Supplementary Material
Open Vocabulary Object Detection with Pseudo Bounding-Box Labels

Mingfei Gao* , Chen Xing*, Juan Carlos Niebles, Junnan Li,
Ran Xu, Wenhao Liu, Caiming Xiong

Salesforce Research, Palo Alto, USA
{mingfei.gao,cxing,jniebles,junnan.li}@salesforce.com
{ran.xu,wenhao.liu,cxiong}@salesforce.com

1 Additional Analysis

We conduct the following analysis to understand more about the performance of our method and our major baseline, *Zareian et al.*, [2].

The effect of data amount for pre-training vision-language model. Both *Zareian et al.* and our method leverage a vision-language (VL) model. Intuitively, a VL model can be improved by training with more data that will potentially improve both our method and baseline. We verify this hypothesis as follows.

First, we attempt to improve *Zareian et al.* by training their VL model with more data. We observe that the performance of [2] drops when pre-training with more data. This observation aligns well with the findings in their original paper (see Table 2 of [2]). Specifically, [2] is originally pre-trained with COCO Caption (the baseline we compared with in our main paper). When pre-trained with COCO Caption, VG and SBU (1M images in total), their performance drops by 4.2% on COCO novel categories. When pre-trained with COCO Caption, VG, SBU and CC3M (4M images in total), their performance drops further by 10.8% compared to our baseline.

Then, we also try to weaken our method by re-training our VL model (ALBEF) with COCO caption data only. The comparison is shown in Table A1. The results suggest that our method still outperforms our baseline largely by 5% even when we do not leverage one main strength of our method: being able to effectively employ more image-caption data without much extra cost.

Also, our method can benefit from more image-caption data, as suggested by our performance in Table A1 (27.8) vs. our performance (30.8) in Table 1 in our main paper.

Why *Zareian et al.* cannot easily take advantage of more powerful VL models? We find that *Zareian et al.* cannot easily incorporate more powerful VL models, e.g., ALBEF [1] due to its core designs. First, it requires the same model architecture for the VL model's visual encoder and the detection backbone, since the main idea of [2] is to initialize their open vocabulary detector using the

[*] Mingfei and Chen contributed equally.

Table A1. Results when our method only utilizes COCO caption to pre-train our VL model

Methods	Pre-train Source	COCO Novel AP
Zareian et al.	COCO Caption	22.8
Our method	COCO Caption	27.8

parameters of visual encoder of the VL model as a way of knowledge transfer. Therefore, their ResNet-based detector makes it impossible to utilize ALBEF's visual encoder because ALBEF's visual encoder is transformer-based. It is also challenging to utilize other SOTA VL models due to the different designs and sizes of visual encoders. In contrast, our framework can utilize most of recent and powerful VL models because they are disentangled from the training of our detectors. This flexibility brings large performance improvement.

References

1. Li, J., Selvaraju, R.R., Gotmare, A.D., Joty, S., Xiong, C., Hoi, S.: Align before fuse: Vision and language representation learning with momentum distillation. In: NeurIPS (2021)
2. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: CVPR. pp. 14393–14402 (2021)