

# OMG-Seg Is One Model Good Enough For All Segmentation?

Xiangtai Li<sup>1</sup> Haobo Yuan<sup>1</sup> Wei Li<sup>1</sup> Henghui Ding<sup>1</sup> Size Wu<sup>1</sup> Wenwei Zhang<sup>1</sup>  
 Yining Li<sup>2</sup> Kai Chen<sup>2</sup> Chen Change Loy<sup>1</sup>

<sup>1</sup>S-Lab, Nanyang Technological University <sup>2</sup>Shanghai Artificial Intelligence Laboratory

Project Page: [https://lxtgh.github.io/project/omg\\_seg](https://lxtgh.github.io/project/omg_seg)

xiangtai94@gmail.com ccloy@ntu.edu.sg

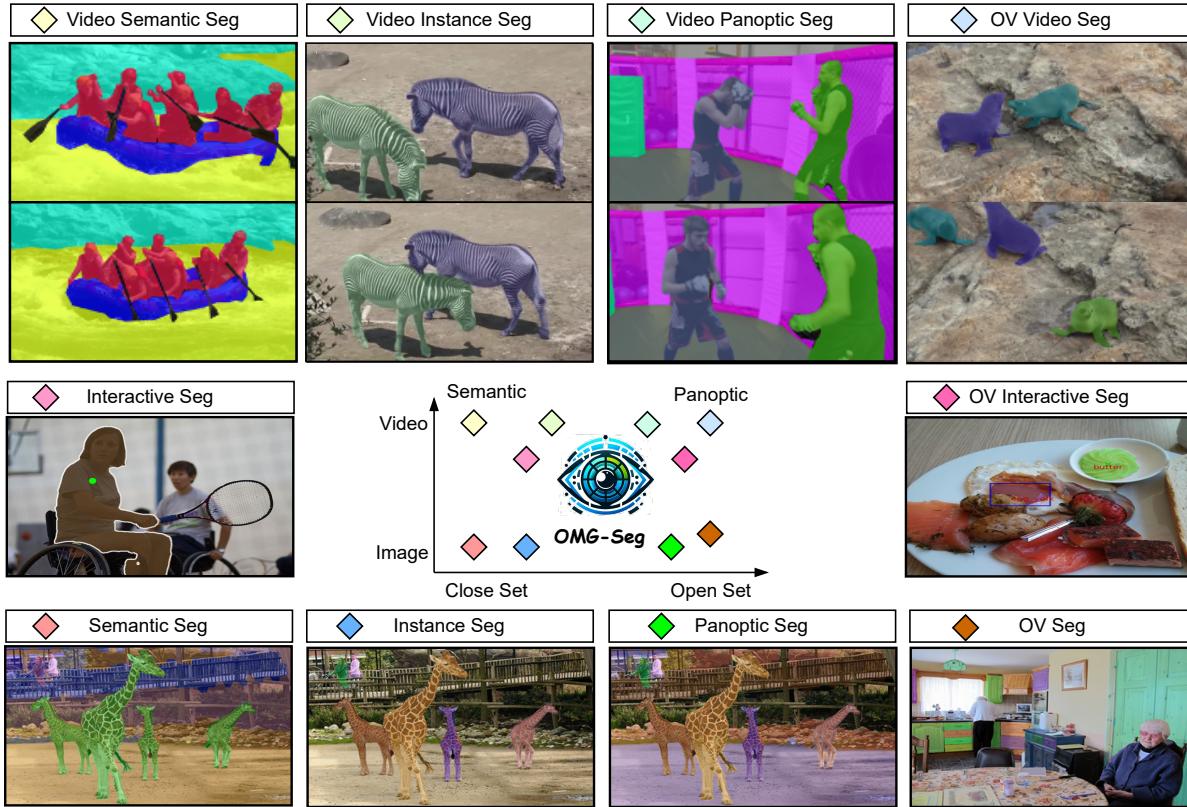


Figure 1. OMG-Seg can handle over ten different segmentation tasks in one framework, including image-level and video-level segmentation tasks, interactive segmentation, and open-vocabulary segmentation. To our knowledge, this is the first model to unify these four directions.

## Abstract

In this work, we address various segmentation tasks, each traditionally tackled by distinct or partially unified models. We propose OMG-Seg, One Model that is Good enough to efficiently and effectively handle all the segmentation tasks, including image semantic, instance, and panoptic segmentation, as well as their video counterparts, open vocabulary settings, prompt-driven, interactive segmentation like SAM, and video object segmentation. To our knowledge, this is the first model to handle

all these tasks in one model and achieve satisfactory performance. We show that OMG-Seg, a transformer-based encoder-decoder architecture with task-specific queries and outputs, can support over ten distinct segmentation tasks and yet significantly reduce computational and parameter overhead across various tasks and datasets. We rigorously evaluate the inter-task influences and correlations during co-training. Code and models are available at <https://github.com/lxtGH/OMG-Seg>.

## 1. Introduction

Visual segmentation that aims to understand semantics at the pixel level has been a longstanding problem [55, 67] in the vision community, fueling advancements in diverse applications such as robotics, autonomous vehicles, and augmented / virtual reality systems. Over the past decade, owing to the tremendous progress in deep learning [10, 11, 36, 45, 46, 48, 60, 100], this fundamental problem has been significantly transformed into a diverse set of tasks for image and video data, including basic semantic object / instance segmentation, panoptic segmentation, and the more recent prompt-driven interactive segmentation [38]. Consequently, a plethora of task-specific deep segmentation models (*e.g.*, Mask-RCNN [30], Mask2Former [18], and SAM [38]), along with different benchmarks, have been proposed. The latest studies [26, 28, 82, 95] strive to extend these standard close-set segmentation models to more dynamic, real-world scenarios. This involves integrating pre-trained vision-language foundation models, *e.g.*, CLIP [66], into deep segmentation frameworks, enabling visual segmentation through open-vocabulary text descriptions.

Most existing deep segmentation models often focus on a single specific task. In many scenarios, a generalizable model capable of handling a broader spectrum of segmentation tasks is highly desired. A unified model of this nature would eliminate the necessity for task-specific designs, while providing a versatile solution to a wide range of segmentation tasks through a single and cohesive architecture. This approach benefits significantly from leveraging large and varied data corpora, which enhances the model’s adaptability and effectiveness across different segmentation tasks.

Unifying diverse segmentation tasks within a single model is non-trivial because each task typically comes with its own unique model design. The emergence of transformers [6, 23, 53] has catalyzed several segmentation models based on the Detection Transformer (DETR) architecture [18, 33, 35, 49, 51, 97], yielding notable successes in performance and task integration. Concurrently, there are also models [2, 27, 75, 86, 89, 91, 103] that employ a similar framework to merge open-vocabulary and multi-dataset segmentation within a unified architecture. Yet, these models often fall short in generalizing to video or interactive segmentation, both essential for broader applications. Some recent studies [71, 76, 77] aim to unify all vision tasks under one single framework with segmentation included. However, these more generalized models still lag behind task-specific segmentation models in terms of performance.

In this study, we demonstrate that *one model is good enough for all segmentation*<sup>1</sup> by introducing OMG-Seg, a

<sup>1</sup>This includes primarily pure visual and 2D segmentation tasks, excluding specific tasks like medical segmentation [7] and referring segmen-

unified segmentation model designed to deliver competitive performance across a broad spectrum of visual segmentation tasks. Unlike previous unified models that typically employ a shared visual backbone but several task-specific branches, OMG-Seg adopts a shared encoder-decoder architecture. In particular, we unify all the task outputs as a unified query representation. One query can represent a mask label, an image or tube mask, a unique ID, and a visual prompt. Then, we can adopt a shared decoder to process all types of queries with their features. This setup facilitates general training and inference processes that unify all visual-only segmentation tasks, capitalizing on the extensive parameter sharing across tasks. Through co-training on combined image and video datasets, OMG-Seg, once trained, is capable of handling up to ten diverse segmentation tasks across different datasets.

OMG-Seg achieves comparable results on image, video, open-vocabulary, and interactive segmentation settings over eight different datasets, including COCO [52], ADE-20k [98], VIPSeg [58], Youtub-VIS-2019 [90], Youtub-VIS-2021, and DAVIS-17 [5], based on one single shared model. To the best of our knowledge, we are the first to achieve four different settings in one single model.

## 2. Related Work

**Universal Image/Video Segmentation.** The advent of vision transformers [6, 23, 53] has led to a wave of innovation in universal segmentation. Recent works [18, 47, 70, 92, 97] have developed mask classification architectures grounded in an end-to-end set prediction approach, outperforming specialized models [8, 20–22, 30, 37, 48, 50] in both image and video segmentation tasks [35, 49, 51]. Despite these advancements, most existing methods still rely on distinct models for different segmentation tasks and datasets. Recently, there has been a shift towards training a single model [27, 33, 88, 89] across diverse datasets and tasks, reaping the benefits of parameter sharing. For instance, OneFormer [33] integrates three image segmentation tasks within a single model, while UNINEXT [89] concentrates on unifying instance-level tasks. Similarly, TarVIS [2] combines various video segmentation tasks using target prompts. However, none of these existing works has thoroughly investigated the joint training of image, video, and prompt-driven data within one comprehensive segmentation model. Our work stands as the first attempt in this direction, stretching the potential of co-training across these domains. For a more in-depth comparison of model capabilities, please refer to Tab. 1.

**Visual Foundation Models.** Recent studies in visual foundation models have exhibited a diversification

tation [56]. Nonetheless, these could be seamlessly integrated into our OMG-Seg framework with appropriate input adaptations.

Table 1. Setting Comparison For Different Models. We include several representative methods here. Our proposed OMG-Seg can perform various segmentation tasks in one model.

Methods	SS	IS	PS	VSS	VIS	VPS	VOS	Open-Set	Multi dataset training	Interactive	Shared model
DeeplabV3+ [11]	✓										
MaskRCNN [30]		✓									
PanopticFPN [36]			✓								
DETR [6]			✓								
DetectorRS [64]	✓	✓									
TCB [58]				✓							
VisTR [78]					✓						
VPSNet [34]						✓					
STM [61]							✓				
K-Net [97]	✓	✓	✓								
Mask2Former [18]	✓	✓	✓								
Video K-Net [51]				✓	✓	✓					
Tube-Link [49]				✓	✓	✓					
TubeFormer [35]				✓	✓	✓					
OneFormer [33]	✓	✓	✓								
TarViS [2]				✓	✓	✓	✓			✓	✓
MSeg [40]	✓							✓			✓
UNINEXT [89]		✓			✓		✓		✓		
OpenSeg [26]	✓						✓	✓			✓
SAM [38]							✓				
Semantic-SAM [42]	✓	✓	✓				✓	✓		✓	✓
SEEM [104]	✓	✓	✓				✓	✓		✓	✓
OPSNet [14]				✓			✓				
FreeSeg [65]	✓	✓	✓				✓				✓
<b>OMG-Seg</b>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

in optimization techniques, encompassing various learning paradigms. These include vision-only pre-training strategies [29, 44, 85], joint vision-language pre-training approaches[24, 43], and multi-modal frameworks that incorporate visual prompting [1, 63, 77]. A notable example, SAM [38], demonstrates the generalizability and scalability of extensive training in achieving general segmentation. Building on this, Semantic-SAM [42] augments the SAM model by adding semantic labels and increased levels of granularity. However, despite their impressive capabilities, these visual foundation models typically fall short in video segmentation tasks, necessitating further refinement for optimal performance in more dynamic contexts.

**Open Vocabulary Segmentation.** This line of visual segmentation research [41, 72] aims to recognize and segment novel objects beyond the limited closed-set visual concepts. Leveraging the transferrable representations offered by vision language models (VLMs), many studies [26, 28, 79, 81, 82, 84, 87, 93–95, 99] explore the alignment between region and text representations during training. At the inference stage, detectors can recognize new classes using the text embeddings derived from VLMs. Our model follows this notion to achieve open vocabulary segmentation. In particular, we use frozen VLMs to serve both as a feature extractor and classifier. This strategy allows for a seamless transition into the open vocabulary setting.

**Unified Modeling.** The adaptable nature of the transformer architecture [23, 69] facilitates the sharing of fundamental modules across various modalities. This versatil-

ity has inspired several research initiatives that use a common transformer framework for different domains. Notably, efforts in the realm of vision generalists have been directed toward unifying disparate tasks within the vision domain. For instance, the Pix2Seq series [12, 13] approach task unification through auto-regressive token prediction. Similarly, Unified-IO [54] implements a sequence-to-sequence pipeline, converting diverse inputs and outputs into discrete token sequences. Furthermore, recent advancements [3, 4, 25, 73, 76, 77] have explored visual in-context learning as the means to combine various vision tasks. These methods predominantly target task unification across domains. However, bridging the performance gap between unified segmentation models and purpose-built segmentation models remains an open problem.

### 3. Methodology

**Motivation and Overview.** Our OMG-Seg is a single yet versatile model—with reduced task-specific customization and maximal parameter sharing—that can support a diverse set of segmentation tasks, making it one model for all segmentation. Our goal is not to pursue state-of-the-art results for each task but to increase the modeling capacity of one generalizable segmentation model while allowing extensive knowledge sharing between tasks.

The main idea of our approach is to leverage object queries for representing distinct entities, encompassing various mask types and their respective video formats. In Sec. 3.1, we begin by reexamining the definitions of im-

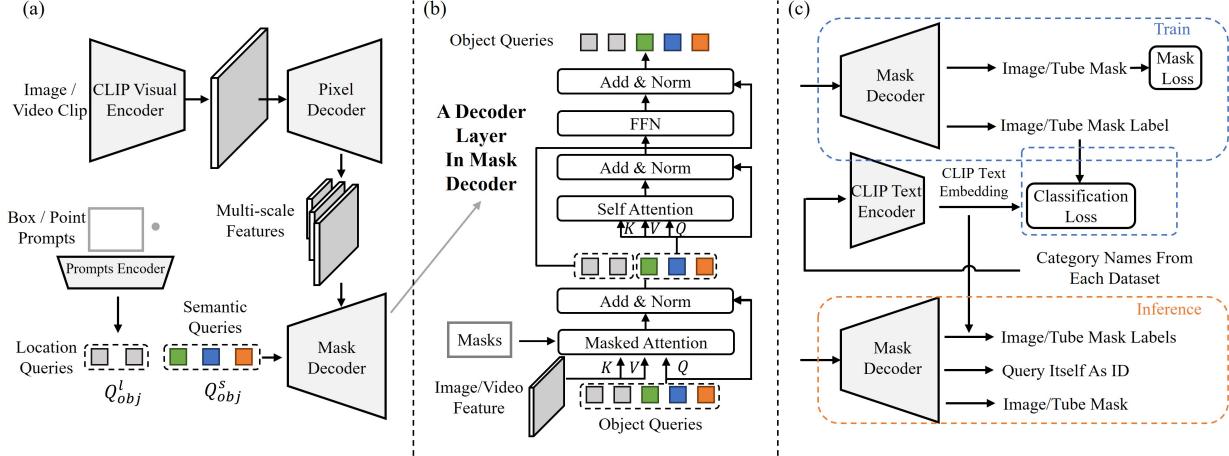


Figure 2. OMG-Seg meta-architecture. (a) OMG-Seg follows the architecture of Mask2Former [18], containing a backbone (CLIP Visual Encoder), a pixel decoder, and a mask decoder. The different parts are a shared mask decoder for both image and video segmentation and a visual prompt encoder. We use two types of mask queries, i.e., semantic queries, for instance/semantic masks or mask tubes, and location queries that encode box or point prompts. (b) One decoder layer in the Mask Decoder. The location queries skip the self-attention operation as they are only conditioned on the image content and the location prompts. (c) The forward pass of OMG-Seg in training and inference. We use CLIP’s text encoder to represent category names and classify masks by calculating cosine similarity between mask features and text embeddings.

age, video, interactive, and open vocabulary segmentation settings. In this exploration, we show that the target outputs of these varied settings can be effectively transformed into a unified query representation. Specifically, a single query can encapsulate a mask label, an image or tube mask, a unique identifier, or a visual prompt.

For example, video segmentation tasks require only an additional ID compared to image segmentation, which can be adapted from the query ID. This allows us to employ a shared decoder to process each query and its associated features in a streamlined manner, with the primary distinction being the specific feature inputs used in cross-attention layers. In the context of image tasks, we follow the established design of Mask2former [18], enabling queries and features to engage in masked-cross attention. For video tasks, we incorporate temporal features with 3D position embeddings and focus on predicting tube masks for objects across short video clips. For interactive segmentation tasks, we employ the same decoder as image tasks but skip the self-attention operation to condition mask prediction only on the visual prompts and image contents, as detailed in Sec. 3.2.

In addition, to circumvent class taxonomy conflicts, we adopt CLIP embeddings for mask classification. We employ the frozen CLIP visual encoder as the backbone, whose features are shared by the pixel decoder and the open-vocabulary mask classification. This design enables efficient open-vocabulary inference without incurring additional costs. The training and inference pipelines built on such a frozen backbone are described in Sec. 3.3.

### 3.1. Unified Task Representation

**Image Segmentation.** Given an input image  $I \in \mathbb{R}^{H \times W \times 3}$ , the goal of image segmentation is to output a group of masks  $\{y_i\}_{i=1}^G = \{(m_i, c_i)\}_{i=1}^G$  where  $c_i$  denotes the class label of the binary mask  $m_i$  and  $G$  is the number of masks,  $H \times W$  are the spatial size. According to the scope of class labels and masks, we report the results of three different segmentation tasks, including semantic segmentation (SS), instance segmentation (IS), and panoptic segmentation (PS). PS is the unification of both SS and IS, which contains countable thing classes and uncountable stuff classes. For all three tasks, we adopt mask classification architecture [19, 97], where each mask corresponds to a semantic label.

**Video Segmentation.** Given a video clip input as  $V \in \mathbb{R}^{T \times H \times W \times 3}$ , where  $T$  represents the frame number, the goal of video segmentation is to obtain a mask tube  $\{y_i\}_{i=1}^N = \{(m_i, c_i, d_i)\}_{i=1}^N$ , where  $N$  is the number of the tube masks  $m_i \in \{0, 1\}^{T \times H \times W}$ .  $c_i$  denotes the class label of the tube mask  $m_i$  while  $d_i$  denotes the instance ID of each tube mask. Each tube mask can be classified into a countable thing class or uncountable stuff class, where the thing classes are also assigned a unique ID. For stuff masks, the tracking is zero by default. When  $N = C$  and the task only contains stuff classes, and all thing classes have no IDs, VPS turns into video semantic segmentation (VSS). If  $\{y_i\}_{i=1}^N$  overlap and  $C$  only contains the thing classes and all stuff classes are ignored, VPS turns into video instance segmentation (VIS). Video Object Segmen-

tation (VOS) aims to track the first framework masks without performing classification. Motivated by image segmentation, we also adopt the tube mask classification architecture [35, 49] to train and link short tubes along the temporal dimension. For VOS, we adopt class-agnostic tube-wised training, which is similar to VPS and VIS.

**Interactive Segmentation.** The interactive segmentation in SAM [38] framework takes both image  $I$  and visual prompts  $P \in \mathbb{R}^{N \times \{2,4\}}$ , such as points and boxes, as inputs, and it outputs the corresponding binary image masks  $\{y_i\}_{i=1}^N = \{m_i \in H \times W\}_{i=1}^N$ .  $N$  is the number of visual prompts. Each visual prompt is encoded into an object query, which naturally can be the input of the decoder, like in [18, 38]. In our experiments, we use the shared decoder for all different task queries.

**Open-Vocabulary and Multi-Dataset Segmentation.** The task formulation is the same as the previous image and video segmentation. However, this setting goes beyond fixed label space. In particular, it requires open-set recognition on various datasets. Meanwhile, multi-dataset segmentation requires one model to segment more concepts under different datasets. As a common practice, we adopt CLIP text embedding as the mask classifier, which avoids taxonomy conflicts and achieves open-set recognition at the same time. As a result, we measure the distance between the visual query feature and class embeddings rather than the learned classifier.

**All the Things are in Queries.** As mentioned above, by combining all different settings, we can represent all the output segmentation entities using the same query-based mask classification framework. In particular, one object query corresponds to one mask  $m_i$ , label  $c_i$ , and ID  $d_i$ . Depending on different task settings, the formats and ranges of  $m_i$ ,  $c_i$ , and  $d_i$  are different. However, the formats and ranges of  $m_i$ ,  $c_i$ , and  $d_i$  are similar. Thus, it is natural to put all these tasks into one shared encoder-and-decoder framework and co-train one model for all segmentation tasks. Thus, it is natural to put all these tasks into one shared encoder-and-decoder framework and co-train one model for all segmentation tasks.

### 3.2. OMG-Seg Architecture

**Overview.** OMG-Seg follows the architecture design of Mask2Former [18]. As shown in Fig. 2, it contains a backbone, a pixel decoder, and a mask decoder. The difference lies in the following aspects, including frozen backbone design, combined object queries which contain both object query and visual prompt, and a shared multi-task decoder. Given different task settings, the decoder outputs corresponding masks and labels. We

**VLM Encoder as Frozen Backbone.** To enable open-vocabulary recognition, for the backbone part, we adopt the frozen CLIP visual model as a feature extractor. We use

the ConvNeXt architecture [53] from the OpenCLIP [32]. Given image/video inputs, the VLM encoder extracts multi-scale frozen feature  $\{F_j^{frozen}\}_{j=1}^3$ , for further process.

**Pixel Decoder as Feature Adapter.** The pixel decoder is the same as Mask2Former, which contains multi-stage deformable attention layers. It transforms the frozen feature  $\{F_j^{frozen}\}_{j=1}^3$ , into the fused feature  $\{F_j^{fuse}\}_{j=1}^3$ , with the same channel dimension, where  $j$  is the layer index of feature.  $j = 3$  is the highest-resolution feature.

**Combined Object Queries.** As analyzed above, each object query represents one type of mask output. However, from the functionality perspective, image, video, and interactive modes represent different properties. For images, object queries focus on object-level localization and recognition. For video, object queries may involve temporal consistency, such as the same object long different frames. For interactive segmentation, object queries are forced to locate specific regions. For image and video input, we adopt object queries to represent image masks or tracked tube masks. Since both need semantic labels. We term them as semantic queries,  $Q_{obj}^s$ . For interactive mode, following SAM [38], we adopt the prompt encoder to encode the various visual prompts into the same shape of object queries. We term them as location queries,  $Q_{obj}^l$ . Thus, we can share the same interface for the transformer decoder.

**Shared Multi-Task Decoder.** Its main operation is cross-attention, which takes in the combined object queries ( $Q_{obj}^s$  and  $Q_{obj}^l$ ) and the image/video feature  $F_j^{fuse}$ , and outputs refined object queries. The final masks are obtained via dot-product of refined queries and high-resolution feature  $F_3^{fuse}$ . For image semantic level tasks, we adopt the same procedure of Mask2Former. In particular,  $Q_{obj}^s$  perform masked cross-attention [18] with multi-scale features  $F_j^{fuse}$ .  $Q_{obj}^s$  is Query while  $F_j^{fuse}$  are the Key and Value. Then, a multi-head self-attention (MHSA) layer is applied to the refined queries. The refined queries and high-resolution features are used to

For video tasks, we adopt the same cross-attention design. The only difference is the pyramid features  $F_j^{fuse}$  are contacted along the temporal dimension with 3D position embeddings, which are the default setting as previous works [16, 49]. The combined video features and refined queries are used to predict the tube mask.

For interactive segmentation, we carry out the same cross-attention design. However, we skip the self-attention to avoid interaction between mask queries in the MHSA layer, since the interactive segmentation only cares about the input visual prompt regions. After obtaining the refined object query, it is passed through a prediction FFN, which typically consists of a 3-layer perceptron with a ReLU activation layer and a linear projection layer. All the queries are supervised by mask classification loss and mask prediction loss. The decoding process is in a cascaded manner, in three

stages for each feature pyramid.

### 3.3. Training and Inference

**Joint Image Video Dataset Co-training.** Rather than first pre-trained on image datasets, our goal is to train all segmentation tasks only once jointly. All training targets are one entity label and mask for all three different cases. The entity can be thing, stuff, class-agnostic masks, and their corresponding labels. Note that the instance masks with the same ID  $d$  form the tube masks. During training, we apply Hungarian matching between the predicted and ground-truth entity masks to assign object queries to video/image entities, and then supervise their predicted masks and classification. The classifier is replaced by CLIP text embedding to avoid cross-dataset taxonomy conflicts. The final loss function is given as  $L = \lambda_{cls}L_{cls} + \lambda_{ce}L_{ce} + \lambda_{dice}L_{dice}$ . Here,  $L_{cls}$  is the Cross-Entropy (CE) loss for mask classification, and  $L_{ce}$  and  $L_{dice}$  are mask Cross Entropy (CE) loss and Dice loss [59, 74] for segmentation, respectively.

**Universal Inference.** For image segmentation, we follow the same inference procedure of Mask2Former [18]. For example, for PS, we merge the things and stuff according to the sorted scores. The scores are generated by CLIP text embedding. For video segmentation tasks, for VIS and VPS, to generate instance ID, following previous work, we use query matching rather than introducing extra tracking components. For VOS tasks, we adopt mask matching between the first frame and the remaining frames. For interactive segmentation tasks, we follow the original SAM [38], by providing box and point prompts, and obtain the binary masks. For open vocabulary segmentation, since we have a frozen CLIP encoder, we merge mask pooled score and learned score with the open-vocabulary embeddings.

**Combining Tasks For More Applications.** Since our model can perform various segmentation tasks, combining interactive, open vocabulary and image/video segmentation tasks can lead to several new applications. For example, we can combine interactive and video segmentation, leading to flexible prompt-driven video object segmentation. Or we can combine interactive segmentation with an open vocabulary setting, which results in open vocabulary interactive segmentation. More examples are provided in Sec. 4 and supplementary.

## 4. Experiments

**Datasets and Metrics.** Unlike regular settings, we aim to explore co-training on multiple datasets as much as possible. In Tab. 2, we use COCO panoptic [52], COCO-SAM, VIPSeg [57], and Youtube-VIS-2019 [90] (YT-VIS-19) as training datasets. In addition to the closed-set testing, we include the open vocabulary (OV) inference by using Youtube-VIS-2021, ADE-20k [98], and DAVIS-2017 datasets [5], where their annotations are not used during

the training. COCO-SAM is created by using the ground truth boxes, and mask center points are visual prompts. The annotations are obtained by COCO panoptic masks. Moreover, we also include the multi-dataset settings in Tab. 3 to verify the effectiveness of multi-dataset co-training of our OMG-Seg. In addition to Tab. 2, we add more datasets, including ADE-20k and YT-VIS21 for joint co-training. We use the corresponding metrics for each dataset, including PQ [37], mask mAP [52], VPQ [34], tube mAP [90], J&F [5], and mIoU [98].

**Implementation Details.** We implement our models and all other baselines in MMDetection [9]. We use the distributed training framework with 32 A100 GPUs. Each mini-batch has one image per GPU. For data augmentation, we adopt large-scale jitter as previous works [18, 49] to build strong baselines. For all models in each table, we adopt the same training steps. We use OpenCLIP [66] to initialize the backbone network and replace learned classifiers with their corresponding text embeddings. For image inputs, we treat them as pseudo videos by concatenating two images and their masks into one. We adopt different sampling rates to balance the training examples for each dataset. We report results of both frozen and trained backbones for reference. We list more details in the supplementary material.

## 4.1. Main Results

**System-level Comparison.** In Tab. 2, we present a comparative analysis of our OMG-Seg against recent methodologies across a variety of settings. A significant highlight of our work is its unique capability to deliver substantial results in all scenarios using a *single* model framework. In the realm of specific image and video segmentation models, OMG-Seg demonstrates performance on par with leading approaches like Mask2Former [18], TubeLink [49], and TarViS [2]. While it exhibits a slight decrease in performance on the COCO image segmentation benchmark, it achieves near state-of-the-art results on the VIPSeg datasets, showcasing its robustness and versatility. Furthermore, when benchmarked against open vocabulary methods such as FCCLIP [91] and ODISE [86], OMG-Seg not only competes favorably but also outperforms ODISE in certain scenarios. This is particularly evident in the realm of open vocabulary video segmentation on YT-VIS-21, as detailed in the 7th column of the table. These findings underscore the effectiveness and adaptability of our OMG-Seg approach in handling a wide array of segmentation challenges.

In addition, our method has been benchmarked against recent unified models, revealing insightful comparisons. When compared with vision generalists such as that described in [76], our approach, OMG-Seg, demonstrates superior performance. However, in comparison with several specialized segmentation models, including

Table 2. Experiment results of OMG-Seg on image, video, open-vocabulary, and SAM-like settings. \* denotes models are pre-trained on the Object365 dataset [68]. We only list representative methods due to the page limit. Refer to the supplementary material for more methods. Our results are the averaged results of five different experiments.

Methods	Backbone	COCO-PS PQ	Cityscapes-PS PQ	COCO-IS mAP	VIPSeg-VPS VPQ	YT-VIS-19 mAP	YT-VIS-21-OV mAP	ADE-OV PQ	DAVIS-17-VOS-OV J&F	COCO-SAM mIoU	Share Model
DetectorRS [64]	ResNet50	-	-	42.1	-	-	-	-	-	-	-
HITC [8]	ResNet50	-	-	38.4	-	-	-	-	-	-	-
STM [62]	ResNet101	-	-	-	-	-	-	-	79.2	-	-
K-Net [97]	ResNet50	47.1	-	38.6	-	-	-	-	-	-	-
MaskFormer [18]	ResNet50	51.9	62.1	43.7	-	-	-	-	-	-	-
MaskFormer [18]	Swin-Large	57.8	66.6	50.1	-	-	-	-	-	-	-
k-Max DeepLab [92]	ResNet50	53.0	64.3	-	-	-	-	-	-	-	-
k-Max DeepLab [92]	ConvNeXt-Large	58.1	68.4	-	-	-	-	-	-	-	-
SeqFormer [80]	ResNet50	-	-	-	-	47.4	-	-	-	-	-
IDOL [83]	Swin-Large	-	-	-	-	64.3	-	-	-	-	-
MinVIS [31]	Swin-Large	-	-	-	-	61.6	-	-	-	-	-
Video K-Net [51]	ResNet50	-	-	-	26.1	40.5	-	-	-	-	-
Tube-Link [49]	ResNet50	-	-	-	41.2	52.8	-	-	-	-	-
Tube-Link [49]	Swin-base	-	-	-	54.5	-	-	-	-	-	-
OneFormer [33]	Swin-Large	58.0	67.2	49.2	-	-	-	-	-	-	✓
TarViS [2]	Swin-Large	-	-	-	48.0	-	-	-	-	-	✓
fc-clip [91]	ConvNeXt-Large	54.4	-	44.6	-	-	-	26.8	-	-	✓
ODISE [86]	ViT-Large	55.4	-	46.0	-	-	-	22.6	-	-	✓
DaTaSeg [27]	ViT-L	53.5	-	-	-	-	-	-	-	-	✓
X-Decoder [103]	DaViT	56.9	-	46.7	-	-	-	21.8	-	-	✓
SEEM [104] *	DaViT	57.5	-	47.7	-	-	-	-	58.9	83.4	✓
UNINEXT [89] *	ConvNeXt-L	-	-	49.6	-	64.3	-	-	77.2	-	✓
HIPIE [75] *	ViT-H	58.0	-	51.9	-	-	-	20.6	-	-	✓
OpenSeeD [96] *	Swin-L	59.5	-	53.2	-	-	-	19.7	-	-	✓
SAM [38]	ViT-H	-	-	-	-	-	-	-	-	55.3	✓
Semantic-SAM [42]	Swin-T	55.2	-	47.4	-	-	-	-	-	53.0	✓
Painter [76]	ViT-L	43.4	-	-	-	-	-	-	-	-	✓
OMG-Seg	ConvNeXt-Large (frozen)	53.8	65.7	44.5	49.8	56.4	50.5	27.9	74.3	58.0	✓
OMG-Seg	ConvNeXt-XX-Large (frozen)	55.4	65.3	46.5	53.1	60.3	55.2	27.8	76.9	59.3	✓

Table 3. Experiment results of OMG-Seg on multiple dataset settings. We use five different datasets for balanced joint co-training for only 12 epochs. We also implement compared baselines in the same codebase.

Methods / Settings	Backbone	COCO-PS	COCO-IS	ADE-PS	VIPSeg-VPS	YT-VIS-19	YT-VIS-21	Params(M)	Share Model
K-Net [97]	ConvNeXt-Large (trained)	50.5	42.3	40.2	-	-	-	-	-
Mask2Former [18]	ConvNeXt-Large (trained)	53.2	45.2	43.2	-	-	-	-	-
Mask2Former-VIS [16]	ConvNeXt-Large (trained)	-	-	-	-	45.8	42.3	-	-
single dataset baseline	ConvNeXt-Large (frozen)	52.5	45.6	41.2	42.3	45.3	44.3	1326	-
OMG-Seg	ConvNeXt-Large (frozen)	52.9	44.3	28.2	46.9	48.8	46.2	221	✓
OMG-Seg	ConvNeXt-Large (trained)	55.0	45.3	36.8	45.8	47.2	45.2	221	✓

UNINEXT [89] and Wang et al. [75], we observe a discernible performance discrepancy in the COCO datasets, notably in panoptic and instance segmentation tasks. This gap, we argue, can be partially attributed to our training regime, which spans only 24 epochs, and also we keep the backbone frozen. Furthermore, the integration of video segmentation and interactive segmentation datasets for joint co-training presents a more formidable challenge compared to previous works. This is primarily because learning spatial-temporal and localization-sensitive features from image data is inherently more complex, given the diversity of the learning targets.

Despite these challenges, it is noteworthy that no other existing models offer the comprehensive segmentation capabilities that OMG-Seg does. This ability to effectively handle all forms of segmentation, despite the small performance gaps noted, reinforces our assertion that OMG-Seg is a robust and versatile model suitable for diverse segmentation scenarios.

**Multi-dataset Setting.** In Tab. 3, we extend our investigation to multi-dataset settings. To ensure a fair compar-

ison in the same setting, we reimplemented two key baselines: K-Net [97] and Mask2Former [18]. Our findings indicate that joint co-training generally enhances performance across most video segmentation datasets, leading to substantial model parameter reduction (from 1326M to 221M). This improvement is consistent across three VPS and VIS datasets, irrespective of whether the backbones are frozen or not. However, it is noteworthy that the performance on the ADE-20k dataset significantly diminishes under joint co-training. We hypothesize that this is largely due to the challenges posed by scale variance and the uneven distribution of classes within the dataset. Interestingly, when using a pre-trained backbone, we observe an uplift in image segmentation performance, albeit at the cost of a minor decline in video segmentation efficacy. This trade-off can be attributed to the unbalanced nature of samples that pursue different optimization objectives, essentially causing a tug-of-war over the representational capacity of the backbone. Such a scenario suggests that incorporating a greater volume of video training examples could potentially address this issue.

Table 4. Ablation on joint co-training. (a), COCO-PS. (b), VIPSeg-VPS. (c). YT-VIS-19.

Setting	COCO-PS	VIPSeg-VPS	YT-VIS-19	ADE-OV	YT-VIS-21-OV
a	53.4	32.2	34.2	25.5	30.3
a + b	52.9	49.0	45.2	26.2	39.6
a + b + c	53.0	48.5	56.8	26.1	50.3

Table 5. Ablation on shared decoder design.

Setting	COCO-PS	VIPSeg-VPS	Param	GFlops
shared	53.0	48.5	221	868
decoupled image/video	53.6	46.2	243	868

**Qualitative Result.** In Fig. 3, we show the effectiveness of our OMG-Seg model using a ConvNeXt-Large model across five different tasks. The first two rows demonstrate the model’s high-quality image segmentation capabilities on the COCO dataset. In the VIS and VPS tasks, OMG-Seg shows proficiency in segmenting and tracking foreground objects. Notably, in the last row, we show an open-vocabulary video instance segmentation on Youtube-VIS, successfully identifying the “lizard” class, which was not included in the training set.

## 4.2. Ablation Study and Analysis

In this section, we use COCO, VIPSeg, and Youtube-VIS-19 for ablation studies of our OMG-Seg. All experiments use frozen ConvNeXt-Large as the backbone and the same data augmentation with 12 epochs training by default.

**Effect of Training Dataset.** In Tab. 4, we evaluate the impact of various datasets on model performance. As indicated in the first row, using only the COCO dataset yields satisfactory zero-shot results across other datasets, largely attributed to the employment of frozen CLIP visual features for zero-shot region feature classification. Upon integration of the VIPSeg dataset, a slight dip in performance on the COCO dataset is observed. However, this is counterbalanced by significant improvements in both the VIPSeg and Youtube-VIS datasets. Incorporating all three datasets, COCO, VIPSeg, and Youtube-VIS, results in an optimal performance balance across all datasets, establishing this combination as our preferred and default configuration.

**Ablation on Shared Decoder Design.** In Tab. 5, we explore the efficacy of a shared decoder design. Employing a separate decoder head for video segmentation tasks results in a slight performance decrease. This outcome is influenced by our use of pseudo-video samples during image dataset training. By sharing the decoder, we align the optimization objectives more closely, which particularly benefits the video datasets with short clips [90].

**Ablation on Extra Adapter.** In Tab. 6, we assess the addition of an extra adapter to the frozen CLIP backbone, enhancing the capacity of OMG-Seg. Our experiments reveal

Table 6. Ablation on whether using extra adapters.

Setting	epoch	COCO-PS	VIPSeg-VPS	Params (M)	GFlops (G)
+ More Pixel Decoder Layer [102]	12	53.0	48.5	221	868
	12	53.5	49.2	+11	+103
	12	53.6	49.4	+21	+60
+ Adapter [15]	36	54.8	50.1	221	868
	36	54.6	49.6	+11	+103
	36	54.7	50.2	+21	+60

Table 7. Ablation on different CLIPs.

Backbone	epoch	COCO-PS	VIPSeg-VPS	ADE-OV	Params(M)	Flops(G)
ResNet50	12	44.8	42.0	18.2	59.5	340
ConvNeXt Large	12	53.0	48.5	26.8	221	868
ConvNeXt XX-Large	12	54.3	53.2	27.2	820	2854
ConvNeXt XX-Large	24	55.5	53.3	27.8	820	2854
ConvNeXt XX-Large	36	56.0	53.0	26.7	820	2854

that the adapter [15] boosts performance with fewer training epochs, but its effectiveness against the baseline disappears in extended training scenarios. In addition, we experiment with increasing the neck capacity by duplicating attention layers in the pixel decoder, observing similar outcomes to the adapter implementation. Consequently, we opt not to incorporate additional adapters, maintaining a cleaner and simpler framework.

**Ablation on Other CLIPs.** In Tab. 7, following the approach of prior open vocabulary research [39, 91], we primarily employ convolution-based CLIP models due to their spatial information handling and adaptability to scale variations across different datasets. As we scale up the CLIP model size and extend training steps, we observe improvements across all three datasets. Notably, model convergence is achieved at 24 epochs, faster than in previous studies [18]. This accelerated convergence may be attributed to the model’s limited capacity, suggesting that larger models could further elevate performance.

## 5. Conclusion

In this study, we introduce the first joint co-training framework for image, video, open-vocabulary, and interactive segmentation. Our solution, OMG-Seg, is a novel yet simple framework that uses a unified query representation and a shared decoder for diverse tasks. For the first time, it is possible to train a single segmentation model capable of performing across ten different tasks with competitive performance compared to task-specific models. This approach significantly reduces both the parameter size and the need for specialized engineering in model design for various applications. We envision that our efficient and versatile framework will serve as a robust baseline for multi-task and multi-dataset segmentation.

## 6. Appendix

**Overview.** In this appendix, we first present more method details in Sec. A. Then, we present more experiment re-



Figure 3. Functional Visualization of OMG-Seg model. We list five different tasks from four datasets as examples. Our method achieves high-quality segmentation, tracking, and as well as interactive segmentation in one shared model.

sults in Sec. B. Finally, we show more image, video, open-vocabulary, and interactive segmentation demos in Sec. C.

## A. More Method Details

**More Detailed Comparison with Recent Works.** Due to the page limitation, we only select several representative works for setting comparison. Compared with specific models [17, 64], our method achieves extreme parameter sharing and performs various tasks that these models cannot perform.

Compared with video segmentation and unified video segmentation [2, 49], our method can also achieve open-vocabulary and interactive segmentation, as well as good enough performance on image segmentation. This is because our model is jointly co-trained on both image and video segmentation datasets without introducing task-specific tuning on video segmentation datasets. In addition, due to the frozen CLIP backbone, our method can also perform video open vocabulary segmentation without any architecture modification.

Compared with recent partial unified models, our method achieves all related visual segmentation in one model. For example, compared with Semantic-SAM [42], our model can achieve both video segmentation (VIS, VSS, VPS) and open-vocabulary segmentation. Compared with UNINEXT [89], our method can perform interactive segmentation, panoptic segmentation (VPS, PS), and open-

vocabulary segmentation. Compared with OneFormer [33], we can achieve video, open-vocabulary, and interactive segmentation. Compared with TarVS [2], we can keep image segmentation without specific fine-tuning. Compared with recent FreeSeg [65], we can achieve both video segmentation and interactive segmentation in one model.

**Implementation Details of OMG-Seg.** We use balanced training for our model. In particular, for two different setting of Tab.2 and Tab.3 in the main paper, we balance each dataset sample according to the COCO dataset size. Then, we choose the same data augmentation as Mask2Former [18]. For the text embedding generation, we follow the standard open-vocabulary detection and segmentation setting [82, 101]. We generate multiple text prompts with the class names and keep the text embedding fixed for both training and inference. In this way, we can achieve multi-dataset and open-vocabulary segmentation.

**More Detailed Inference Process.** Our model has various inference modes. For image segmentation on various datasets, we simply follow the Mask2Former to obtain the corresponding mask and labels. For video segmentation, we adopt simple query matching [31, 51] without learning the extra tracking query embedding. We believe adding such components will improve the video segmentation. For open-vocabulary segmentation, we fuse the frozen CLIP visual scope and predicted scope to boost the novel class segmentation. For interactive segmentation, we mainly use the

Table 8. Results using ResNet50 backbone.

Method	Backbone	COCO-PS	VIPSeg-VPS	Youtube-VIS-2019
Mask2Former [18]	ResNe50	52.0	-	-
Mask2Former-VIS [16]	ResNe50	-	-	46.4
OMG-Seg	ResNe50	49.9	42.3	46.0
OMG-Seg	ConvNext-L	54.5	50.5	56.2

Table 9. Results using ViT backbone.

Backbone	COCO-PS	Youtube-VIS-2019	VIP-Seg
ViT-L (frozen)	34.5	23.2	34.5
ViT-L (learned)	52.2	54.3	48.2
ConvNext-L (frozen)	54.5	56.2	50.5

Table 10. Ablation on self-attention mode for interactive segmentation tasks. We use ResNet50 as the backbone. The masks are used to filter out the correlation of each query during self-attention.

Setting	COCO-PS	COCO-SAM
Self Attention without masks	45.2	40.7
Self Attention with masks	49.9	52.2

point prompts to evaluate despite the box prompts, which can also be used as SAM [38]. Moreover, since our model adopts the frozen CLIP features, we can freely label the prompt-driven segmentation masks, where we can achieve open-vocabulary interactive segmentation. The GFlops of the main paper are calculated with  $1200 \times 800$  by default.

## B. More Experiment Results

In addition to the main paper, we also provide more ablation studies and experiment results here.

**Results Using ResNe50 backbone.** In Tab. 8, we report our model using ResNet50 backbone. We jointly co-train our model with 24 epochs. Compared with specific Mask2Former for 50 epoch training, our model can achieve considerable results but with less parameter costs.

**Exploration on ViT-based CLIP backbone.** In Tab. 9, we explore the CLIP-ViT backbone. We find using frozen CLIP-ViT leads to inferior results. This is because the position embedding of ViT is fixed (224 by default), and a simple bilinear upsampling operation hurts the origin representation. Thus, in the second row, we adopt the learned architecture. However, we still find performance gaps with convolution-based CLIP. Moreover, since there is no frozen CLIP and the open-vocabulary ability is lost during the fine-tuning.

**Interactive Segmentation with Masked Self-Attention.** In interactive mode, we set the query invisible (achieve this by masking) to each other during the cross-attention process. If not, as shown in Tab. 10, we find a significant performance drop for both COCO-SAM and COCO-PS. This is because, for interactive segmentation, the local features

are good enough, while introducing the global information will bring noise to the query learning.

## C. More Visualization Example

**More Visual Results on More Tasks.** In Fig. 4, we present more visual examples for two additional tasks. One is open-vocabulary panoptic segmentation on ADE-20k. As shown in the top row, our method can achieve good zero-shot segmentation quality. In the second row, we also provide interactive segmentation on the ImageNet-1k dataset. We add the class labels that are from the simple CLIP score. To this end, we achieve open-vocabulary interactive segmentation.

**Limitation and Future Work.** One limitation of our work is the capacity of our model. Since we use the frozen architecture to keep the open-vocabulary ability, which leads to inferior results for one specific dataset or task. However, we believe adding more dataset co-training [38] with the learned backbone will improve our model performance. With the aid of more text-image pairs or classification datasets, we also achieve open-vocabulary segmentation ability while keeping the performance improved on close sets. This is our future work to scale up our model. Moreover, we can also add a text path to support language-driven segmentation tasks, such as referring image/video segmentation or even with large language models (LLMs) to perform joint reasoning and segmentation in one framework.

## References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 3
- [2] Ali Athar, Alexander Hermans, Jonathon Luiten, Deva Ramanan, and Bastian Leibe. Tarvis: A unified architecture for target-based video segmentation. In *CVPR*, 2023. 2, 3, 6, 7, 9
- [3] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. *arXiv preprint arXiv:2312.00785*, 2023. 3
- [4] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. In *NeurIPS*, 2022. 3
- [5] Sergi Caelles, Alberto Montes, Kevis-Kokitsi Maninis, Yuhua Chen, Luc Van Gool, Federico Perazzi, and Jordi Pont-Tuset. The 2018 davis challenge on video object segmentation. *arXiv preprint arXiv:1803.00557*, 2018. 2, 6
- [6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2, 3
- [7] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou.



Figure 4. More functional Visualization of OMG-Seg model. In addition to five different tasks of the main paper, we also visualize the open-vocabulary segmentation results: open-vocabulary panoptic segmentation results on ADE-20k, open-vocabulary interactive segmentation results on ImageNet 1k dataset.

- Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2
- [8] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. In *CVPR*, 2019. 2, 7
- [9] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint*, 2019. 6
- [10] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv:1706.05587*, 2017. 2
- [11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 2, 3
- [12] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021. 3
- [13] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J. Fleet, and Geoffrey Hinton. A unified sequence interface for vision tasks. *arXiv preprint arXiv:2206.07669*, 2022. 3
- [14] Xi Chen, Shuang Li, Ser-Nam Lim, Antonio Torralba, and Hengshuang Zhao. Open-vocabulary panoptic segmentation with embedding modulation. *ICCV*, 2023. 3
- [15] Zhe Chen, Yuchen Duan, Wenhui Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 8
- [16] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint*, 2021. 5, 7, 10
- [17] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for

- bottom-up panoptic segmentation. In *CVPR*, 2020. 9
- [18] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 2, 3, 4, 5, 6, 7, 8, 9, 10
- [19] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. 4
- [20] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*, 2023. 2
- [21] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. MOSE: A new dataset for video object segmentation in complex scenes. In *ICCV*, 2023.
- [22] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. VLT: Vision-language transformer and query generation for referring segmentation. *IEEE TPAMI*, 2023. 2
- [23] 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. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3
- [24] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022. 3
- [25] Zhongbin Fang, Xiangtai Li, Xia Li, Joachim M Buhmann, Chen Change Loy, and Mengyuan Liu. Explore in-context learning for 3d point cloud understanding. *arXiv preprint arXiv:2306.08659*, 2023. 3
- [26] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, 2022. 2, 3
- [27] Xiuye Gu, Yin Cui, Jonathan Huang, Abdullah Rashwan, Xuan Yang, Xingyi Zhou, Golnaz Ghiasi, Weicheng Kuo, Huizhong Chen, Liang-Chieh Chen, et al. DaTaseg: Taming a universal multi-dataset multi-task segmentation

- model. In *NeurIPS*, 2023. 2, 7
- [28] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2021. 2, 3
- [29] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 3
- [30] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2, 3
- [31] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. In *NeurIPS*, 2022. 7, 9
- [32] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hanneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Open-clip, July 2021. 5
- [33] Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. OneFormer: One Transformer to Rule Universal Image Segmentation. In *CVPR*, 2023. 2, 3, 7, 9
- [34] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *CVPR*, 2020. 3, 6
- [35] Dahun Kim, Jun Xie, Huiyu Wang, Siyuan Qiao, Qi-hang Yu, Hong-Seok Kim, Hartwig Adam, In So Kweon, and Liang-Chieh Chen. Tubeformer-deeplab: Video mask transformer. In *CVPR*, 2022. 2, 3, 5
- [36] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *CVPR*, 2019. 2, 3
- [37] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, 2019. 2, 6
- [38] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 2, 3, 5, 6, 7, 10
- [39] Weicheng Kuo, Yin Cui, Xiuye Gu, A. J. Piergiovanni, and Anelia Angelova. F-VLM: Open-vocabulary object detection upon frozen vision and language models. In *ICLR*, 2023. 8
- [40] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. Mseg: A composite dataset for multi-domain semantic segmentation. In *CVPR*, 2020. 3
- [41] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022. 3
- [42] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*, 2023. 3, 7, 9
- [43] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 3
- [44] Wei Li, Jiahao Xie, and Chen Change Loy. Correlational image modeling for self-supervised visual pre-training. In *CVPR*, 2023. 3
- [45] Xiangtai Li, Henghui Ding, Wenwei Zhang, Haobo Yuan, Guangliang Cheng, Pang Jiangmiao, Kai Chen, Ziwei Liu, and Chen Change Loy. Transformer-based visual segmentation: A survey. *arXiv pre-print*, 2023. 2
- [46] Xiangtai Li, Xia Li, Li Zhang, Guangliang Cheng, Jianping Shi, Zhouchen Lin, Shaohua Tan, and Yunhai Tong. Improving semantic segmentation via decoupled body and edge supervision. In *ECCV*, 2020. 2
- [47] Xiangtai Li, Shilin Xu, Yibo Yang, Guangliang Cheng, Yunhai Tong, and Dacheng Tao. Panoptic-partformer: Learning a unified model for panoptic part segmentation. In *ECCV*, 2022. 2
- [48] Xiangtai Li, Ansheng You, Zhen Zhu, Houlong Zhao, Maoke Yang, Kuiyuan Yang, and Yunhai Tong. Semantic flow for fast and accurate scene parsing. In *ECCV*, 2020. 2
- [49] Xiangtai Li, Haobo Yuan, Wenwei Zhang, Guangliang Cheng, Jiangmiao Pang, and Chen Change Loy. Tube-link: A flexible cross tube baseline for universal video segmentation. In *ICCV*, 2023. 2, 3, 5, 6, 7, 9
- [50] Xiangtai Li, Jiangning Zhang, Yibo Yang, Guangliang Cheng, Kuiyuan Yang, Yu Tong, and Dacheng Tao. Sfnet: Faster and accurate domain agnostic semantic segmentation via semantic flow. *IJCV*, 2023. 2
- [51] Xiangtai Li, Wenwei Zhang, Jiangmiao Pang, Kai Chen, Guangliang Cheng, Yunhai Tong, and Chen Change Loy. Video k-net: A simple, strong, and unified baseline for video segmentation. In *CVPR*, 2022. 2, 3, 7, 9
- [52] 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 *ECCV*, 2014. 2, 6
- [53] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 2, 5
- [54] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. In *ICLR*, 2023. 3
- [55] Jitendra Malik, Serge Belongie, Thomas Leung, and Jianbo Shi. Contour and texture analysis for image segmentation. *IJCV*, 2001. 2
- [56] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 2
- [57] Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *CVPR*, 2022. 6
- [58] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *CVPR*, 2021. 2, 3
- [59] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016. 6
- [60] Shervin Minaee, Yuri Y Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *PAMI*, 2021. 2

- [61] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019. 3
- [62] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019. 7
- [63] Zhihang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 3
- [64] Siyuan Qiao, Liang-Chieh Chen, and Alan Yuille. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In *CVPR*, 2021. 3, 7, 9
- [65] Jie Qin, Jie Wu, Pengxiang Yan, Ming Li, Ren Yuxi, Xuefeng Xiao, Yitong Wang, Rui Wang, Shilei Wen, Xin Pan, et al. Freeseg: Unified, universal and open-vocabulary image segmentation. In *CVPR*, 2023. 3, 9
- [66] 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. In *ICML*, 2021. 2, 6
- [67] Florian Schroff, Antonio Criminisi, and Andrew Zisserman. Object class segmentation using random forests. In *BMVC*, 2008. 2
- [68] 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 *ICCV*, 2019. 7
- [69] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 3
- [70] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *CVPR*, 2021. 2
- [71] Wenhui Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023. 2
- [72] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *ICCV*, 2021. 3
- [73] Xinshun Wang, Zhongbin Fang, Xia Li, Xiangtai Li, Chen Chen, and Mengyuan Liu. Skeleton-in-context: Unified skeleton sequence modeling with in-context learning. *arXiv preprint arXiv:2312.03703*, 2023. 3
- [74] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *ECCV*, 2020. 6
- [75] Xudong Wang, Shufan Li, Konstantinos Kallidromitis, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Hierarchical open-vocabulary universal image segmentation. In *NeurIPS*, 2023. 2, 7
- [76] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *CVPR*, 2023. 2, 3, 6, 7
- [77] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. In *ICCV*, 2023. 2, 3
- [78] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *CVPR*, 2021. 3
- [79] Zhenyu Wang, Yali Li, Xi Chen, Ser-Nam Lim, Antonio Torralba, Hengshuang Zhao, and Shengjin Wang. Detecting everything in the open world: Towards universal object detection. In *CVPR*, 2023. 3
- [80] Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance segmentation. In *ECCV*, 2022. 7
- [81] Jianzong Wu, Xiangtai Li, Henghui Ding, Xia Li, Guanglei Cheng, Yunhai Tong, and Chen Change Loy. Betrayed by captions: Joint caption grounding and generation for open vocabulary instance segmentation. In *ICCV*, 2023. 3
- [82] Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, Bernard Ghanem, and Dacheng Tao. Towards open vocabulary learning: A survey. *arXiv pre-print*, 2023. 2, 3, 9
- [83] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *ECCV*, 2022. 7
- [84] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *arXiv preprint arXiv:2310.01403*, 2023. 3
- [85] Jiahao Xie, Wei Li, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Masked frequency modeling for self-supervised visual pre-training. In *ICLR*, 2023. 3
- [86] 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 *CVPR*, 2023. 2, 6, 7
- [87] Shilin Xu, Xiangtai Li, Size Wu, Wenwei Zhang, Ying Li, Guanglei Cheng, Yunhai Tong, Kai Chen, and Chen Change Loy. Dst-det: Simple dynamic self-training for open-vocabulary object detection. *arXiv preprint arXiv:2310.01393*, 2023. 3
- [88] Shilin Xu, Haobo Yuan, Qingyu Shi, Lu Qi, Jingbo Wang, Yibo Yang, Yining Li, Kai Chen, Yunhai Tong, Bernard Ghanem, Xiangtai Li, and Ming-Hsuan Yang. Rap-sam: Towards real-time all-purpose segment anything. *arXiv preprint*, 2024. 2
- [89] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *CVPR*, 2023. 2, 3, 7, 9
- [90] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 2, 6, 8
- [91] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. In *NeurIPS*, 2023. 2, 6, 7, 8
- [92] Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. k-means mask transformer. In *ECCV*, 2022. 2, 7

- [93] Haobo Yuan, Xiangtai Li, Chong Zhou, Yining Li, Kai Chen, and Chen Change Loy. Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively. *arXiv preprint*, 2024. 3
- [94] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *ECCV*, 2022.
- [95] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021. 2, 3
- [96] 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 *ICCV*, 2023. 7
- [97] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. In *NeurIPS*, 2021. 2, 3, 4, 7
- [98] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. In *CVPR*, 2017. 2, 6
- [99] Hao Zhou, Tiancheng Shen, Xu Yang, Hai Huang, Xiangtai Li, Lu Qi, and Ming-Hsuan Yang. Rethinking evaluation metrics of open-vocabulary segmentaion. *arXiv preprint arXiv:2311.03352*, 2023. 3
- [100] Tianfei Zhou, Fatih Porikli, David J Crandall, Luc Van Gool, and Wenguan Wang. A survey on deep learning technique for video segmentation. *PAMI*, 2023. 2
- [101] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 9
- [102] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2020. 8
- [103] Xueyan Zou\*, Zi-Yi Dou\*, Jianwei Yang\*, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee\*, and Jianfeng Gao\*. Generalized decoding for pixel, image and language. In *CVPR*, 2023. 2, 7
- [104] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. In *NeurIPS*, 2023. 3, 7