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

Supplementary Material

In this supplementary material, we provide more experimental results and benchmark details:

- Sec. A: Web interface for manual annotation.
- Sec. B: Implementation details.
- Sec. C: Additional experimental results.
- Sec. D: Additional benchmark details.
- Sec. E: Additional related work.
- Sec. F: Limitations and future work.
- Sec. G: Broader impact.

A Web Interface for Manual Annotation

We implement a web interface for manual annotation of the visual linguistic aspect (*e.g.*, *material*), as shown in Figure A. Annotators are shown an interactive 3D mesh of a scene, a list of target objects, and a list of attributes. Users can control the 3D mesh from different viewpoints interactively by rotating, zooming in, zooming out, and panning to observe the scene from various viewpoints. When users select a 3D mesh by clicking the mouse in the scene, the target object will be highlighted and the corresponding object ID and object class will be displayed. The annotation process requires annotators to first select a target object in the 3D mesh (*e.g.*, table of ID 2) and then select a primary attribute that belongs to the target object (*e.g.*, stone). Finally, annotators click the confirm button to submit and store the annotations. Once the selected objects are annotated and confirmed, the corresponding object in the object list will show a check mark symbol. Additionally, to address visual ambiguity issues in 3D object appearance, our annotation process allows users to review the scene's video sequences, providing contextual visual cues to resolve uncertainties about target objects' visual attributes during annotation.

B Implementation Details

B.1 Open-Vocabulary 3D Scene Understanding (OV-3D) Baselines

We report implementation details of the OV-3D models (Takmaz et al. 2023; Yin et al. 2024; Yan et al. 2024; Nguyen et al. 2024; Peng et al. 2023; Ding et al. 2023; Yang et al. 2024) as follows:

OpenMask3D. In the class-agnostic mask proposal module, we employ the Mask3D (Schult et al. 2023) architecture trained on the ScanNet200 (Rozenberszki et al. 2022) training set. For 2D mask proposals, we use SAM (Kirillov et al. 2023) with ViT-H as the backbone. We utilize the pre-trained CLIP (Radford et al. 2021) visual encoder of ViT-L/14 at a 336 pixel resolution to extract image features with 768 dimensions. We set the number of queries to 150, following the implementation of OpenMask3D (Takmaz et al. 2023) and Mask3D (Schult et al. 2023), to ensure a sufficient number of mask proposals for the GOV-3D task.

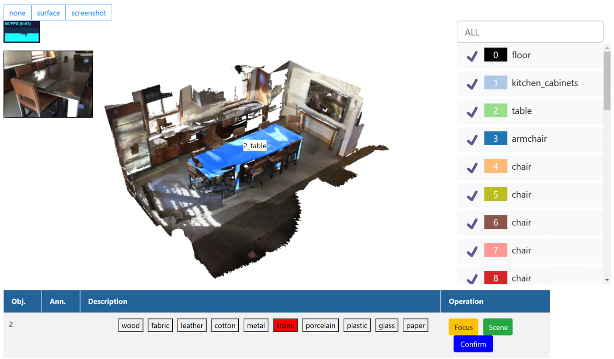
SAI3D. We employ Semantic-SAM (Li et al. 2024) with Swin-L as the backbone to generate 2D mask proposals. The number of queries is set to 150 to ensure sufficient mask proposals for the GOV-3D task.

MaskClustering. We utilize CropFormer (Qi et al. 2023) as a 2D mask predictor. For 2D mask proposals, we use CLIP (Radford et al. 2021) visual encoder of ViT-H/14 to extract image features. We follow MaskClustering (Yan et al. 2024) to adopt the post-processing approach from OVIR-3D (Lu et al. 2023) to refine the output 3D instances. Specifically, we employ the DBSCAN (Ester et al. 1996) algorithm to partition disconnected point clusters.

Open3DIS. We utilize the class-agnostic 3D proposal network ISBNet (Ngo, Hua, and Nguyen 2023) trained on the ScanNet200 (Rozenberszki et al. 2022) training set as 3D proposal. We employ the 2D-Guided-3D Instance Proposal Module in Open3DIS (Nguyen et al. 2024). For 2D mask proposals, we adopt Grounded-SAM (Ren et al. 2024) as 2D segmentor, which incorporates a Swin-T-based Grounding-DINO (Liu et al. 2024) decoder and SAM (Kirillov et al. 2023) with ViT-H as the backbone.

OpenScene. We employ OpenSeg (Ghiasi et al. 2022) for image feature extraction and a 2D-3D ensemble model in OpenScene (Peng et al. 2023). We utilize MinkowskiNet18A (Choy, Gwak, and Savarese 2019) as the 3D back-

[*] Jiaying Lin and Rynson Lau are the corresponding authors.



Obj.	Ann.	Description	Operation
2		wood fabric leather cotton metal stone porcelain plastic glass paper	Focus Scene Confirm

Figure A: Web interface for manual annotation that allows users to view the 3D scene from multiple viewpoints and select the target object by clicking.

bone during 3D distillation.

PLA. We utilize a model trained on the ScanNet (Dai et al. 2017) partition of B15/N4, where B15/N4 indicates 15 base and 4 novel categories. We adopt a SparseUNet16 architecture based on sparse convolutions UNet (Graham, Engelcke, and Van Der Maaten 2018) as our 3D encoder for semantic segmentation and integrate the CLIP (Radford et al. 2021) text encoder as the final classifier.

RegionPLC. We utilize a model trained on the ScanNet (Dai et al. 2017) partition of B15/N4, where B15/N4 represents 15 base and 4 novel categories. We employ a sparse-convolution-based UNet (Graham, Engelcke, and Van Der Maaten 2018) of SparseUNet16 as the 3D encoder for semantic segmentation, leveraging the CLIP (Radford et al. 2021) text encoder as the final classifier.

B.2 Evaluation Protocol

We evaluate the OV-3D baselines on 312 scenes following the validation split of ScanNet200 (Rozenberszki et al. 2022) dataset. Our evaluation includes all 341 attributes across 8 linguistic aspects.

C Additional Experimental Results

C.1 Results of Radar Charts.

Figure B presents radar charts of the main results on our OpenScan benchmark. The charts compare performance across AP, AP$_{50}$, and AP$_{25}$ for OpenMask3D (Takmaz et al. 2023), SAI3D (Yin et al. 2024), MaskClustering (Yan et al. 2024), and Open3DIS (Nguyen et al. 2024). Our results demonstrate that Open3DIS achieves the strongest performance across eight linguistic aspects, particularly in terms of AP and AP$_{50}$. Meanwhile, MaskClustering exhibits competitive performance in AP$_{50}$ and AP$_{25}$, with notable strengths in the *synonym* and *requirement* aspects.

C.2 Results Without Query Templates.

In this paper, we adopt query templates (*e.g.*, "*this term is made of wood*") as the default experimental configuration. For template-free evaluation on the GOV-3D benchmark (*e.g.*, "*wood*"), detailed results are provided in Table A. 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 instance segmentation. Our experiments show that Open3DIS achieves the highest AP, AP$_{50}$, and AP$_{25}$ scores across every linguistic aspect. This performance aligns with its strong performance in the GOV-3D task with query templates.

C.3 Results of Visual Attributes

We present comparative visual attribute results for the *material* aspect on our OpenScan benchmark in Table B. Our evaluation involves OpenMask3D (Takmaz et al. 2023), SAI3D (Yin et al. 2024), MaskClustering (Yan et al. 2024), and Open3DIS (Nguyen et al. 2024) across 10 *material* attributes. It demonstrates that these OV-3D models perform strongly on the "porcelain" material, indicating that the visual information of the "porcelain" material in 3D objects (*e.g.*, "toilet" and "bathtub") is more distinguishable than that of other materials. However, these OV-3D models struggle to accurately segment the "stone" material. This difficulty stems from the fact that stone is commonly associated with large 3D regions (*e.g.*, "wall" and "floor"), which are often neglected following the common practice (Schult et al. 2023; Takmaz et al. 2023; Yin et al. 2024) during 3D segmentation. These OV-3D models cannot correctly segment these large 3D areas, resulting in low results of the "stone" material. Notably, Open3DIS shows impressive results on each material compared to other OV-3D models, aligning with its strong performance in the classic OV-3D task.

C.4 Results of Upper Bound

During the 3D prediction in the GOV-3D task, we query the attributes to obtain the attribute-related 3D mask predictions. For annotating the OpenScan benchmark, object classes are associated with corresponding attributes using the ConceptNet (Speer, Chin, and Havasi 2017) database. Conversely, each attribute query can also be associated with the corresponding object classes. Therefore, we can replace the attribute queries with the ground truth attribute-related object classes from ConceptNet to finalize 3D mask results. We exclude the *material* aspect since the related object classes of *material* in the ConceptNet database are limited. We serve this setting as our upper bound performance. Table C shows the comparison of baseline methods OpenMask3D (Takmaz et al. 2023), SAI3D (Yin et al. 2024), MaskClustering (Yan et al. 2024), and Open3DIS (Nguyen et al. 2024) with their upper bounds, highlighting significant performance gaps that underscore the potential for attribute-aware 3D reasoning.

C.5 Results of Introducing LLM for Attribute Understanding

In our failure case analysis in the main paper, we observe that the OV-3D model Open3DIS (Nguyen et al. 2024) can identify the object classes (*e.g.*, "piano") in the OV-3D task but fails to recognize the associated object attributes (*e.g.*, "this term has 88 keys") in the GOV-3D task. This suggests a promising direction for improving GOV-3D performance by

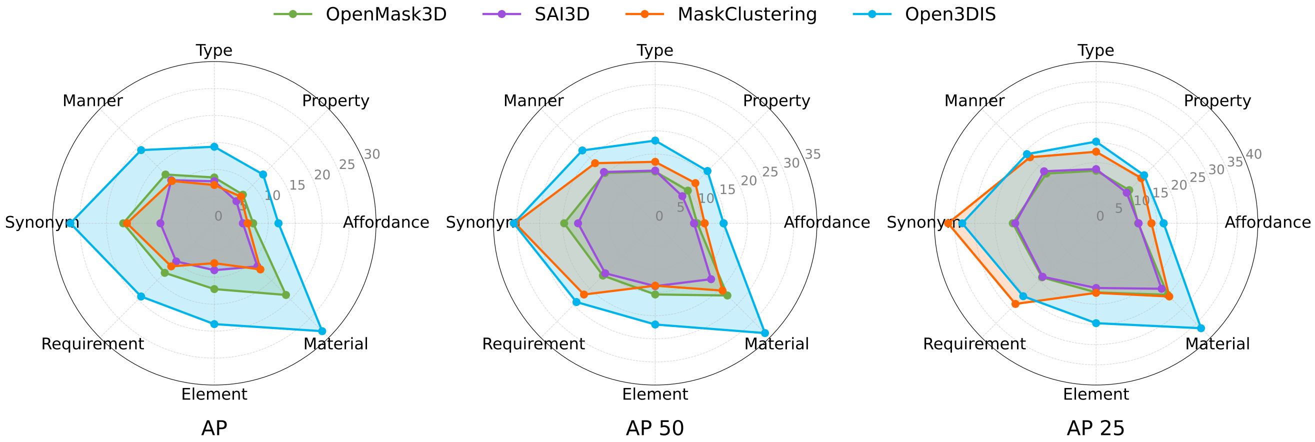


Figure B: Radar charts of AP, AP_{50}, and AP_{25} results for eight linguistic aspects on our OpenScan benchmark.

Method	Affordance	Property	Type	Manner	Synonym	Requirement	Element	Material	Mean
					AP				
OpenMask3D (Takmaz et al. 2023)	7.7	8.2	8.7	14.2	16.3	12.0	10.0	14.7	9.7
SAI3D (Yin et al. 2024)	4.0	5.5	7.4	10.6	8.6	8.8	7.3	8.3	6.7
MaskClustering (Yan et al. 2024)	6.0	6.2	4.1	8.0	11.7	10.3	9.3	8.1	6.8
Open3DIS (Nguyen et al. 2024)	**11.5**	**17.6**	**14.5**	**18.5**	**27.7**	**18.6**	**16.6**	**23.4**	**15.6**
					AP_{50}				
OpenMask3D (Takmaz et al. 2023)	9.6	10.7	11.2	18.1	18.4	15.2	12.8	17.8	12.2
SAI3D (Yin et al. 2024)	6.4	8.2	10.6	14.9	13.0	13.8	11.5	13.3	10.1
MaskClustering (Yan et al. 2024)	10.3	11.5	7.2	13.5	23.1	18.7	16.1	14.0	12.0
Open3DIS (Nguyen et al. 2024)	**14.3**	**22.2**	**18.3**	**22.6**	**32.9**	**23.1**	**19.7**	**28.5**	**19.2**
					AP_{25}				
OpenMask3D (Takmaz et al. 2023)	10.8	12.4	13.9	20.4	18.9	18.0	14.4	20.5	14.1
SAI3D (Yin et al. 2024)	8.5	11.1	13.1	18.6	15.9	17.5	14.7	18.6	12.8
MaskClustering (Yan et al. 2024)	12.3	13.5	9.4	16.9	25.1	24.1	19.0	18.1	14.6
Open3DIS (Nguyen et al. 2024)	**16.2**	**24.0**	**20.2**	**24.8**	**35.9**	**24.9**	**22.7**	**31.9**	**21.3**

Table A: 3D instance segmentation results without query template on our OpenScan benchmark.

leveraging large language models (LLMs) to perform high-level reasoning, transforming the GOV-3D attribute queries (*e.g.*, "this term has 88 keys") back to the OV-3D class queries (*e.g.*, "piano"). To this end, we design an experiment where an LLM (*i.e.*, Vicuna-7B (Chiang et al. 2023)) is prompted to map object attributes to corresponding object classes. Given an attribute [attribute], the LLM is prompted as:

Q: Given an object's attribute [attribute], please output the related object's classes in the indoor scene separated by commas.

The LLM will generate a list of object classes corresponding to the input attribute [attribute]. We then format the object classes into a sentence as a query for evaluation. We utilize the baseline methods OpenMask3D (Takmaz et al. 2023), SAI3D (Yin et al. 2024), MaskClustering (Yan et al. 2024), and Open3DIS (Nguyen et al. 2024) to compare their performance on whether introducing LLM for attribute understanding. As shown in Table D, introducing LLM can improve the attribute understanding performance across most linguistic aspects of the GOV-3D task. However, performance declines in the *material* aspect, as LLMs rely exclusively on linguistic inputs and lack visual context required to differentiate material properties (*e.g.*"wooden chair" and "plastic chair"). Additionally, the LLM's output can be noisy and inconsistent, occasionally producing object classes unrelated to the input attribute, which degrades performance in some linguist aspects.

C.6 Results of Weighted Mean

In the OpenScan benchmark, we observe disparities in the attribute annotations for linguistic aspects (*e.g.*, "affordance" and "synonym"). To address the imbalance in attribute annotations, we introduce a weighted mean score (w-Mean) metric that normalizes contributions based on annotation counts. For linguist aspects with annotation counts $L = \{l_k\}_{k=1}^{H}$ and cor-

Method	Wood	Fabric	Leather	Cotton	Metal	Stone	Porcelain	Plastic	Glass	Paper	Mean
						AP					
OpenMask3D (Takmaz et al. 2023)	19.1	12.7	28.0	26.5	9.1	0.1	41.8	16.9	23.1	10.7	18.8
SAI3D (Yin et al. 2024)	13.1	10.3	19.2	5.6	6.5	0.1	19.7	12.8	14.9	10.4	11.3
MaskClustering (Yan et al. 2024)	12.8	18.7	26.3	10.4	6.3	0.3	25.1	13.2	4.8	3.5	12.1
Open3DIS (Nguyen et al. 2024)	**32.9**	**27.1**	**35.4**	**33.0**	**24.6**	**2.4**	**43.1**	**33.8**	**29.9**	**20.9**	**28.3**
						AP$_{50}$					
OpenMask3D (Takmaz et al. 2023)	23.9	16.0	30.4	30.3	11.6	0.1	44.5	20.0	29.7	15.2	22.1
SAI3D (Yin et al. 2024)	19.6	15.7	26.9	9.5	10.2	0.1	31.8	18.4	22.6	16.2	17.1
MaskClustering (Yan et al. 2024)	23.8	31.6	**41.0**	19.7	12.0	0.5	37.0	23.2	9.3	7.6	20.6
Open3DIS (Nguyen et al. 2024)	**40.9**	**32.7**	39.0	**38.6**	**30.7**	**3.5**	**46.3**	**38.8**	**37.6**	**27.8**	**33.6**
						AP$_{25}$					
OpenMask3D (Takmaz et al. 2023)	27.4	19.1	32.5	33.5	13.3	0.1	47.0	21.7	34.8	21.4	25.0
SAI3D (Yin et al. 2024)	26.4	20.5	32.2	18.5	13.7	0.2	41.0	23.4	31.4	22.0	22.9
MaskClustering (Yan et al. 2024)	31.3	**38.4**	**49.7**	23.0	16.7	0.6	41.6	28.2	16.1	10.9	25.6
Open3DIS (Nguyen et al. 2024)	**44.7**	35.3	42.6	**42.2**	**33.5**	**5.1**	**48.3**	**42.4**	**41.8**	**31.6**	**36.7**

Table B: 3D instance segmentation results for the *material* aspect on our OpenScan benchmark.

Method	Affordance UpB ✗ / ✓	Property UpB ✗ / ✓	Type UpB ✗ / ✓	Manner UpB ✗ / ✓	Synonym UpB ✗ / ✓	Requirement UpB ✗ / ✓	Element UpB ✗ / ✓	Mean UpB ✗ / ✓
				AP				
OpenMask3D	7.2 / 19.3 $_{+12.1}$	7.5 / 26.7 $_{+19.2}$	8.5 / 18.5 $_{+10.0}$	12.8 / 27.7 $_{+14.9}$	16.9 / 29.4 $_{+12.5}$	13.0 / 26.5 $_{+13.5}$	12.2 / 22.1 $_{+9.9}$	9.7 / 21.6 $_{+11.9}$
SAI3D	5.3 / 14.6 $_{+9.3}$	5.8 / 17.7 $_{+11.9}$	7.8 / 17.4 $_{+9.6}$	11.3 / 18.9 $_{+7.6}$	10.0 / 19.3 $_{+9.3}$	10.0 / 19.2 $_{+9.2}$	8.7 / 16.8 $_{+8.1}$	7.6 / 16.8 $_{+9.2}$
MaskClustering	6.2 / 15.4 $_{+9.2}$	7.0 / 21.1 $_{+14.1}$	7.1 / 13.9 $_{+6.8}$	11.1 / 18.0 $_{+6.9}$	16.2 / 18.0 $_{+1.8}$	11.3 / 26.3 $_{+15.0}$	7.4 / 17.6 $_{+10.2}$	7.9 / 16.9 $_{+9.0}$
Open3DIS	**11.9 / 23.9** $_{+12.0}$	**12.8 / 36.0** $_{+23.2}$	**14.2 / 24.5** $_{+10.3}$	**19.2 / 34.9** $_{+15.7}$	**26.7 / 36.1** $_{+9.4}$	**19.2 / 35.0** $_{+15.8}$	**18.7 / 28.0** $_{+9.3}$	**15.4 / 27.7** $_{+12.3}$
				AP$_{50}$				
OpenMask3D	9.1 / 24.3 $_{+15.2}$	10.0 / 34.2 $_{+24.2}$	11.2 / 24.3 $_{+13.1}$	15.4 / 35.1 $_{+19.7}$	19.7 / 34.7 $_{+15.0}$	16.0 / 34.4 $_{+18.4}$	15.4 / 27.2 $_{+11.8}$	12.2 / 27.4 $_{+15.2}$
SAI3D	8.4 / 22.0 $_{+13.6}$	8.3 / 27.7 $_{+19.4}$	11.4 / 26.0 $_{+14.6}$	15.7 / 28.4 $_{+12.7}$	16.7 / 28.3 $_{+11.6}$	15.3 / 30.1 $_{+14.8}$	13.6 / 25.1 $_{+11.5}$	11.5 / 25.4 $_{+13.9}$
MaskClustering	10.7 / 28.6 $_{+17.9}$	12.3 / 38.5 $_{+26.2}$	13.3 / 26.8 $_{+13.5}$	18.4 / 33.5 $_{+15.1}$	30.3 / 34.2 $_{+3.9}$	21.8 / **49.7** $_{+27.9}$	13.5 / 32.5 $_{+19.0}$	14.4 / 31.6 $_{+17.2}$
Open3DIS	**14.8 / 29.9** $_{+15.1}$	**16.0 / 43.8** $_{+27.8}$	**17.9 / 30.6** $_{+12.7}$	**22.3 / 43.1** $_{+20.8}$	**30.6 / 42.6** $_{+12.0}$	**24.1** / 43.8 $_{+19.7}$	**21.9 / 33.5** $_{11.6}$	**18.9 / 34.1** $_{+15.2}$
				AP$_{25}$				
OpenMask3D	10.4 / 27.6 $_{+17.2}$	11.6 / 37.8 $_{+26.2}$	13.0 / 27.4 $_{+14.4}$	17.4 / 38.6 $_{+21.2}$	20.6 / 37.4 $_{+16.8}$	18.9 / 39.4 $_{+20.5}$	17.1 / 31.2 $_{+14.1}$	13.9 / 30.9 $_{+17.0}$
SAI3D	10.5 / 28.3 $_{+17.8}$	10.7 / 35.4 $_{+24.7}$	13.4 / 33.2 $_{+19.8}$	18.2 / 36.0 $_{+17.8}$	20.0 / 32.9 $_{+12.9}$	18.7 / 39.5 $_{+20.8}$	16.0 / 32.4 $_{+16.4}$	13.8 / 32.4 $_{+18.6}$
MaskClustering	13.7 / **37.2** $_{+23.5}$	15.8 / **48.8** $_{+33.0}$	17.7 / **35.0** $_{+17.3}$	23.1 / 44.8 $_{+21.7}$	**36.6 / 45.0** $_{+8.4}$	**28.2 / 61.6** $_{+33.4}$	17.2 / **42.7** $_{+25.5}$	18.5 / **41.0** $_{+22.5}$
Open3DIS	**16.7** / 32.7 $_{+16.0}$	**16.8** / 47.1 $_{+30.3}$	**20.2** / 34.5 $_{+14.3}$	**24.2 / 46.0** $_{+21.8}$	33.1 / 44.9 $_{+11.8}$	25.5 / 47.2 $_{+21.7}$	**24.7** / 38.6 $_{+13.9}$	**20.9** / 37.6 $_{+16.7}$

Table C: 3D instance segmentation results with the upper bound on our OpenScan benchmark, where "UpB" denotes upper bound.

responding scores $S = \{s_k\}_{k=1}^{H}$, the w-Mean is computed as:

$$\text{w-Mean} = \frac{\sum_{k=1}^{H} s_k l_k}{\sum_{k=1}^{H} l_k} \tag{1}$$

As shown in Table E, applying the w-Mean metric improves the performance of baseline methods, including OpenMask3D (Takmaz et al. 2023), SAI3D (Yin et al. 2024), MaskClustering (Yan et al. 2024), and Open3DIS (Nguyen et al. 2024). This improvement steams from the w-Mean metric's ability to normalize the contribution of each linguistic aspect based on its annotation count, thus mitigating biases from attributes and enhancing the robustness of performance evaluation across linguistic aspects in the GOV-3D task.

C.7 Additional Failure Cases Analysis

As illustrated in Figure C, the Open3DIS model(Nguyen et al. 2024) for OV-3D exhibits limitations in the GOV-3D task under specific conditions. Specifically, the model struggles to generate accurate 3D masks when: (a) the attribute query requires complex commonsense knowledge (*e.g.*, "this term requires water and sun"), resulting in failure to predict 3D masks; (b) the target 3D object contains noisy geometry, such as 3D holes or irregular 3D structures, leading to partially incorrect 3D masks; and (c) the target object is small, providing insufficient geometric detail for segmentation, causing the model to fail in predicting 3D masks. In contrast, Open3DIS correctly predicts 3D masks for attribute-related class queries in the OV-3D task under these scenarios, underscoring the challenge of the GOV-3D task.

Method	Affordance LLM ✗ / ✓	Property LLM ✗ / ✓	Type LLM ✗ / ✓	Manner LLM ✗ / ✓	Synonym LLM ✗ / ✓	Requirement LLM ✗ / ✓	Element LLM ✗ / ✓	Material LLM ✗ / ✓	Mean LLM ✗ / ✓
					AP				
OpenMask3D	7.2 / 10.8 $_{+3.6}$	7.5 / 13.9 $_{+6.4}$	8.5 / 8.5 $_{+0}$	12.8 / 14.1 $_{+1.3}$	16.9 / 25.7 $_{+8.8}$	13.0 / 13.4 $_{+0.4}$	12.2 / 12.1 $_{-0.1}$	18.8 / 10.7 $_{-8.1}$	9.9 / 11.7 $_{+1.8}$
SAI3D	5.3 / 7.2 $_{+1.9}$	5.8 / 9.9 $_{+4.1}$	7.8 / 7.5 $_{-0.3}$	11.3 / 14.5 $_{+3.2}$	10.0 / 16.1 $_{+6.1}$	10.0 / 9.3 $_{-0.7}$	8.7 / 7.9 $_{-0.8}$	11.3 / 6.6 $_{-4.7}$	7.7 / 8.6 $_{+0.9}$
MaskClustering	6.2 / 8.2 $_{+2.0}$	7.0 / 8.5 $_{+1.5}$	7.1 / 7.9 $_{+0.8}$	11.1 / 9.1 $_{-2.0}$	16.2 / 11.1 $_{-5.1}$	11.3 / 16.1 $_{+4.8}$	7.4 / 11.6 $_{+4.2}$	12.1 / 8.9 $_{-3.2}$	8.1 / 9.5 $_{+1.4}$
Open3DIS	**11.9 / 15.1** $_{+3.2}$	**12.8 / 22.7** $_{+9.9}$	**14.2 / 15.4** $_{+1.2}$	**19.2 / 21.6** $_{+2.4}$	**26.7 / 31.0** $_{+4.3}$	**19.2 / 21.3** $_{+2.1}$	**18.7 / 17.6** $_{-1.1}$	**28.3 / 18.7** $_{-9.6}$	**15.8 / 17.8** $_{+2.0}$
					AP$_{50}$				
OpenMask3D	9.1 / 13.5 $_{+4.4}$	10.0 / 18.7 $_{+8.7}$	11.2 / 11.0 $_{-0.2}$	15.4 / 17.9 $_{+2.5}$	19.7 / 30.7 $_{+11.0}$	16.0 / 17.3 $_{+1.3}$	15.4 / 15.0 $_{-0.4}$	22.1 / 12.6 $_{-9.5}$	12.5 / 14.7 $_{+2.2}$
SAI3D	8.4 / 10.4 $_{+2.0}$	8.3 / 15.7 $_{+7.4}$	11.4 / 10.9 $_{-0.5}$	15.7 / 20.7 $_{+5.0}$	16.7 / 24.0 $_{+7.3}$	15.3 / 15.1 $_{-0.2}$	13.6 / 12.6 $_{-1.0}$	17.1 / 10.0 $_{-7.1}$	11.6 / 12.8 $_{+1.2}$
MaskClustering	10.7 / 14.3 $_{+3.6}$	12.3 / 16.3 $_{+4.0}$	13.3 / 15.1 $_{+1.8}$	18.4 / 16.1 $_{-2.3}$	30.3 / 21.9 $_{-8.4}$	21.8 / **30.5** $_{+8.7}$	13.5 / **21.2** $_{+7.7}$	20.6 / 15.7 $_{-4.9}$	14.6 / 17.5 $_{+2.9}$
Open3DIS	**14.8 / 18.7** $_{+3.9}$	**16.0 / 28.6** $_{+12.6}$	**17.9 / 18.8** $_{+0.9}$	**22.3 / 26.2** $_{+3.9}$	**30.6 / 36.9** $_{+6.3}$	**24.1** / 26.7 $_{+2.6}$	**21.9** / 21.0 $_{-0.9}$	**33.6 / 21.8** $_{-11.8}$	**19.3 / 21.7** $_{+2.4}$
					AP$_{25}$				
OpenMask3D	10.4 / 15.3 $_{+4.9}$	11.6 / 21.1 $_{+9.5}$	13.0 / 14.3 $_{+1.3}$	17.4 / 19.9 $_{+2.5}$	20.6 / 33.2 $_{+12.6}$	18.9 / 20.3 $_{+1.4}$	17.1 / 17.0 $_{-0.1}$	25.0 / 14.1 $_{-10.9}$	14.2 / 17.1 $_{+2.9}$
SAI3D	10.5 / 13.2 $_{+2.7}$	10.7 / 20.2 $_{+9.5}$	13.4 / 14.6 $_{+1.2}$	18.2 / 23.3 $_{+5.1}$	20.0 / 28.3 $_{+8.3}$	18.7 / 18.9 $_{+0.2}$	16.0 / 15.8 $_{-0.2}$	22.9 / 13.4 $_{-9.5}$	14.1 / 16.2 $_{+2.1}$
MaskClustering	13.7 / 17.7 $_{+4.0}$	15.8 / 20.4 $_{+4.6}$	17.7 / 18.5 $_{+0.8}$	23.1 / 22.4 $_{-0.7}$	**36.6** / 26.7 $_{-9.9}$	**28.2 / 36.0** $_{+7.8}$	17.2 / **25.5** $_{+8.3}$	25.6 / 19.9 $_{-5.7}$	18.7 / 21.5 $_{+2.8}$
Open3DIS	**16.7 / 20.4** $_{+3.7}$	**16.8 / 30.2** $_{+13.4}$	**20.2 / 20.4** $_{+0.2}$	**24.2 / 28.2** $_{+4.0}$	33.1 / **39.2** $_{+6.1}$	25.5 / 28.5 $_{+3.0}$	**24.7** / 23.4 $_{-1.3}$	**36.7 / 24.3** $_{-12.4}$	**21.4 / 23.6** $_{+2.2}$

Table D: 3D instance segmentation results with LLM for attribute understanding on our OpenScan benchmark.

Method	Mean			w-Mean		
	AP	AP$_{50}$	AP$_{25}$	AP	AP$_{50}$	AP$_{25}$
OpenMask3D	9.9	12.5	14.2	12.3	15.0	17.1
SAI3D	7.7	11.6	14.1	8.6	13.0	16.4
MaskClustering	8.1	14.6	18.7	9.0	16.0	20.3
Open3DIS	**15.8**	**19.3**	**21.4**	**19.1**	**23.1**	**25.5**

Table E: 3D instance segmentation results for weighted-mean score (w-Mean) on our OpenScan benchmark.

D Additional Benchmark Details

D.1 Does OpenScan Represent 200 Object Classes From ScanNet200 Well Enough?

During the annotation of our OpenScan benchmark, object classes from ScanNet200 (Rozenberszki et al. 2022) are labeled with attributes using the ConceptNet (Speer, Chin, and Havasi 2017) database and manual annotation. Figure D shows the number of attributes per object class from ScanNet200 in our OpenScan benchmark. Notably, most object classes from ScanNet200 are annotated with more than one attribute in our OpenScan, indicating that our OpenScan benchmark adequately represents object classes from ScanNet200. Besides, the object class has up to seven attributes (*i.e.*, "*bicycle*", and "*ball*") in our OpenScan benchmark.

D.2 Number of Attributes per Object.

Figure E summarizes the distribution of attributes per object in our OpenScan benchmark. The majority of objects have 1–6 attributes.

D.3 Number of Attributes per Scene.

Figure F presents the distribution of attributes per scene in our OpenScan benchmark. It demonstrates that the attributes in 3D scenes are semantically rich.

D.4 Attribute Verification in Benchmark Annotation

During the attribute annotation process, we leverage a knowledge graph to automatically generate object-related attributes. However, the initial attribute set often contains noise, including attributes that are irrelevant, ambiguous, or semantically inconsistent with the related object classes. To address this, we conduct a meticulous manual verification process to refine the attribute set, ensuring semantic consistency and coherence in our OpenScan benchmark.

As shown in Figure G, our OpenScan benchmark generates 528 attributes initially. After attribute verification, 341 attributes are retained, resulting in an overall reduction of 35.4% of noisy attributes. Notably, the *affordance* aspect exhibits high noise level, with 42.5% of attributes being filtered out, suggesting that the *affordance* attributes are particularly prone to ambiguity due to the diverse nature of affordance candidates within the knowledge graph. In contrast, all *synonym* attributes are retained during verification. This robustness is attributed to the high semantic similarity between synonym attributes and their corresponding object classes, ensuring reliable alignment during the generation process.

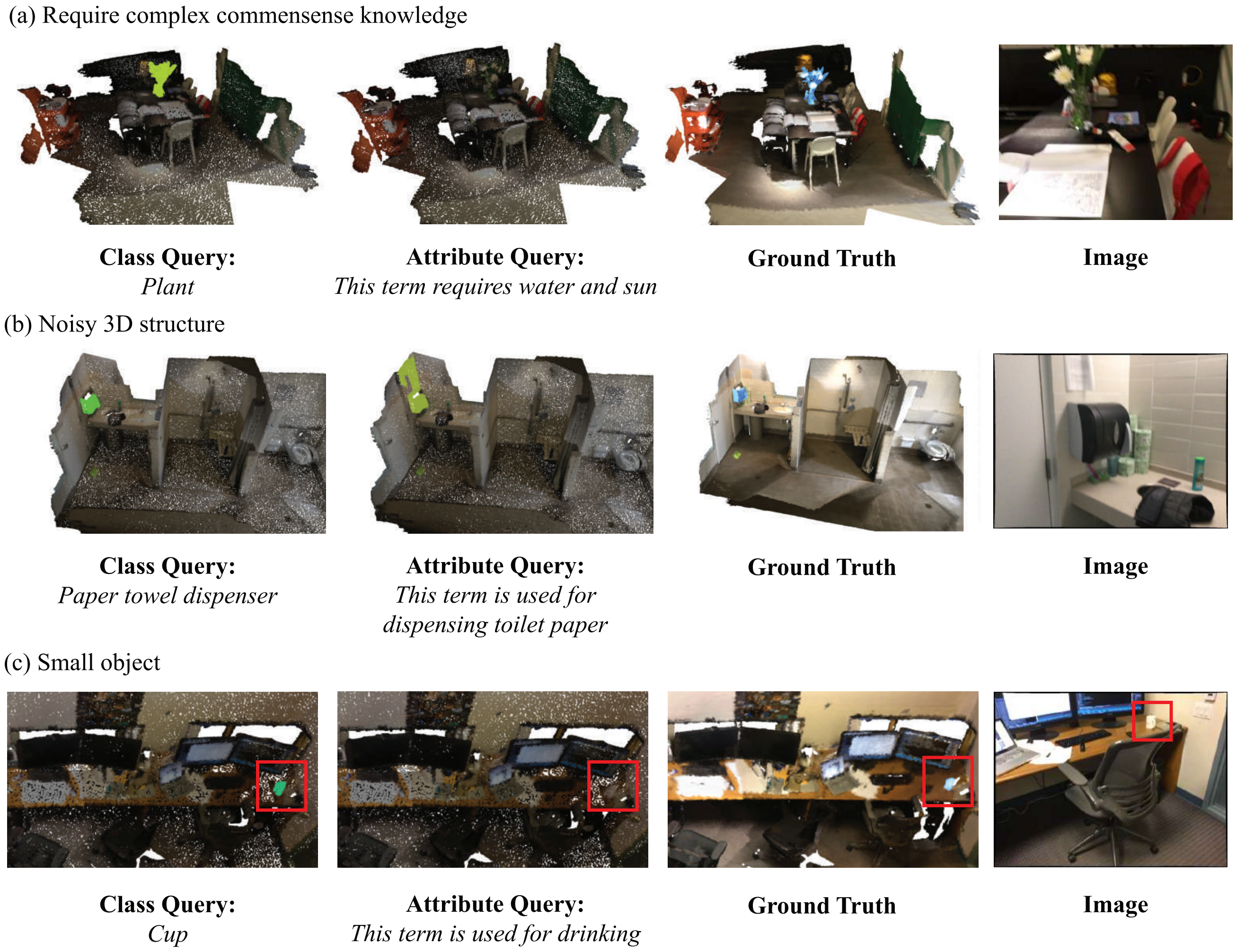
D.5 Additional Benchmark Samples

We provide additional samples of our OpenScan benchmark. Figure H presents the examples of objects and their corresponding attributes. Figure J displays the *affordance*, *property*, *type*, and *manner* aspects, while Figure K shows the *synonym*, *requirement*, *element*, and *material* aspects.

D.6 Benchmark Formats

Figure I shows an example of our OpenScan benchmark formats. Our OpenScan is formatted in the JSON file. Each target 3D object is annotated with the following items:

- ***Scene ID***: indicates the scene in which the target object is located.



(a) Require complex commensense knowledge

Class Query: *Plant*

Attribute Query: *This term requires water and sun*

Ground Truth

Image

(b) Noisy 3D structure

Class Query: *Paper towel dispenser*

Attribute Query: *This term is used for dispensing toilet paper*

Ground Truth

Image

(c) Small object

Class Query: *Cup*

Attribute Query: *This term is used for drinking*

Ground Truth

Image

Figure C: Visualization of the Open3DIS failure cases. The ground truth objects and outputs are highlighted in color. Best view with zoom in.

- ***Object ID***: identifies the target object's unique ID within the scene.
- ***Object Name***: specifies the object class of the target object.

In addition, each object is annotated with eight linguistic aspects (*affordance*, *property*, *type*, *manner*, *synonym*, *requirement*, *element*, and *material*). If the target 3D object does not contain an attribute of a specific linguistic aspect, it is marked as "other".

D.7 Benchmark Details

We construct our OpenScan benchmark based on ScanNet200 (Rozenberszki et al. 2022) across eight linguistic aspects. We present all attributes and their corresponding query templates in our OpenScan benchmark: Table F displays the *affordance* and *property* aspects; Table G shows the *type*, *manner*, and *synonym* aspects; and Table H presents the *requirement*, *element*, and *material* aspects. The "object" in the query template is replaced with "this term" in our experiments.

E Additional Related Work

Open-Vocabulary 2D Understanding Benchmarks. Open-vocabulary 2D understanding refers to the task of detecting or segmenting novel object classes that are not present in the training dataset. For the object detection task, COCO (Lin et al. 2014) and LVIS (Gupta, Dollar, and Girshick 2019) are two widely used datasets. For the image segmentation task, popular datasets include COCO (Lin et al. 2014), ADE20k (Zhou et al. 2019), PASCAL-VOC (Everingham et al. 2015), and Cityscapes (Cordts et al. 2016). However, these benchmarks primarily evaluate the model's open-

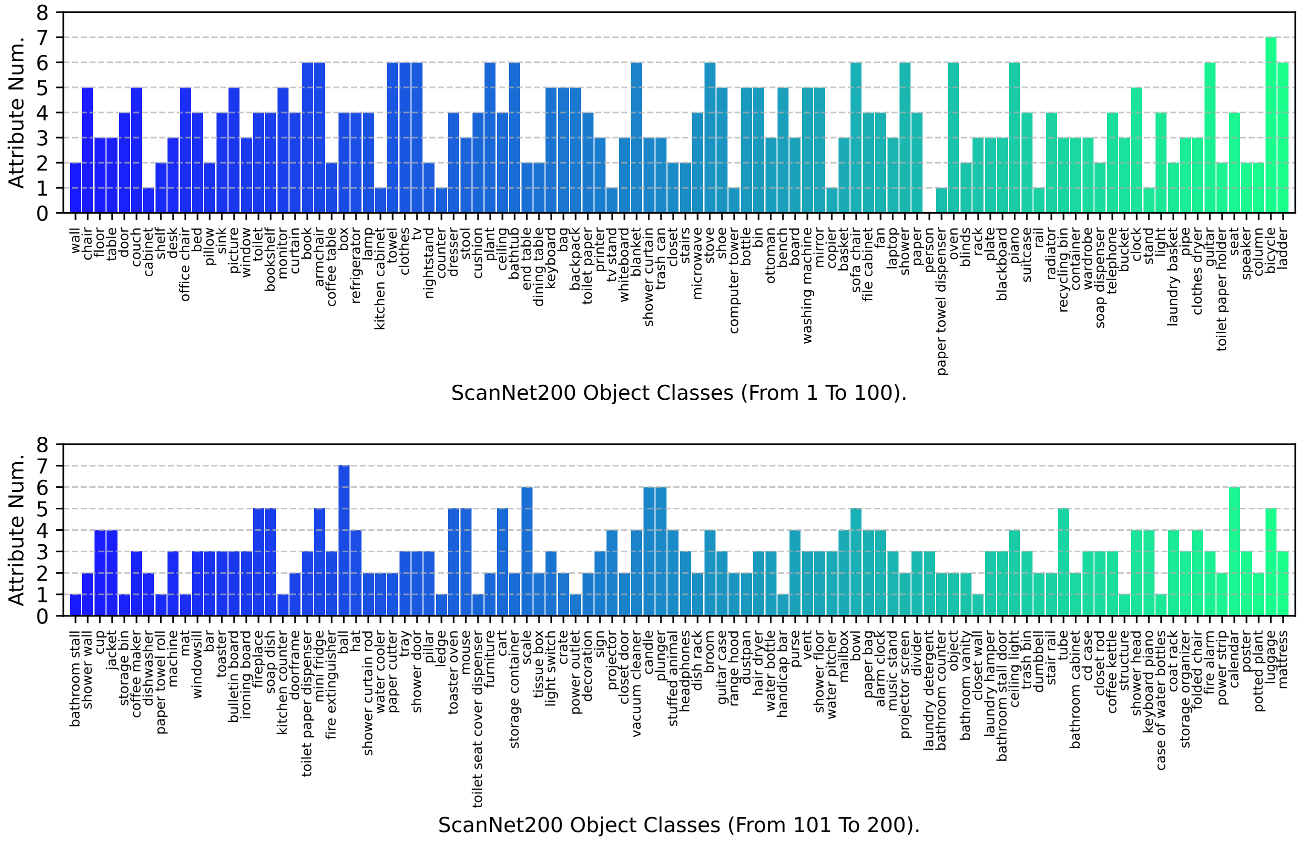


Figure D: Number of attributes per object class from ScanNet200 in our OpenScan benchmark.

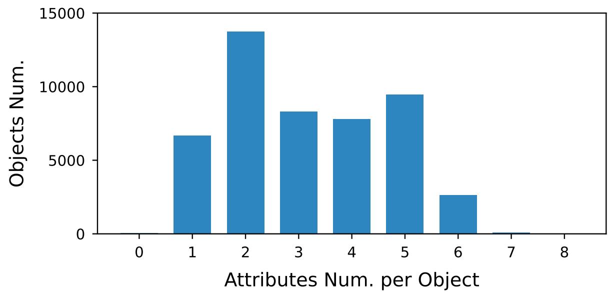


Figure E: Number of attributes per object in our OpenScan benchmark and corresponding number of objects.

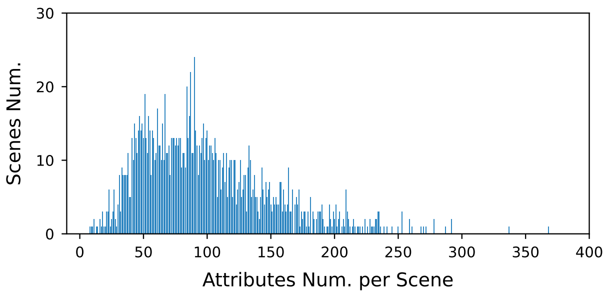


Figure F: Number of attributes per scene in our OpenScan benchmark and corresponding number of scenes.

vocabulary ability but do not explicitly assess its capability to recognize specific object characteristics. PACO (Ramanathan et al. 2023) introduces a 2D segmentation benchmark that focuses on parts and attributes of common objects. Inspired by PACO (Ramanathan et al. 2023), FG-OVD (Bianchi et al. 2024) presents a challenging task and benchmark for fine-grained open-vocabulary object detection to evaluate the ability of open-vocabulary detectors to discern extrinsic object properties. Similarly, OVDEval (Yao et al. 2024) introduces an open-vocabulary detection benchmark to evaluate the performance on linguistic aspects using complex language prompts. Our work is different from them (Ramanathan et al. 2023; Bianchi et al. 2024; Yao et al. 2024) since we focus on the understanding of object attributes on 3D data, which poses greater challenges compared to the understanding in 2D images due to the limited annotations in 3D benchmarks.

F Limitations and Future Work

Our benchmark is currently constructed solely on the ScanNet200 benchmark with limited 3D indoor scene. It would

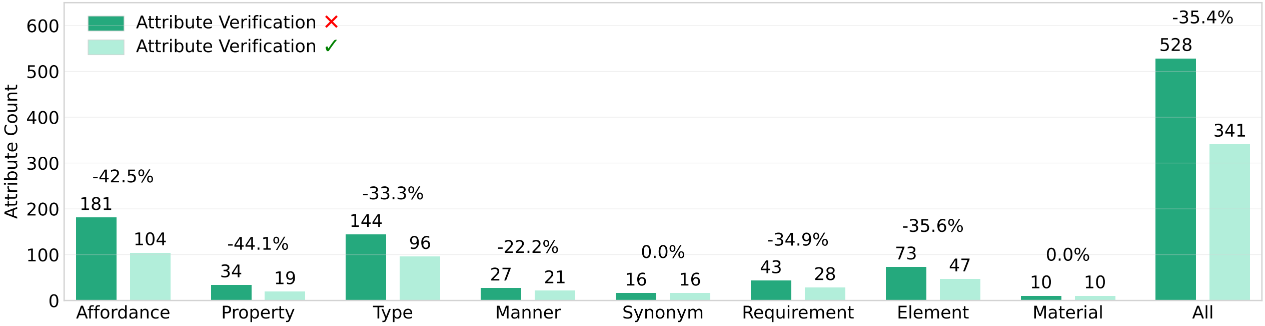


Figure G: OpenScan benchmark statistics of attributes during attribute verification.

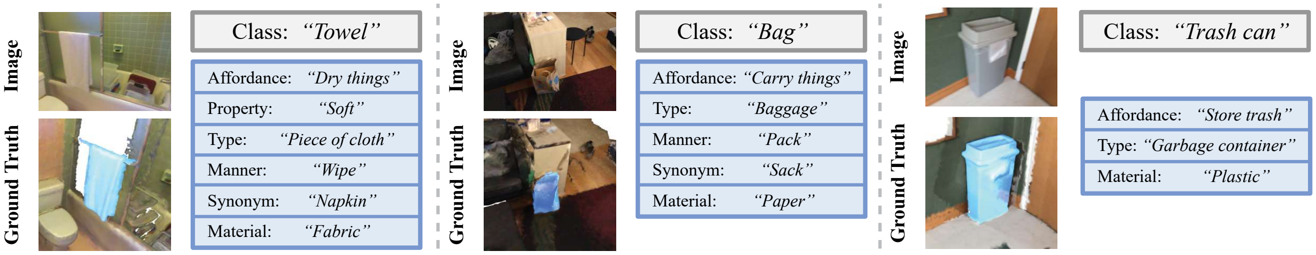
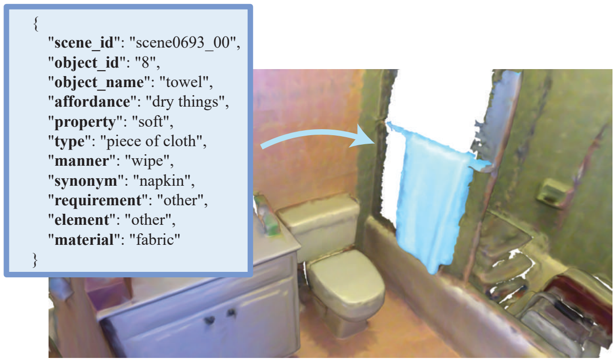


Figure H: Examples of objects and corresponding attributes in our OpenScan benchmark.



{
"**scene_id**": "scene0693_00",
"**object_id**": "8",
"**object_name**": "towel",
"**affordance**": "dry things",
"**property**": "soft",
"**type**": "piece of cloth",
"**manner**": "wipe",
"**synonym**": "napkin",
"**requirement**": "other",
"**element**": "other",
"**material**": "fabric"
}

Figure I: OpenScan benchmark format. The target object is highlighted in blue.

be beneficial to increase the scale of our benchmark to include a wider variety of 3D scenes and objects. In future work, we plan to extend our OpenScan benchmark to encompass more diverse scenes by incorporating indoor 3D datasets such as ScanNet++ (Yeshwanth et al. 2023) and Matterport3D (Chang et al. 2017). Our mature annotation procedures can be readily adapted to these datasets. Moreover, we aim to evaluate current OV-3D models on our GOV-3D task, particularly examining performance variations when using higher point resolutions in ScanNet++ (Yeshwanth et al. 2023) and larger scene areas in Matterport3D (Chang et al. 2017).

G Broader Impact

Our approach does not introduce any negative societal impacts. All experiments are performed on publicly available datasets, with no use of private data. Although our benchmark is constructed exclusively from public data, we recognize the potential for unintended consequences if the data is applied without appropriate safeguards. We urge readers to ensure that the application of this research remains lawful and ethical, strictly adhering to established regulations and guidelines.

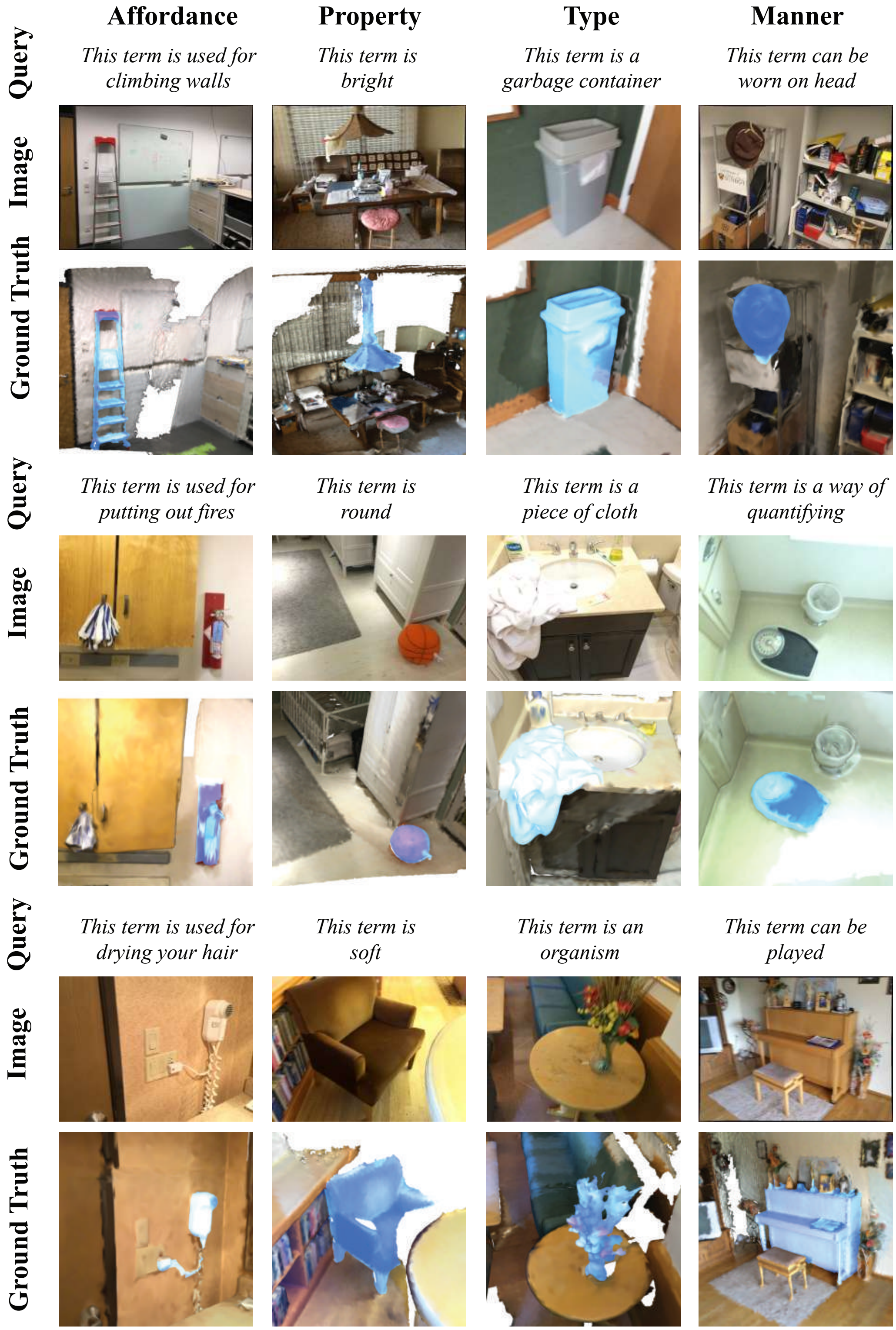

Affordance
Property
Type
Manner
Query
This term is used for climbing walls
This term is bright
This term is a garbage container
This term can be worn on head
Image
Ground Truth
Query
This term is used for putting out fires
This term is round
This term is a piece of cloth
This term is a way of quantifying
Image
Ground Truth
Query
This term is used for drying your hair
This term is soft
This term is an organism
This term can be played
Image
Ground Truth

Figure J: Additional OpenScan benchmark samples of *affordance*, *property*, *type*, and *manner* aspects. Target objects are highlighted in blue.

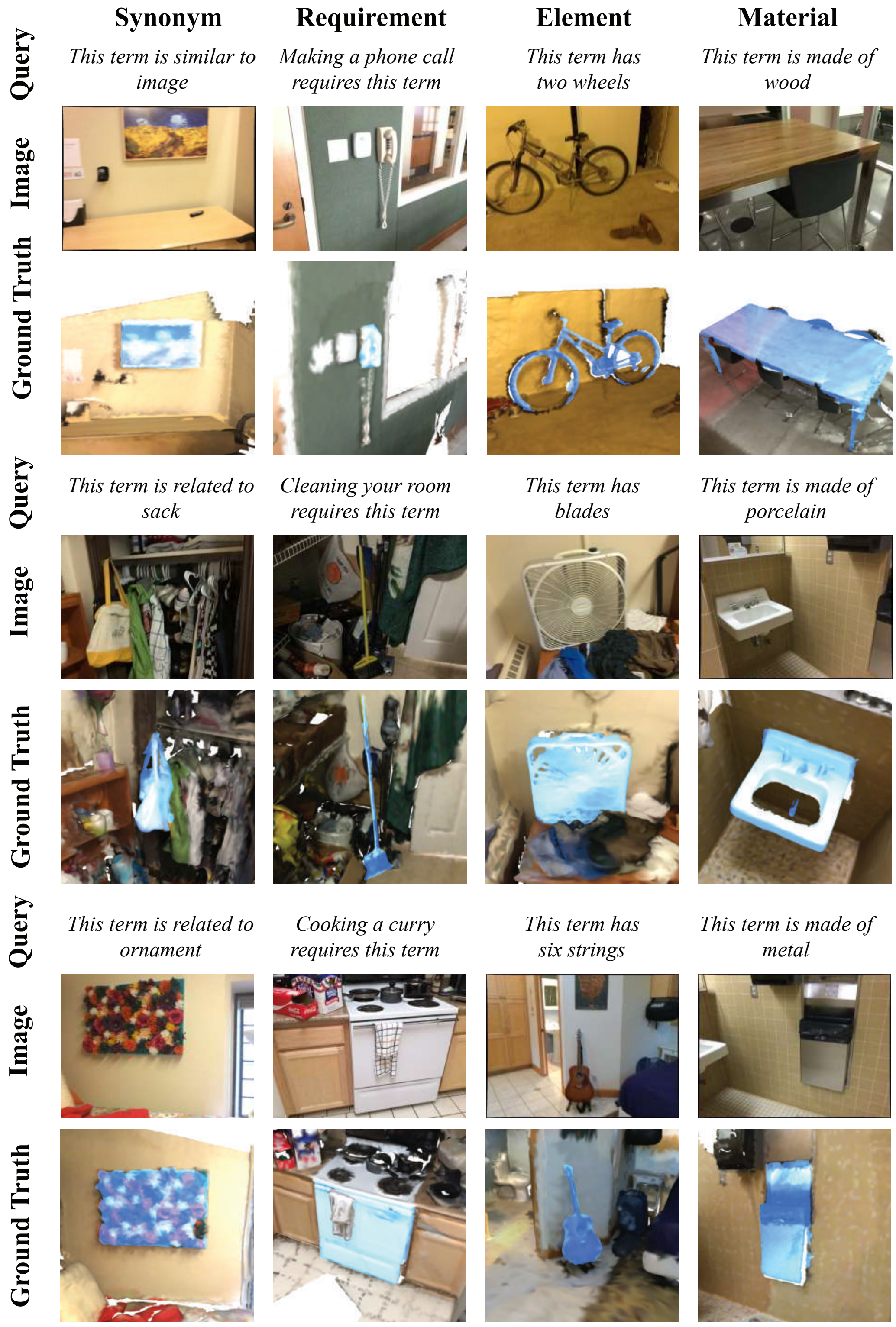

Synonym
Requirement
Element
Material
Query
This term is similar to image
Making a phone call requires this term
This term has two wheels
This term is made of wood
Image
Ground Truth
Query
This term is related to sack
Cleaning your room requires this term
This term has blades
This term is made of porcelain
Image
Ground Truth
Query
This term is related to ornament
Cooking a curry requires this term
This term has six strings
This term is made of metal
Image
Ground Truth

Figure K: Additional OpenScan benchmark samples of *synonym*, *requirement*, *element*, and *material* aspects. Target objects are highlighted in blue.

	Attribute	**Template**	**Attribute**	**Template**
Affordance	carry things	[object] is used for carrying things	holding up a roof	[object] is used for holding up a roof
	rest	[object] is used for resting	sit	[object] is used for sitting
	keep food cold	[object] is used for keeping food cold	place coffee	[object] is used for placing coffee
	work	[object] is used for working	look outside	[object] is used for looking outside
	bath	[object] is used for bathing	cover a window	[object] is used for covering a window
	stand	[object] is used for standing	wash dishes	[object] is used for washing dishes
	measure weight	[object] is used for measuring weight	store trash	[object] is used for storing trash
	display images	[object] is used for displaying images	sleep	[object] is used for sleeping
	poop	[object] is used for pooping	tell time	[object] is used for telling time
	bake toaster	[object] is used for baking toaster	perform music	[object] is used for performing music
	making toast	[object] is used for making toast	heat food	[object] is used for heating food
	separate rooms	[object] is used for separating rooms	close the top of a room	[object] is used for closing the top of a room
	ride	[object] is used for riding	store books	[object] is used for storing books
	see yourself	[object] is used for seeing yourself	store guitar	[object] is used for storing guitar
	dry things	[object] is used for drying things	put your feet on	[object] is used for putting your feet on
	storage dirty clothes	[object] is used for storaging dirty clothes	hold up the roof	[object] is used for holding up the roof
	represent	[object] is used for representing	hang clothes	[object] is used for hanging clothes
	heat the room	[object] is used for heating the room	make coffee	[object] is used for making coffee
	presenting information	[object] is used for presenting information	grow in a garden	[object] is used for growing in a garden
	cool a person	[object] is used for cooling a person	foot protection	[object] is used for foot protection
	heat a room	[object] is used for heating a room	illuminate an area	[object] is used for illuminating an area
	protecting your head	[object] is used for protecting your head	print documents	[object] is used for printing documents
	store liquids	[object] is used for storing liquids	keep out light from houses	[object] is used for keeping out light from houses
	transport things	[object] is used for transporting things	collect recyclable plastics	[object] is used for collecting recyclable plastics
	communicate	[object] is used for communicating	pack clothes for a trip	[object] is used for packing clothes for a trip
	carrying money	[object] is used for carrying money	wear	[object] is used for wearing
	learning	[object] is used for learning	store things	[object] is used for storing things
	carry liquids	[object] is used for carrying liquids	turn on a light	[object] is used for turning on a light
	write ideas and terms on	[object] is used for writing ideas and terms on	store file	[object] is used for storing file
	make a captured voice become audible	[object] is used for making a captured voice become audible	type	[object] is used for typing
	eat dinner	[object] is used for eating dinner	bake cookies	[object] is used for baking cookies
	furnish	[object] is used for furnishing	detect fire	[object] is used for detecting fire
	have privacy	[object] is used for having privacy	hold toilet paper	[object] is used for holding toilet paper
	blow your nose	[object] is used for blowing your nose	store water	[object] is used for storing water
	bounce	[object] is used for bouncing	cover a bed	[object] is used for covering a bed
	organize books	[object] is used for organizing books	hold trash	[object] is used for holding trash
	climb	[object] is used for climbing	store clothes	[object] is used for storing clothes
	drink	[object] is used for drinking	listen to music	[object] is used for listening to music
	hold sheet music	[object] is used for holding sheet music	unblocking a toilet	[object] is used for unblocking a toilet
	hang clothes	[object] is used for hanging clothes	entertain a child	[object] is used for entertaining a child
	control a computer	[object] is used for controlling a computer	dispense toilet paper	[object] is used for dispensing toilet paper
	keep clothes	[object] is used for keeping clothes	entry and exit to the shower	[object] is used for entry and exit to the shower
	climb walls	[object] is used for climbing walls	hold soap	[object] is used for holding soap
	hold things	[object] is used for holding things	get drunk	[object] is used for getting drunk
	putting out fires	[object] is used for putting out fires	carry something	[object] is used for carrying something
	hang coat	[object] is used for hanging coat	spray water	[object] is used for spraying water
	hold food	[object] is used for holding food	dry your hair	[object] is used for drying your hair
	show movies	[object] is used for showing movies	dry clothes	[object] is used for drying clothes
	wash clothes	[object] is used for washing clothes	mark that special date	[object] is used for marking that special date
	vacumming	[object] is used for vacumming	ironing clothes	[object] is used for ironing clothes
	decorating your room	[object] is used for decorating your room	sweeping	[object] is used for sweeping
	receiving letters	[object] is used for receiving letters	hold cd	[object] is used for holding cd
Property	useful for camping	[object] is useful for camping	soft	[object] is soft
	opaque and closed	[object] is opaque and closed	essential for privacy	[object] is essential for privacy
	helpful in making comparisons	[object] is helpful in making comparisons	analog or digital	[object] is analog or digital
	hot	[object] is hot	one kind of stringed instrument	[object] is one kind of stringed instrument
	open or closed	[object] is open or closed	horizontal	[object] is horizontal
	fun to ride	[object] is fun to ride	reflective	[object] is reflective
	alive	[object] is alive	bright	[object] is bright
	hollow	[object] is hollow	round	[object] is round
	useful for unblocking a toilet	[object] is useful for unblocking a toilet	shaped like a shell	[object] is shaped like a shell
	convex down	[object] is convex down		

Table F: OpenScan benchmark attributes of *affordance* and *property* aspects.

	Attribute	Template	Attribute	Template
Type	baggage	[object] is a baggage	seat	[object] is a seat
	table were someone works	[object] is a table were someone works	plumbing fixture	[object] is a plumbing fixture
	window covering	[object] is a window covering	land	[object] is a land
	measuring instrument	[object] is a measuring instrument	garbage container	[object] is a garbage container
	a way to relax	[object] is a way to relax	a good place to lie	[object] is a good place to lie
	vanity	[object] is a vanity	kitchen appliance	[object] is a kitchen appliance
	basket	[object] is a basket	box	[object] is a box
	string instrument	[object] is a string instrument	rack	[object] is a rack
	appliances	[object] is an appliance	movable barrier	[object] is a movable barrier
	upper surface	[object] is an upper surface	a two wheel vehicle	[object] is a two wheel vehicle
	reflector	[object] is a reflector	container	[object] is a container
	piece of cloth	[object] is a piece of cloth	representation	[object] is a representation
	clue	[object] is a clue	organism	[object] is an organism
	a cooling device	[object] is a cooling device	footwear	[object] is a footwear
	heater	[object] is a heater	source of illumination	[object] is a source of illumination
	a form of clothing	[object] is a form of clothing	refrigerator	[object] is a refrigerator
	dispenser	[object] is a dispenser	a long seat with no backrest	[object] is a long seat with no backrest
	a vehicle	[object] is a vehicle	bin	[object] is a bin
	a communication device	[object] is a communication device	handbag	[object] is a handbag
	coat	[object] is a coat	an excellent source of information	[object] is an excellent source of information
	tube	[object] is a tube	switch	[object] is a switch
	sill	[object] is a sill	door	[object] is a door
	board	[object] is a board	cabinet	[object] is a cabinet
	portable computer	[object] is a portable computer	display	[object] is a display
	computer device	[object] is a computer device	shaft	[object] is a shaft
	alarm	[object] is an alarm	curtain	[object] is a curtain
	paper	[object] is a paper	bottle	[object] is a bottle
	an instrument of music	[object] is an instrument of music	a toy	[object] is a toy
	bedclothes	[object] is a bedclothes	cutting implement	[object] is a cutting implement
	shelf	[object] is a shelf	table	[object] is a table
	supporter	[object] is a supporter	railing	[object] is a railing
	trophy	[object] is a trophy	audio device	[object] is an audio device
	vessel	[object] is a vessel	a tool to unclog toilets	[object] is a tool to unclog toilets
	rod	[object] is a rod	padding	[object] is a padding
	bag	[object] is a bag	toy animal	[object] is a toy animal
	a container for clothes	[object] is a container for clothes	hole	[object] is a hole
	stairs	[object] is a stairs	storage device	[object] is a storage device
	firefighting equipment	[object] is a firefighting equipment	fitness equipment	[object] is a fitness equipment
	device for spraying	[object] is a device for spraying	counter	[object] is a counter
	clock	[object] is a clock	kettle	[object] is a kettle
	hood	[object] is a hood	beauty device	[object] is a beauty device
	optical device	[object] is a optical device	dryer	[object] is a dryer
	detergent	[object] is a detergent	machine	[object] is a machine
	screen	[object] is a screen	drafting instrument	[object] is a drafting instrument
	pad	[object] is a pad	pitcher	[object] is a pitcher
	electronic piano	[object] is an electronic piano	time list	[object] is a time list
	household cleaning tool	[object] is a household cleaning tool	sign	[object] is a sign
	receptacle container	[object] is a receptacle container	a container for letters	[object] is a container for letters
Manner	pack	[object] is a way of packing	observe	[object] is a way of observing
	bathe	[object] is a way of bathing	quantify	[object] is a way of quantifying
	cook	[object] is a way of cooking	steered by handlebars	[object] can be steered by handlebars
	wipe	[object] is a way of wiping	wear	[object] is a way of wearing
	worn on a head	[object] can be worn on a head	transport things	[object] is a way of transporting things
	written on	[object] can be written on	played	[object] can be played
	played with	[object] can be played with	cover bed	[object] is a way of covering bed
	used in a toilet	[object] can be used in a toilet	manipulate computer	[object] is a way of manipulating computer
	climbed to reach some place high	[object] can be climbed to reach some place high	store	[object] is a way of storing
	produce	[object] is a way of producing	lit with a match	[object] can be lit with a match
	schedule	[object] is a way of scheduling		
Synonym	weight	[object] is related to weight	news	[object] is related to news
	bedside table	[object] is related to bedside table	napkin	[object] is related to napkin
	image	[object] is similar to image	sack	[object] is related to sack
	reading	[object] is related to reading	pipe	[object] is related to pipe
	power bar	[object] is related to power bar	round	[object] is related to round
	ornament	[object] is related to ornament	suction cup	[object] is similar to suction cup
	dress	[object] is related to dress	houseplant	[object] is related to houseplant
	suitcase	[object] is related to suitcase	almanac	[object] is related to almanac

Table G: OpenScan benchmark attributes of *type*, *manner*, and *synonym* aspects.

	Attribute	Template	Attribute	Template
Requirement	sit down	sitting down requires [object]	be unpluged	[object] does not desire to be unpluged
	have a bath	having a bath requires [object]	using a VCR	using a VCR requires [object]
	wake up in the morning	waking up in the morning requires [object]	playing a guitar	playing a guitar requires [object]
	balance to ride	[object] requires balance to ride	grooming	grooming requires [object]
	get warm	getting warm requires [object]	water and sun	[object] requires water and sun
	print	printing requires [object]	drink	drinking requires [object]
	buying food	buying food requires [object]	make a phone call	making a phone call requires [object]
	bring suit	bringing suit requires [object]	go on the internet	going on the internet requires [object]
	write	writing requires [object]	type	typing requires [object]
	cook a curry	cooking a curry requires [object]	play the piano	playing the piano requires [object]
	playing soccer	playing soccer requires [object]	eating breakfast in bed	eating breakfast in bed requires [object]
	paint a house	painting a house requires [object]	going on a vacation	going on a vacation requires [object]
	a goldfish	a goldfish requires [object]	washing your clothes	washing your clothes requires [object]
	cleaning clothing	cleaning clothing requires [object]	cleaning your room	cleaning your room requires [object]
Element	water	[object] has water	news	[object] has news
	urine	[object] has urine	twelve numbers	[object] has twelve numbers
	toaster	[object] has toaster	six strings	[object] has six strings
	two wheels	[object] has two wheels	doorway	doorway has [object]
	legs	[object] has legs	an art show	an art show has [object]
	fire	[object] has fire	ecosystem	ecosystem has [object]
	blades	[object] has blades	foot	[object] has foot
	heating system	heating system has [object]	money	[object] has money
	knowledge	[object] has knowledge	six sides	[object] has six sides
	circuit	circuit has [object]	window frame	window frame has [object]
	bathroom	bathroom has [object]	a document folder	[object] has a document folder
	screen	[object] has screen	keys	[object] has keys
	food	[object] has food	88 keys	[object] has 88 keys
	books	[object] has books	trash	[object] has trash
	tack	[object] has tack	clothes	[object] has clothes
	sofa	sofa has [object]	computer	computer has [object]
	toilet paper	[object] has toilet paper	air passage	air passage has [object]
	rundle	rundle has [object]	soap	[object] has soap
	beer	[object] has beer	a coat	[object] has a coat
	a shower stall	a shower stall has [object]	table	table has [object]
	the movies	the movies have [object]	clothing	[object] has clothing
	bed	bed has [object]	a wick	[object] has a wick
	the date	[object] has the date	mail	[object] has mail
	a cd	[object] has a cd		
Material	wood	[object] is made of wood	fabric	[object] is made of fabric
	leather	[object] is made of leather	cotton	[object] is made of cotton
	metal	[object] is made of metal	stone	[object] is made of stone
	porcelain	[object] is made of porcelain	plastic	[object] is made of plastic
	glass	[object] is made of glass	paper	[object] is made of paper

Table H: OpenScan benchmark attributes of *requirement*, *element*, and *material* aspects.

References

Bianchi, L.; Carrara, F.; Messina, N.; Gennaro, C.; and Falchi, F. 2024. The devil is in the fine-grained details: Evaluating open-vocabulary object detectors for fine-grained understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22520–22529.

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*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Choy, C.; Gwak, J.; and Savarese, S. 2019. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3075–3084.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3213–3223.

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, 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.

Ester, M.; Kriegel, H.-P.; Sander, J.; and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, 226–231. AAAI Press.

Everingham, M.; Eslami, S. A.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2015. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111: 98–136.

Ghiasi, G.; Gu, X.; Cui, Y.; and Lin, T.-Y. 2022. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, 540–557. Springer.

Graham, B.; Engelcke, M.; and Van Der Maaten, L. 2018. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9224–9232.

Gupta, A.; Dollar, P.; and Girshick, R. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5356–5364.

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.

Li, F.; Zhang, H.; Sun, P.; Zou, X.; Liu, S.; Li, C.; Yang, J.; Zhang, L.; and Gao, J. 2024. Segment and recognize anything at any granularity. In *European Conference on Computer Vision*, 467–484. Springer.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 740–755. Springer.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 38–55. Springer.

Lu, S.; Chang, H.; Jing, E. P.; Boularias, A.; and Bekris, K. 2023. Ovir-3d: Open-vocabulary 3d instance retrieval without training on 3d data. In *Conference on Robot Learning*, 1610–1620. PMLR.

Ngo, T. D.; Hua, B.-S.; and Nguyen, K. 2023. ISBNet: A 3D Point Cloud Instance Segmentation Network With Instance-Aware Sampling and Box-Aware Dynamic Convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13550–13559.

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.

Qi, L.; Kuen, J.; Shen, T.; Gu, J.; Li, W.; Guo, W.; Jia, J.; Lin, Z.; and Yang, M.-H. 2023. High Quality Entity Segmentation. In *2023 IEEE/CVF International Conference on Computer Vision*, 4024–4033.

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.

Ramanathan, V.; Kalia, A.; Petrovic, V.; Wen, Y.; Zheng, B.; Guo, B.; Wang, R.; Marquez, A.; Kovvuri, R.; Kadian, A.; et al. 2023. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7141–7151.

Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv:2401.14159*.

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.

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*.

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.

Yao, Y.; Liu, P.; Zhao, T.; Zhang, Q.; Liao, J.; Fang, C.; Lee, K.; and Wang, Q. 2024. How to Evaluate the Generalization of Detection? A Benchmark for Comprehensive Open-Vocabulary Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6630–6638.

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.

Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127: 302–321.