Hierarchical Cross-Modal Alignment for Open-Vocabulary 3D Object Detection Supplementary Material

Anonymous submission

Analysis of Generalization Ability on ScanNet200 Dataset

To verify the open-vocabulary capabilities of our method, we conduct evaluations on large-scale vocabularies using the ScanNet200 (Rozenberszki, Litany, and Dai 2022) dataset. The evaluation provides valuable insights into the robustness and scalability of our method, further validating its potential for practical applications in real-world 3D object detection scenarios.

Specifically, we train our HCMA on ScanNet and test it on the vocabularies in ScanNet200 without fine-tuning. Compared with our training vocabularies, ScanNet200 contains 53 overlapping vocabularies. This indicates that there is a small degree of vocabulary overlap between our training data and the ScanNet200 dataset. Hence, it can be leveraged to verify the open-vocabulary ability of HCMA.

We compare our HCMA with OV-3DET (Lu et al. 2023), as shown in Table 1. Our HCMA exhibits superior performance compared to OV-3DET (Lu et al. 2023), as measured by mAP_{25} and mAP_{50}. This remarkable performance demonstrates the robustness and effectiveness of HCMA in handling large-scale open vocabularies. This generalization ability is significant as it demonstrates the potential of our model in real-world applications.

Method	mAP_{25}	mAP_{50}
OV-3DET	2.39	0.84
HCMA (Ours)	**3.10**	**1.03**

Table 1: Result on ScanNet200 in terms of mAP_{25} and mAP_{50}.

More Implementation Details

In this section, We provide comparison of computational overhead with baseline methods OV-3DET (Lu et al. 2023) and CoDA (Cao et al. 2024) in Table 2. We conduct our experiment on a single RTX4090 GPU, while OV-3DET and CoDA require 8 GPUs for the experiments. In addition, our training time is short than the baseline methods. These indicate that HCMA is not only more efficient in terms of computational resources but also faster in terms of model training.

Method	Computational Resource	Training Time Phase 1	Training Time Phase 2
OV-3DET	8×2080Ti	48 hours	24 hours
CoDA	8×V100	2-3 days	1-2 days
HCMA(Ours)	1×4090	40 hours	20 hours

Table 2: Comparison of computational overhead.

More Quantitative Comparison

To compare our HCMA with the latest OV-3DOD method FM-OV3D (Zhang et al. 2024), we adapt HCMA following the experimental setting of FM-OV3D (Zhang et al. 2024). In this evaluation, we compare HCMA with OV-3DETIC (Lu et al. 2022) and FM-OV3D (Zhang et al. 2024) on ScanNet (Dai et al. 2017) and SUN RGB-D (Song, Lichtenberg, and Xiao 2015) datasets. FM-OV3D* denotes the model trained in an annotation-free setting, and FM-OV3D represents the model trained only utilizing knowledge blending. As Tables 3 and 4 show, our HCMA outperforms the latest OV-3DOD method on both ScanNet and SUN RGB-D datasets by a large margin, demonstrating the superior performance of HCMA.

Implementation of the Alignment Loss Function

Table 5 shows the implementation of the alignment loss function in Eq. 11. In this function, $\mathbb{1}$ is the indicator function that controls the specific form of $\mathcal{L}_{align}$, which is decided by different hierarchies utilized in the training stage.

Vocabulary Details

In our proposed HCMA, the use of vocabulary can be divided into two stages, including 1) pseudo-3D bounding box generation and cross-modal alignment in the training stage, 2) object detection in the testing stage. Specifically, we use categories from LVIS (Gupta, Dollar, and Girshick 2019) as training vocabularies in the training stage. The baseline

Method	mAP_{25}	toilet	bed	chair	sofa	dresser	table	cabinet	bookshelf	pillow	sink
OV-3DETIC (Lu et al. 2022)	12.65	48.99	2.63	7.27	18.64	2.77	14.34	2.35	4.54	3.93	21.08
FM-OV3D* (Zhang et al. 2024)	14.34	2.17	41.11	27.91	33.25	0.67	12.60	2.28	8.47	9.08	5.83
FM-OV3D (Zhang et al. 2024)	21.53	62.32	41.97	22.24	31.80	1.89	10.73	1.38	0.11	12.26	30.62
HCMA (Ours)	**31.63**	**72.88**	**50.64**	**37.28**	**56.83**	**3.11**	**15.30**	**3.10**	**11.79**	**20.91**	**44.49**

Table 3: Results on ScanNet in terms of AP_{25}. We report the average value of all 10 categories.

Method	mAP_{25}	toilet	bed	chair	bathtub	sofa	dresser	scanner	fridge	lamp	desk
OV-3DETIC (Lu et al. 2022)	13.03	43.97	6.17	0.89	45.75	2.26	**8.22**	0.02	8.32	0.07	**14.60**
FM-OV3D* (Zhang et al. 2024)	16.98	32.40	18.81	27.82	15.14	35.40	7.53	**1.95**	9.67	13.57	7.47
FM-OV3D (Zhang et al. 2024)	21.47	55.00	38.80	19.20	41.91	23.82	3.52	0.36	5.95	**17.40**	8.77
HCMA (Ours)	**32.35**	**68.81**	**72.87**	**41.63**	**49.90**	**43.18**	3.62	0.05	**16.71**	14.52	12.26

Table 4: Results on SUN RGB-D in terms of AP_{25}. We report the average value of all 10 categories.

HDI	$\mathcal{L}_m^o$	$\mathcal{L}_m^v$	$\mathcal{L}_m^s$	$\mathcal{L}_m^l$	$\mathcal{L}_m^g$	$\mathcal{L}_m^a$
O	✓					
O V		✓		✓		
O S			✓		✓	
O V S		✓	✓			✓

Table 5: Components of the alignment loss. O, V, and S denote object, view, and scene levels, respectively.

Testing Benchmark	Total	Overlap	Open
ScanNet	20	12	8
ScanNet200	200	53	147
SUN RGB-D	20	14	6

Table 6: Relationship between training and testing vocabularies.

methods use the same vocabulary set during training for the fair comparison. In the testing stage, we sample 20 common categories as testing vocabularies in the evaluation on the ScanNet (Dai et al. 2017) and SUN RGB-D (Song, Lichtenberg, and Xiao 2015) benchmarks, respectively. As for ScanNet200 (Rozenberszki, Litany, and Dai 2022), we follow the default vocabulary setting in this benchmark. The relationship of the vocabulary set between training and the testing phase is shown in Table 6. Open vocabulary refers to vocabularies used in the testing phase but not in the training phase, while overlapping vocabulary includes vocabularies that are both used in the training and testing phases.

Results on Open Vocabulary

To further verify the open-vocabulary ability of our HCMA, we conduct additional experiments involving 3D object detection on the ScanNet (Dai et al. 2017) and SUN RGB-D (Song, Lichtenberg, and Xiao 2015) datasets, focusing solely on open vocabularies. The number of open vocabularies used in this experiment is presented in Table 6. The evaluation results, shown in Table 7, indicate that HCMA outperforms the baseline method OV-3DET (Lu et al. 2023). It demonstrates the capabilities of HCMA on open-vocabulary 3D object detection. However, we observe that there is still a large gap in 3D object detection results between the open-vocabulary and seen-vocabulary. This observation demonstrates the significant potential for further development in 3D object detection involving open-vocabulary scenarios.

Method	ScanNet	SUN RGB-D
OV-3DET	7.80	6.29
HCMA (Ours)	**8.09**	**6.60**

Table 7: Results of open vocabularies on ScanNet and SUN RGB-D in terms of mAP_{25}.

Analysis of the Upper Bound Performance

Since we utilize 3DETR (Misra, Girdhar, and Joulin 2021) as our 3D detector, we use 3DETR as our upper bound. 3DETR (Misra, Girdhar, and Joulin 2021) is trained with ground truth 3D annotations in a fully supervised manner, while HCMA performs OV-3DOD without 3D annotations. Results are shown in Tables 8 and 9. It indicates that there is still a large gap between fully-supervised 3D detection and open-vocabulary 3D detection. While advancements have been made in the field of 2D open-vocabulary understanding, achieving comparable performance to fully-supervised methods remains a challenging task in the 3D domain.

More Qualitative Results

In this section, we provide more qualitative results on ScanNet (Dai et al. 2017) in Figure 1. It shows that our HCMA can correctly predict the location, size, and orientation of the

Method	mAP_{25}	toilet	bed	chair	sofa	dresser	table	cabinet	bookshelf	pillow	sink	bathtub	refrigerator	desk	night stand	counter	door	curtain	box	lamp	bag
3DETR	46.95	90.75	62.97	67.56	68.42	34.29	53.48	38.90	44.22	43.66	62.56	77.27	43.67	52.54	50.61	28.15	44.63	36.53	11.96	17.66	9.13
HCMA (Ours)	21.77	72.85	50.61	37.26	56.82	3.11	15.19	3.10	11.78	20.89	44.51	50.13	9.49	12.73	24.59	0.10	13.56	0.19	3.27	3.28	1.86

Table 8: Results on ScanNet in terms of AP_{25}. We report the average value of all 20 categories.

Method	mAP_{25}	toilet	bed	chair	bathtub	sofa	dresser	scanner	fridge	lamp	desk	table	stand	cabinet	counter	bin	bookshelf	pillow	microwave	sink	stool
3DETR	39.50	89.62	82.08	65.91	74.20	57.06	24.49	12.49	24.43	24.94	28.17	49.74	59.71	18.18	28.86	43.76	31.58	19.45	10.07	31.61	13.55
HCMA (ours)	21.53	68.50	72.81	40.59	49.65	43.20	3.32	0.03	17.38	14.34	11.73	28.61	19.48	0.56	0.12	11.33	1.51	10.34	1.34	31.56	4.10

Table 9: Results on SUN RGB-D in terms of AP_{25}. We report the average value of all 20 categories.

3D bounding box. In contrast, the baseline OV-3DET (Lu et al. 2023) may exhibit inaccuracies in predicting the size and number of 3D bounding boxes, such as the chair in sample a and sample b. The baseline OV-3DET may miss the object in prediction. For example, the chair in sample c is an example of a target object missed by the baseline method, while HCMA can detect the chair in the 3D scene. The observed discrepancy between HCMA and OV-3DET can be attributed to the differences in their ability to comprehend the overall structure of the target objects. OV-3DET may struggle to understand the holistic information of the objects, leading to the generation of small and redundant 3D bounding boxes. In contrast, HCMA leverages scene context from diverse hierarchies, enabling a better understanding of the comprehensive information regarding the 3D objects. This enhanced contextual understanding allows HCMA to generate more accurate and correct 3D bounding boxes.

Failure Cases

We observe that HCMA may sometimes miss the target object in the 3D scene. For example, both HCMA and OV-3DET cannot detect the curtain and table in sample d. The transparent glass table makes it very difficult to be detected since point clouds often struggle to accurately represent transparent objects. Additionally, detecting objects under open vocabularies, such as curtains, can also present difficulties. Open-vocabulary objects may not be sufficiently presented in the training data, resulting in limited recognition capabilities for these objects during testing.

Limitations

Our proposed HCMA demonstrates promising results in the OV-3DOD task. However, it still inherits the limitations of current open-vocabulary 3D methods, such as the requirement for pre-defined vocabularies. As a future work, we would like to explore designing an OV-3DOD method that does not depend on pre-defined vocabularies.

Broader Impacts

Our paper does not have direct societal impact. We train our method on public databases, with no private data used. While we do not foresee any negative societal impact from our paper, it may be leveraged in personal 3D data, resulting in leakage risks that raise privacy concerns. We urge readers to limit the usage of this work to legal use cases.

References

Cao, Y.; Yihan, Z.; Xu, H.; and Xu, D. 2024. Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection. *Advances in Neural Information Processing Systems*, 36.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5828–5839.

Gupta, A.; Dollar, P.; and Girshick, R. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5356–5364.

Lu, Y.; Xu, C.; Wei, X.; Xie, X.; Tomizuka, M.; Keutzer, K.; and Zhang, S. 2022. Open-vocabulary 3d detection via image-level class and debiased cross-modal contrastive learning. *arXiv preprint arXiv:2207.01987*.

Lu, Y.; Xu, C.; Wei, X.; Xie, X.; Tomizuka, M.; Keutzer, K.; and Zhang, S. 2023. Open-vocabulary point-cloud object detection without 3d annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1190–1199.

Misra, I.; Girdhar, R.; and Joulin, A. 2021. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2906–2917.

Rozenberszki, D.; Litany, O.; and Dai, A. 2022. Language-grounded indoor 3d semantic segmentation in the wild. In *European Conference on Computer Vision*, 125–141. Springer.

Song, S.; Lichtenberg, S. P.; and Xiao, J. 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 567–576.

Zhang, D.; Li, C.; Zhang, R.; Xie, S.; Xue, W.; Xie, X.; and Zhang, S. 2024. FM-OV3D: Foundation Model-Based

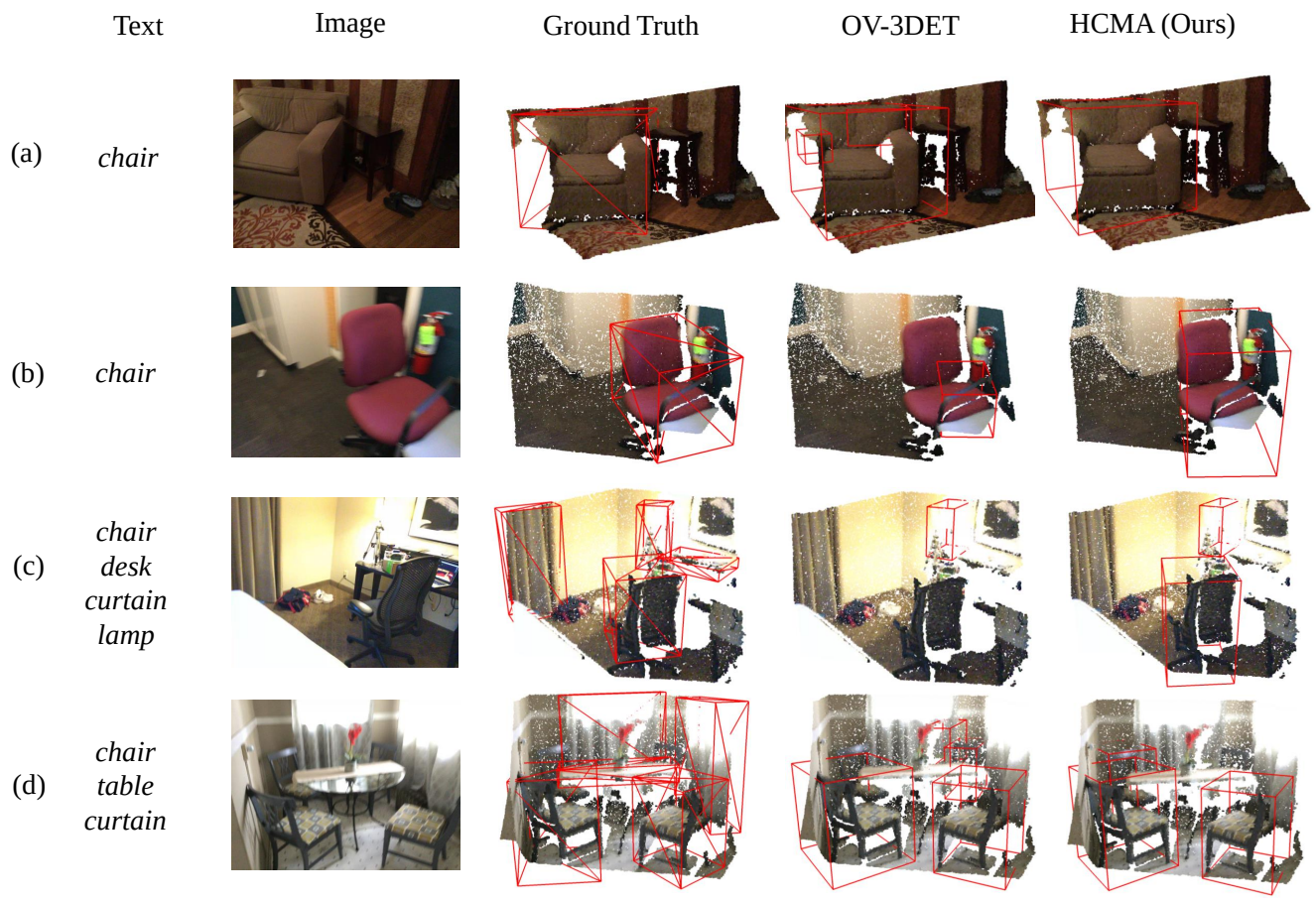

Figure 1: More qualitative comparisons with OV-3DET (Lu et al. 2023). For each case, the detection text prompts are shown on the left. Note that HCMA only utilize coordinate information of the 3D point cloud and do not include color information. For better visualization, the colors of the 3D point clouds are displayed.

Cross-Modal Knowledge Blending for Open-Vocabulary 3D Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16723–16731.