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

1. Appendix

Overview. In this appendix, we first present more method details in Sec. A. Then, we present more experiment results 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 [3, 11], our method achieves extreme parameter sharing and performs various tasks that these models cannot perform.

Compared with video segmentation and unified video segmentation [1, 9], 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 [8], our model can achieve both video segmentation (VIS, VSS, VPS) and open-vocabulary segmentation. Compared with UNINEXT [14], our method can perform interactive segmentation, panoptic segmentation (VPS, PS), and open-vocabulary segmentation. Compared with OneFormer [6], we can achieve video, open-vocabulary, and interactive segmentation. Compared with TarVS [1], we can keep image segmentation without specific fine-tuning. Compared with recent FreeSeg [12], 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 [4]. For the text embedding generation, we follow the standard open-vocabulary detection and segmentation setting [13, 15]. 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 [5, 10] 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 point prompts to evaluate despite the box prompts, which can also be used as SAM [7]. 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 1200×800 by default.

Table 1. Results using ResNet50 backbone.

Method	Backbone	COCO-PS	VIPSeg-VPS	Youtube-VIS-2019
Mask2Former [4]	ResNe50	52.0	-	-
Mask2Former-VIS [2]	ResNe50	-	-	46.4
OMG-Seg	ResNe50	49.9	42.3	46.0
OMG-Seg	ConvNext-L	54.5	50.5	56.2

Table 2. 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

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. 1, we report our model using ResNet50 backbone. We jointly co-

Table 3. Ablation on self-attention mode for interactive segmentation tasks. We use ResNet50 as the backbone. The masks 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

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. 2, 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. 3, 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. 1, 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 [7] 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] Ali Athar, Alexander Hermans, Jonathon Luiten, Deva Ramanan, and Bastian Leibe. Tarvis: A unified architecture for target-based video segmentation. In *CVPR*, 2023. 1

[2] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint*, 2021. 1

[3] 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. 1

[4] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 1

[5] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. In *NeurIPS*, 2022. 1

[6] Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. OneFormer: One Transformer to Rule Universal Image Segmentation. In *CVPR*, 2023. 1

[7] 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. 1, 2

[8] 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. 1

[9] 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. 1

[10] 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. 1

[11] Siyuan Qiao, Liang-Chieh Chen, and Alan Yuille. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In *CVPR*, 2021. 1

[12] 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. 1

[13] 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. 1

[14] 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. 1

[15] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 1

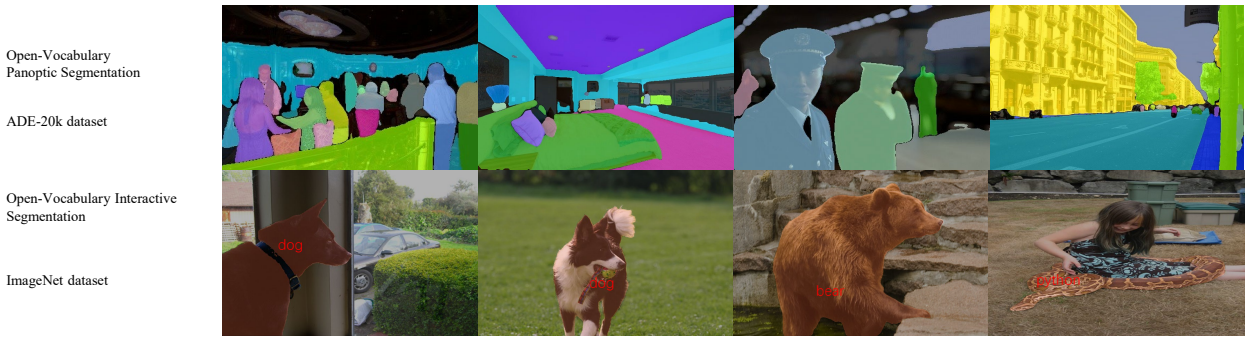

Figure 1. More functional Visualization of OMG-Seg model. In addition to five different tasks of the main paper, we also visualize the open-vocabulary segmentation results: open-vocabulary panoptic segmentation results on ADE-20k, open-vocabulary interactive segmentation results on ImageNet 1k dataset.