentation tasks. Unlike previous segmentation methods [18, 20, 56] that train all parameters of models, COSINE fixes the foundation models in the Model Pool and only trains the relatively lightweight SegDecoder. Our approach enables lower training overhead while unleashing the potential of the foundation models for open-world perception.

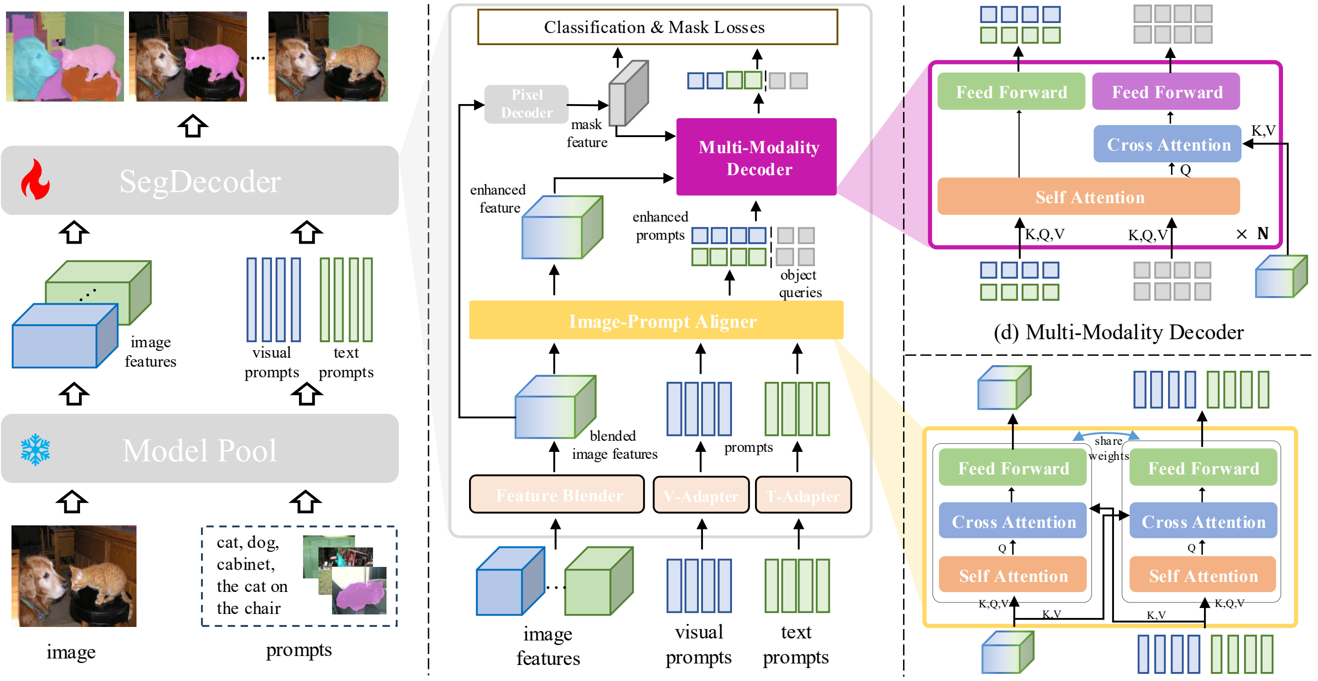
Model Pool. Model Pool includes different vision models (DINOv2 and CLIP vision encoder) and language models (CLIP text encoder). The inputs of Model Pool include a target image, its image prompts and text prompts. For the target image, we use all vision models to encode it into the image features $\mathcal{F} = \{\mathbf{F}_i\}_i^P$. Then, we use DINOv2 to encode the images of visual prompts into image features and use the in-context masks to pool the features into prompt tokens $\mathcal{V} = \{\mathbf{v}_i\}_i^M$. The language model is used to extract the text prompt features $\mathcal{T} = \{\mathbf{t}_i\}_i^N$. The image and prompt representations extracted by different foundation models contain rich multi-modal information, and the modalities complement each other, facilitating efficient transfer to downstream perception tasks.

3.2. SegDecoder

The architecture of SegDecoder is shown in Fig. 2 (b), which include a set of adapters, an Image-Prompt Aligmenter, a Pixel decoder and a Multi-Modality Decoder.

We deploy different adapters, including Feature Blender, V-Adapter and T-Adapter, for different features with the following objectives: (1) blending different image features, (2) aligning the feature dimensions of the image and various modality prompts, and (3) mapping the representations from different modalities to a shared space. Feature Blender consists of two convolutional layers that take the concatenated image features as input and outputs the blended image feature $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$. Similarly, V-Adapter (T-Adapter) maps image (text) prompts into $\mathbf{V} \in \mathbb{R}^{M\times C}$ ($\mathbf{T} \in \mathbb{R}^{N\times C}$).

Image-Prompt Aligner. The image features integrate knowl-



(a) Overview of COSINE

(b) SegDecoder

(c) Image-Prompt Aligner

(d) Multi-Modality Decoder

Figure 2. The architecture of COSINE. (a) COSINE consists a Model Pool (*e.g.*, DINOv2 and CLIP) used to extract image and prompt features and a SegDecoder used for unified open-world segmentation tasks. (b) SegDecoder consists a set of adapters, an Image-Prompt Aligmenter, a Pixel decoder and a Multi-Modality Decoder for modality alignment between the image and prompts, effectively enhancing open-world perception modeling. (c) and (d) show the details of Image-Prompt Aligmenter and Multi-Modality Decoder.

edge from multiple models, while each prompt only encode knowledge from a single modality, leading to a disparity between them. To mitigate the disparity, we employ an Image-Prompt Aligner to align and harmonize these heterogeneous representations. As shown in Fig. 2(c), the Image-Prompt Aligner is implemented by a set of self-attention, cross-attention, and feed-forward networks (FFN). Through the shared network structure, the Aligner aligns the image and different modality prompts, enhancing their representations in the multi-modal space, enabling high-quality segmentation results using prompts from any modality. Specifically, the blended image feature $\mathbf{F}$ and the prompt features $\mathbf{V}$ and $\mathbf{T}$ are aligned in the multi-modal representation space through the Image-Prompt Aligner module:

$$\left\langle \mathbf{F}^{'}, \mathbf{V}^{'}, \mathbf{T}^{'} \right\rangle = Alignment\left(\mathbf{F}, \mathbf{V}, \mathbf{T}; \theta\right), \qquad (1)$$

where $\mathbf{F}^{'} \in \mathbb{R}^{C \times H \times W}$ is the enhanced image feature, and $\mathbf{V}^{'} \in \mathbb{R}^{M \times C}$ and $\mathbf{T}^{'} \in \mathbb{R}^{N \times C}$ are the enhanced prompt features. θ is the parameters of the Image-Prompt Aligner.

Pixel Decoder. The aim of the Pixel Decoder is to output high-resolution mask features. For single-scale image feature, the Pixel Decoder is implemented as a small network consisting of two transpose convolution layers. It takes the blended image feature $\mathbf{F}$ as input, performs $4\times$ upsampling, and generates the mask feature $\mathbf{F}_{mask} \in \mathbb{R}^{C \times H^{'} \times W^{'}}$.

Multi-Modality Decoder. As show in Fig. 2(d), we introduce a Multi-Modality Decoder to refine the different modality prompts and object queries $\mathbf{Q} \in \mathbb{R}^{K \times C}$. Following [27], the Multi-Modality Decoder adopts a dual-path design, utilizing both self-attention and cross-attention [3] to facilitate interaction between object queries, different modality prompts, and the image features, while simultaneously refining the object queries and prompts. The process of the Multi-Modality Decoder can be summarized as follows:

$$\langle \mathbf{Q}_r, \mathbf{V}_r, \mathbf{T}_r \rangle = Decoder\left(\mathbf{Q}, \mathbf{V}^{'}, \mathbf{T}^{'}, \mathbf{F}^{'}, \mathbf{F}_{mask}; \phi\right), \qquad (2)$$

where $\mathbf{Q}_r$, $\mathbf{V}_r$ and $\mathbf{T}_r$ are refined object queries, visual and text prompts. ϕ is the parameters of the Multi-Modality Decoder. By using $\mathbf{V}_r$ and $\mathbf{T}_r$ as classifiers, COSINE can obtain the classification scores $\mathbf{S}_v$ and $\mathbf{S}_t$ for in-context segmentation and open-vocabulary segmentation. This can be illustrated in the following equation:

$$\begin{aligned} \mathbf{S}_v &= \mathbf{Q}_r {\mathbf{V}_r}^{\top}, \;\; \mathbf{S}_v \in \mathbb{R}^{K \times M}, \\ \mathbf{S}_t &= \mathbf{Q}_r {\mathbf{T}_r}^{\top}, \;\; \mathbf{S}_t \in \mathbb{R}^{K \times N}. \end{aligned} \qquad (3)$$

The mask results $\mathbf{M}$ can be obtained via the following operation:

$$\mathbf{M} = MLP\left(\mathbf{Q}_r\right)\mathbf{F}_{mask}, \;\; \mathbf{M} \in \mathbb{R}^{K \times H^{'} \times W^{'}}. \qquad (4)$$

Note that $\mathbf{S}_v$ and $\mathbf{S}_t$ are not computed independently. In practice, the vision and text prompts are concatenated and jointly processed to produce a unified score matrix. The separate notation of $\mathbf{S}_v$ and $\mathbf{S}_t$ is adopted purely for clarity of presentation.

Multi-Scale Feature Injection. Similar to previous approaches [22, 55], we can obtain multi-scale image features from these foundation models to enhance the performance of SegDecoder. We employ a multi-scale deformable attention Transformer [64] as the Pixel Decoder and use these multi-scale image features as inputs to both the Image-Prompt Aligner and the Multi-Modality Decoder.

3.3. Training and Inference

Image and Text Prompts Co-training. During the training phase, we experientially maintain a 1:1 sample ratio between image and text prompts for each batch, ensuring that the model effectively balances in-context and open-vocabulary segmentation tasks within an open-world framework. For in-context segmentation, we employ two distinct sampling strategies for image prompts: (1) sampling diverse images of the same class to serve as both the target image and the image prompt, thereby enhancing the model's capability to discern target semantics, and (2) randomly cropping a single image to generate multiple views of the same object, which strengthens the model's ability to recognize objects across varying perspectives. For open-vocabulary segmentation, we sample a set of class names that encompass the target image classes as prompts. The architecture of the SegDecoder is designed to process both in-context and open-vocabulary segmentation tasks uniformly, facilitating the use of a shared loss function for optimization. In line with prior DETR-based approaches [2, 3, 27, 64], we use bipartite matching to match predictions with ground-truth. Subsequently, we optimize the model using cross-entropy loss for classification, along with binary mask loss and Dice loss [33] for mask prediction.

Multi-Modal Prompts Collaborative Inference. The SegDecoder architecture empowers COSINE to seamlessly process inputs from single-modality prompts, enabling it to tackle both open-vocabulary segmentation and in-context segmentation tasks. Furthermore, COSINE supports collaborative inference by leveraging prompts from multiple modalities. By integrating image and text prompts, COSINE generates mask predictions based on blended instructions. Notably, in addition to utilizing single-modality prompts, we introduce a straightforward averaging fusion mechanism to blend image prompts $\mathbf{V}$ and text prompts $\mathbf{T}$. This fusion strategy allows complementary information from different modalities to enhance COSINE's generalization capabilities in open-world scenarios.

4. Experiments

4.1. Experiments Setting

Training Data. We train our model with publicly available academic segmentation datasets, encompassing semantic, instance, and panoptic segmentation tasks. Specifically, we leverage datasets including: COCO [24] is a widely used dataset that supports multiple tasks, including object detection, instance segmentation, and panoptic segmentation. It comprises 80 "things" categories and 53 "stuff" categories, with a total of 118K training images and 5K validation images. Objects365 [40] is a large-scale, high-quality dataset designed for object detection. It includes 365 object categories, 638K images, and approximately 10 million bounding boxes. Following the approach in [27], we utilize Objects365-SAM, an enhanced version of the Objects365 dataset where instance segmentation annotations are extended with SAM [18]. In addition, we incorporate referring segmentation datasets to endow COSINE with the capability of understanding referring expressions. Specifically, following LISA [19], we utilize refCLEF, refCOCO, refCOCO+ [16], and refCOCOg [31], which are widely adopted benchmarks for referring expression comprehension and segmentation. More implementation details are provided in the Appendix B.

4.2. Main Results

We simultaneously evaluate COSINE's ability to process prompts in both the visual and textual modalities. For the visual modality, we select few-shot (semantic/instance) segmentation and video object segmentation tasks, while for the textual modality, we choose open-vocabulary (semantic/panoptic) segmentation and referring segmentation tasks. We compare COSINE against task-specific expert models as well as several general-purpose segmentation models. The results, presented in Table 1, demonstrate that COSINE exhibits a significant advantage in most tasks.

Few-Shot Semantic Segmentation. Following the evaluation protocol of [26], we evaluate COSINE on the LVIS [12] dataset for few-shot semantic segmentation. Compared with the specialized few-shot model DiffewS [63] and the in-context model, SINE [27], COSINE achieves better performance in both one-shot and few-shot settings. COSINE decouples low-level mask features from high-level image and prompt features. Using a query-based decoder, it achieves strong contextual understanding and accurate masks. COSINE outperforms SegGPT [44], demonstrating its effective contextual reasoning. As shown in the first column of Table 1, COSINE even outperforms the SAM-based model, Matcher [26], despite being trained on only a small amount of segmentation data.

Few-Shot Instance Segmentation. We evaluate the few-shot instance segmentation performance of COSINE on the

		few-shot sem.		few-shot ins.		open-voc. pano.					open-voc. sem.	
Methods	Venue	LVIS-92^{i}		LVIS		ADE20K			Cityscapes		A-847	PC-459
		one-shot	few-shot	AP	APr	PQ	AP	mIoU	PQ	mIoU	mIoU	mIoU
few-shot model												
HSNet [34]	ICCV'21	17.4	22.9	-	-	-	-	-	-	-	-	-
VAT [15]	ECCV'22	18.5	22.7	-	-	-	-	-	-	-	-	-
DiffewS [63]	NeurIPS'24	31.4	35.4	-	-	-	-	-	-	-	-	-
in-context model												
SegGPT [44]	ICCV'23	18.6	25.4	-	-	-	-	-	-	-	-	-
PerSAM-F [57]	ICLR'24	18.4	-	-	-	-	-	-	-	-	-	-
Matcher [26]	ICLR'24	33.0	40.0	-	-	-	-	-	-	-	-	-
SINE [27]	NeurIPS'24	31.2	35.5	8.6	7.1	-	-	-	-	-	-	-
open-vocabulary model												
ODISE [48]	CVPR'23	-	-	-	-	23.4	13.9	28.7	23.9	-	11.1	14.5
FC-CLIP [55]	NeurIPS'23	-	-	-	-	26.8	16.8	34.1	44.0	56.2	14.8	18.2
HIPIE [42]	NeurIPS'23	-	-	-	-	22.9	19.0	29.0	-	-	9.7	14.4
SED [46]	CVPR'24	-	-	-	-	-	-	-	-	-	13.9	22.6
universal model												
X-Decoder [65]	CVPR'23	-	-	-	-	21.8	13.1	29.6	38.1	52.0	9.2	16.1
UNINEXT* [51]	CVPR'23	-	-	-	-	8.9	14.9	6.4	-	-	1.8	5.8
OpenSeeD [56]	ICCV'23	-	-	-	-	19.7	15.0	23.4	41.4	47.8	-	-
DINOv [20]	CVPR'24	-	-	15.4	14.5	23.2	15.1	25.3	-	-	-	-
OMG-Seg [21]	CVPR'24	-	-	-	-	27.9	-	-	-	-	-	-
PSALM [58]	ECCV'24	-	-	-	-	-	13.9	24.4	-	-	-	14.0
COSINE†	this work	34.2	39.1	17.4	23.3	28.1	16.7	35.2	37.1	53.4	15.2	19.6
COSINE		35.2	40.7	20.3	25.8	31.0	21.1	35.7	42.0	56.1	15.6	19.2

Table 1. Results of different open world segmentation tasks including few-shot semantic segmentation, open-vocabulary panoptic segmentation and semantic segmentation. * We report the performance evaluated in [42]. † indicates the single-scale variant of COSINE.

long-tailed dataset LVIS [12], which contains over 1,000 categories. For each category, we randomly select 10 samples (or all available samples if fewer than 10 are present). We report the AP for all categories and the APr for rare categories. As shown in Table 1, COSINE achieves an AP of 20.3, significantly outperforming the universal model DINOv. Furthermore, COSINE achieves an APr of 25.8, demonstrating stronger generalization to rare categories compared to previous methods. These results highlight COSINE's superior adaptability to open-world scenarios.

Open-Vocabulary Panoptic Segmentation. Following ODISE [48] and FC-CLIP [55], we evaluate COSINE on the ADE20K [61] and Cityscapes [7] datasets for open-vocabulary panoptic segmentation. We compare COSINE with specialized open-vocabulary models, including ODISE, FC-CLIP, and HIPIE [42], as well as general-purpose models such as X-Decoder [65] and OpenSeeD [56]. As shown in Table 1, COSINE achieves outstanding performance on ADE20K, reaching a PQ of 31.0 and an AP of 21.1. Additionally, on Cityscapes, COSINE attains performance comparable to state-of-the-art methods, further demonstrating its effectiveness in open-vocabulary panoptic segmentation.

Open-Vocabulary Semantic Segmentation. For open-vocabulary semantic segmentation, COSINE outperforms all universal models and the majority of open-vocabulary models, including the LLM-based model PSALM [58]. As shown in Table 1, COSINE achieves a mIoU of 15.6 on the A-847 dataset [60] and 19.2 on the PC-459 dataset [35], demonstrating its ability to recognize and segment novel categories effectively. Notably, while COSINE slightly lags behind specialized open-vocabulary semantic segmentation models such as SED [46], its key advantage lies in its versatility, enabling flexible support for a wide range of segmentation tasks.

Video Object Segmentation. We evaluate COSINE on video object segmentation using DAVIS 2017 [37] and YouTube-VOS 2019 [50] datasets. As shown in Table 3, we compare COSINE with both video-trained models (gray rows) and models that do not use explicit video data. Video-trained models, such as Cutie [6] and XMem [4], achieve strong performance by leveraging temporal information and explicitly training on video datasets. COSINE, despite not using video data, achieves competitive performance without requiring video supervision, making it a more generaliz-

Method	Venue	refCOCO			refCOCO+			refCOCOg	
		val	testA	testB	val	testA	testB	val(U)	test(U)
MAttNet [54]	CVPR'18	56.5	62.4	51.7	46.7	52.4	40.1	47.6	48.6
MCN [30]	CVPR'20	62.4	64.2	59.7	50.6	55.0	44.7	49.2	49.4
VLT [9]	ICCV'21	67.5	70.5	65.2	56.3	61.0	50.1	55.0	57.7
LAVT [53]	CVPR'22	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
CRIS [45]	CVPR'22	70.5	73.2	66.1	62.3	68.1	53.7	59.9	60.4
ReLA [25]	CVPR'23	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0
X-Decoder [65]	CVPR'23	-	-	-	-	-	-	64.6	-
SEEM [66]	NeurIPS'23	-	-	-	-	-	-	65.7	-
LISA [19]	CVPR'24	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6
COSINE	this work	77.2	80.7	71.1	66.4	73.2	56.4	67.4	68.5

Table 2. Results of referring segmentation on refCOCO, refCOCO+ and RefCOCOg. We report the metric of cIoU.

Methods	Venue	DAVIS 2017	YT-VOS 2019
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{G}$
with video data			
AOT [52]	NeurIPS'21	85.4	85.3
XMem [4]	ECCV'22	87.7	85.5
DEVA [5]	ICCV'23	86.8	85.5
Cutie [6]	CVPR'24	88.8	86.1
without video data			
Painter [43]	CVPR'23	34.6	20.6
SegGPT [44]	ICCV'23	75.6	73.1
SEEM [66]	NeurIPS'23	58.9	-
DINOv [20]	CVPR'24	73.3	52.0
PerSAM-F [57]	ICLR'24	76.1	46.6
SINE [27]	NeurIPS'24	77.0	66.4
COSINE	this work	76.7	66.0
COSINE-FT		80.2	70.0

Table 3. Results of video object segmentation on DAVIS 2017, and YouTube-VOS 2019. Gray indicates the model is trained on target datasets with video data.

Prompt		LVIS-92^i		ADE20K		
vision	text	1-shot	5-shot	PQ	AP	mIoU
✓		24.5	27.8	-	-	-
	✓	-	-	13.2	7.6	30.2
✓	✓	27.7	32.1	17.7	8.1	30.4

Table 4. Effect of the interaction between visual and textual branches during Training. All models are trained for 10k steps.

Prompt		LVIS-92^i		ADE20K		
vision	text	1-shot	5-shot	PQ	AP	mIoU
✓		35.2	40.7	23.8	15.8	26.3
	✓	37.8	-	31.0	21.1	35.7
✓	✓	43.1	45.9	31.4	21.3	36.3

Table 5. Effect of the interaction between visual and textual branches during inference.

able and scalable approach. Compared with general-purpose segmentation models, COSINE achieves state-of-the-art performance among methods that do not use video data. On DAVIS 2017, COSINE obtains a $\mathcal{J}\&\mathcal{F}$ score of 76.7, significantly outperforming SEEM [66] and DINOv [20], and achieving comparable performance to PerSAM-F [57] and SINE [27]. We also introduce a fine-tuned variant, COSINE-FT, which improves the overall performance across both datasets by further tuning on the pure image prompt dataset. This demonstrates that COSINE's segmentation capabilities can be significantly enhanced through fine-tuning, making it a versatile and scalable approach adaptable to both static and dynamic scene understanding.

Referring Segmentation. We test COSINE on the referring segmentation task following the evaluation protocol of LISA [19], which requires further fine-tuning on the referring segmentation dataset. As shown in Table 2, COSINE not only surpasses all traditional methods and general-purpose segmentation models but also achieves performance comparable to the LLM-based model LISA [19]. This further validates COSINE's capability in handling long text descriptions and complex semantic understanding.

4.3. Exploration of Visual-Textual Interaction

As mentioned earlier, COSINE consists of a textual branch that processes text prompts and a visual branch that processes visual prompts. Few works have explored the interaction between these two modalities. COSINE is the first to investigate their interplay from two perspectives: during training and inference.

Training Phase. We aim to investigate whether jointly training both branches with multimodal data can enhance the performance of each individual branch. Specifically, we train the model for 10k steps using data from a single modality and multi-modality data, respectively. As shown in Table 4, we demonstrate that introducing the textual branch during training significantly improves the performance of the visual branch, and conversely, the visual branch also enhances the textual branch's performance. This finding highlights the mutual benefits of multimodal training in segmentation.

#ID	Model	LVIS-92^i		ADE20K		
		1-shot	5-shot	PQ	AP	mIoU
0	COSINE (Full Model)	27.7	32.1	17.7	8.1	30.4
1	only DINOv2 Encoder	27.7	32.0	9.7	4.2	18.6
2	only CLIP Encoder	24.4	28.0	11.4	3.3	27.0
3	w/o Feature Blending	24.4	29.0	9.4	2.6	26.3
4	w/o Image-Prompt Aligner	26.3	31.3	12.0	6.4	30.1
5	w/o Prompt Refining	27.0	31.2	16.5	6.6	30.9

Table 6. Ablation study. All models are trained for 10k steps.

Inference Phase. We further investigate whether providing both visual and textual prompts during inference can improve model performance. We validate this hypothesis on few-shot semantic segmentation and open-vocabulary panoptic segmentation tasks, as shown in Table 5. Our results reveal that for tasks originally supported by the visual branch, such as LVIS-92^i, incorporating an additional text prompt during inference leads to significant performance gains. Similarly, for tasks primarily supported by the textual branch, such as ADE20K, introducing a visual prompt during inference also enhances performance. These findings confirm that COSINE can flexibly process prompts from multiple modalities, resulting in more robust segmentation. This insight also provides valuable inspiration for future research on universal segmentation models.

4.4. Ablation Study

As shown in Table 6, we validate the impact of each design within our model, including the Model Pool(#1, #2), the Feature Blender(#3), the Image-Prompt Aligner(#4), and the Multi-Modality Decoder(#5). Specifically, a comparison with model #1 and #2 reveals that different pre-trained models are informationally complementary, enhancing both in-context and open-vocabulary segmentation performance. The comparison with model #3 demonstrates that merely concatenating different foundation models is insufficient. An explicit blending of representations is necessary to learn a unified multi-modal representation. Furthermore, comparisons with models #4 and #5 indicate that the introduced Aligner and Decoder are both simple and effective.

4.5. Visualization Analysis

Qualitative Results. As shown in Fig. 3, we visualize the segmentation results under both open-vocabulary and in-context settings, as well as the results obtained by combining both types of prompts. Additionally, we provide visualizations of referring segmentation and VOS results. These results demonstrate that COSINE achieves highly accurate predictions across various modalities and granularities, highlighting its strong potential for open-world generalization.

Prompt Synergy. As shown in Fig. 4, COSINE can generalize to real-world domains such as industrial inspection, medical imaging. The industrial inspection case shows effective collaboration between visual and textual prompts, accurately segmenting targets that would produce incorrect masks under a single modality. These results highlight the potential of multimodal prompt synergy in COSINE for complex segmentation. By integrating visual and textual cues, COSINE captures fine-grained details that are difficult to express with a single modality.

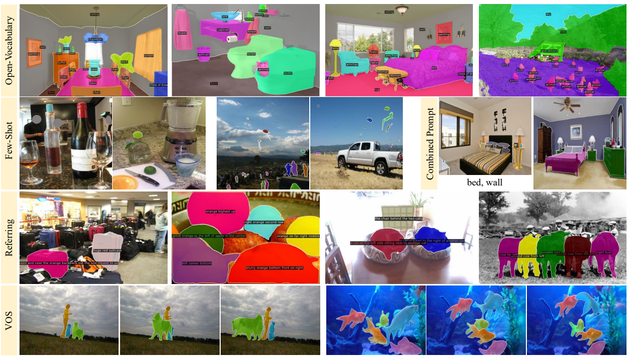


Figure 3. Qualitative results. COSINE can perform various open-world segmentation tasks with different modal prompts (image and text). For few-shot segmentation, the left image is the example image and the right is the result.

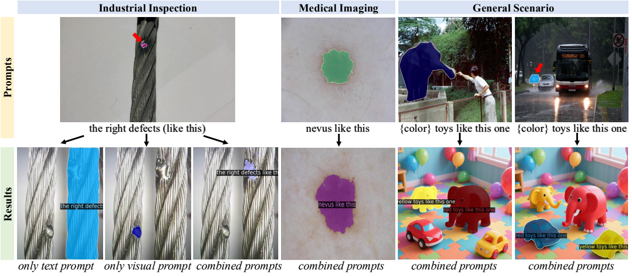


Figure 4. Visualization of prompt synergy. The top row shows the input prompts, the bottom row presents the corresponding outputs.

5. Conclusion

In this work, we present COSINE, a unified open-world segmentation model that unifies open-vocabulary Segmentation and in-context segmentation. COSINE supports diverse modalities of input, such as images and text, offering powerful open-world perception capabilities. Our exploratory analysis highlights that the synergistic collaboration between visual and textual branches enhances generalization in open-world segmentation, providing valuable insights for the research community. More discussion and limitations are provided in the Appendix A.

Acknowledgments

This work was supported by the National Key R&D Program of China (No. 2022ZD0160101), the Ningbo Science and Technology Bureau (No. 2024Z291), and the National Natural Science Foundation of China (No. 62206244).

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 3

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis.*, 2020. 5

[3] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2, 4, 5

[4] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *Eur. Conf. Comput. Vis.*, 2022. 6, 7

[5] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *Int. Conf. Comput. Vis.*, 2023. 7

[6] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 6, 7

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 6

[8] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation with a discriminative loss function. *arXiv preprint arXiv:1708.02551*, 2017. 2

[9] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Int. Conf. Comput. Vis.*, 2021. 7

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Int. Conf. Learn. Represent.*, 2020. 2, 3

[11] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *Eur. Conf. Comput. Vis.*, 2022. 2, 3

[12] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 5, 6

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Int. Conf. Comput. Vis.*, 2017. 2

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2, 3

[15] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *Eur. Conf. Comput. Vis.*, 2022. 6

[16] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. 2014. 5

[17] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1, 2

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Int. Conf. Comput. Vis.*, 2023. 2, 3, 5

[19] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 5, 7

[20] Feng Li, Qing Jiang, Hao Zhang, Tianhe Ren, Shilong Liu, Xueyan Zou, Huaizhe Xu, Hongyang Li, Jianwei Yang, Chunyuan Li, et al. Visual in-context prompting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 2, 3, 6, 7

[21] Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation? In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 6

[22] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *Eur. Conf. Comput. Vis.*, 2022. 5

[23] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 2, 3

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, 2014. 1, 2, 5

[25] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 7

[26] Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. *arXiv preprint arXiv:2305.13310*, 2023. 2, 3, 5, 6

[27] Yang Liu, Chenchen Jing, Hengtao Li, Muzhi Zhu, Hao Chen, Xinlong Wang, and Chunhua Shen. A simple image segmentation framework via in-context examples. *arXiv preprint arXiv:2410.04842*, 2024. 2, 3, 4, 5, 6, 7

[28] Yang Liu, Xinlong Wang, Muzhi Zhu, Yue Cao, Tiejun Huang, and Chunhua Shen. Masked channel modeling for bootstrapping visual pre-training. *Int. J. Comput. Vis.*, 2024. 3

[29] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 2

[30] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 7

[31] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 5

[32] Lingchen Meng, Shiyi Lan, Hengduo Li, Jose M Alvarez, Zuxuan Wu, and Yu-Gang Jiang. Segic: Unleashing the emergent correspondence for in-context segmentation. In *Eur. Conf. Comput. Vis.*, 2024. 2, 3

[33] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 2016. 5

[34] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 6

[35] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014. 6

[36] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 3

[37] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 6

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. 2021. 2, 3

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2

[40] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Int. Conf. Comput. Vis.*, 2019. 5

[41] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Solo: A simple framework for instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021. 2

[42] Xudong Wang, Shufan Li, Konstantinos Kallidromitis, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Hierarchical open-vocabulary universal image segmentation. In *Adv. Neural Inform. Process. Syst.*, 2023. 6

[43] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 7

[44] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. In *Int. Conf. Comput. Vis.*, 2023. 2, 3, 5, 6, 7

[45] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 7

[46] Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 6

[47] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 3

[48] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 2, 3, 6

[49] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 2, 3

[50] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 6

[51] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 6

[52] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. In *Adv. Neural Inform. Process. Syst.*, 2021. 7

[53] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 7

[54] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 7

[55] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. In *Adv. Neural Inform. Process. Syst.*, 2023. 2, 3, 5, 6

[56] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *Int. Conf. Comput. Vis.*, 2023. 2, 3, 6

[57] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*, 2023. 2, 3, 6, 7

[58] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. In *Eur. Conf. Comput. Vis.*, 2024. 6

[59] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2

[60] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 6

[61] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *Int. J. Comput. Vis.*, 2019. 1, 2, 6

[62] Muzhi Zhu, Hengtao Li, Hao Chen, Chengxiang Fan, Weian Mao, Chenchen Jing, Yifan Liu, and Chunhua Shen. Segprompt: Boosting open-world segmentation via category-level prompt learning. In *Int. Conf. Comput. Vis.*, 2023. 2, 3

[63] Muzhi Zhu, Yang Liu, Zekai Luo, Chenchen Jing, Hao Chen, Guangkai Xu, Xinlong Wang, and Chunhua Shen. Unleashing the potential of the diffusion model in few-shot semantic segmentation. *arXiv preprint arXiv:2410.02369*, 2024. 5, 6

[64] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 5

[65] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 6, 7

[66] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023. 7