

# Supplementary Materials: Sniffing Threatening Open-World Objects in Autonomous Driving by Open-Vocabulary Models

Anonymous Authors

## 1 SUPPLEMENTARY MATERIALS

We introduce the guidance of code implementation in Sec. 1.1, and our codes are available in the zip file. In Tab. 1, we provide the complete results of AD-oriented OWO benchmark using BDD100K. In Sec. 1.2, we conduct more discussions on the proposed U-ARecall and compare it with mAP. In Sec. 1.3, we visualize the results of coarse and fine vocabulary generation. Finally, we introduce the prospects and further works in Sec. 1.4.

### 1.1 Guidance of Code Implementation

We provide our codes in the zip file of the supplemental materials. The “auto-gd” folder contains the classic object detection methods (see “od\_configs”), fine-tuning methods (see “finetune” folder in configs/auto\_driving\_grounding\_dino), and our method (see configs/auto\_driving\_grounding\_dino). In the “owod” folder, it includes the classic OWO methods, *i.e.*, UnSniffer, VOS, OW-DETR, and PROB. The “tools” folder consists of some data conversion codes, some visualization codes, and some evaluation codes. In summary, we provide the implementation codes of our method and the proposed AD-oriented OWO benchmark, including the codes for all compared methods and the evaluation protocol (more details see the README.md).

### 1.2 More Discussions on ARecall

To evaluate the performance of detecting unknown objects, we propose a novel evaluation metric called U-ARecall, aiming to amplify the penalties associated with false positives. In this section, we detail the reasons for not employing AP (average precision) to evaluate the performance of detecting unknown objects. Generally speaking, AP is a default criterion to evaluate the detection accuracy. As shown in Fig. 1 (a), AP equals the area under precision-recall (PR) curve, representing a trade-off between precision and recall. However, due to the inclusion of precision, AP requires that there are almost no missing annotations, otherwise the accuracy of the assessment will be significantly compromised. Therefore, AP demands strict annotations with almost no missing object. This requirement is not suitable for open world object detection (OWOD), as unknown objects encompass diverse classes, increasing the likelihood of missing labels. While, as shown in Fig. 1 (b), we redisplay the calculation of ARecall as the area of recall and numerical ratio curve. Unlike AP, ARecall does not involve precision, it is not as strict as AP on annotations. However, ARecall can still evaluate accuracy by penalizing false positives within predictions.

### 1.3 More Visualization

We compare the experimental results of coarse and fine vocabulary generation in Tab. 4 of this paper. In this section, we offer additional visualizations for these two generation methods, as shown in Fig. 2. It is evident that the fine vocabulary generation yields more false

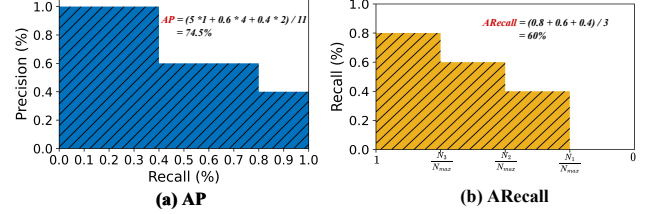


Figure 1: Examples of calculation process for (a) AP and (b) the proposed ARecall. For simplicity, we sample three values of  $N$  in ARecall for demonstration.

Table 1: Complete performance comparison using BDD100K for the AD-oriented OWO benchmark. U-R denotes U-Recall. FR and R-OVM mean FasterRCNN and Raw OVM.

Method	U-R <sup>10</sup>	U-R <sup>20</sup>	U-R <sup>30</sup>	U-ARecall	K-mAP	UK-Mean
FR[9]	0.	0.	0.	0.	53.3	26.6
D-DETR[11]	0.	0.	0.	0.	51.8	25.9
OW-DETR[3]	4.0	5.9	7.8	5.9	47.4	26.6
PROB[12]	4.0	10.1	15.4	9.8	48.4	29.1
VOS[2]	11.8	20.6	25.5	19.2	48.5	33.8
UnSniffer[7]	11.3	21.3	27.3	20.0	55.3	37.6
R-OVM[8]	55.2	71.5	78.8	<b>68.5</b>	42.2	55.3
FFT	34.5	40.4	43.6	39.5	60.3	49.9
LP	34.6	41.5	45.1	40.4	<b>60.7</b>	50.5
LP-FT[6]	30.0	37.0	39.9	35.6	60.6	48.1
Adapter[4]	37.0	44.7	48.8	43.5	60.1	51.8
LoRA[5]	44.9	61.5	68.8	58.4	56.4	<u>57.4</u>
<b>Ours</b>	54.9	71.3	78.8	<u>68.3</u>	60.5	<b>64.4</b>

positives, including some irrelevant objects (traffic signal and rumble strip) and misidentified objects (car). Especially, some cars are misidentified as various types of vehicle (hybrid vehicle and electric vehicle), which significantly undermines the performance of unknown object detection. These results suggest that finer unknown object classes do not necessarily lead to better performance.

### 1.4 Prospects and Future Works

**Prospects.** We adapt OWO to the auto-driving (AD) scenario and name this task as AD-oriented OWO, which is more practical than classic OWO. We further discuss the prospects of AD-oriented OWO in the research of AD. In AD, occupancy prediction (OP) [10] has recently been proposed to address unknown objects by constructing a 3D voxel space and requiring voxel-level labels. OP is a 3D space-centric task, which requires multiple cameras (typically 6) or additional radar sensors. On the other hand, AD-oriented OWO is a 2D semantic-centric task, which needs only one image from a camera and shows great potential for application to edge-side devices. For instance, recently, Yolo-World [1] has been proposed to

marry open-vocabulary models (OVMs) with YOLO, empowering OVMs real-time capabilities. Therefore, the computational cost of AD-oriented OWO is much lighter than that of OP. Besides, the labeling cost of AD-oriented OWOD is also much cheaper than that of OP. Nevertheless, the predicted 3D space of OP is more reliable than 2D bounding boxes of AD-oriented OWOD, as it provides additional depth information. In summary, we believe that AD-oriented OWOD and OP tasks are complementary, with the former being 2D semantic-centric and the latter being 3D space-centric. Together, their use can further enhance driving safety.

**Future Works.** In this work, we concretize the semantics of “threatening objects” by a general vocabulary bag. However, these semantics remain fixed and unlearnable. Therefore, it is worthwhile to explore the possibility of learning generic representations of threatening objects with limited annotations in future research.

## REFERENCES

- [1] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. 2024. YOLO-World: Real-Time Open-Vocabulary Object Detection. *arXiv preprint arXiv:2401.17270* (2024).
- [2] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. 2021. VOS: Learning What You Don’t Know by Virtual Outlier Synthesis. In *International Conference on Learning Representations*.
- [3] Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. 2022. Ow-detr: Open-world detection transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

- 9235–9244.
- [4] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*. PMLR, 2790–2799.
- [5] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.
- [6] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. 2021. Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. In *International Conference on Learning Representations*.
- [7] Wenteng Liang, Feng Xue, Yihao Liu, Guofeng Zhong, and Anlong Ming. 2023. Unknown Sniffer for Object Detection: Don’t Turn a Blind Eye to Unknown Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3230–3239.
- [8] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (2023).
- [9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).
- [10] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. 2024. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems* 36 (2024).
- [11] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2021. Deformable detr: Deformable transformers for end-to-end object detection. *Proceedings of the International Conference on Learning Representations (ICLR)* (2021).
- [12] Orr Zohar, Kuan-Chieh Wang, and Serena Yeung. 2023. Prob: Probabilistic objectness for open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11444–11453.

175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232

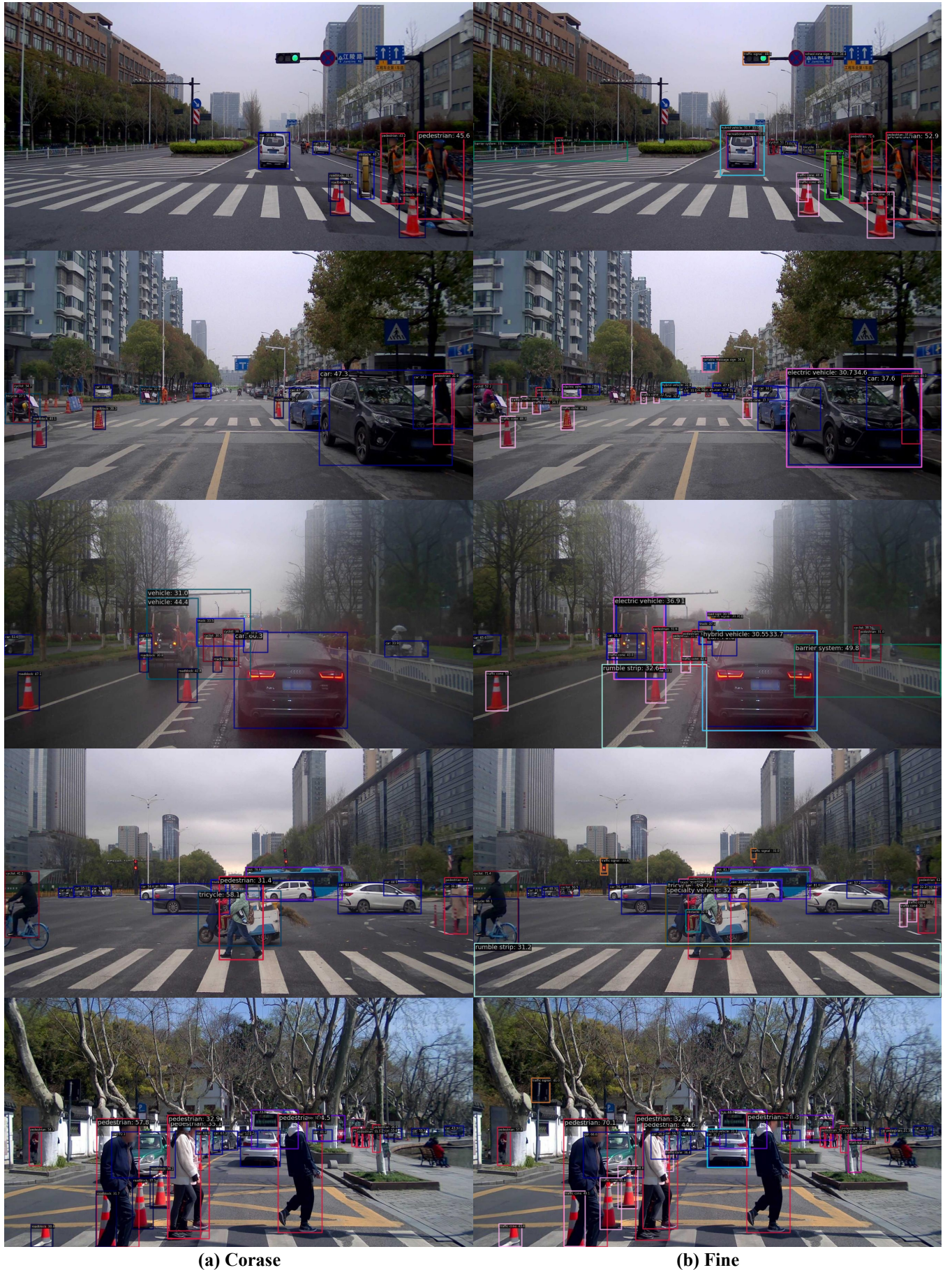


Figure 2: Visualization comparison of coarse and fine vocabulary generation for a vocabulary bag.