Few-Shot Incremental 3D Object Detection in Dynamic Indoor Environments

Supplementary Material

A Overview

In this supplementary material, we first provide a detailed description of our training details (§ B), including base training details (§ B.1) and incremental learning details (§ B.2). We then describe the detailed category split information of ScanNet V2 and SUN RGB-D (§ C). Next, we present additional experiments (§ D), including the performance of unknown object learning on fully incremental 3D object detection (§ D.1), the performance of our method when adopting different VLM-based models (§ D.2), more qualitative visualizations (§ D.3), and the results under alternative category split settings (§ D.4). Finally, we discuss the limitations and future work (§ E) of our proposed approach.

B Training Details

B.1 Base Training Details

In this section, we provide a detailed description of how our framework leverages unknown objects during the base class training stage. Although these objects do not have semantic labels and therefore cannot contribute to the classification loss, their location and feature representation provide valuable cues to enhance the model's generalization ability to novel categories. We incorporate unknown objects through three auxiliary supervisory signals: foreground supervision, feature supervision, and regression supervision.

Foreground Supervision. Given a pseudo box $\mathcal{B}_j$, we assign each point p_e a continuous foreground score $w_e \in [0, 1]$ with Sigmoid function if it falls inside this pseudo box. This target score $w_e = w_j^{\text{box}} w_{e,j}^{\text{point}}$ is determined by its spatial proximity to the box center and the internal feature consistency of the box, as described in the main paper. Let o_e denote the predicted objectness probability. We supervise the objectness branch exclusively on unknown-object regions using a combination of binary cross-entropy and Dice loss [6]:

$$\mathcal{L}_{\text{obj}} = \text{BCE}(o_e, w_e) + \text{Dice}(o_e, w_e). \tag{1}$$

This supervision is independent of semantic categories and therefore naturally supports incremental learning, enabling the model to acquire foreground awareness for unseen objects even without manual labels.

Feature Supervision. To further enhance semantic understanding of unknown objects, we introduce feature supervision for points inside pseudo boxes. As mentioned in the main paper, each unknown object has a pseudo 3D box $\mathcal{B}_j$ and an instance feature $\mathcal{F}_j^{2D}$. For all points falling inside $\mathcal{B}_j$, we enforce cosine directional alignment:

$$\mathcal{L}_{\text{feat}} = \frac{1}{Z_j} \sum_{p_e \in \mathcal{B}_j} \left(1 - \cos(\hat{f_e^{2D}}, f_j^{2D})\right) w_e, \tag{2}$$

where w_e is the soft weight / target score and cos is the cosine similarity, which measures the directional alignment between two feature vectors. This loss encourages the internal points of a pseudo box to form consistent semantic embeddings, allowing the model to inherit semantic priors for novel classes during the base training.

Regression Supervision. Beyond the above parts, we also guide the model to learn the geometric structure of unknown objects. Since the regression branch does not involve any semantic information, we directly share it with the base class regression head, enabling the geometric priors learned by the network to naturally generalize to unknown categories. For point p_e inside pseudo box $\mathcal{B}_j$, let $\hat{r}_e$ denote the predicted bounding box and r_e^* the geometric parameters of the pseudo 3D box. We apply a soft-weighted regression loss with DIOU loss [11]:

$$\mathcal{L}_{\text{reg}}^{\text{unk}} = \frac{1}{Z_j} \sum_{p_e \in \mathcal{B}_j} \left(1 - \text{DIOU}(\hat{r}_e, r_j^*)\right) w_e. \tag{3}$$

This formulation provides the model with localization cues for unknown objects during base training, enabling it to acquire spatial awareness of novel categories.

Overall. By combining the above objectives, we obtain the total auxiliary loss associated with unknown objects:

$$\mathcal{L}_{\text{aux}} = \mathcal{L}_{\text{obj}} + \mathcal{L}_{\text{feat}} + \mathcal{L}_{\text{reg}}^{\text{unk}}. \tag{4}$$

B.2 Incremental Learning Details

As introduced in the main paper, during the incremental learning stage, the fusion weights $\alpha^{\text{3D}} \in \mathbb{R}^{N\times 1}$ and $\alpha^{\text{2D}} \in \mathbb{R}^{N\times 1}$ control the modality-specific contributions of the 3D and 2D branches, while $\gamma \in \mathbb{R}^{N\times C_{\text{novel}}}$ re-balances per-class responses to mitigate overconfident predictions from other categories. To update these weighting parameters, we introduce an incremental loss $\mathcal{L}_{\text{inc}}$. Let s_e^{fuse} denote the fused prediction after applying the modality- and class-wise weights, and let y_e be the one-hot target for the novel categories. The incremental loss is defined using a simple positive-negative supervision:

$$\mathcal{L}_{\text{inc}} = (1 - s_e^{\text{fuse}})\, y_e + s_e^{\text{fuse}}(1 - y_e). \tag{5}$$

This supervision enables the model to gradually form discriminative boundaries for novel categories.

C Dataset Split Details

In this section, we provide more category split information for ScanNet V2 [3] and SUN RGB-D [8] mentioned in the main paper, including both batch-incremental and sequence-incremental settings.

ScanNet V2 [3] contains 1,201 training samples and 312 validation samples, annotated with 18 object categories: *bathtub, bed, bookshelf, cabinet, chair, counter, curtain, desk, door, garbagebin, picture, refrigerator, showercurtain, sink, sofa, table, toilet, and window* in alphabetical order. For batch incremental settings, we design four few-shot incremental detection settings to evaluate the model's generalization ability:

- **1-way 1-shot:** 17 base classes (*bathtub–toilet*), 1 novel class (*window*) with 1 labeled sample.
- **1-way 5-shot:** Same as above, but with 5 labeled samples for the novel class.
- **9-way 1-shot:** 9 base classes (*bathtub–door*), 9 novel classes (*garbagebin–window*) with 1 labeled sample per novel class.
- **9-way 5-shot:** Same as above, but with 5 labeled samples per novel class.

For the sequential incremental setting, we initialize the model with 9 base classes (*bathtub–door*) and introduce 3 novel classes at each incremental step, each with 5 labeled samples per novel class, resulting in three tasks, namely:

- **Task 1**: *garbagebin, picture, refrigerator*.
- **Task 2**: *showercurtain, sink, sofa*.
- **Task 3**: *table, toilet, window*.

SUN RGB-D [8] consists of 5,285 training samples and 5,050 validation samples, annotated with 10 object categories: *bathtub, bed, bookshelf, chair, desk, dresser, night_stand, sofa, table, and toilet* in alphabetical order. Similar to ScanNet V2, we design four few-shot batch incremental detection settings to evaluate the model:

- **1-way 1-shot:** 9 base classes (*bathtub–table*), 1 novel class (*toilet*) with 1 labeled sample.
- **1-way 5-shot:** Same as above, but with 5 labeled samples for the novel class.
- **5-way 1-shot:** 5 base classes (*bathtub–desk*), 5 novel classes (*dresser–toilet*) with 1 labeled sample per novel class.
- **5-way 5-shot:** Same as above, but with 5 labeled samples per novel class.

For the sequential incremental setting, we initialize the model with 5 base classes (*bathtub–desk*) and introduce 3 novel classes an 2 novel classes sequentially, each with 5 labeled samples per novel class, resulting in two tasks:

- **Task 1**: *dresser, night_stand, sofa*.
- **Task 2**: *table, toilet*.

Table 1. Batch incremental 3D object detection results with fully labeled novel objects on ScanNet V2. All models are based on the TR3D baseline. Results are reported under 1-way and 9-way configurations. "Base" denotes base classes, "Novel" denotes novel classes, and "All" indicates the overall mean AP@0.25.

	1-way			9-way		
Method	Base	Novel	All	Base	Novel	All
Baseline	71.47	-	-	72.77	-	-
Ours* [9]	72.71	55.52	71.75	74.61	69.63	72.12
Ours*+UOL	**74.26**	**59.76**	**73.45**	**75.40**	**71.91**	**73.66**

D More Experiments

D.1 Results on Fully Incremental Settings

We extend our VLM-guided Unknown object learning (denoted as UOL) to the fully incremental setting, where the model has access to all annotated novel categories during the incremental stage. To ensure a clean evaluation, we implement a simplified version based on TR3D that retains only the pseudo labeling mechanism (denoted as Ours*) to highlight the contribution of our base training strategy.

As shown in Tab. 1, when integrating our proposed VLM-guided base training strategy (denoted as Ours* + UOL) into Ours*, the model achieves consistent and stable improvements across the Base, Novel, and All metrics under both the 1-way and 9-way incremental configurations. In the 1-way setting, the Novel mAP increases from 55.52 to 59.76, and the All mAP improves from 71.75 to 73.45. In the 9-way setting, the Novel mAP similarly rises from 69.63 to 71.91, and the All mAP improves from 72.12 to 73.66. These results demonstrate the generalization ability of our base training strategy in fully incremental scenarios.

D.2 Results based on Other VLMs

In this experiment, we replace the VLM backbone to further validate the generalization of our framework shown in Tab. 2. Specifically, we substitute GroundingDINO [5] with YOLO-World [1], which also supports open-vocabulary detection but requires explicit category prompts to perform inference. To ensure a comparison, we provide YOLO-World with a comprehensive prompt set containing 50 common indoor categories, including:

"chair", "table", "sofa", "bed", "desk", "cabinet", "shelf", "lamp", "door", "window", "television", "refrigerator", "washing machine", "microwave", "fan", "air conditioner", "sink", "toilet", "bathtub", "shower", "mirror", "carpet", "pillow", "blanket", "curtain", "picture", "vase", "clock", "books", "laptop", "keyboard", "mouse", "monitor", "printer", "trash bin", "cup", "plate", "bottle", "kettle", "knife", "wardrobe",

"shoe", "bag", "clothes", "towel", "plant", "cushion", "stool", "nightstand", and "drawer".

Moreover, to preserve the incremental learning protocol, no category information produced by the VLM is retained during base training. As shown in Tab. 2, both VLM-based variants outperform the baseline by a large margin. Among them, GroundingDINO achieves the best overall performance on both ScanNet V2 and SUN RGB-D, particularly in novel class detection (*e.g.*, +3.85 and +2.84 mAP over YOLO-World under 9-way and 5-way 5-shot settings, respectively). These findings demonstrate that our proposed framework can flexibly integrate different VLMs while maintaining strong incremental detection capability.

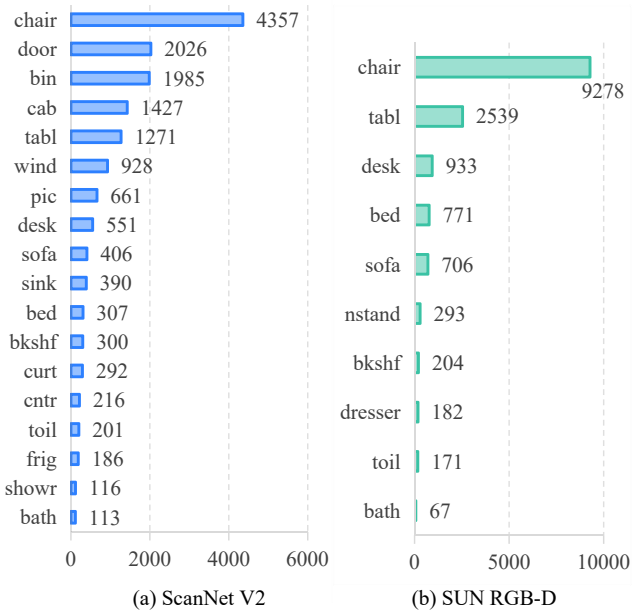
D.3 More Quantitative Results

In this section, we provide additional quantitative results on ScanNet V2 [3] and SUN RGB-D [8]. The predicted bounding boxes on these two datasets are shown in Fig. 2 and Fig. 3. In the qualitative results, the red dashed circles highlight novel object categories, including "*sofa*", "*refrigerator*", and "*window*" in ScanNet V2, as well as "*table*" and "*dresser*" in SUN RGB-D. Compared with Baseline and VLM-vanilla, which often miss or inaccurately localize these novel categories, our FI3Det produces more accurate and stable detections that closely match the ground truth, demonstrating stronger generalization to novel classes under few-shot incremental settings.

D.4 Results on Alternative Category Splits

In the main paper, we follow the setting of [10], where novel categories for few-shot incremental 3D object detection are selected based on alphabetical order. In this section, we further explore alternative category split strategies to verify the robustness of our method. As shown in Fig. 1, both ScanNet v2 and SUN RGB-D datasets contain a large number of object instances, but their category distributions are highly imbalanced, exhibiting a clear long-tailed property. Rather than randomly selecting novel categories, which can bias the evaluation, we divide the base and novel categories based on the number of instances per class. We design four incremental detection settings to evaluate the model's generalization ability based on ScanNet V2 [3] and SUN RGB-D [8]:

- **9-way 1-shot:** 9 base classes (*chair–sofa*), 9 novel classes (*sink–bathtub*) with 1 labeled sample per novel class on ScanNet V2.
- **9-way 5-shot:** Same as above, but with 5 labeled samples per novel class on ScanNet V2.
- **5-way 1-shot:** 5 base classes (*chair–sofa*), 5 novel classes (*night_stand–bathtub*) with 1 labeled sample per novel class on SUN RGB-D.



(a) ScanNet V2 (b) SUN RGB-D

Figure 1. Statistical analysis of the number of instances for each category in ScanNet V2 and SUN RGB-D.

- **5-way 5-shot:** Same as above, but with 5 labeled samples per novel class on SUN RGB-D.

Tab. 3 presents the results on ScanNet V2 and SUN RGB-D. For ScanNet V2, results are reported under 9-way 1-shot and 9-way 5-shot configurations, while for SUN RGB-D, results are provided under 5-way 1-shot and 5-way 5-shot configurations. As shown in the table, our proposed FI3Det achieves consistently superior performance across all configurations. In particular, it significantly improves detection on novel categories while maintaining strong performance on base categories.

E Limitations

This work leverages vision-language models (VLMs) to learn general semantic representations during the base class training stage, enabling the detector to achieve adaptability when encountering novel objects. Although we assume that the robot has a basic exploration of the environment before task switching, this setting is reasonable in most indoor scenarios (*e.g.*, homes or offices) but may present limitations in more complex or dynamic environments.

In the future, we plan to enhance the robustness of the network through improved architectural designs, enabling more stable learning in real-world embodied perception tasks. Moreover, although our method is capable of handling indoor environments, large-scale outdoor autonomous driving scenarios remain a relatively underexplored domain, which we plan to investigate further in our future work.

Table 2. Batch few-shot incremental 3D object detection results with different Vision-Language Models (VLMs) on ScanNet V2 and SUN RGB-D. Results are reported under 9-way/5-way and 1-shot/5-shot configurations. "Base" denotes base classes, "Novel" denotes novel classes, and "All" indicates the overall mean AP@0.25.

Method	ScanNet V2						SUN RGB-D					
	9-way 1-shot			9-way 5-shot			5-way 1-shot			5-way 5-shot		
	Base	Novel	All	Base	Novel	All	Base	Novel	All	Base	Novel	All
Baseline	**72.77**	6.52	39.64	**72.77**	7.10	39.94	61.58	4.70	33.14	61.58	4.32	32.95
YOLO-World [1]	72.43	28.09	50.76	72.44	26.38	49.91	61.77	16.24	39.01	61.77	23.97	42.87
GroundDINO (ours) [5]	72.27	**30.81**	**51.54**	72.28	**30.23**	**51.26**	**62.49**	**15.27**	**38.88**	**62.49**	**26.81**	**44.65**

Table 3. Batch few-shot incremental 3D object detection performance on ScanNet V2 [3] and SUN RGB-D [8]. Results are reported under 9-way/5-way and 1-shot/5-shot configurations. "Base" denotes base classes, "Novel" denotes novel classes, and "All" indicates the overall mean AP@0.25. The base and novel categories are divided according to the number of object instances in each class.

Methods	ScanNet V2						SUN RGB-D					
	9-way 1-shot			9-way 5-shot			5-way 1-shot			5-way 5-shot		
	Base	Novel	All	Base	Novel	All	Base	Novel	All	Base	Novel	All
Imprinting [7]	66.84	4.33	38.90	66.84	10.94	38.90	66.69	0.86	33.77	66.69	0.62	31.10
IL-DETR [4]	61.36	6.50	33.93	57.64	20.12	38.88	65.65	0.09	32.87	63.36	0.19	31.77
SDCOT++ [10]	48.26	5.38	23.82	27.38	19.41	23.40	60.86	0.02	30.44	51.51	0.06	25.78
AIC3DOD [2]	66.82	5.14	35.98	66.97	10.09	38.53	66.72	0.05	33.39	66.51	0.07	33.29
VLM-vanilla	67.58	9.06	44.61	67.56	21.65	44.61	67.19	4.18	35.67	67.19	3.75	35.47
FI3Det (ours)	**67.59**	**24.18**	**45.89**	**67.63**	**38.63**	**53.10**	**68.15**	**8.92**	**38.54**	**68.16**	**20.46**	**44.31**

References

[1] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. YOLO-World: Real-Time Open-Vocabulary Object Detection. In *CVPR*, 2024. 2, 4

[2] Zhongyao Cheng, Fang Wu, Peisheng Qian, Ziyuan Zhao, and Xulei Yang. AIC3DOD: Advancing Indoor Class-Incremental 3D Object Detection with Point Transformer Architecture and Room Layout Constraints. In *WACV*, 2025. 4

[3] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *CVPR*, 2017. 2, 3, 4

[4] Na Dong, Yongqiang Zhang, Mingli Ding, and Gim Hee Lee. Incremental-DETR: Incremental Few-Shot Object Detection via Self-Supervised Learning. In *AAAI*, 2023. 4

[5] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. In *ECCV*, 2024. 2, 4

[6] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for columetric medical image segmentation. In *3DV*, 2016. 1

[7] Hang Qi, Matthew Brown, and David G Lowe. Low-shot Learning with Imprinted Weights. In *CVPR*, 2018. 4

[8] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. In *CVPR*, 2015. 2, 3, 4, 5

[9] Na Zhao and Gim Hee Lee. Static-Dynamic Co-teaching for Class-Incremental 3D Object Detection. In *AAAI*, 2022. 2

[10] Na Zhao, Peisheng Qian, Fang Wu, Xun Xu, Xulei Yang, and Gim Hee Lee. SDCoT++: Improved Static-Dynamic Co-Teaching for Class-Incremental 3D Object Detection. *IEEE Transactions on Image Processing*, 2025. 3, 4

[11] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-IoU loss: Faster and Better Learning for Bounding Box Regression. In *AAAI*, 2020. 1

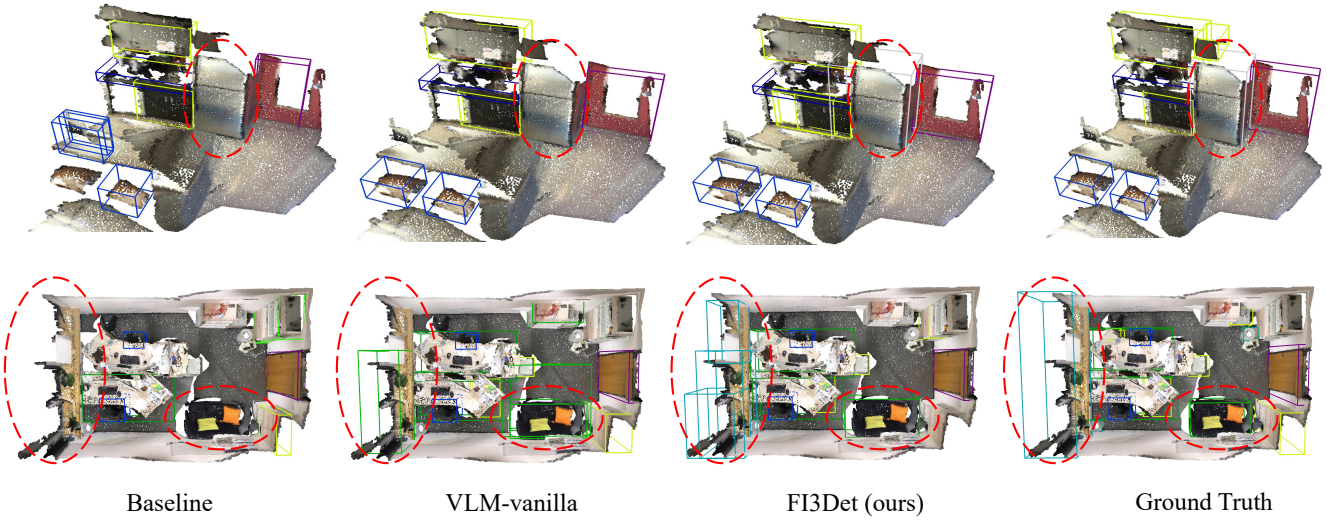


Figure 2. Qualitative comparison on the ScanNet V2 [8]. The red dashed circles highlight novel object categories "*sofa*", "*refrigerator*", and "*window*".

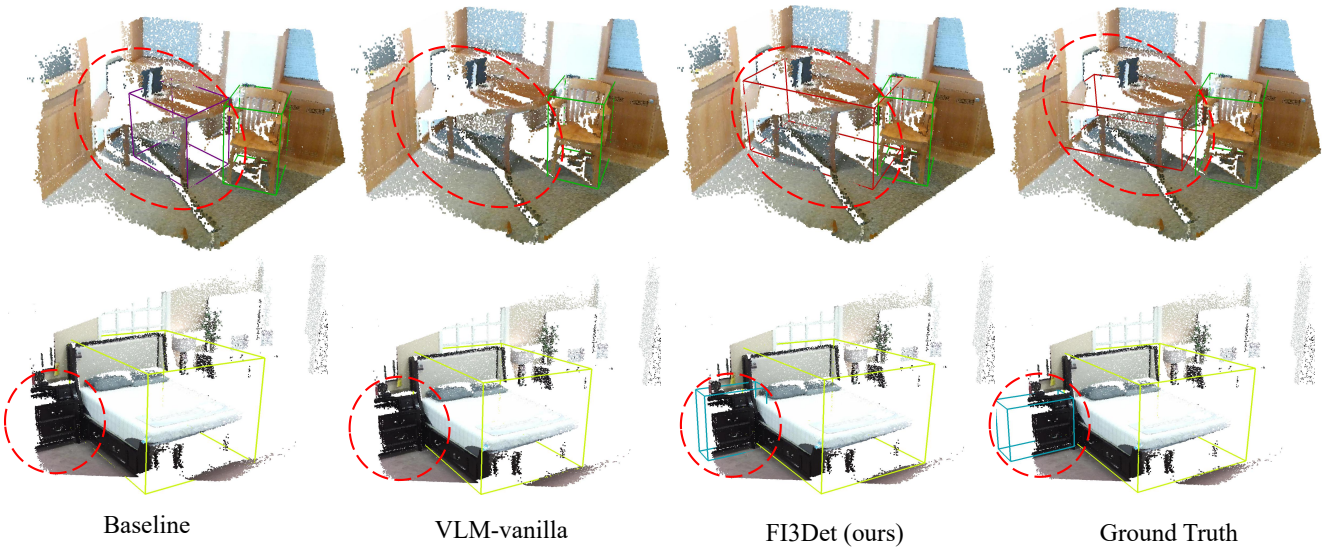


Figure 3. Qualitative comparison on the SUN RGB-D [8]. The red dashed circles highlight novel object categories"*table*" and "*dresser*".