RegionPLC: Regional Point-Language Contrastive Learning for Open-World 3D Scene Understanding

Supplementary Material

Outline

In this supplementary file, we provide more experimental results and details not elaborated in our main paper due to page length limits:

- Sec. S1: Implementation details of RegionPLC and baseline methods.
- Sec. S2: Additional experimental results on per-class performance, zero-shot domain transfer results, and error-bar analysis.
- Sec. S3: Qualitative results of RegionPLC.
- Sec. S4: Prompts for LLM in RegionGR.
- Sec. S5: Limitation and open problems.

S1. Implementation Details

Here, we present the implementation details of our RegionPLC, dataset category partition and the implementation of baseline methods.

S1.1. Implementation Details of RegionPLC

Network Architecture. The network architecture of RegionPLC is the same as PLA [5] (*i.e.* SparseUNet16), ensuring a fair comparison on ScanNet [4]. For ScanNet200 [13], we augment the base hidden dimension (*i.e.* the hidden dimension of the bottleneck) of sparse UNet from 16 to 32 (*i.e.* SparseUNet32), yielding better performance on this complex, long-tail dataset. In nuScenes [2], we adopt the same backbone used in ScanNet200 with 5 residual blocks and a base hidden dimension of 32. In addition, in the base-annotated open-world setting, we follow PLA [5] to employ the binary encoder with binary loss, classification head with semantic segmentation loss and instance head with instance loss on base categories.

Training Schedule. We train 512 epochs on ScanNet and ScanNet200, and 50 epochs on nuScenes for semantic segmentation, and 640 epochs for ScanNet instance segmentation. The initial learning rate is set as 0.004 for ScanNet and ScanNet200 and 0.01 for nuScenes. The learning rate decays in cosine and polynomials on ScanNet and nuScenes, respectively.

Regional 3D-Language pairs. Regarding vision-language (VL) models that generate captions, we use OFA [15] for generating $\mathbf{t}^{\text{sw}}$ and $\mathbf{t}^{\text{det-c}}$ as the caption model. As for the object proposal in $\mathbf{t}^{\text{det-t}}$ and $\mathbf{t}^{\text{det-c}}$, we use Detic [18] with LVIS [7] vocabulary space. The prompt template of $\mathbf{t}^{\text{det-t}}$ is the same as CLIP [12]. Other VL foundation models are also feasible, and a more robust VL foundation model should enhance our methods' performance through providing higher-quality object proposals and language descriptions. By default, we use 125K frames in the ScanNet dataset and all images in the nuScenes dataset to extract their regional captions, respectively.

Inference Cost Analysis. In the annotation-free setting of the main paper, we evaluate the efficiency of open-world models in terms of training hours (on 8 NVIDIA A100 GPUs), extra storage usage and inference latency (on a single NVIDIA 2080Ti GPU), since they impose a significant overhead during training or inference. The extra storage for RegionPLC and PLA [5] is to store 3D-language pairs, while for OpenScene [10] is to save fused 2D features.

Category Prompt in nuScenes. We use a category prompt for nuScenes to replace ambiguous words within the category names such as "manmade" and "driveable_surface". The concrete category mapping is illustrated in Table S1.

Category Name	Category Prompts
barrier	barrier or fence
bicycle	bicycle or bike or cycle
bus	bus
car	car
construction_vehicle	construction vehicle or bulldozer or excavator or concrete mixer or crane or dump truck
motorcycle	motorcycle or motorbike
pedestrian	person or people or man or woman
traffic_cone	traffic cone
trailer	trailer
truck	truck
driveable_surface	road or street
other flat	other flat
sidewalk	sidewalk
terrain	grass or rolling hills or soil or gravel
manmade	building or wall or fence or pole or sign or traffic light
vegetation	bushes or plants or trees or potted plants

Table S1. The category prompt for nuScenes.

S1.2. Category Partition for Base-annotated Results

In the base-annotated open-world setting, we divide all categories into base and novel. As for ScanNet [4], we fol-

Partition	Base Categories	Novel Categories
B12/N3	barrier, bicycle, bus, car, construction_vehicle, trailer, truck, diveable_surface, sidewalk, terrain, manmade, vegetation	traffic_cone, motorcycle, pedestrian
B10/N5	bicycle, bus, car, construction_vehicle, trailer, truck, driveable_surface, terrain, manmade, vegetation	barrier, motorcycle, pedestrian, traffic_cone, sidewalk

Table S2. Category partitions for open-world semantic segmentation on nuScenes.

Partition	Novel Categories
B170/N30	pillow, box, clothes, counter, dresser, keyboard, backpack, printer, shower curtain, bin, copier, sofa chair, recycling bin, clock, guitar, set, ladder, cup, toaster, ironing board, toilet seat cover dispenser, furniture, cart, projector, shower floor, laundry detergent, bathroom stall door, dumbbell, folded chair, mattress
B150/N50	couch, window, bookshelf, coffee table, kitchen cabinet, clothes, counter, end table, bag, backpack, printer, microwave, shoe, bin, washing machine, sofa chair, paper, blinds, radiator, recycling bin, soap dispenser, bucket, stand, light, pipe, bathroom stall, cup, storage bin, coffee maker, machine, fireplace, mini fridge, hat, cart, light switch, decoration, plunger, stuffed animal, dish rack, broom, range hood, water pitcher, paper bag, bathroom vanity ceiling light, trash bin, stair rail, coat rack, calendar, poster

Table S3. Novel Categories for open-world semantic segmentation on ScanNet200. We only present novel categories here as there are too many base categories to show here. The partition of base categories can be easily obtained by carrying on the set difference between all categories and novel categories.

low the category partition of PLA [5]. As for nuScenes [2], we discard the ambiguous category "otherflat" and split the remaining 15 categories as illustrated in Table S2. We also randomly split 30 and 50 novel categories for ScanNet200 [13], as shown in Table S3. Notably, 11 categories absent from the ScanNet200 validation set are consistently partitioned into base categories (*i.e.* training set) to guarantee sufficient novel categories for validation. These 11 train-only categories in ScanNet200 are "bicycle", "storage container", "candle", "guitar case", "purse", "alarm clock", "music stand", "cd case", "structure", "storage organizer" and "luggage".

S1.3. Implementation of Baseline Methods

We re-produce the baseline methods including MaskCLIP [17], PointCLIP-Seg [16] and OpenScene [10] for annotation-free open-world semantic segmentation in ScanNet. Details are as follows.

PointCLIP-Seg and MaskCLIP. To apply MaskCLIP [17] on 3D segmentation, we assemble its predictions on multi-view images and back-project them to 3D space as [5]. As for PointCLIP [16], it cannot be directly utilized for the semantic segmentation task, so we extend a segmentation version by modifying the attentive pooling layer of CLIP [12], as per the method used in MaskCLIP [17]. It is named as PointCLIP-Seg. The major distinction between PointCLIP-Seg and MaskCLIP lies in that PointCLIP-Seg uses depth images rather than RGB images for extracting 2D features.

OpenScene. We use the official fused feature released by OpenScene [10] and its prompt engineering techniques to obtain OpenScene-2D results. To ensure a fair comparison, we train OpenScene-3D using the same training schedule and 3D backbone as our RegionPLC. This allows us to compare performance under the same conditions and analyze the results more accurately.

PLA. As for PLA [5] in the annotation-free open-world setting, we only carry on the point-language contrastive learning and discard its binary encoder as there is no annotated base category in the training set.

S1.4. Comparisons of 3D Open-world Scene Understanding Methods

As shown in Table S4, we compare our RegionPLC to other three cutting-edge 3D open-world scene understanding methods: ConceptFusion [8], OpenScene [10] and PLA [5]. ConceptFusion [8] relies on a multi-view fusion of image predictions during its inference phase. However, its inability to learn from 3D point clouds makes it difficult to extract 3D geometric information. On the other hand, OpenScene-3D [10] can learn directly from the 3D point cloud, but this approach necessitates significant additional storage for saving fused 2D features, making it unsuitable for handling large-scale datasets. Furthermore, the ceiling of its performance is limited by the 2D semantic feature and distillation strategy, making it harder to integrate with more advanced 3D backbones. PLA [5], while only requiring minimal additional storage and being scalable due to only 3D-language supervisions during training, is restricted in its performance by the sparseness and roughness of its language supervisions. In contrast, our RegionPLC inherits all the strengths of PLA [5] and incorporates more advanced

Method	2D models	Multi-view inference	Learning in 3D	Scale up with better 3D backbone	Extra storage	Supervision
ConceptFusion [8]	Mask2Former [3] & CLIP [12]	✓	×	×	×	×
OpenScene-3D [10]	LSeg [9] & OpenSeg [6]	×	✓	×	high	Pixel-aligned 2D features
PLA [5]	VIT-GPT2 [1]	×	✓	✓	low	Sparse language supervision
RegionPLC	OFA [15] & Detic [18] & Kosmos-2 [11]	×	✓	✓	low	Dense language supervision

Table S4. Comparison between different 3D open-world scene understanding methods.

Method	Partition	wall	floor	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	desk	curtain	fridge	shower c.	toilet	sink	bathtub
PLA [5]	B15/N4	84.6	95.0	64.9	81.1	87.9	75.9	72.2	61.9	62.1	69.5	30.9	60.1	46.5	70.7	50.5	66.1	56.8	59.0	81.7
	B12/N7	84.7	95.1	65.3	57.8	44.2	75.9	34.5	62.5	62.3	62.1	20.5	57.8	61.4	72.4	47.9	64.9	85.9	28.4	69.6
	B10/N9	83.8	95.2	64.3	80.9	88.0	78.5	73.2	60.6	61.5	68.6	17.7	23.4	51.3	70.6	25.7	38.2	51.3	27.3	61.7
RegionPLC	B15/N4	84.2	95.1	66.6	81.2	88.2	81.3	72.6	61.4	60.7	75.3	30.4	57.7	53.4	70.6	46.1	64.6	72.6	59.4	84.0
	B12/N7	84.9	95.1	65.2	76.3	79.5	75.8	64.3	60.0	64.3	77.9	31.1	56.7	65.7	72.7	49.5	65.6	83.4	55.5	81.9
	B10/N9	84.3	95.2	65.5	80.6	89.2	82.7	73.8	59.6	62.0	79.7	25.0	47.7	56.3	69.8	38.0	53.2	74.4	46.6	78.9

Table S5. Per-class results of base-annotated open-world 3D semantic segmentation on ScanNet in terms of IoU. Performance on novel categories is marked in blue.

2D models for regional 3D-language association, thereby boosting its open-world capability.

S2. More Experimental Results

In this section, we present some supplementary experimental results, in addition to the ones provided in our main paper. This part consists of a detailed analysis of the per-class performance, an error-bar analysis and the zero-shot domain transfer experiments.

S2.1. Per-category Results

Here, we show the per-category performance comparison between PLA [5] and RegionPLC for base-annotated open-world 3D semantic segmentation on ScanNet [4]. As shown in Table S5, our RegionPLC obtains improvements on all novel categories across different partitions, which demonstrates its effectiveness.

S2.2. Error Bar

Here, we provide an error bar for our open-world 3D scene understanding framework on both base-annotated and annotation-free settings by reproducing each experiment 3 times. As shown in Table S6, the performance of RegionPLC is generally stable on ScanNet open-world segmentation, demonstrating its robustness.

S2.3. Zero-shot Domain Transfer

We study the zero-shot domain generalization capability of open-world methods by transferring the ScanNet-trained model to S3DIS without fine-tuning. As shown in Table S7, RegionPLC enjoys 6.8% $\sim$ 37.1% boosts compared to PLA [5] in mIoU† on different splits. Notice that more base categories on ScanNet can hinder the generalization on S3DIS, indicating that dataset-specific annotation penalizes the model's transferability. In contrast, solely learning from semantic-rich caption supervision achieves great out-of-domain generalization ability.

S3. Qualitative Results for Annotation-free Open World

Here, we provide more qualitative results of RegionPLC in the most challenging annotation-free open-world scenario. As shown in Figure S1, our RegionPLC can distinguish different semantics with remarkable segmentation results covering a wide range of categories.

On the other hand, we also explore the potential of our RegionPLC to discover tail and rare categories in real-world scenarios. As shown in Figure S2, we visualize the heat maps of the point-wise response given a text query. Our RegionPLC can discover a lot of tail categories such as "trash can", "shoe" and "nightstand" without any human annotation. These results demonstrate the effectiveness of our regional point-language contrastive learning framework in solving open-world 3D scene understanding problems.

S4. Prompts for RegionGR

As highlighted in the main paper, our RegionPLC is capable of incorporating large language models (LLM), such as

Round	Base-annotated ScanNet [4]			Annotation-free ScanNet [4]
	B15/N4	B12/N7	B10/N9	
1	69.4 / 68.2 / 70.7	68.2 / 69.9 / 66.6	64.3 / 76.3 / 55.6	59.6 (77.5)
2	69.5 / 68.6 / 70.4	67.6 / 69.8 / 65.4	63.9 / 76.4 / 54.9	59.2 (78.0)
3	69.7 / 68.6 / 70.8	67.7 / 69.4 / 66.1	63.8 / 76.1 / 54.9	59.1 (76.6)

Table S6. Repeated results for base-annotated and annotation-free open-world 3D semantic segmentation on ScanNet. Base-annotated results are measured in hIoU / mIoU$^{\mathcal{B}}$ / mIoU$^{\mathcal{N}}$, while annotation-free experiments are measured in mIoU† (mAcc†).

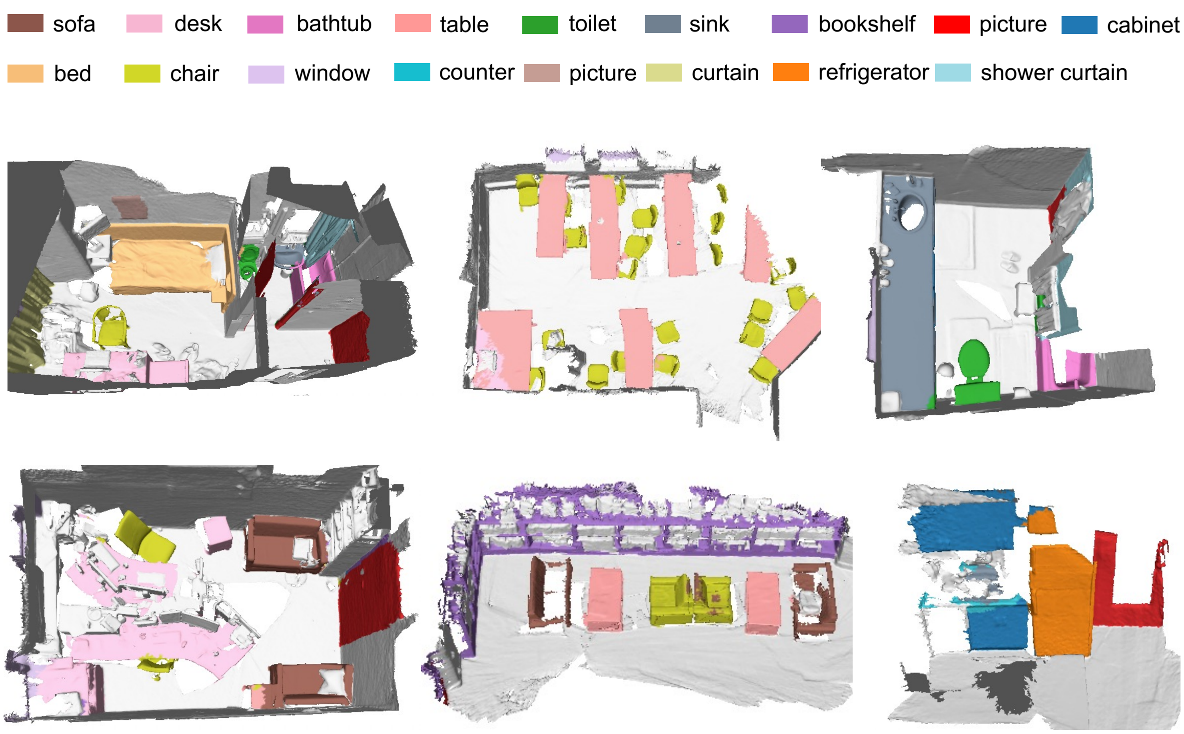

Figure S1. Qualitative results of annotation-free semantic segmentation on ScanNet.

ScanNet partition	S3DIS Semantic Segmentation		
	OVSeg-3D [5]	PLA [5]	RegionPLC
B15/N4	31.1 (46.6)	39.1 (56.2)	**52.2 (64.5)**
B12/N7	23.6 (42.7)	35.4 (60.4)	**45.0 (61.5)**
B10/N9	36.0 (50.9)	43.7 (60.4)	**50.5 (63.2)**
B0/N17	01.7 (11.2)	13.4 (25.1)	**50.5 (67.6)**

Table S7. Zero-shot domain transfer results for semantic segmentation in items of mIoU† (mAcc†) on ScanNet → S3DIS.

GPT-3.5 [14], to execute grounded 3D reasoning, a pipeline we refer to as RegionGR. The LLM is given human queries and regional captions for the purpose of reasoning. Note that if a human query pertains to a particular 3D region, we will filter captions, retaining only those that show significant overlap with the specified 3D region as the input. The prompt example we used is as follows.

```
[Role]
You are a household manager.
Your job is to understand human
instructions, and you should give
step-by-step suggestions according
to the provided environmental
context.

[Task]
Your task is to give a suitable
response to the <question> according
to the <env_context>; if possible,
respond in detail with clear logic.
Both the question and env context
are given, delimited by triple
quotes.

[Env Context]
Here is the env context, containing
some words, phrases, or short
```

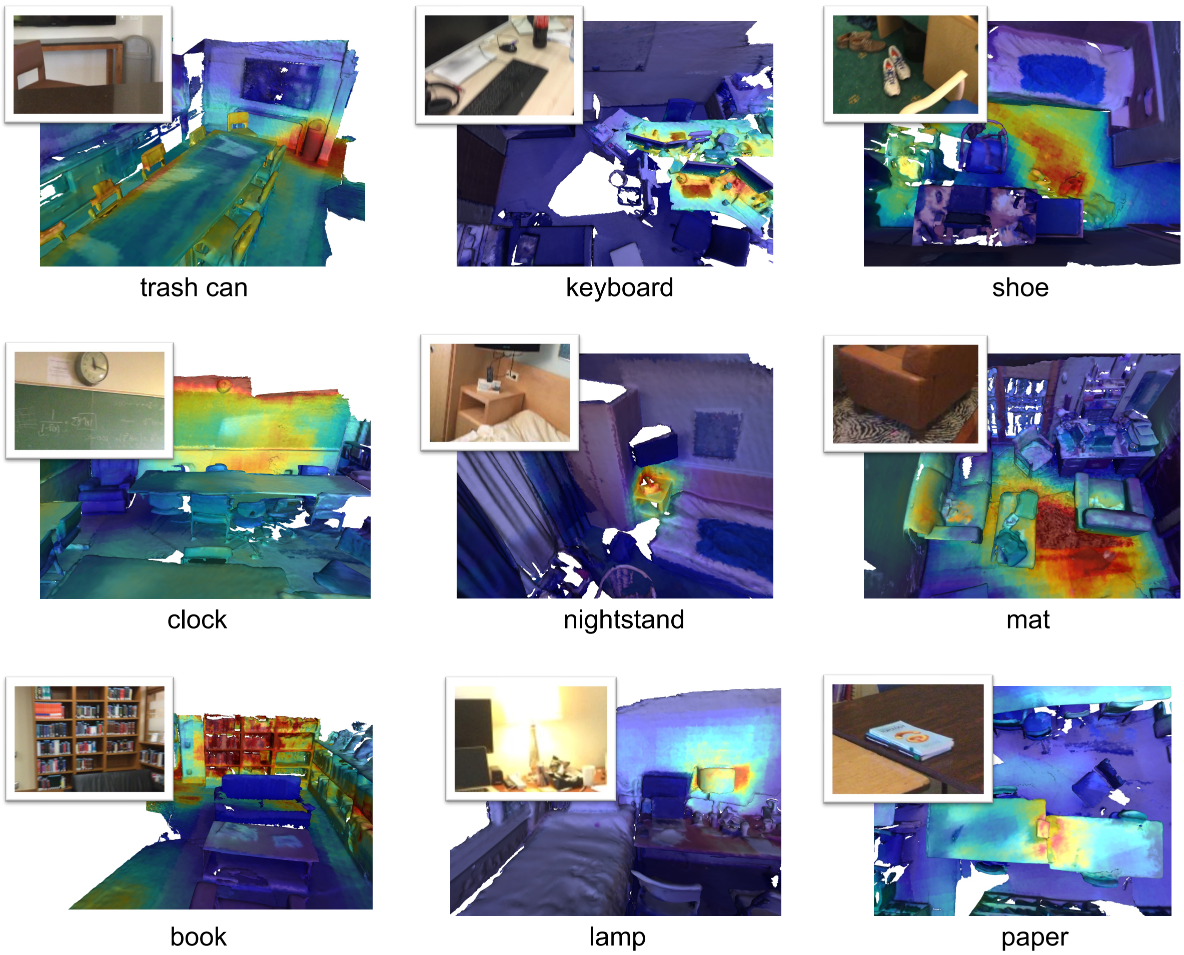

Figure S2. Visualization in heat map of tail classes with the annotation-free model on ScanNet.

```
sentences describing contents in a
3D room.  Answer the user's request
based on this.
<{env}>

[Rules]
Return answers closely related to
the provided information, especially
the objects mentioned in the
provided context.  Keep the final
answer simple and short, within 30
words.  Use natural language like
humans do in daily life.

[Steps]
According to the query, understand
the intention behind "What do I
want/need to do?" Find the objects
related to my question from the env
context.  To give the final answer,
you should tell me the operation
I need to do and the object I need
to interact with.  The answer needs
to be realistic, and the objects in
your answer need to be based on the
provided env context.

[Dialog Style]
You should respond in a polite,
kind, and natural language tone.
Try to talk like a human, but
keep it short.

Begin Task

The question:  <{question}>
```

S5. Limitation and Future Works

Although our RegionPLC has yielded impressive results in 3D open-world scene understanding with a broad spectrum of unseen categories, certain limitations and potential avenues for enhancement remain. On the one hand, the promising results obtained by the combination of RegionPLC and OpenScene [10] demonstrate the strong potential to introduce 2D image features as auxiliary supervision for training RegionPLC. The current loss combination is straightforward, and we believe that more advanced combination strategies that integrate language, 3D and image features can lead to better performance.

Another aspect warranting improvement is our utilization of visual prompts, which are pre-defined prior to training and remain unchanged throughout the process. Better and more adaptive visual prompting techniques might improve the quality of language supervision. Moving forward, we are interested in further developing an open-world 3D scene understanding framework that addresses these two limitations.

References

[1] Vit-gpt2 image captioning. https://huggingface.co/nlpconnect/vit-gpt2-image-captioning/discussions. 3

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1, 2

[3] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 3

[4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1, 3, 4

[5] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Language-driven open-vocabulary 3d scene understanding. *arXiv preprint arXiv:2211.16312*, 2022. 1, 2, 3, 4

[6] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, 2021. 3

[7] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 1

[8] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, et al. Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*, 2023. 2, 3

[9] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 3

[10] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. *arXiv preprint arXiv:2211.15654*, 2022. 1, 2, 3, 6

[11] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 3

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3

[13] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 125–141. Springer, 2022. 1, 2

[14] John Schulman, Barret Zoph, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe Ceron Uribe, Liam Fedus, Luke Metz, Michael Pokorny, et al. Chatgpt: Optimizing language models for dialogue. *OpenAI blog*, 2022. 4

[15] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, abs/2202.03052, 2022. 1, 3

[16] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8552–8562, 2022. 2

[17] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[18] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 1, 3