Appendix for CapDet: Unifying Dense Captioning and Open-World Detection Pretraining

A. Detailed Experimental Settings

The detailed architecture parameters for different modules of CapDet are shown in Table 1. For the learning rate scheduler, we assign a base learning rate and then linearly warm it up to the peak learning rate according to the effective total batch size by a square root strategy, $lr_{peak} = lr_{base} \times \sqrt{\text{batchsize}/16}$, *e.g.*, we set image encoder base learning rate to 1×10^{-4} and it automatically scales to 1.4×10^{-4}. The training hyperparameters used for CapDet are shown in Table 2.

Image Encoder	Value
backbone	swin-t
neck	fpn
input resolution	1333×800
Text Encoder	**Value**
width	512
heads	8
layers	12
Cross-Modal Decoder	**Value**
width	512
heads	12
layers	12

Table 1. Detailed architecture parameters for different module.

Hyperparameter	Value(%)
Image encoder lr	1.4×10^{-4}
Text encoder lr	1.4×10^{-5}
Crossmodal decoder lr	1.4×10^{-5}
Learning policy	CosineAnnealing
warmup ratio	0.0001
warmup iters	1000
batchsize	32
weight decay	0.05
w_c	1
w_d	1

Table 2. The training hyperparameters used for CapDet.

B. Fine-tuning Results on LVIS

We provide the fine-tuning results on LVIS in Table 3 below. We observe that CapDet outperforms the baseline DetCLIP with 1.2% AP on average and 6.5% AP on rare classes. Besides, though pre-trained with fewer data and tasks, CapDet shows a competitive performance compared with the GLIPv2.

Model	Backbone	Pre-Train Data	Images Number	LVIS AP	LVIS AP_r / AP_c / AP_f
DetCLIP-T(C)* [11]	Swin-T	O365, VG	0.73M	45.6	33.6 / 45.8 / 47.5
GLIPv2-T [8]	Swin-T+DH+F	O365, GoldG, Cap4M	5.43M	50.6	- / - / -
CapDet	Swin-T	**O365, VG**	**0.73M**	**47.2**	**40.1 / 46.9 / 48.7**

Table 3. Fine-tuning performance on LVIS [4] MiniVal5k datasets. AP_r/AP_c/AP_f indicate the AP values for rare, common, and frequent categories. 'DH' and 'F' in GLIP [8] baselines stand for the dynamic head [2] and cross-modal fusion.

C. Open-World Detection Results on LVIS Full Validation Set

Table 4 reports our zero-shot object detection performance on LVIS [4] full validation set. Following [8, 11], we take the class names with additional manually designed prompts as input of text encoder. Comparing the 5th row and 6th row, our CapDet still outperforms DetCLIP-T(C) on the same data scale and backbone with an extra simple caption head. The zero-shot performance surpasses the previous methods with the same backbone by a large margin on rare classes, *e.g.*, CapDet trained on fewer data outperforms GLIP-T [8] by 10.8% on AP_r.

Model	Backbone	Pre-Train Data	Images Number	LVIS Val Full	
				AP	AP_r / AP_c / AP_f
GLIP-T(A) [8]	Swin-T+DH+F	O365	0.66M	12.3	6.00 / 8.00 / 19.4
GLIP-T [8]	Swin-T+DH+F	O365,GoldG,Cap4M	5.43M	17.2	10.1 / 12.5 / 25.2
DetCLIP-T(A) [11]	Swin-T	O365	0.66M	22.1	18.4 / 20.1 / 26.0
DetCLIP-T(C) [11]	Swin-T	O365, VG	0.73M	23.5	18.4 / 21.6 / 27.9
CapDet	Swin-T	O365, VG	0.73M	**26.1**	**20.9 / 24.4 / 30.2**

Table 4. Zero-shot transfer performance on LVIS [4] full validation dataset. AP_r/AP_c/AP_f indicates the AP values for rare, common, and frequent categories. 'DH' and 'F' in GLIP [8] baselines stand for the dynamic head [2] and cross-modal fusion.

D. Analysis of the Improvements on OVD

We attribute the improvements on OVD to the reason that the incorporation of captioning head brings more generalizability for the region features, which in turn helps the learning of OVD task. Specifically, the dense captioning task is essentially a sequential classification task with a large enough class space (*i.e.*, word tokens), while alignment task is a single-step classification task with a limited class space. Therefore, training with dense captioning tasks will bring the region feature into a more proper location in feature space rather than simply pulling them together via only detection task. As shown in Table 5, we further conduct the experiments to demonstrate the effectiveness of pre-training under captioning. By comparing the row 2 and 5, we observe that even with only dense captioning data (VG data), pre-training with the dense captioning paradigm also brings a significant improvement.

Model	Pre-Train Data	LVIS	
		AP	AP_r / AP_c / AP_f
DetCLIP-T [11]	O365	28.8	26.0 / 28.0 / 30.0
	VG	10.3	8.6 / 10.1 / 10.8
	O365, VG	31.5	27.5 / 30.6 / 33.0
CapDet	O365	28.5	25.2 / 27.5 / 29.9
	VG	11.4	10.2 / 11.1 / 11.8
	O365, VG	**33.8**	**29.6 / 32.8 / 35.5**

Table 5. Zero-shot performance on LVIS [4] MiniVal5k datasets. AP_r / AP_c / AP_f indicate the AP values for rare, common, and frequent categories, respectively. "DH" and "F" in GLIP [9] baselines stand for the dynamic head [2] and cross-modal fusion, respectively.

E. 'Real' Open-world Object Detection Deployment Strategy

In this paper, the detection and dense captioning task are illustrated separately for better understanding and comparison with other methods, since no benchmark has considered combining these two tasks. For the practical deployment, we propose a simple two-stage ensemble way to stay true to the motivation. Specifically, in the first stage, we execute detection on images among the pre-defined categories list and treat the proposals with maximum alignment scores among all classes less than a threshold as 'unknown' objects. Then in the second stage, we generate the captions for the 'unknown' objects. To demonstrate the effectiveness of the proposed strategies, We conduct detection on the images with 80 categories of COCO and regenerate captions for the 'unknown' objects. As shown in the Figure 1 , our proposed strategy expands the semantic space of the limited categories list and shows reasonable results.

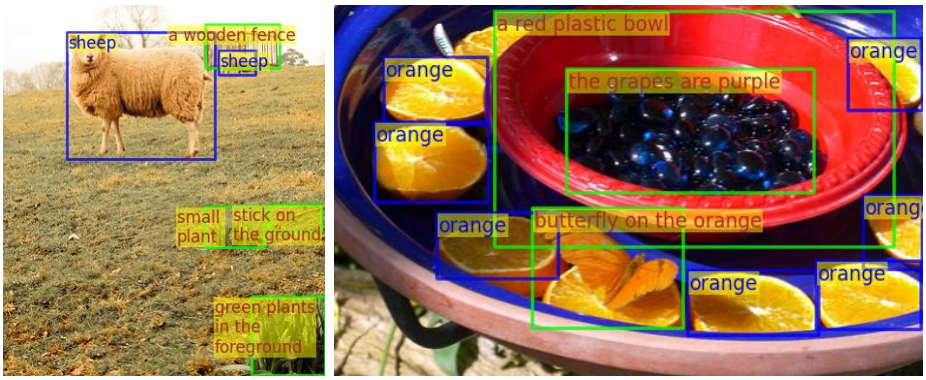


Figure 1. Deployment results.

F. More Ablation Studies

Ablations on Pre-trained Language Model Table 6 reports the effect of different tokenizers and pre-trained language models loaded for text encoder. We ablate two kinds of pre-trained language models and corresponding tokenizers for our text encoder. For dense captioning head, we construct the same decoder as BLIP [7] decoder and keep the tokenizer the same as the text encoder. The results indicate the FILIP [12] encoder with byte pair encoding performs a better generalization, since it is pre-trained on a larger scale of data, *i.e.*, 300M in FILIP [12] *vs.* 128M in BLIP [7].

Pre-trained Model	Tokenizer	Vocab Size	DC Head	LVIS	
				AP	AP_r / AP_c / AP_f
BLIP [7]	WordPiece	30524	✗	30.4	26.7 / 29.4 / 32.0
			✓	32.4	27.4 / 31.8 / 33.9
FILIP [12]	BPE	49408	✗	31.5	27.5 / 30.6 / 33.0
			✓	33.8	29.6 / 32.8 / 35.5

Table 6. Effect of different tokenizers and language models. 'DC Head' and 'BPE' stand for the integration of Dense Captioning Head and Byte Pair Encoding.

Ablations on the Weighting Factor of Dense Captioning Loss We study the effect of weights of detection loss and dense captioning loss during pre-training. We set the weighting factor of detection loss w_d to 1.0. Table 7 provides the ablations of the weighting factor of dense captioning loss w_c. We choose $w_c = 1$ for CapDet, since the result of overall AP is the best.

w_c	LVIS	
	AP	AP_r / AP_c / AP_f
0.5	33.6	31.0 / 32.8 / 34.9
1.0	**33.8**	29.6 / 32.8 / 35.5
1.5	33.5	32.0 / 32.1 / 35.0

Table 7. Effect of weighting factor of dense captioning loss.

G. More Qualitative Results

Open-World Detection Results Figure 2 illustrates more detection results on LVIS [4] dataset from our CapDet. We highlight the detected rare classes's text in red.

Dense Captioning Results Figure 3 shows more captioning results on VisualGenome [5] dataset. Our model CapDet locates not only "object" such as "bicycle" but also "region" such as "a shadow on the ground". We also explored the zero-

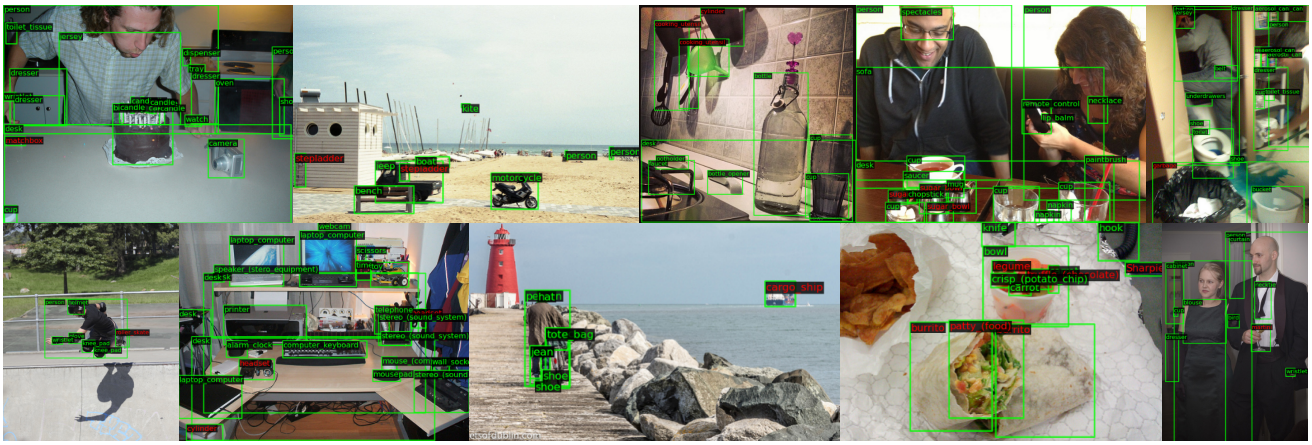


Figure 2. Qualitative visualizations on LVIS.

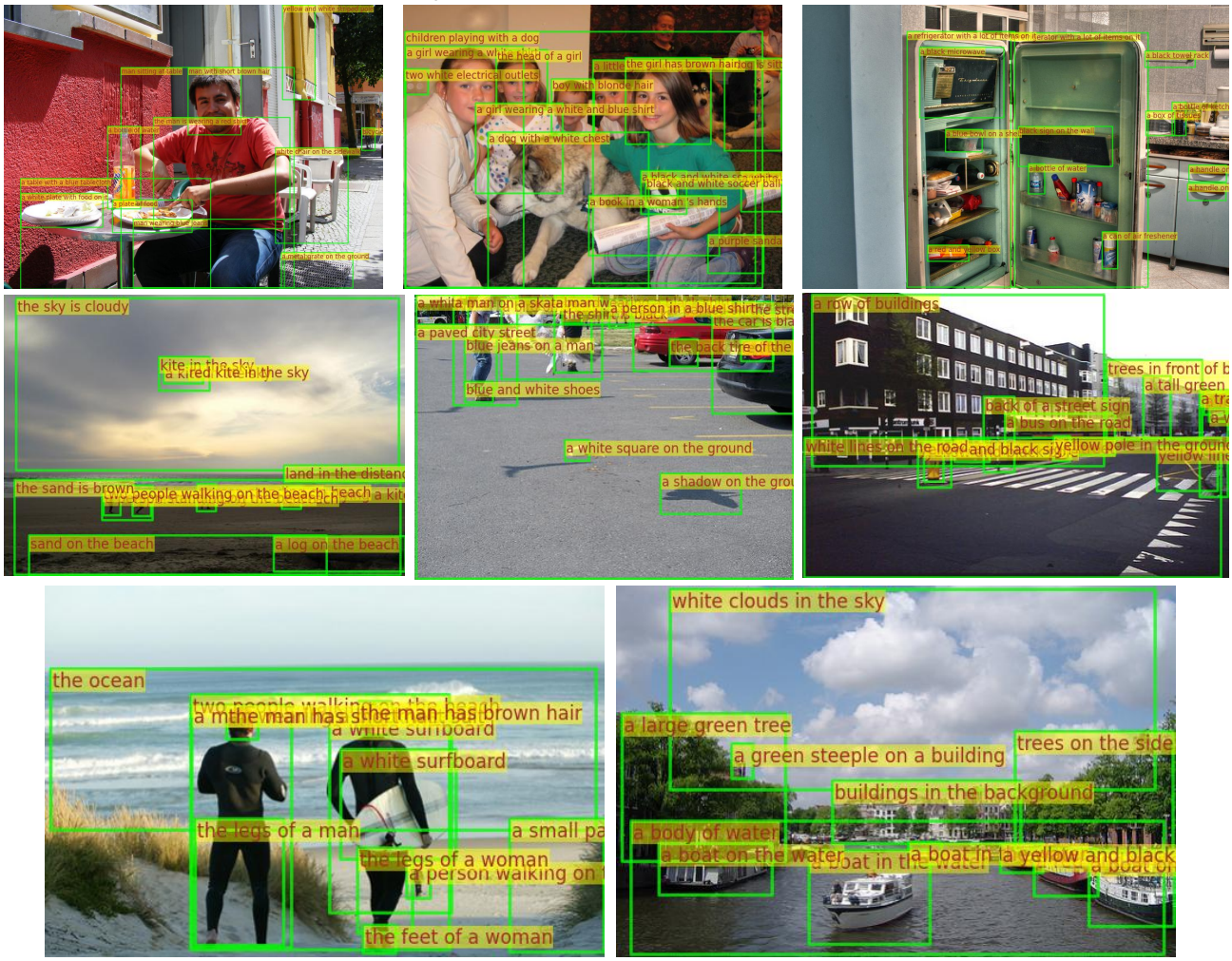
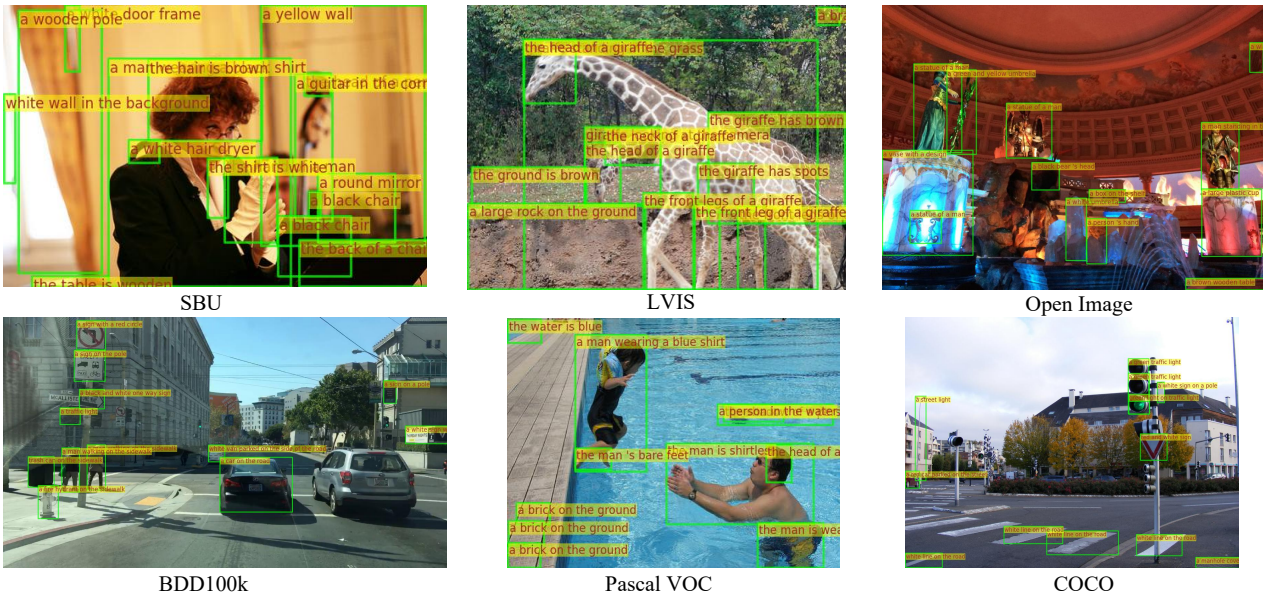


Figure 3. Qualitative visualizations on VG.

shot generalization ability of CapDet. We directly use our model to do the zero-shot dense captioning task without finetuning on serveral datasets, which include SBU [10], LVIS [4], Open Image [6], BDD100K [13], Pascal VOC [3] and COCO [1]. As shown in Figure 4, CapDet can accurately locate objects and generate corresponding region-grounded captions.



SBU LVIS Open Image

BDD100k Pascal VOC COCO

Figure 4. Qualitative visualizations on several datasets.

References

[1] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 4

[2] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7373–7382, 2021. 1, 2

[3] Mark Everingham, SM Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 4

[4] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 1, 2, 3, 4

[5] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 3

[6] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 4

[7] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 3

[8] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. *arXiv preprint arXiv:2112.03857*, 2021. 1, 2

[9] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2

[10] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. 4

[11] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *arXiv preprint arXiv:2209.09407*, 2022. 1, 2

[12] Lewei Yao, Runhu Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *ArXiv*, abs/2111.07783, 2022. 3

[13] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 4