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

Youjun Zhao*, Jiaying Lin*, Rynson W.H. Lau†

Department of Computer Science, City University of Hong Kong
youjun.zhao@my.cityu.edu.hk, jiayinlin5-c@my.cityu.edu.hk, Rynson.Lau@cityu.edu.hk

Abstract

Open-vocabulary 3D object detection (OV-3DOD) aims at localizing and classifying novel objects beyond closed sets. The recent success of vision-language models (VLMs) has demonstrated their remarkable capabilities to understand open vocabularies. Existing works that leverage VLMs for 3D object detection (3DOD) generally resort to representations that lose the rich scene context required for 3D perception. To address this problem, we propose in this paper a hierarchical framework, named **HCMA**, to simultaneously learn local object and global scene information for OV-3DOD. Specifically, we first design a Hierarchical Data Integration (HDI) approach to obtain coarse-to-fine 3D-image-text data, which is fed into a VLM to extract object-centric knowledge. To facilitate the association of feature hierarchies, we then propose an Interactive Cross-Modal Alignment (ICMA) strategy to establish effective intra-level and inter-level feature connections. To better align features across different levels, we further propose an Object-Focusing Context Adjustment (OFCA) module to refine multi-level features by emphasizing object-related features. Extensive experiments demonstrate that the proposed method outperforms SOTA methods on the existing OV-3DOD benchmarks. It also achieves promising OV-3DOD results even without any 3D annotations.

Code and Extended version —
https://youjunzhao.github.io/HCMA/

Introduction

3D object detection (3DOD) aims at localizing and classifying objects in 3D scenes. As a fundamental task in 3D scene understanding, it has various potential applications, *e.g.*, robotic (Zeng et al. 2018) and autonomous driving (Bojarski et al. 2016). However, existing 3DOD methods are limited to a small set of categories due to the lack of annotated 3D datasets. Most 3DOD works focus on detecting seen categories that exist in the training data and cannot be easily generalized to real-world data with unseen categories. To address this limitation, open-vocabulary 3D object detection (OV-3DOD) aims to detect novel objects in 3D scenes that are outside the limited categories of the training dataset.

Recently, vision-language models (VLMs) such as CLIP (Radford et al. 2021a) and ALIGN (Jia et al. 2021) have made tremendous progress by training on large-scale image-text data from the Internet. The ability of VLMs to learn the rich object-level context of image-text embedding has inspired open-vocabulary 2D understanding tasks, such as object detection (Gu et al. 2022; Ma et al. 2022; Bangalath et al. 2022; Wang et al. 2023) and semantic segmentation (Ding et al. 2022; Zhou, Loy, and Dai 2022; Xu et al. 2023; Han et al. 2023). However, it is not straightforward to pre-train 3D-text models due to the lack of large-scale 3D-text data. Recent attempts (Ding et al. 2023; Chen et al. 2023b; Peng et al. 2023) leverage rich linguistic representations of VLMs in 3D open-vocabulary understanding tasks by directly projecting 3D data onto the 2D image space. This motivates us to leverage VLMs for the OV-3DOD task. By doing so, we can utilize the rich object-centric context of linguistic descriptions in the 3D domain, allowing us to connect 3D point clouds with text representation for OV-3DOD.

However, directly applying 2D VLMs to address the 3DOD task is not feasible due to the significant contextual differences between 2D and 3D data. First, pre-trained 2D VLMs (Chen et al. 2023a) are typically trained on images containing mostly a single object. In contrast, 3D scenes often comprise multiple objects of various sizes. Connecting multiple objects with textual descriptions in a 3D scene is more challenging. Second, 3D scenes have more data dimensions than 2D images. This discrepancy may lead to the negligence of important spatial information associated with 3D objects. Third, unlike the 2D images used for pre-training VLMs, in which objects are usually located at the center, objects in a 3D scene may be located in various corners or even partially outside the 3D scene. This variability of 3D scenes makes it difficult to learn the model. The above three challenges pose a significant need for a holistic understanding of 3D scenes in the 3DOD task.

For effective 3D detection, our insight is that it is essential to consider not only the object features, but also the surrounding 3D scene context, in order to accurately locate the object. Hence, in this paper, we propose a hierarchical framework, named **HCMA**, to simultaneously learn 3D objects as well as scene contexts for the OV-3DOD task. Our HCMA framework has three technical contributions. First, unlike the state-of-the-art OV-3DOD method (Lu et al.

* These authors contributed equally.
† Corresponding author.

2023; Cao et al. 2024), which focuses solely on object-level contexts and aligns cross-modal features at the object level, we propose a novel Hierarchical Data Integration (HDI) approach to construct multi-level data, capturing hierarchical contexts in 3D scenes that incorporate object-level, view-level, and scene-level supervisions. By introducing the global scene representation, the multi-level data enables a comprehensive understanding of 3D point cloud scenes. Second, we design an Interactive Cross-Modal Alignment (ICMA) strategy to establish connections between the point cloud, image, and text representations in different hierarchies. ICMA contains two alignment approaches: intra-level cross-modal alignment and inter-level cross-modal alignment. The intra-level cross-modal alignment facilitates the integration of semantically related information within each hierarchy, while the inter-level cross-modal alignment compensates for the interaction between diverse hierarchies, enabling a coarse-to-fine understanding of the 3D scene. To further strengthen these connections across modalities, we employ a novel contrastive learning method to minimize the discrepancy among the representations derived from features in different modalities. Third, we propose an Object-Focusing Context Adjustment (OFCA) module to achieve accurate and object-centric feature alignment across different hierarchies. Instead of directly utilizing the features obtained from pre-trained VLMs (Lu et al. 2023; Zhang et al. 2024; Cao et al. 2024), our OFCA takes an object-centric approach, leveraging object-related features and refining cross-level information to enable seamless context integration for objects across various levels and modalities. This helps alleviate the semantic bias inherited from VLMs and improves the robustness of the hierarchical features.

In summary, the main contributions of this paper include:

- We propose **HCMA**, a hierarchical OV-3DOD framework to simultaneously learn object and 3D scene contexts to facilitate accurate 3D object detection.
- We propose a Hierarchical Data Integration (**HDI**) approach to obtain coarse-to-fine cross-modal data for multi-level supervision. It provides a comprehensive understanding of contexts in 3D scenes.
- We design an effective Interactive Cross-Modal Alignment (**ICMA**) strategy to facilitate a coarse-to-fine feature interaction across different levels and modalities. To improve hierarchical feature alignments in ICMA, we further propose an Object-Focusing Context Adjustment (**OFCA**) module to facilitate the effective integration of contexts across different levels.
- We conduct extensive experiments on the OV-3DOD benchmarks. Our method achieves superior performances compared to existing state-of-the-art approaches, demonstrating its effectiveness for the OV-3DOD task.

Related Work

3D Object Detection. This task is to accurately locate all objects of interest in a given 3D scene. VoteNet (Qi et al. 2019) first encodes point clouds into object volumetric representations and then connects to an RPN to generate detection results. Following (Qi et al. 2019), H3DNet (Zhang et al. 2020) integrates a set of geometric primitives to detect 3D objects. Recently, the Transformer (Vaswani et al. 2017) is also found to be suitable for 3D point cloud modeling since it is permutation-invariant and can capture long-range context. 3DETR (Misra, Girdhar, and Joulin 2021) proposes an end-to-end Transformer model for 3D point cloud object detection. Unlike these works, which only utilize closed-set data to perform 3DOD, our work focuses on the open-vocabulary setting and proposes a hierarchical framework to localize and classify novel 3D objects.

Open-Vocabulary 3D Scene Understanding. Open-vocabulary 3D scene understanding includes tasks such as open-vocabulary 3D segmentation and open-vocabulary 3D object detection (OV-3DOD). Compared with open-vocabulary 2D scene understanding (Cheng et al. 2024; Yang et al. 2024), research on open-vocabulary 3D scene understanding is limited, as 3D datasets are much more labor-intensive to create and therefore much smaller in scale. PLA (Ding et al. 2023) associates 3D point clouds with text by generating text descriptions from captioning multi-view 3D scene images. It can perform open-vocabulary 3D segmentation in a 2D manner with the aligned 3D point cloud and text features. CLIP2Scene (Chen et al. 2023b) utilizes the 2D vision-language model CLIP (Radford et al. 2021a) for 3D scene understanding. It leverages cross-modal knowledge from CLIP to extract semantic text features to achieve annotation-free 3D semantic segmentation. Similar to CLIP2Scene, OpenScene (Peng et al. 2023) also infers CLIP features to obtain 3D-text co-embedding for open-vocabulary semantic segmentation.

Open-vocabulary 3D object detection (OV-3DOD) is another task under 3D scene understanding. OV-3DET (Lu et al. 2023) is one of the first works to extend the open-vocabulary object detection task into the 3D area. It leverages rich object context extracted from a 2D detector pre-trained on large-scale image datasets and CLIP (Radford et al. 2021a) to detect novel 3D objects. CoDA (Cao et al. 2024) is a closer work to OV-3DET (Lu et al. 2023), but it does not rely on an additional 2D detector to help localize the 3D bounding box. Instead, it can simultaneously perform 3D object localization and classification. More recently, FM-OV3D (Zhang et al. 2024) tackles the OV-3DOD problem by blending knowledge from multiple pre-trained foundation models. Our work is closer to (Lu et al. 2023), but we do not rely solely on the object-centered features of point clouds for object detection. Our work can simultaneously learn hierarchical 3D scene contexts to align different modalities for discovering unseen 3D object patterns.

Method

Our method, named HCMA, first learns to localize 3D objects from a pre-trained 2D object detector. To do this, we generate 2D bounding boxes B_{2D} of the view-level paired images I^v from the pre-trained 2D detector. Prompted by n training classes $C = \{C_1, C_2, ..., C_n\}$, the pre-trained 2D detector can generate 2D bounding boxes $B_{2D} \in \mathbb{R}^4$ of the training classes C on the view-level paired images I^v as:

$$B_{2D} = \text{2DDetector}(I^v, C). \quad (1)$$

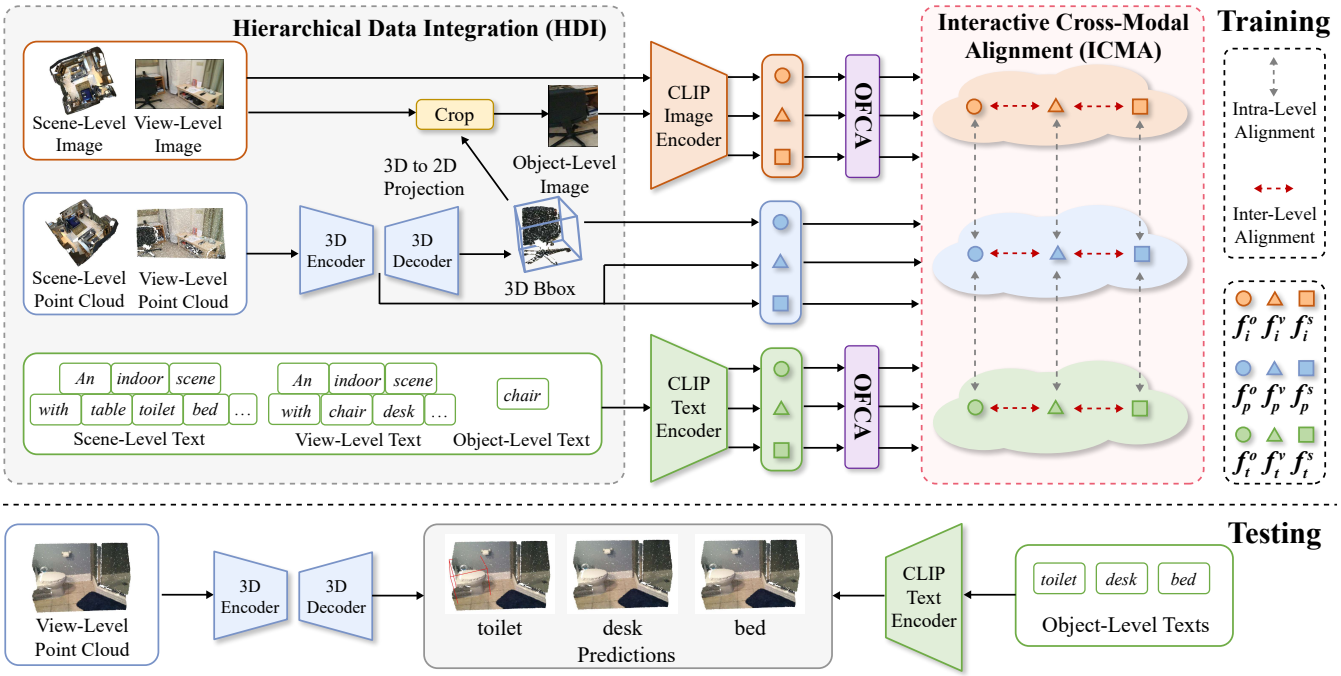

Figure 1: Our HCMA framework is a three-stream network containing pre-trained CLIP image and text encoders, a 3D detector, and our proposed Object-Focusing Context Adjustment (OFCA) module. Each stream processes three types of hierarchical semantics. The input data of each stream is derived from our Hierarchical Data Integration (HDI) approach, while the output semantics of each stream is associated by our Interactive Cross-Modal Alignment (ICMA) strategy.

Given predicted 2D bounding boxes B_{2D}, we then back-project them into the 3D space as 3D bounding boxes $B_{3D} \in \mathbb{R}^7$, and perform clustering to group points belonging to the same object. This step removes unrelated points and improves the accuracy of the resulting 3D bounding boxes. Subsequently, we generate 3D bounding boxes based on the remaining points as:

$$B_{3D} = \text{Cluster}(B_{2D} \cdot M^{-1}), \tag{2}$$

where M is the projection matrix provided in the training dataset. We utilize 3D bounding boxes B_{3D} to supervise the localization of 3D objects by computing a regression loss function between the predicted 3D bounding boxes $\hat{B}_{3D}$ and the target 3D bounding boxes B_{3D} as:

$$\mathcal{L}_{loc} = \mathcal{L}_{3D}(\hat{B}_{3D}, B_{3D}). \tag{3}$$

HCMA next learns to classify 3D localization results from text prompts with cross-modal contrastive learning. In this way, our method can perform 3D object detection without 3D annotations. Figure 1 shows the overall pipeline of our method. During training, our method takes seven types of data as input, including three hierarchies and three modalities. During inference, it takes only the raw view-level point clouds P^v and object-level texts T^o as input. Our method is a three-stream network consisting of a 3D point cloud object detector $\mathbf{H}_p$, an image encoder $\mathbf{H}_i$, a text encoder $\mathbf{H}_t$, and our proposed OFCA module $\mathbf{H}_o$. For each 3D-image-text tuple (P, I, T) as input, we extract the 3D-image-text features (f_p, f_i, f_t) from the three-stream network as:

$$f_p = \mathbf{H}_p(P), \quad f_i = \mathbf{H}_o(\mathbf{H}_i(I)), \quad f_t = \mathbf{H}_o(\mathbf{H}_t(T)). \tag{4}$$

Finally, we utilize hierarchical cross-modal contrastive learning to align f_p, f_i, and f_t. Note that we only update the weights of the 3D detector and the OFCA module since we utilize a pair of frozen CLIP image and text encoders in the training stage.

The HDI Approach

We first introduce the detailed process of our Hierarchical Data Integration (HDI) approach, as shown in Figure 2. Given a view-level 3D-image pair (P^v, I^v), we can obtain object-level point cloud predictions from the 3D detector $\mathbf{H}_p$, including m 3D bounding boxes and an object-level point cloud feature sequence, $f_p^o = (f_{p1}^o, f_{p2}^o, ..., f_{pm}^o)$.

We then project the predicted 3D bounding boxes into 2D bounding boxes with the camera calibration parameter. By cropping the view-level image I^v with 2D bounding boxes, we can obtain an object-level image set, $I^o = (I_1^o, I_2^o, ..., I_m^o)$, which is then sent into the image encoder to get the object-level image features f_i^o. The object-level text set T^o is sent into the text encoder to obtain the object-level text features f_t^o. These three features form the object-level 3D-image-text features (f_p^o, f_i^o, f_t^o).

For m object-level text prompts, $T^o = (T_1^o, T_2^o, ..., T_m^o)$, we concatenate them to form a view-level text caption T^v. Since a 3D view point cloud P^v contains m 3D objects point cloud P^o, we provide a view-level text caption T^v by merging and removing duplicates of these m object-level texts. Given a 3D view point cloud P^v containing m object point

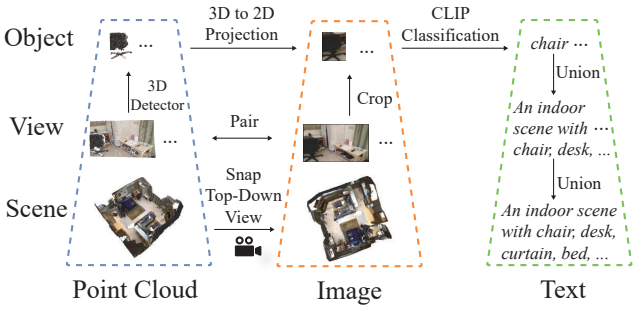

Figure 2: Illustration of the Hierarchical Data Integration (HDI) approach. HDI introduces object-level, view-level, and scene-level hierarchies to associate point clouds, images, and texts.

clouds P^o, we generate a view-level text caption T^v:

$$T^v = (T_1^o \cup T_2^o \cup ... \cup T_m^o), \tag{5}$$

where $\cup$ represents the set union operation. It is worth noting that T^v is constructed by consolidating the textual information from all object-level texts, ensuring the removal of any redundancies. It can also provide a more fine-grained and accurate description of the view-level image I^v. We then feed the view-level 3D-image-text (P^v, I^v, T^v) into their corresponding encoders and the OFCA module. In this way, we can generate the view-level 3D-image-text features (f_p^v, f_i^v, f_t^v).

In addition, we incorporate scene-level data to provide the coarse-grained context of the 3D scene. Despite the absence of scene-level images in the 3D detection dataset, we address this limitation by generating a top-down image of the 3D scene as a scene-level image I^s based on a scene-level 3D point cloud P^s. Since a 3D scene point cloud P^s contains n 3D views point cloud P^v, we provide a scene-level text caption T^s by merging and removing duplicates of these n view-level texts T^v. Similarly, we have:

$$T^s = (T_1^v \cup T_2^v \cup ... \cup T_n^v). \tag{6}$$

Finally, we also feed the scene-level 3D-image-text (P^s, I^s, T^s) into their corresponding encoders and the OFCA module to generate the scene-level 3D-image-text features (f_p^s, f_i^s, f_t^s).

It is worth mentioning that we only utilize point clouds and text in the inference stage. With object-level text T^o as input, our method can detect 3D objects from the view-level point cloud P^v.

The ICMA Strategy

Our Interactive Cross-Modal Alignment (ICMA) strategy consists of two parts: intra-level cross-modal alignment and inter-level cross-modal alignment.

Intra-Level Cross-Modal Alignment. Here, we introduce the details of the intra-level cross-modal alignment in our proposed HCMA. As mentioned above, we have obtained multi-level features from object detector $\mathbf{H}_p$, text encoder $\mathbf{H}_t$, image encoder $\mathbf{H}_i$, and OFCA module $\mathbf{H}_o$. For the

Figure 3: Construction of the positive and negative samples based on the scene label.

object-level features (f_p^o, f_i^o, f_t^o), we first classify them with a category $\mathbb{C}^o$ from the text prompt with pre-trained CLIP:

$$\mathbb{C}^o = \text{argmax}(\text{Softmax}(f_i^o \cdot f_t^o)). \tag{7}$$

In this way, (f_p^o, f_t^o) and (f_p^o, f_i^o) with the same category $\mathbb{C}^o$ are used as two positive pairs. Conversely, they are considered two negative pairs if they do not match the category. In addition, there is no need to align (f_i^o, f_t^o) again since they are already pre-aligned in the pre-trained CLIP model.

For the view-level features (f_p^v, f_i^v, f_t^v), we classify them based on scene labels, as shown in Figure 3. Since an indoor scene can produce multiple view-level point clouds and images from different perspectives, it is natural that these view-level point clouds and images from the same scene are more similar to each other than to those from different scenes. Specifically, if (f_p^v, f_t^v) and (f_p^v, f_i^v) have the same scene label, they form two positive pairs. Otherwise, they are considered as two negative pairs.

Similarly, for the scene-level features (f_p^s, f_i^s, f_t^s), classification is also performed based on scene labels. We divide each of (f_p^s, f_t^s) and (f_p^s, f_i^s) into positive pairs and negative pairs according to their scene labels. This classification strategy based on scene labels allows for the grouping of similar view-level and scene-level features, enhancing the ability of our method to detect nearby objects in the same scene.

Inter-Level Cross-Modal Alignment. To further improve the alignment precision and enhance the integration of information from multiple sources, we introduce the inter-level cross-modal alignment here. It enables coarse-to-fine feature interaction and fusion across different levels and modalities. Object-level input can provide the fine-grained context of the 3D object, while view-level and scene-level inputs are more coarse-grained. We propose three alignment strategies to capture the coarse-to-fine context of 3D objects, including local alignment, global alignment, and all alignment. These strategies aim to provide comprehensive contextual information to improve 3D object understanding.

For local alignment, we concatenate object-level features and view-level features into local features (f_p^l, f_i^l, f_t^l). Local features contain both object-level and view-level features. We then classify the local features with the text prompt category from the pre-trained CLIP, similar to Eq. 7, forming positive pairs and negative pairs from the 3D-text local features as well as from the 3D-image local features. Similar to local alignment, global alignment concatenates object-level features and scene-level features into global features (f_p^g, f_i^g, f_t^g). Global features contain object-level and scene-level features. Again, we form positive pairs and negative pairs from the 3D-text global features as well as from the 3D-image global features according to the text category

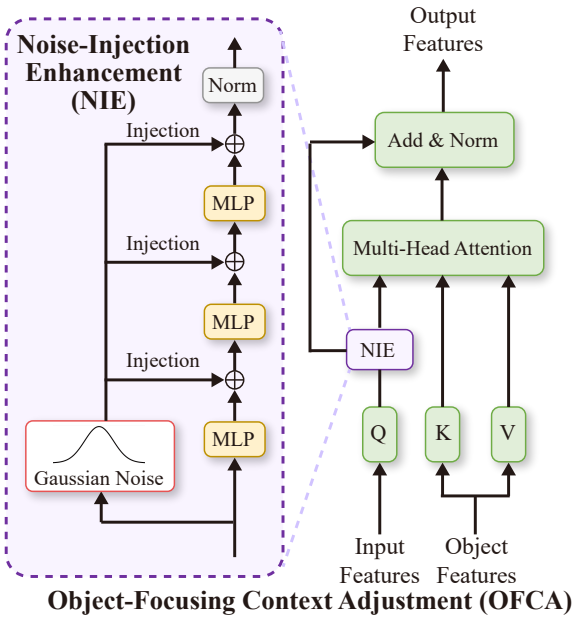

Figure 4: Structure of the proposed Object-Focusing Context Adjustment (OFCA) module.

from the pre-trained CLIP, similar to Eq. 7. In addition, we can also perform all alignment by concatenating object-level features, view-level features, and scene-level features into all features (f_p^a, f_i^a, f_t^a). All features contain object-level, view-level, and scene-level features. Finally, we form positive pairs and negative pairs from the 3D-text all features as well as from the 3D-image all features according to the text category from the pre-trained CLIP, similar to Eq. 7.

The OFCA Module

To better utilize the open-vocabulary capabilities of pre-trained VLMs in the OV-3DOD task, we introduce the Object-Focusing Context Adjustment (OFCA) module, which helps refine the features before cross-modal alignment. As illustrated in Figure 4, each OFCA module consists of a Noise-Injection Enhancement (NIE) block and a multi-head attention layer. The hierarchical semantic derived from object-centric VLMs inherit and amplify the object-level information during alignment in the ICMA module. The NIE block employs Gaussian noise to introduce perturbation to the semantic, which helps reduce potential semantic bias and improve robustness for information enhancement, ultimately improving the model's capability to generalize effectively to unseen data. The NIE block refines input features and transforms them into the query for the attention layer, while the corresponding object features serve as the key and value. NIE employs Gaussian noise by gradually injecting it into the input features. Specifically, the NIE block is constructed by stacking n identical MLP blocks. With the input features x_0 and the output features x_m from the m^{th} MLP block, the m^{th} Gaussian Noise injection operation can be written as:

$$x_m^{'} = \alpha(x_0 + \beta\frac{1}{\sqrt{2\pi}}e^{-\frac{x_0^2}{2}}) + (1-\alpha)x_m, \tag{8}$$

where α and β control the weight of the injection operation. Note that we only input view-level and scene-level features into the NIE block and strategically exclude object-level features from the NIE block, as object-level features are often more reliable compared to other-level features from the pre-trained VLMs.

Hierarchical Cross-Modal Contrastive Learning

Using the obtained 3D-text features (f_p, f_t) and 3D-image features (f_p, f_i), we can guide the 3D detector $\mathbf{H}_p$ to detect novel objects by learning from coarse-to-fine supervisions. We introduce a general contrastive learning method to achieve consistency between 3D, visual, and linguistic representations of hierarchical semantics. The main contrastive loss function (Oord, Li, and Vinyals 2018) is given by:

$$\mathcal{L}_c = -\frac{1}{B}\sum_{b=1}^{B} log\frac{\sum_{i=0}^{m} e^{s_b^T s_i/\tau}}{\sum_{j=0}^{B} e^{s_b^T s_j/\tau}}, \tag{9}$$

where τ is a temperature parameter, s denotes the samples in contrastive learning, B is the number of samples in each batch, and m is the number of positive samples.

For 3D-text pairs (f_p, f_t) and 3D-image pairs (f_p, f_i), we can compute the contrastive loss separately. The cross-modal contrastive loss is given by:

$$\mathcal{L}_m = \mathcal{L}_c(f_p, f_t) + \mathcal{L}_c(f_p, f_i). \tag{10}$$

Subsequently, we employ the positive and negative samples mentioned earlier to construct six coarse-to-fine supervisions. These include object-level $\mathcal{L}_m^o$, view-level $\mathcal{L}_m^v$, and scene-level $\mathcal{L}_m^s$ derived from inter-level cross-modal alignment. Additionally, we incorporate local $\mathcal{L}_m^l$, global $\mathcal{L}_m^g$, and all $\mathcal{L}_m^a$ supervisions through inter-level cross-modal alignment. The alignment loss is given by:

$$\mathcal{L}_{align} = \mathbb{1} \cdot \left[\mathcal{L}_m^o, \mathcal{L}_m^v, \mathcal{L}_m^s, \mathcal{L}_m^l, \mathcal{L}_m^g, \mathcal{L}_m^a\right], \tag{11}$$

where $\mathbb{1}$ is the indicator function that controls the specific form of $\mathcal{L}_{align}$, which is decided by different hierarchies utilized in the training stage. The overall training objective can be written as:

$$\mathcal{L} = \mathcal{L}_{align} + \mathcal{L}_{loc}, \tag{12}$$

where $\mathcal{L}_{loc}$ is also used to supervise the 3D detector.

Experiments

Datasets and Metrics

ScanNet (Dai et al. 2017) is a widely used 3D object detection dataset. As our method requires the view-level and scene-level 3D-image pairs as input, we utilize the raw view-level point clouds, view-level image, and scene-level point clouds from ScanNet. We then generate the scene-level images by snapping the top-down view of the scene-level point clouds. SUN RGB-D (Song, Lichtenberg, and Xiao 2015) is another popular 3D object detection dataset. Since SUN RGB-D dataset lacks scene-level point cloud data, we leverage the available raw view-level point clouds and view-level images from SUN RGB-D for 3D detection. We have conducted our experiment on these two datasets to validate 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
OV-PointCLIP	0.74	1.04	1.85	4.79	1.18	0.19	1.61	0.41	0.03	0.40	0.29	0.51	0.10	1.66	0.16	0.02	0.24	0.04	0.15	0.03	0.05
OV-3DETIC	14.99	53.26	24.88	15.77	31.36	11.54	9.14	2.10	9.39	17.00	29.21	27.45	19.96	13.68	0.01	0.00	0.00	17.73	4.80	3.04	9.51
OV-CLIP-3D	12.68	44.78	23.84	17.52	12.62	4.92	13.24	1.95	3.97	11.37	17.64	32.24	14.87	11.38	2.37	0.51	14.46	8.58	7.45	5.14	4.70
OV-3DET	18.02	57.29	42.26	27.06	31.50	8.21	14.17	2.98	5.56	23.00	31.60	56.28	10.99	19.72	0.77	0.31	9.59	10.53	3.78	2.11	2.71
CoDA	19.32	68.09	44.04	28.72	44.57	3.41	20.23	5.32	0.03	27.95	45.26	50.51	6.55	12.42	15.15	0.68	7.95	0.01	2.94	0.51	2.02
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 1: 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
OV-PointCLIP	1.04	4.76	0.91	4.41	0.07	4.11	0.15	0.05	0.19	0.08	1.11	2.13	0.10	0.24	0.02	0.96	0.06	0.43	0.03	0.91	0.05
OV-3DETIC	12.99	47.68	41.35	4.46	24.41	18.58	10.42	3.72	5.74	12.60	4.89	1.54	0.00	1.24	0.00	19.33	4.58	12.30	23.78	22.02	1.17
OV-CLIP-3D	11.80	38.05	34.45	16.26	20.07	12.72	8.03	2.61	14.62	10.02	5.26	12.31	4.02	1.26	0.09	26.50	7.78	6.52	4.28	10.00	1.19
OV-3DET	20.46	72.64	66.13	34.80	44.74	42.10	11.52	0.29	12.57	14.64	11.21	23.31	2.75	3.40	0.75	23.52	9.83	10.27	1.98	18.57	4.10
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 2: Results on SUN RGB-D in terms of AP_{25}. We report the average value of all 20 categories.

effectiveness of our method in 20 vocabularies. As for the evaluation metrics, we use average precision (AP) and mean average precision (mAP) at IoU thresholds of 0.25 and 0.5, denoted as AP_{25}, mAP_{25}, and mAP_{50}, respectively.

Implementation Details

Our experimental setup follows that of OV-3DET (Lu et al. 2023) for fair comparison. We train our model using the AdamW optimizer with a cosine learning rate scheme. The base learning rate and the weight decay are set to 10^{-4} and 0.1, respectively. The temperature parameter τ is set to 0.1 in contrastive learning. We adopt 3DETR (Misra, Girdhar, and Joulin 2021) as our 3D detector backbone. The number of object queries for 3DETR is set to 128. Experiments are conducted on a single RTX4090 GPU. Our training epoch is the same as the baseline method OV-3DET.

Main Results

We evaluate the performance of our method by comparing it with previous approaches on both ScanNet and SUN RGB-D benchmarks. We conduct evaluations on 20 common classes to assess the effectiveness of our method. Since open-vocabulary 3D object detection is still a new task, very few works can be used for direct comparison. We first directly compare our method with the reported results by OV-3DET (Lu et al. 2023), which is a strong baseline method for OV-3DOD. Following OV-3DET (Lu et al. 2023), we evaluate our method by adapting some existing works on open-vocabulary 2D object detection and 3D object classification into our setting. The baseline methods include PointCLIP (Zhang et al. 2022), 3DETIC (Zhou et al. 2022), and CLIP-3D (Radford et al. 2021b) to conduct open-vocabulary 3D object detection, denoted as OV-PointCLIP (Zhang et al. 2022), OV-3DETIC (Zhou et al. 2022), and OV-CLIP-3D (Radford et al. 2021b). We also compare our method with the reported results by CoDA (Cao et al. 2024), which is the state-of-the-art method for OV-3DOD. Results are presented in Table 1 and Table 2. Our method shows superior performances, outperforming the previous methods on both ScanNet and SUN RGB-D datasets in the OV-3DOD task.

Qualitative Results. Figure 5 shows the qualitative results of our method. Our HCMA framework can generate more accurate bounding boxes compared with the baseline OV-3DET (Lu et al. 2023). HCMA can predict the location, size, and orientation of 3D bounding boxes more precisely. Specific instances, such as the sofa in sample b, highlight the superior performance of HCMA. This is because HCMA can provide coarse-gained context around the 3D object, allowing it to perceive a more complete object structure. Another advantage of our HCMA framework, as compared with the baseline method OV-3DET, is the ability to detect a wider range of objects with diverse vocabularies in 3D scenes. For example, the nightstand in sample a is an occluded object missed by the baseline method. The hierarchical structure of our HCMA facilitates the understanding of objects in the same 3D scene by learning the contextual information, thus enhancing the overall object detection performance.

Ablation Study

Overall Analysis. We conduct an ablation experiment on the ScanNet dataset to analyze the effectiveness of our proposed components. Starting with the base method that leverages solely object-level features and intra-level cross-modal alignment in our framework, we gradually incorporate our proposed modules. The results are summarized in Table 3 and demonstrate that each component contributes to the final result. By employing the HDI approach with both object-level and view-level features, the ICMA strategy with both intra-level and inter-level alignment, and the OFCA module with the NIE block, we observe improvements respectively in terms of mAP_{25}, and mAP_{50}.

Analysis of the HDI Approach. We discuss the impact of the HDI approach through ablation studies on ScanNet. Specifically, we have three coarse-to-fine hierarchical supervisions during the training stage, including object (O), view (V), and scene (S) levels. Since the target of 3D ob-

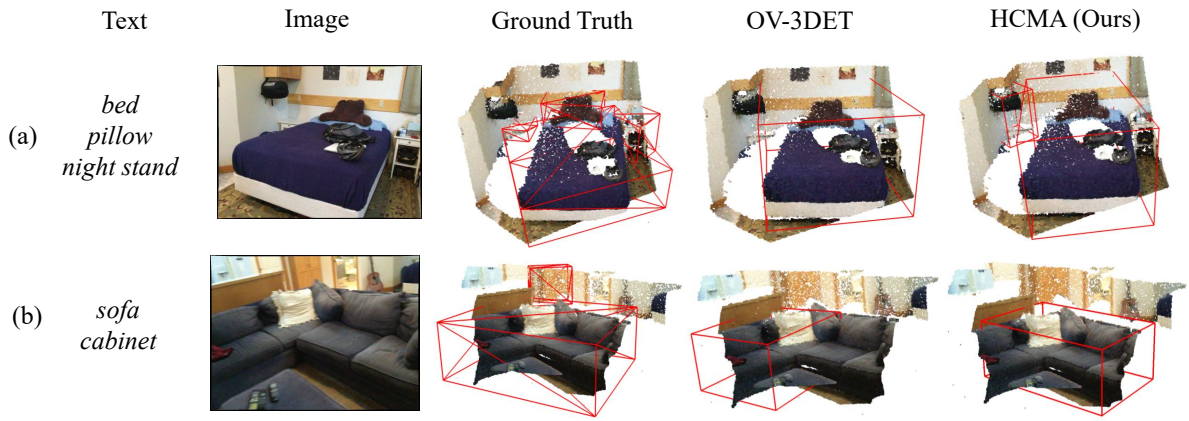

Figure 5: Qualitative comparison with OV-3DET. Our HCMA framework can perform more accurate OV-3DOD and can detect a wider vocabulary of objects in 3D scenes. For each case, the detection text prompts are shown on the left.

ject detection is to detect object-level point clouds in a 3D scene, object-level supervision is naturally the basis of object detection. Hence, we preserve object-level hierarchy to explore the impact of view-level and scene-level hierarchies in the ablation study. We employ four training strategies in our framework and analyze the results in Table 4. The results demonstrate that view-level (V), and scene-level (S) hierarchies can bring significant improvement individually. The coarse-gained cross-modal correspondence is crucial for 3D detection. By leveraging the hierarchical structure and incorporating cross-modal information, our framework benefits from a more comprehensive understanding of the 3D scenes.

Analysis of the ICMA Strategy. We investigate the effectiveness of our ICMA strategy through an ablation study on ScanNet, which includes intra-level and inter-level cross-modal alignment. The results are shown in Table 5. Specifically, we compare our method with three strategies: intra-level, inter-level, and both. The experiments are conducted on both object-level and view-level hierarchies. Results show that using both intra-level and inter-level cross-modal alignments outperforms the other two strategies in terms of mAP_{25} and mAP_{50}, which indicates that both intra-level and inter-level alignment can contribute to cross-modal connections. The intra-level alignment provides hierarchical information to the inter-level alignment, while the inter-level interaction can compensate for the lack of coarse-to-fine context in the intra-level alignment. This complementary relationship enhances the overall performance by incorporating both local and global contextual information.

Analysis of the OFCA Module. To assess the effectiveness of the OFCA module, we conduct ablation studies on the ScanNet dataset as shown in Table 6. Experiments are conducted on both object-level and view-level hierarchies. The results show that the performance improves when employing the OFCA module, as evidenced by higher mAP_{25} and mAP_{50} compared to the baseline without the OFCA module ("w/o OFCA"). This suggests that the proposed OFCA effectively learns more accurate features by adjusting the cross-level features. We further analyze the impact of the NIE block within the OFCA module. Including the NIE block leads to further performance gains compared to using OFCA alone ("OFCA w/o NIE"). This indicates that the NIE block in the OFCA module contributes to performance improvement by enhancing the robustness at the feature level.

HDI	ICMA	OFCA	mAP_{25}	mAP_{50}
			19.37	5.18
✓			19.71	5.80
✓	✓		20.18	6.11
✓	✓	✓	**20.84**	**6.53**

Table 3: Ablation study of our designs.

O	V	S	mAP_{25}	mAP_{50}
✓			19.37	5.18
✓	✓		20.18	6.11
✓		✓	20.20	6.18
✓	✓	✓	**20.30**	**6.39**

Table 4: Ablation study of the HDI approach.

Intra	Inter	mAP_{25}	mAP_{50}
✓		19.71	5.80
	✓	19.68	5.82
✓	✓	**20.18**	**6.11**

Table 5: Ablation study of the ICMA strategy.

Method	mAP_{25}	mAP_{50}
w/o OFCA	20.18	6.11
OFCA w/o NIE	20.55	6.33
OFCA	**20.84**	**6.53**

Table 6: Ablation study of the OFCA module.

Conclusion

In this paper, we have introduced HCMA, a hierarchical framework to address fundamental detection issues of OV-3DOD. HCMA is designed to enhance the accuracy of cross-modal alignment of 3D, image, and text modalities. We first introduce the Hierarchical Data Integration (HDI) approach to obtain hierarchical context from object, view, and scene levels. This coarse-to-fine supervision enables HCMA to obtain more accurate results. In addition, we design an effective Interactive Cross-Modal Alignment (ICMA) strategy, along with the Object-Focusing Context Adjustment (OFCA) module, to align features of 3D, images, and texts to achieve more robust performances. Extensive experimental results demonstrate the superiority of HCMA in OV-3DOD.

References

Bangalath, H.; Maaz, M.; Khattak, M. U.; Khan, S.; and Shahbaz Khan, F. 2022. Bridging the gap between object and image-level representations for open-vocabulary detection. In *Proceedings of the Advances in Neural Information Processing Systems*, 33781–33794.

Bojarski, M.; Del Testa, D.; Dworakowski, D.; Firner, B.; Flepp, B.; Goyal, P.; Jackel, L. D.; Monfort, M.; Muller, U.; Zhang, J.; et al. 2016. End to end learning for self-driving cars. *arXiv:1604.07316*.

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. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 36.

Chen, D.; Wu, Z.; Liu, F.; Yang, Z.; Zheng, S.; Tan, Y.; and Zhou, E. 2023a. ProtoCLIP: Prototypical Contrastive Language Image Pretraining. *IEEE Transactions on Neural Networks and Learning Systems*.

Chen, R.; Liu, Y.; Kong, L.; Zhu, X.; Ma, Y.; Li, Y.; Hou, Y.; Qiao, Y.; and Wang, W. 2023b. CLIP2Scene: Towards Label-efficient 3D Scene Understanding by CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7020–7030.

Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; and Shan, Y. 2024. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16901–16911.

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.

Ding, J.; Xue, N.; Xia, G.-S.; and Dai, D. 2022. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11583–11592.

Ding, R.; Yang, J.; Xue, C.; Zhang, W.; Bai, S.; and Qi, X. 2023. PLA: Language-Driven Open-Vocabulary 3D Scene Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7010–7019.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2022. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. In *Proceedings of the International Conference on Learning Representations*.

Han, K.; Liu, Y.; Liew, J. H.; Ding, H.; Liu, J.; Wang, Y.; Tang, Y.; Yang, Y.; Feng, J.; Zhao, Y.; et al. 2023. Global knowledge calibration for fast open-vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 797–807.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International conference on Machine Learning*, 4904–4916.

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.

Ma, Z.; Luo, G.; Gao, J.; Li, L.; Chen, Y.; Wang, S.; Zhang, C.; and Hu, W. 2022. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14074–14083.

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.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv:1807.03748*.

Peng, S.; Genova, K.; Jiang, C.; Tagliasacchi, A.; Pollefeys, M.; Funkhouser, T.; et al. 2023. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 815–824.

Qi, C. R.; Litany, O.; He, K.; and Guibas, L. J. 2019. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9277–9286.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021a. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 8748–8763. PMLR.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021b. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 8748–8763.

Song, S.; Lichtenberg, S.; 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.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 30.

Wang, L.; Liu, Y.; Du, P.; Ding, Z.; Liao, Y.; Qi, Q.; Chen, B.; and Liu, S. 2023. Object-Aware Distillation Pyramid for Open-Vocabulary Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11186–11196.

Xu, M.; Zhang, Z.; Wei, F.; Hu, H.; and Bai, X. 2023. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2945–2954.

Yang, Z.; Liu, Y.; Lin, J.; Hancke, G.; and Lau, R. W. 2024. Boosting weakly-supervised referring image seg-

mentation via progressive comprehension. *arXiv preprint arXiv:2410.01544*.

Zeng, A.; Song, S.; Welker, S.; Lee, J.; Rodriguez, A.; and Funkhouser, T. 2018. Learning synergies between pushing and grasping with self-supervised deep reinforcement learning. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4238–4245.

Zhang, D.; Li, C.; Zhang, R.; Xie, S.; Xue, W.; Xie, X.; and Zhang, S. 2024. FM-OV3D: Foundation Model-Based Cross-Modal Knowledge Blending for Open-Vocabulary 3D Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16723–16731.

Zhang, R.; Guo, Z.; Zhang, W.; Li, K.; Miao, X.; Cui, B.; Qiao, Y.; Gao, P.; and Li, H. 2022. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8552–8562.

Zhang, Z.; Sun, B.; Yang, H.; and Huang, Q. 2020. H3dnet: 3d object detection using hybrid geometric primitives. In *Proceedings of the European Conference on Computer Vision*, 311–329.

Zhou, C.; Loy, C. C.; and Dai, B. 2022. Extract free dense labels from clip. In *Proceedings of the European Conference on Computer Vision*, 696–712.

Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; and Misra, I. 2022. Detecting twenty-thousand classes using image-level supervision. In *Proceedings of the European Conference on Computer Vision*, 350–368.