—Supplementary Materials—
A Simple Framework for Open-Vocabulary Segmentation and Detection

Table 1: *OpenSeeD*(L) model with 4-scale image features compared with X-Decoder (L) on ADE20K.

Method	Training Data			ADE			
	SEG	DET	ITP	PQ	mask AP	box AP	mIoU
X-Decoder (L) [4]	✓	✗	✓	21.8	13.1	–	29.6
OpenSeeD (L)	✓	✓	✗	20.3	**15.0**	**18.3**	23.6

Table 2: Ablation of the effectiveness of pseudo annotations in offline mask assistance for our open-vocabulary model. We evaluate the model performance on ADE20K and COCO. ”-anno” denotes with annotations and ”w/o anno” denotes without annotations.

Method	ADE				COCO			
	PQ	mask AP	box AP	mIoU	PQ	mask AP	box AP	mIoU
OpenSeeD-SwinT w/o anno	19.8	14.1	17.0	22.9	55.2	47.3	51.9	63.7
OpenSeeD-SwinT-anno	**20.4**(+0.6)	**14.8**(+0.7)	**17.1**	**24.0**(+1.1)	55.5	**47.9**(+0.6)	**52.4**(+0.5)	63.9

Overview

This supplementary material presents more details and additional results not included in the main paper due to page limitation. The list of items included are:

- More experimental results in Sec. A.
- Visualization of the predictions of *OpenSeeD* in Sec. B.
- More implementation details in Sec. C.

A. More Experimental Results

A.1. SwinL 4-scale results

Our Swin-L results in Table 2 adopts 5 scales of image features. In order to compare with other methods more thoroughly, we show the performance of our model with Swin-L and 4 scales of image features in Table 1.

A.2. Offline Mask Guidance Ablation with SwinT

In order to be coherent with Table 2, we also show the ablation of offline mask assistance in Table 2, which verifies the effectiveness of our pseudo-annotations.

B. Visualization

In this section, we show a visualization of *OpenSeeD* for open-vocabulary segmentation on ADE20K and Objects365 we also show the conditioned segmentation ability of *OpenSeeD*. Note that all experiments here utilize the model jointly trained on COCO panoptic segmentation and Objects365 detection without fine-tuning.

B.1. Three open-vocabulary segmentation tasks on ADE20K

In Fig. 1, we show a visualization of *OpenSeeD* for open-vocabulary instance segmentation and detection, panoptic and semantic segmentation on ADE20K dataset without finetuning.

B.2. Segmentation in Objects365 Categories

In Fig. 2, we show the instance segmentation on Objects365 where unseen concepts are listed under each image. Unseen concepts are the categories that do not exist in COCO, which means they are not trained with segmentation annotations. Our model can segment instances from the unseen categories well although it is only trained with detection task on these categories.

B.3. Conditioned Segmentation

With the help of dn groups proposed in [1, 3], the model has the ability to predict masks given boxes. In Fig. 3, we show the conditioned segmentation ability of *OpenSeeD*. When we give different conditioned boxes and text, we can obtain the corresponding masks.

C. More Implementation Details

Online Conditioned Mask as the Guidance. Given that our model can generate reasonably good masks with GT boxes as the condition, we seek to use them to better align detection with segmentation. A straightforward way is directly using masks for supervision on the fly. This requires high mask quality during the whole training phase, which however is not true, especially in the early training stage. Therefore, we alternatively use the generated mask as the additional guidance to find the matched foreground queries with the GT concept and box in detection data. As shown in Fig. 4 (a), detection during training fully ignores the predicted mask quality when finding the matched foreground queries, which is different from segmentation in

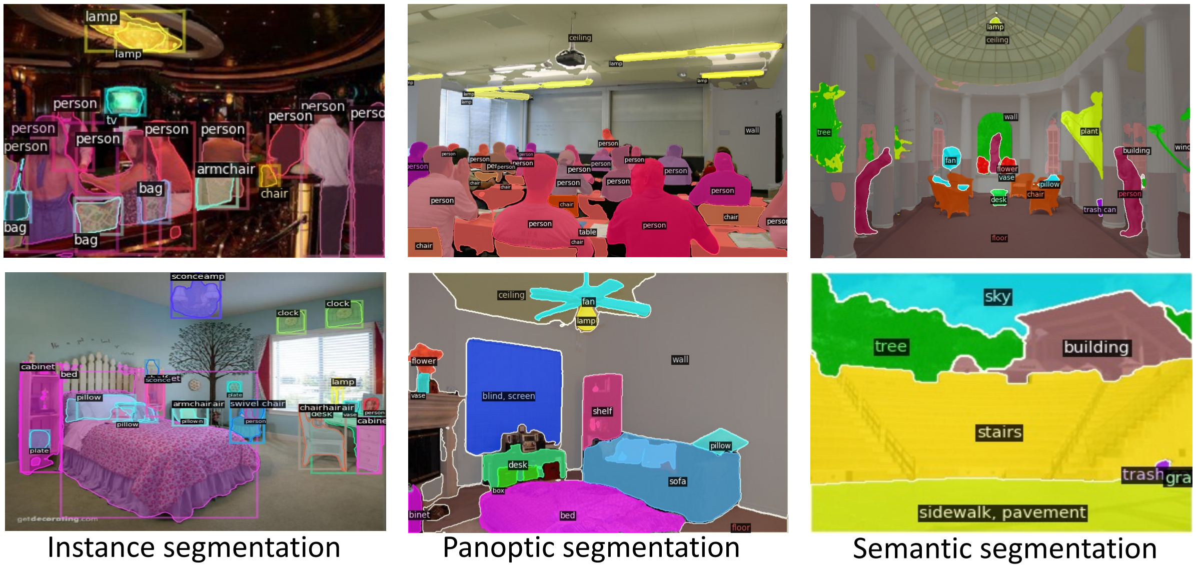

Figure 1: Visualizations of *OpenSeeD* for open-vocabulary instance segmentation and detection, panoptic and semantic segmentation.

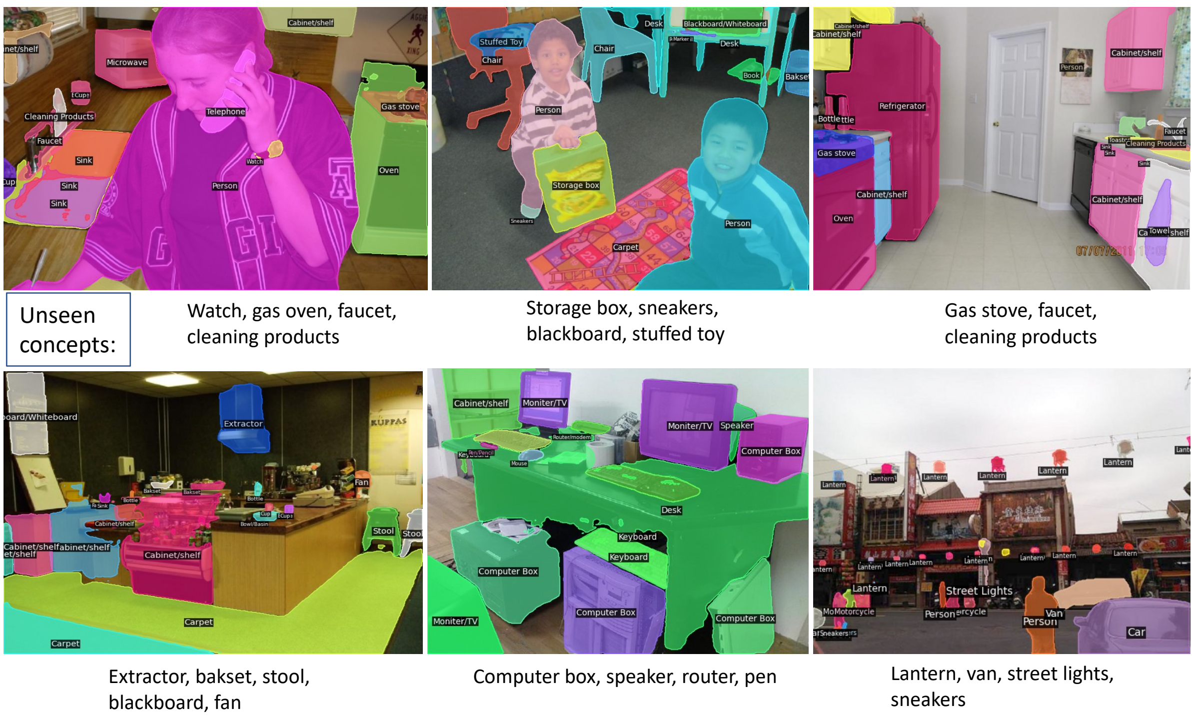

Figure 2: Visualizations of *OpenSeeD* for open-vocabulary instance segmentation and detection, panoptic and semantic segmentation.

Fig. 4 (b). This ignorance may lead to a biased matching toward box quality for our model which needs to produce high-quality boxes and masks simultaneously. Therefore, we use the generated masks from GT boxes and labels as the guidance for better label assignment, as shown in Fig. 4 (c).

Offline Conditioned Mask as the Supervision. Another way to better use the generated mask is to train two models. For example, we can train a large conditioned mask decoding model first and then use it to generate pseudo masks for all the detection data. Afterward, we can train the second model on the annotated detection dataset with mask supervision. However, though our model can generate fairly good masks for novel categories, these masks still has inferior quality compared to the trained categories. Considering that 85 categories in Objects365 [2] are in the COCO foreground category set, we treat the generated annotations on these categories as *golden annotations* because they are trained with mask annotations on COCO. Other annotations are *coarse annotations*. We adopt different strategies for the two types of annotations, where *golden annotations* are used for mask supervision while *coarse annotations* are only used for matching (similar to our online guidance).

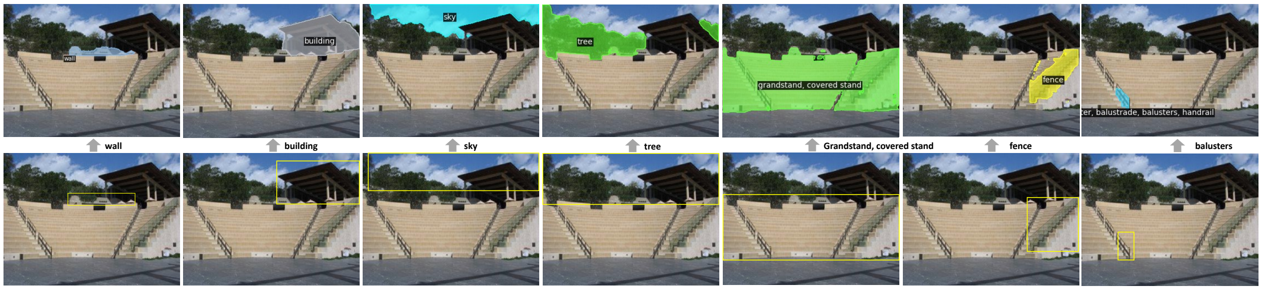

Figure 3: Visualizations of *OpenSeeD* for conditioned segmentation. The lower row is the original image with GT boxes and labels as conditions. The upper row is conditioned segmentation.

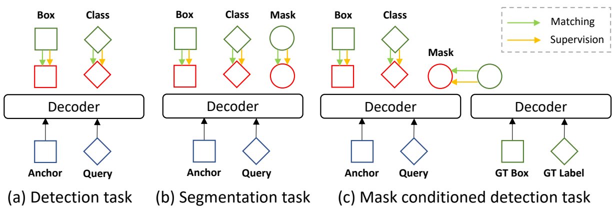

Figure 4: (a)(b) Standard detection and segmentation. They differ in that the segmentation task has mask supervision and matching. (c) In our online training, the detection task is assisted by conditioned-generated masks when matching.

References

[1] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13619–13627, 2022. 1

[2] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 2

[3] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022. 1

[4] 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. *arXiv preprint arXiv:2212.11270*, 2022. 1