OpenScan: A Benchmark for Generalized Open-Vocabulary 3D Scene Understanding

Youjun Zhao[1], **Jiaying Lin**[1,2,*], **Shuquan Ye**[1,3], **Qianshi Pang**[4], **Rynson W.H. Lau**[1,*]

[1] City University of Hong Kong
[2] The Hong Kong University of Science and Technology
[3] The Chinese University of Hong Kong [4] South China University of Technology

Abstract

Open-vocabulary 3D scene understanding (OV-3D) aims to localize and classify novel objects beyond the closed set of object classes. However, existing approaches and benchmarks primarily focus on the open vocabulary problem within the context of object classes, which is insufficient in providing a holistic evaluation to what extent a model understands the 3D scene. In this paper, we introduce a more challenging task called Generalized Open-Vocabulary 3D Scene Understanding (GOV-3D) to explore the open vocabulary problem beyond object classes. It encompasses an open and diverse set of generalized knowledge, expressed as linguistic queries of fine-grained and object-specific attributes. To this end, we contribute a new benchmark named *OpenScan*, which consists of 3D object attributes across eight representative linguistic aspects, including affordance, property, and material. We further evaluate state-of-the-art OV-3D methods on our OpenScan benchmark and discover that these methods struggle to comprehend the abstract vocabularies of the GOV-3D task, a challenge that cannot be addressed simply by scaling up object classes during training. We highlight the limitations of existing methodologies and explore promising directions to overcome the identified shortcomings.

Code, Datasets, and Extended version — https://youjunzhao.github.io/OpenScan/

Introduction

Open-vocabulary 3D scene understanding (OV-3D) involves recognizing object classes that are not included in the training set. It is important to applications like autonomous driving (Bojarski et al. 2016) and robotics (Zeng et al. 2018). Recently, vision-language models (VLMs), *e.g.*, CLIP (Radford et al. 2021), have achieved significant progress by leveraging large-scale image-text datasets with semantically rich captions. The impressive capability of VLMs in capturing rich contexts between images and texts has inspired further exploration in open-vocabulary tasks in both 2D (Gu et al. 2022; Zhong et al. 2022) and 3D (Takmaz et al. 2023; Peng et al. 2023) domains.

For AI systems, the capability to comprehend diverse linguistic aspects of object-related attributes and their association with corresponding objects is as important as the identification of the objects themselves. Consequently, the field of open-vocabulary 3D scene understanding should ideally extend beyond specific object classes to encompass complex object-related attributes, such as affordance and material, articulated through natural languages. However, the generalization ability of existing OV-3D methods (Peng et al. 2023; Takmaz et al. 2023; Yan et al. 2024; Yin et al. 2024; Nguyen et al. 2024) to object-related attributes has not been thoroughly and systematically explored. Besides, evaluating the ability of an OV-3D model to recognize specific object attributes is difficult due to the shortage of large-scale OV-3D attribute benchmarks. Existing OV-3D benchmarks, such as ScanNet (Dai et al. 2017) and ScanNet200 (Rozenberszki et al. 2022), primarily focus on object classes, and do not include annotations of object attributes for evaluating the generalized ability of OV-3D methods.

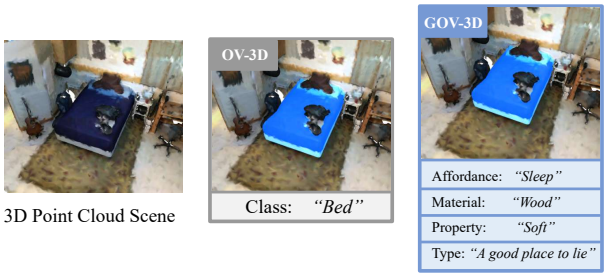

Figure 1: The proposed Generalized Open-Vocabulary 3D Scene Understanding (GOV-3D) task expands the vocabulary types of the classic 3D Scene Understanding (OV-3D) task. While OV-3D only supports queries of object classes, GOV-3D supports queries of object-related abstract attributes.

In this paper, we aim to study how well current OV-3D methods can generalize their understanding beyond 3D object classes to open-set object attribute vocabularies. Specifically, we introduce a more challenging task called Generalized Open-Vocabulary 3D Scene Understanding (GOV-3D). GOV-3D takes a 3D point cloud scene and a text query as input to predict a corresponding 3D mask of the best matching object, which is the same as OV-3D. However, unlike OV-3D which only supports object classes as the input text query, GOV-3D supports abstract vocabularies that specify the attribute of the target object in the input text query, as shown in Figure 1. This requires a comprehensive understanding of both 3D objects

* Jiaying Lin and Rynson Lau are the corresponding authors.

and 3D scenes, making the GOV-3D task more challenging in practical scenarios.

Existing 3D scene understanding benchmarks, such as ScanNet (Dai et al. 2017), ScanNet200 (Rozenberszki et al. 2022), and ScanNet++ (Yeshwanth et al. 2023), only provide annotations for object classes. To address this limitation of existing benchmarks, we construct a new benchmark, named *OpenScan*, for the GOV-3D task. OpenScan is constructed based on the ScanNet200 (Rozenberszki et al. 2022) benchmark. It expands the single category of object classes in ScanNet200 into eight linguistic aspects of object-related attributes, including *affordance*, *property*, *type*, *manner*, *synonym*, *requirement*, *element*, and *material*. This allows each object to be associated with some generalized knowledge beyond object classes. With our OpenScan benchmark, it becomes possible to comprehensively evaluate existing OV-3D models from various aspects, enabling a quantitative assessment of their generalization capabilities in understanding abstract object attributes.

We have compared seven strong baseline methods under the GOV-3D task, on our OpenScan benchmark. Experimental results demonstrate that the current state-of-the-art OV-3D models excel in understanding basic object classes, but significantly degrade in their ability to understand object attributes, such as affordance and material. This highlights the importance of establishing a comprehensive and reliable benchmark to identify the weaknesses of OV-3D models. The key contributions of this work can be summarized as:

- We introduce a challenging task of Generalized Open-Vocabulary 3D Scene Understanding (GOV-3D) to extend the classic OV-3D task for a more general understanding of 3D scenes.
- We provide a novel benchmark named OpenScan for the GOV-3D task, which facilitates comprehensive evaluation of the generalization ability of OV-3D segmentation models on abstract object attributes.
- We conduct extensive experiments with existing OV-3D segmentation models on our OpenScan benchmark, showing that even the latest methods struggle to understand the abstract object attributes beyond object classes.

Related Work

Open-Vocabulary 3D Scene Understanding. The study of open-vocabulary 3D scene understanding (Zhao, Lin, and Lau 2025) has been relatively limited compared to open-vocabulary 2D understanding. This is primarily due to the complexity and difficulty in obtaining 3D datasets. OpenMask3D (Takmaz et al. 2023) introduces the open-vocabulary 3D instance segmentation task. It proposes the first approach for the task in a zero-shot setting. OpenScene (Peng et al. 2023) also proposes a zero-shot method for open-vocabulary 3D scene understanding. Beyond the object class, it is able to utilize arbitrary text queries for semantic segmentation. Previous methods have mainly focused on object context for 3D scene understanding. PLA (Ding et al. 2023) and RegionPLC (Yang et al. 2024) extend the context to a more coarse-to-fine semantic representation to provide a more comprehensive supervision. Recently, OpenIns3D (Huang et al. 2024), Open3DIS (Nguyen et al. 2024), and SAI3D (Yin et al. 2024) utilize powerful 2D segmentation models to generate 2D instances and then merge them into 3D instances. Instead of utilizing accurate 2D masks from 2D segmentation models, MaskClustering (Yan et al. 2024) leverages clustering algorithms to perform zero-shot 3D segmentation. Recently, UniSeg3D (Xu et al. 2024) proposes a unified framework for 3D scene understanding. However, these methods only provide qualitative results for object attributes and lack a thorough evaluation of performance beyond object classes. This motivates us to conduct a quantitative evaluation that encompasses a wider range of object attributes.

3D Scene Understanding Benchmark. Existing open-vocabulary 3D scene understanding benchmarks, *e.g.*, ScanNet (Dai et al. 2017), ScanNet200 (Rozenberszki et al. 2022), S3DIS (Armeni et al. 2016), and Matterport3D (Chang et al. 2017), utilize RGB-D cameras, while ARKitScenes (Baruch et al. 2021), Replica (Straub et al. 2019), and ScanNet++ (Yeshwanth et al. 2023) leverage high-resolution laser scanners to capture high-fidelity 3D data for 3D reconstructions. Our proposed OpenScan benchmark expands the object class annotations of the open-vocabulary 3D scene understanding benchmarks (*i.e.*, ScanNet200) to object-related attributes. Similar to the 3D referring (Chen, Chang, and Nießner 2020) and 3D reasoning (Huang et al. 2025) benchmarks, our OpenScan introduces new annotations for existing 3D scans to locate 3D objects via text queries. Unlike these tasks, which assume the queried object exists in the scene, our introduced GOV-3D task requires discriminative capabilities to determine whether the query presents in the scene. MMScan (Lyu et al. 2024) provides a benchmark for visual attribute understanding but lacks commonsense-related attribute annotations (*e.g.*, "*synonym*" and "*requirement*") included in our OpenScan. Recently, SceneFun3D (Delitzas et al. 2024) provides a large-scale 3D dataset with annotations for functionality and affordance interactions in 3D scenes. However, our OpenScan differs from SceneFun3D by considering a broader range of attributes. Specifically, while SceneFun3D focuses on function or affordance understanding, our OpenScan covers eight linguistic aspects, with affordance representing just one of them. Besides, while SceneFun3D focuses on element-level human-scene interaction (*e.g.*, "*door handle*"), our OpenScan is designed for object-level scene understanding (*e.g.*, "*door*").

Task Setting and Benchmark

Task Formulation

OV-3D. Let $P = \{p_n\}_{n=1}^{N} \in \mathbb{R}^{N\times 3}$ represent 3D scenes with N points, $I = \{i_x\}_{x=1}^{X} \in \mathbb{R}^{H\times W\times 3}$ denote X RGB image frames, and $V = \{c_t\}_{t=1}^{T}$ is a vocabulary set of T text sentences, each describing the object class c_t that we aim to detect. An OV-3D model, $\mathbb{M}$, generates predictions with high confidence scores, $Q = \mathbb{M}(P, I, V)$. Predictions Q are then compared with the GT label G for evaluation.

GOV-3D. The existing 3D scene understanding benchmark, denoted as $\mathcal{D} = \{(o_k, c_k)\}_{k=1}^{K}$, comprises a collection of K object-label pairs. Each pair consists of an object o_k represented as a 3D mask and its corresponding class la-

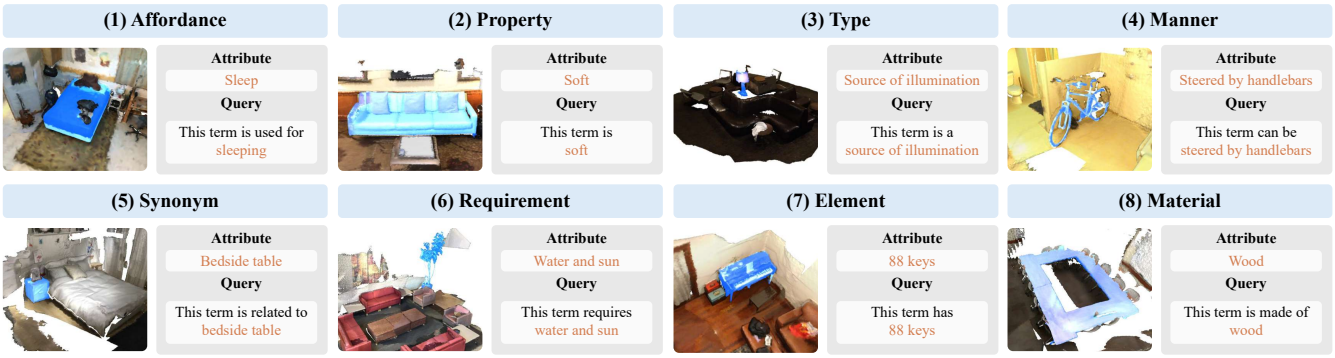

Figure 2: OpenScan benchmark samples. The target objects are highlighted in blue.

bel c_k. The benchmark is composed of multiple 3D scenes $P = \{p_n\}_{n=1}^{N} \in \mathbb{R}^{N\times 3}$ with N points, and X RGB image frames $I = \{i_x\}_{x=1}^{X} \in \mathbb{R}^{H\times W\times 3}$. Building upon this, GOV-3D extends the class label c_k to object attribute a_k. Accordingly, the attribute set for 3D scenes P is defined as $A = \{a_k\}_{k=1}^{H}$, composed of H text sentences, each describing a specific attribute a_k that we aim to detect. A GOV-3D model, $\mathbb{N}$, produces predictions with high confidence scores, $Q = \mathbb{N}(P, I, A)$. The evaluation of the GOV-3D task involves comparing the predictions Q with the GT label G.

Benchmark Description

The OpenScan benchmark is constructed based on the ScanNet200 (Rozenberszki et al. 2022) benchmark, which consists of 200 object classes with more than 1,500 3D scans. Since the ScanNet200 benchmark contains only an object-level class annotation for each object, it is not suitable for our GOV-3D task. To perform the GOV-3D task, we construct the OpenScan benchmark by leveraging the object annotations of the ScanNet benchmark. Our OpenScan provides attribute annotations for each object, expanding the single category of object classes in ScanNet200 into eight linguistic aspects of object-related attributes, including *affordance*, *property*, *type*, *manner*, *synonym*, *requirement*, *element*, and *material*. Figure 2 shows an example from our OpenScan benchmark. The target objects in our OpenScan are annotated with eight linguistic aspects of object attributes. The explanations of these eight object attributes are as follows:

- ***Affordance***: is the object function or usage (*e.g.*, *"sit"* for a chair).
- ***Property***: indicates the object characteristic (*e.g.*, *"soft"* for a pillow).
- ***Type***: indicates the object category or group (*e.g.*, *"a communication device"* for a telephone).
- ***Manner***: indicates the object behavior (*e.g.*, *"worn on a head"* for a hat).
- ***Synonym***: is a term with a similar meaning (*e.g.*, *"image"* for a picture).
- ***Requirement***: indicates an essential condition that an object should possess to fulfill a specific need (*e.g.*, "*balance to ride*" for a bicycle).
- ***Element***: indicates an individual component or part that constitutes the object (*e.g.*, "*two wheels*" for a bicycle).
- ***Material***: indicates the type of material of the object (*e.g.*, "*plastic*" for a bottle).

Benchmark Annotation

Figure 3 illustrates the annotation process of our OpenScan benchmark. We first leverage the knowledge graph to establish the association between objects and various attributes. We also conduct manual annotations to label the visual attributes of each object. Finally, we classify and verify these attributes to ensure semantic consistency.

Object-Attribute Association with Knowledge Graph. We associate each object with various attributes using knowledge graphs, as illustrated in Figure 3. Let $\mathcal{D} = \{(p_k, c_k)\}_{k=1}^{K}$ denotes the existing 3D scene understanding benchmark, *e.g.*, ScanNet200 (Rozenberszki et al. 2022) in our implementation, where p_k is a target object, c_k is the corresponding class label, and K denotes the number of target object and label pairs. The benchmark is composed of multiple 3D scenes $P \in \mathbb{R}^{N\times 3}$ with N points. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denotes the knowledge graph, where $\mathcal{V}$ is the node set and $\mathcal{E}$ is the edge set. The nodes $v \in \mathcal{V}$ are natural language words and phrases, and the edges $e \in \mathcal{E}$ are relation knowledge connecting them. Each edge e is directional, and can be represented as a tuple (v_m, r, w, v_n), where $v_m, v_n \in \mathcal{V}$ are the names of the head node and the tail node, r is the relation, and w is the importance weight of this relation. We extract the relation knowledge from the popular and high-quality NLP knowledge base ConceptNet (Speer, Chin, and Havasi 2017). An example of relation knowledge from it is:

$$e = (\text{“}bed\text{”}, \text{“}is\ used\ for\text{”}, 2.0, \text{“}sleep\text{”}). \tag{1}$$

We query a set of relation knowledge $\{e\}_i$ linked to object class c_i from the knowledge graph $\mathcal{G}$. Formally, for each edge within it, the head node name is the same as the input object class, *i.e.*, $v_m = c_i$. The query process is defined as:

$$\{e\}_i = \{(v_m, r, w, v_n) \in \mathcal{E} | v_m = c_i\}. \tag{2}$$

Attribute Selection. In the set of relation knowledge $\mathcal{E}$, we keep the attribute with the highest weight w in the same relation r. Given a relation knowledge $e_i \in \{e\}$, we have:

$$\{e\}_i^{'} = \{e_i | r_j = r_i \wedge \forall e_j \in \{e\} : w_j \leq w_i\}. \tag{3}$$

These object-attribute pairs form the basic annotations of OpenScan, which is useful in the GOV-3D task. Finally, each

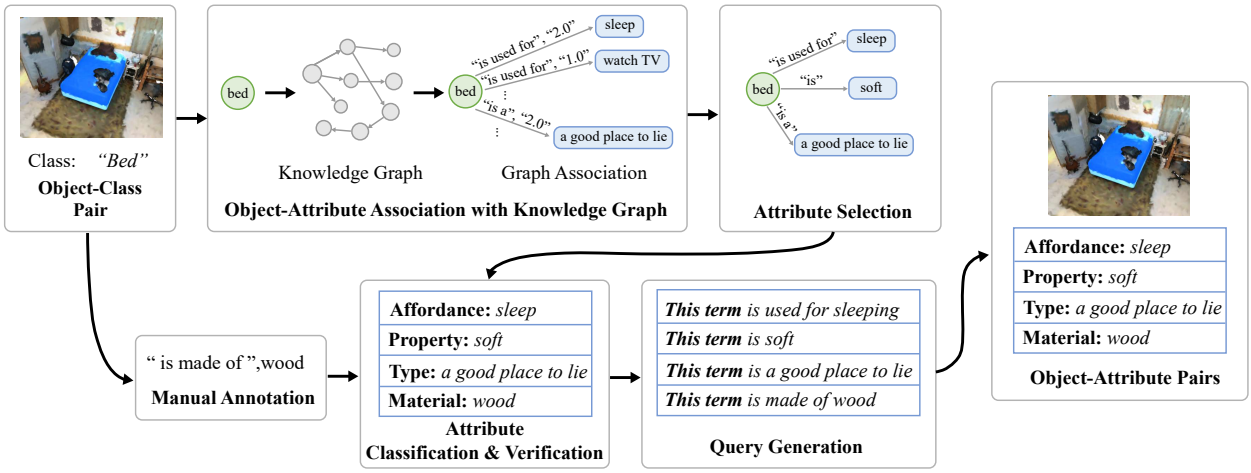

Figure 3: Illustration of the data generation process for our OpenScan benchmark.

3D object p_i is assigned a relation knowledge e_i through I annotations, serving as commonsense knowledge $\mathcal{Y}_c$ as:

$$\mathcal{Y}_c = \{(p_i, e_i), e_i \in \{e\}' | v_m = c_i\}_i^I. \quad (4)$$

Manual Annotation. For the visual attribute that cannot be inferred without human perception, we manually annotate each 3D object following a rigorous protocol. We create a web interface for annotators to select each object's visual attribute. Specifically, for each scene, annotators view the 3D point cloud and the target's corresponding 2D image. Taking the *material* attribute as an example, annotators are tasked with identifying the primary material composition of the target object. Any 3D object with an ambiguous appearance is carefully identified through different camera views of the scene and the corresponding image frames around the object. Finally, each 3D object p_j is assigned with a relation $r_j =$ "$is\ made\ of$" and a visual attribute like *material* v_n through J annotations, serving as visual appearance $\mathcal{Y}_m$ in:

$$\mathcal{Y}_m = \{(p_j, (r_j, v_n)\}_j^J. \quad (5)$$

After obtaining the attribute annotations based on commonsense knowledge $\mathcal{Y}_c$ and visual appearance $\mathcal{Y}_m$ of the 3D objects, we use the combination of these two categories of attributes as the whole annotations $\mathcal{Y}$ for our OpenScan.

Attribute Classification. To better organize our benchmark, we manually group each attribute into eight linguistic aspects based on relation r and attribute v_n. This process involves considering the nature of the relation r and attribute v_n, and how they align with each linguistic aspect. Subsequently, each attribute v_n is assigned to a linguistic aspect.

Attribute Verification. After the initial attribute classification, we manually verify each 3D object p_k with its corresponding attribute v_n and linguistic aspect. If a 3D object p_k contains multiple attributes v_n within a single linguistic aspect, we manually assign the most representative attribute to ensure evaluation consistency. This attribute is uniquely tied to the object class of p_k to eliminate cross-class ambiguity. We also filter out similar attributes v_n (*e.g.*, "store things" and "store somethings"), preserving only one attribute to ensure consistency. During the verification process, the attributes across eight linguistic aspects are reduced from 528 to 341.

Query Generation. A practical GOV-3D query should incorporate attribute names but exclude object names, requiring a query strategy that focuses on attributes rather than exposing object identities (*i.e.*, object classes). To achieve this, we perform query generation by hiding the object class v_m of the object p_k. We first replace the object classes v_m with a substitution term $t =$ "*this term*". The substitution term t, the relation r, and the corresponding attribute v_n are then concatenated to form the text query q as:

$$q = Concate(t, r, v_n). \quad (6)$$

In this way, we generate text queries q that correspond to object-attribute annotations $\mathcal{Y}$. We then perform manual verification again on text queries. With text queries as input, we can conduct evaluations on existing OV-3D models.

Benchmark Statistics

Table 1 shows the statistics of our OpenScan benchmark. We have collected eight linguistic aspects of attributes, providing a total of 153,644 attribute annotations across 341 attributes for 1,513 scenes in ScanNet200 (Rozenberszki et al. 2022). In these aspects, the visual aspect of *material* is annotated manually, while other attributes are automatically generated via knowledge graph. There are 101.55 attribute annotations per scene in 1,513 3D indoor scenes. Besides, each object is annotated with an average of 3.15 attributes, indicating that most objects in ScanNet200 receive multiple attribute labels and are comprehensively represented. While certain linguistic aspects such as *manner* and *synonym* encompass a limited number of attributes, others like *affordance* and *type* consist of a wide range of attributes. We follow the training and validation split settings of ScanNet200.

Evaluation Metrics

We employ commonly used OV-3D metrics to evaluate our GOV-3D task. For semantic segmentation, we follow (Peng et al. 2023; Ding et al. 2023; Yang et al. 2024) to apply mean IoU (mIoU) and mean accuracy (mAcc). For instance segmentation, we follow (Takmaz et al. 2023; Yin et al. 2024; Yan et al. 2024; Nguyen et al. 2024) to apply average precision (AP) at IoU scores of 25% (AP_{25}), 50% (AP_{50}), and the

Statistics	Affordance	Property	Type	Manner	Synonym	Requirement	Element	Material	All
Attributes	104	19	96	21	16	28	47	10	341
Attribute Annotations	37,362	8,591	28,293	4,925	2,937	9,695	13,505	48,336	153,644
Attribute Annotations per Object	0.77	0.18	0.58	0.10	0.06	0.20	0.28	0.99	3.15
Attribute Annotations per Scene	24.69	5.68	18.70	3.26	1.94	6.41	8.93	31.95	101.55

Table 1: OpenScan benchmark statistics of object-related attributes for the eight linguistic aspects.

Model	Training	Pre-Trained 3D Proposal	Pre-Trained 2D Proposal
OpenMask3D	-	Mask3D	SAM
SAI3D	-	-	Semantic-SAM
MaskClustering	-	-	CropFormer
Open3DIS	-	ISBNet	Grounded-SAM
OpenScene	-	-	-
PLA	ScanNet	-	-
RegionPLC	ScanNet	-	-

Table 2: The detailed information of the OV-3D models.

mean of AP from 50% to 95% at 5% steps. We compute the mean score (Mean) across all attributes to obtain the overall performance.

Experiments

We conduct our experiments on the validation set of our OpenScan across eight linguistic aspects using the publicly available OV-3D models. All OV-3D models are evaluated in a zero-shot setting without training on the OpenScan benchmark. For 3D instance segmentation, we evaluate OpenMask3D (Takmaz et al. 2023), SAI3D (Yin et al. 2024), MaskClustering (Yan et al. 2024), and Open3DIS (Nguyen et al. 2024). For 3D semantic segmentation, we evaluate OpenScene (Peng et al. 2023), PLA (Ding et al. 2023), and RegionPLC (Yang et al. 2024). Table 2 summarizes the information of these models: the training set, pre-trained 3D proposal, and pre-trained 2D proposal. All experiments are conducted on one NVIDIA RTX 4090 GPU.

Main Results

3D Instance Segmentation. We evaluate OpenMask3D (Takmaz et al. 2023), SAI3D (Yin et al. 2024), MaskClustering (Yan et al. 2024), and Open3DIS (Nguyen et al. 2024) across 341 attributes from our OpenScan and 198 object classes from ScanNet200 (Rozenberszki et al. 2022). Table 3 show that the performance of these OV-3D models on our GOV-3D benchmark, OpenScan, are significantly lower than those on the classic OV-3D dataset, ScanNet200. This gap underscores that our proposed GOV-3D task is a more challenging extension of the traditional OV-3D task.

When comparing the results of each OV-3D model across different linguistic aspects, we observe higher performances in the *synonym* and *material* aspects but struggle in the *affordance* and *property* aspects. The high performance in the *synonym* aspect can be attributed to the close similarity between attributes in this aspect and object classes, making recognition easier compared to the more abstract *affordance* and *property* aspects. An example of these closely related terms is shown in Figure 2, where the corresponding *synonym* aspect of the object class "*nightstand*" is "*bedside table*". The high performance in the *material* aspect highlights the ability of these OV-3D models to recognize visual patterns. By utilizing CLIP (Radford et al. 2021) for 3D scene understanding, these models benefit from its visual patterns, including material and color from image-text pretraining , enhancing their comprehension of visual attributes beyond other attributes.

When comparing the results of each linguistic aspect in our OpenScan to those of the object class in ScanNet200, we notice that certain aspects like *synonym* and *material* perform even better than the object class. This can be attributed to the smaller number of attributes in these two aspects when compared to the broader and more diverse set of object classes. A smaller set of classes can increase the model's confidence in its predictions, facilitating more accurate predictions without the complexity of distinguishing among a large number of categories. Notably, Open3DIS shows impressive results in various linguistic aspects compared to other OV-3D models, aligning with its strong performances in the classic OV-3D task (*i.e.*, evaluating only on object classes).

3D Semantic Segmentation. We evaluate OpenScene (Peng et al. 2023), PLA (Ding et al. 2023), and RegionPLC (Yang et al. 2024), reporting the average score of all attributes in our OpenScan and that of all object classes in ScanNet (Dai et al. 2017). Table 4 shows that although these OV-3D models perform well in recognizing object classes, they exhibit poor performances on linguistic aspects with low mIoU and mAcc metrics. The methods for semantic segmentation suffer from a more significant performance drop on OpenScan when compared with those for instance segmentation. This drop can be caused by several factors. First, there is a significant discrepancy in the vocabulary size between ScanNet and our OpenScan. A larger vocabulary size implies a more diverse set of semantic concepts that the model needs to comprehend, making our OpenScan more challenging and practical in real-world scenarios. In addition, the arbitrary nature of object attributes in contrast to object classes adds complexity to the GOV-3D task. Besides, the lack of both robust 3D proposals (*e.g.*, Mask3D (Schult et al. 2023)) and 2D proposals (*e.g.*, SAM (Kirillov et al. 2023)) for class-agnostic masks can also be attributed to the drop. Conversely, instance segmentation models like OpenMask3D (Takmaz et al. 2023) leverage strong instance-level knowledge, *e.g.*, proposals extracted from Mask3D and SAM, to effectively segment novel 3D objects, leading to higher performances on the GOV-3D task.

The Impact of the Pre-trained Vocabulary Size

To study the impact of the pre-trained vocabulary size (*i.e.*, the number of pre-trained object classes) on the GOV-3D task, we conduct experiments based on RegionPLC (Yang et al. 2024) for 3D semantic segmentation. Figure 4 reports the mIoU and mAcc scores under different pre-training vocabulary sizes $S \in \{10, 12, 15, 150, 170\}$, on the ScanNet (Dai et al. 2017) and ScanNet200 (Rozenberszki et al. 2022)

Method	OpenScan									ScanNet200
	Affordance	Property	Type	Manner	Synonym	Requirement	Element	Material	Mean	Object Class
AP										
OpenMask3D (Takmaz et al. 2023)	7.2	7.5	8.5	12.8	16.9	13.0	12.2	18.8	9.9	15.4
SAI3D (Yin et al. 2024)	5.3	5.8	7.8	11.3	10.0	10.0	8.7	11.3	7.7	12.7
MaskClustering (Yan et al. 2024)	6.2	7.0	7.1	11.1	16.2	11.3	7.4	12.1	8.1	12.0
Open3DIS (Nguyen et al. 2024)	**11.9**	**12.8**	**14.2**	**19.2**	**26.7**	**19.2**	**18.7**	**28.3**	**15.8**	**23.7**
AP_{50}										
OpenMask3D (Takmaz et al. 2023)	9.1	10.0	11.2	15.4	19.7	16.0	15.4	22.1	12.5	19.9
SAI3D (Yin et al. 2024)	8.4	8.3	11.4	15.7	16.7	15.3	13.6	17.1	11.6	18.8
MaskClustering (Yan et al. 2024)	10.7	12.3	13.3	18.4	30.3	21.8	13.5	20.6	14.6	23.3
Open3DIS (Nguyen et al. 2024)	**14.8**	**16.0**	**17.9**	**22.3**	**30.6**	**24.1**	**21.9**	**33.6**	**19.3**	**29.4**
AP_{25}										
OpenMask3D (Takmaz et al. 2023)	10.4	11.6	13.0	17.4	20.6	18.9	17.1	25.0	14.2	23.1
SAI3D (Yin et al. 2024)	10.5	10.7	13.4	18.2	20.0	18.7	16.0	22.9	14.1	24.1
MaskClustering (Yan et al. 2024)	13.7	15.8	17.7	23.1	**36.6**	**28.2**	17.2	25.6	18.7	30.1
Open3DIS (Nguyen et al. 2024)	**16.7**	**16.8**	**20.2**	**24.2**	33.1	25.5	**24.7**	**36.7**	**21.4**	**32.8**

Table 3: 3D instance segmentation results on our OpenScan benchmark.

Method	OpenScan		ScanNet	
	mIoU	mAcc	mIoU	mAcc
OpenScene (Peng et al. 2023)	**0.45**	1.87	47.5	70.7
PLA (Ding et al. 2023)	0.01	**2.37**	66.6	77.5
RegionPLC (Yang et al. 2024)	0.07	2.36	**68.7**	**78.7**

Table 4: 3D semantic segmentation results on OpenScan.

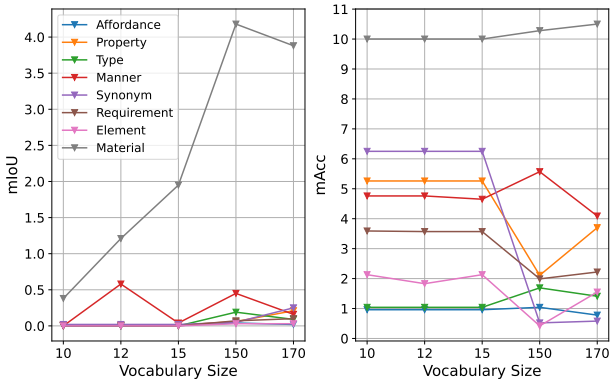

Figure 4: Impact of different pre-training vocabulary size.

datasets. Results show that the majority of the linguistic aspects of object attributes do not exhibit a notable enhancement as S increases, reflected by both mIoU and mAcc scores, aligning with our expectations. Some linguistic aspects of object attributes show relatively low performances and exhibit random jitters. Among the eight linguistic aspects, the aspect *material* illustrates an enhancement in mIoU and a marginal improvement in mAcc as S increases. This improvement can be attributed to the framework adopted by RegionPLC, which associates 3D objects with language through explicit visual image captioning models, providing detailed descriptions of visual attributes like material and color for each 3D object. Therefore, as the vocabulary size S increases, more objects are processed by the image captioning model to produce visual descriptions that ultimately improve the semantic segmentation results for the aspect *material*.

Method	Template	AP	AP_{50}	AP_{25}
OpenMask3D (Takmaz et al. 2023)	-	9.7	12.2	14.1
	✓	**9.9**	**12.5**	**14.2**
SAI3D (Yin et al. 2024)	-	6.7	10.1	12.8
	✓	**7.7**	**11.6**	**14.1**
MaskClustering (Yan et al. 2024)	-	6.8	12.0	14.6
	✓	**8.1**	**14.6**	**18.7**
Open3DIS (Nguyen et al. 2024)	-	15.6	19.2	21.3
	✓	**15.8**	**19.3**	**21.4**

Table 5: Effects of query form on our OpenScan benchmark.

This observation suggests that simply increasing the size of the object vocabulary during training may not effectively enhance the generalization capability of OV-3D models. This limitation can be attributed to existing OV-3D benchmarks, like ScanNet (Dai et al. 2017), ScanNet200 (Rozenberszki et al. 2022), and ScanNet++ (Yeshwanth et al. 2023), which primarily focus on object classes and lack object-related attributes. While increasing the size of the object vocabulary during training can improves the OV-3D performance, as demonstrated by the results from PLA (Ding et al. 2023) and RegionPLC (Yang et al. 2024), this approach is not suitable for the more challenging GOV-3D task. The significant performance gap between the two tasks cannot be resolved simply by transferring the OV-3D technique into GOV-3D.

The Impact of the Query Form

In benchmark annotation, we generate queries linking attributes to object classes. An ideal query should contain an attribute name and the relation between the attribute and the corresponding object class. Table 5 shows the effect of using a query template (*e.g.*, "*this term is made of wood*") versus a plain term (*e.g.*, "*wood*") in GOV-3D. We evaluate the 3D instance segmentation of OpenMask3D (Takmaz et al. 2023), SAI3D (Yin et al. 2024), MaskClustering (Yan et al. 2024), and Open3DIS (Nguyen et al. 2024), reporting the mean score of all attributes in OpenScan. Note that, as expected, using the query template improves model performance, as shown

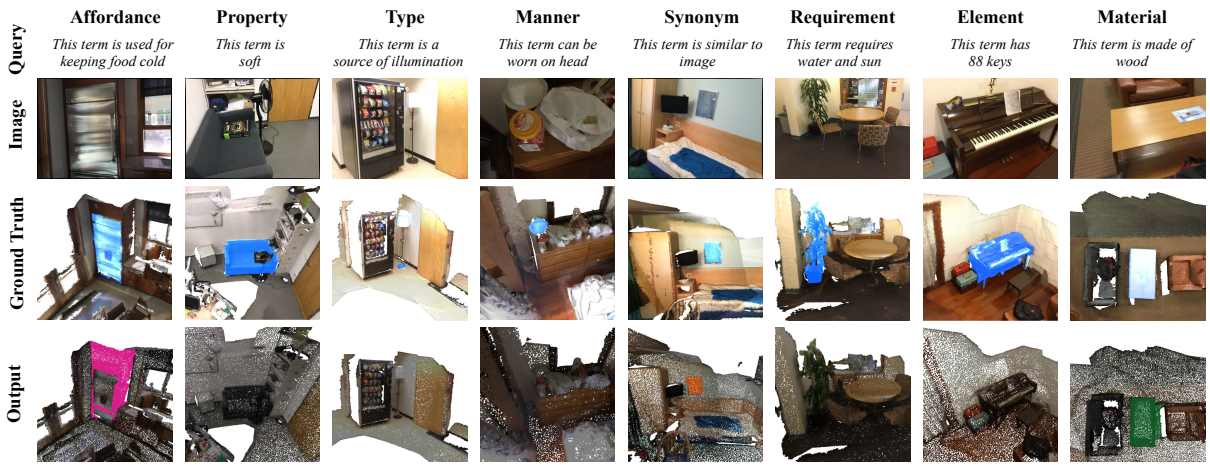

Figure 5: Qualitative results of Open3DIS on our OpenScan benchmark. The GT objects and outputs are highlighted in color.

by higher AP, AP_{50}, and AP_{25}. SAI3D and MaskClustering are more sensitive to query templates, while OpenMask3D and Open3DIS show greater robustness.

These results stem from the fact that VLMs, like CLIP (Radford et al. 2021), struggle with attribute classification in GOV-3D when minor commonsense knowledge is required, as stated in (Ye et al. 2023). Since most OV-3D models rely on VLMs like CLIP (Radford et al. 2021) for open-vocabulary comprehension, they inherit VLMs' commonsense limitations. Thus, incorporating query templates that link attributes to object classes as commonsense knowledge improves OV-3D models' performances in GOV-3D.

Qualitative Results

We present qualitative results from Open3DIS (Nguyen et al. 2024) on our OpenScan benchmark. We evaluate Open3DIS across eight linguistic aspects, as shown in Figure 5. It demonstrates that Open3DIS can comprehend specific linguistic aspects such as *synonym* and *material*. When exploring the *affordance* aspect by querying "*keep food cold*" for the target object, Open3DIS can successfully identify the "*refrigerator*" as the target object but struggles to generate a correct 3D mask. Additionally, Open3DIS fails to generate predictions for other linguistic aspects. These observations align with the quantitative results in Table 3.

Failure Cases Analysis

Figure 6 shows a failure case of Open3DIS (Nguyen et al. 2024) when applied to GOV-3D, despite its strong performance on OV-3D. While Open3DIS correctly identifies the object class (*e.g.*, "*piano*"), it fails to recognize the associated object attribute (*e.g.*, "*this term has 88 keys*"). To investigate this discrepancy, we analyze the CLIP (Radford et al. 2021) image-text similarity scores in Open3DIS, given that most OV-3D models rely on VLMs like CLIP (Radford et al. 2021) for 3D predictions. Our analysis reveals that CLIP assigns lower image-text similarity scores to the object attribute compared to the object class, suggesting that its intrinsic attribute knowledge is limited. This observation demonstrates that GOV-3D presents greater challenges compared to OV-3D. A promising direction for GOV-3D involves integrating attribute knowledge into VLMs like CLIP, which may improve image-text alignment and enable more reliable predictions.

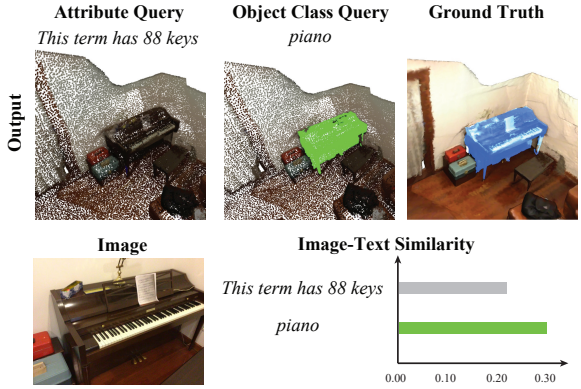

Figure 6: Visualization of the Open3DIS failure case with the corresponding CLIP image-text similarity scores.

Conclusion

In this paper, we address the constraints of the classic Open-Vocabulary 3D Scene Understanding (OV-3D) task, which is limited in handling object attributes beyond object classes. We introduce a more challenging task, called Generalized Open-Vocabulary 3D Scene Understanding (GOV-3D), to comprehensively evaluate the generalization capability of OV-3D models. To facilitate research on the GOV-3D task, we construct a large-scale benchmark named OpenScan, which consists of 341 attributes across 8 linguistic aspects. We systematically evaluate the OV-3D models on the OpenScan benchmark, revealing their challenges in understanding attributes beyond object classes. We also conduct experiments to investigate the impact of the pre-trained vocabulary size and query form, demonstrating that the generalization ability can be enhanced by utilizing query templates rather than scaling up the vocabulary size during training. We further explore a promising direction for the GOV-3D task by integrating attribute knowledge into VLMs of the OV-3D models. We believe our OpenScan benchmark can facilitate future research on improving the generalization capability of OV-3D models.

References

Armeni, I.; Sener, O.; Zamir, A. R.; Jiang, H.; Brilakis, I.; Fischer, M.; and Savarese, S. 2016. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1534–1543.

Baruch, G.; Chen, Z.; Dehghan, A.; Dimry, T.; Feigin, Y.; Fu, P.; Gebauer, T.; Joffe, B.; Kurz, D.; Schwartz, A.; and Shulman, E. 2021. ARKitScenes - A Diverse Real-World Dataset for 3D Indoor Scene Understanding Using Mobile RGB-D Data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

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*.

Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Niessner, M.; Savva, M.; Song, S.; Zeng, A.; and Zhang, Y. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. *International Conference on 3D Vision*.

Chen, D. Z.; Chang, A. X.; and Nießner, M. 2020. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European Conference on Computer Vision*, 202–221. Springer.

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.

Delitzas, A.; Takmaz, A.; Tombari, F.; Sumner, R.; Pollefeys, M.; and Engelmann, F. 2024. Scenefun3d: Fine-grained functionality and affordance understanding in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14531–14542.

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 *International Conference on Learning Representations*.

Huang, K.-C.; Li, X.; Qi, L.; Yan, S.; and Yang, M.-H. 2025. Reason3d: Searching and reasoning 3d segmentation via large language model. In *International Conference on 3D Vision 2025*.

Huang, Z.; Wu, X.; Chen, X.; Zhao, H.; Zhu, L.; and Lasenby, J. 2024. Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation. In *European Conference on Computer Vision*, 169–185. Springer.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollar, P.; and Girshick, R. 2023. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Lyu, R.; Lin, J.; Wang, T.; Mao, X.; Chen, Y.; Xu, R.; Huang, H.; Zhu, C.; Lin, D.; and Pang, J. 2024. Mmscan: A multi-modal 3d scene dataset with hierarchical grounded language annotations. *Advances in Neural Information Processing Systems*, 37: 50898–50924.

Nguyen, P.; Ngo, T. D.; Kalogerakis, E.; Gan, C.; Tran, A.; Pham, C.; and Nguyen, K. 2024. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4018–4028.

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.

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

Rozenberszki, D.; Litany, O.; Dai, A.; and Dai, A. 2022. Language-Grounded Indoor 3D Semantic Segmentation in the Wild. In *Proceedings of the European Conference on Computer Vision*.

Schult, J.; Engelmann, F.; Hermans, A.; Litany, O.; Tang, S.; and Leibe, B. 2023. Mask3D: Mask Transformer for 3D Semantic Instance Segmentation. In *2023 IEEE International Conference on Robotics and Automation*, 8216–8223.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Straub, J.; Whelan, T.; Ma, L.; Chen, Y.; Wijmans, E.; Green, S.; Engel, J. J.; Mur-Artal, R.; Ren, C.; Verma, S.; et al. 2019. The Replica dataset: A digital replica of indoor spaces. *arXiv:1906.05797*.

Takmaz, A.; Fedele, E.; Sumner, R. W.; Pollefeys, M.; Tombari, F.; and Engelmann, F. 2023. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *Advances in Neural Information Processing Systems*.

Xu, W.; Shi, C.; Tu, S.; Zhou, X.; Liang, D.; and Bai, X. 2024. A Unified Framework for 3D Scene Understanding. In *Advances in Neural Information Processing Systems*.

Yan, M.; Zhang, J.; Zhu, Y.; and Wang, H. 2024. Maskclustering: View consensus based mask graph clustering for open-vocabulary 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28274–28284.

Yang, J.; Ding, R.; Deng, W.; Wang, Z.; and Qi, X. 2024. Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19823–19832.

Ye, S.; Xie, Y.; Chen, D.; Xu, Y.; Yuan, L.; Zhu, C.; and Liao, J. 2023. Improving commonsense in vision-language models via knowledge graph riddles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2634–2645.

Yeshwanth, C.; Liu, Y.-C.; Nießner, M.; and Dai, A. 2023. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12–22.

Yin, Y.; Liu, Y.; Xiao, Y.; Cohen-Or, D.; Huang, J.; and Chen, B. 2024. Sai3d: Segment any instance in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3292–3302.

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 *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 4238–4245. IEEE.

Zhao, Y.; Lin, J.; and Lau, R. W. 2025. Hierarchical Cross-Modal Alignment for Open-Vocabulary 3D Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10501–10509.

Zhong, Y.; Yang, J.; Zhang, P.; Li, C.; Codella, N.; Li, L. H.; Zhou, L.; Dai, X.; Yuan, L.; Li, Y.; et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16793–16803.