Supplementary Materials for PLA: Language-Driven Open-Vocabulary 3D Scene Understanding

Runyu Ding[1*†] Jihan Yang[1*] Chuhui Xue[2] Wenqing Zhang[2] Song Bai[2‡] Xiaojuan Qi[1‡]
[1]The University of Hong Kong [2]ByteDance

Outline

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

- Sec. S1: Details of our open-vocabulary scene understanding benchmark.
- Sec. S2: Limitation analysis of PointCLIP for scene understanding tasks.
- Sec. S3: Additional experimental results on re-partition results, per-class results, error bar results, fully-supervised results with caption supervision and combination of caption supervisions.
- Sec. S4: Examples of image-caption pairs and hierarchical point-caption pairs.
- Sec. S5: Qualitative results of open-vocabulary scene understanding.
- Sec. S6: Limitation and open problems.

S1. Implementation Details

Here, we present the implementation details of dataset category partition, network modifications, baseline setups, hyper-parameter configurations and usage of images.

S1.1. Dataset Category Partition

As mentioned in Sec. 4.1 of the main paper, we build a 3D open-vocabulary benchmark on ScanNet [5] and S3DIS [2] with multiple base/novel partitions. ScanNet [5] consists of 1,613 scenes (1,201 scenes for training, 312 scenes for validation and 100 for testing) densely annotated in 20 classes. We discard the 'otherfurniture' class and partition the rest 19 classes into three partitions for semantic segmentation as shown in Table S1. Note that the B15/N4 partition adheres to the 3DGenZ [11] partitioning scheme. As for instance segmentation, we follow SoftGroup [15] to ignore two background classes (*i.e.* wall and floor) and obtain corresponding partitions (see Table S2).

S3DIS [2] contains 271 scans across 6 building areas along with 13 categories. Following previous work [12], we treat the 5th area as the validation split and other areas as the training split. We discard the 'clutter' class and partition the rest 12 classes into two partitions for both semantic segmentation and instance segmentation as demonstrated in Table S3.

S1.2. Network Modifications

In this section, we elaborate on how to extend a close-set network to an open-vocabulary learner for semantic segmentation and instance segmentation. We employ sparse-convolution-based UNet [6] with a base hidden dimension of 16 as our backbone $\mathrm{F}_{3\mathrm{D}}$.

First, as illustrated in Fig. S1 (a), the close-set network contains a learnable semantic head $\mathrm{F}_{\mathrm{sem}}$ that classifies a fixed number of categories. As discussed in Sec. 3.2 in the main paper, to obtain an open-vocabulary model, we replace the semantic head $\mathrm{F}_{\mathrm{sem}}$ with a vision-language (VL) adapter F_{θ} and the category embedding $\mathbf{f}^{l}$ encoded by a fixed text encoder $\mathrm{F}_{\mathrm{text}}$. Note that the category embedding $\mathbf{f}^{l}$ can be treated as replacing the weights of the classifier. The category embedding $\mathbf{f}^{l}$ encodes semantic attributes of base classes in the training stage and encodes any desired categories during inference to achieve open-vocabulary semantic segmentation.

Further, as we follow SoftGroup [15] to develop instance head $\mathrm{F}_{\mathrm{ins}}$, we modify the close-set designs in SoftGroup to obtain an open-vocabulary instance head. First, as shown in Fig. S2, the seg head and the score head that produce per-class confidence in the vector form are modified to class-agnostic modules that produce a single scalar for each generated instance proposal. In this way, we can train these two heads without needing to know novel categories. Second, the learnable cls head that predicts the classification scores of generated proposals is replaced by the proposal-level pooling of semantic scores $\mathbf{s}$, which can be extended to arbitrary categories. Finally, the class statistics, such as

*Equal contribution: {ryding, jhyang}@eee.hku.hk
†Part of the work is done during an internship at ByteDance AI Lab.
‡Corresponding authors: song.site@gmail.com, xjqi@eee.hku.hk

Partition	Base Categories	Novel Categories
B15/N4	wall, floor, cabinet, bed, chair, table, door, window, picture, counter, curtain, refrigerator, showercurtain, sink, bathtub	sofa, bookshelf, desk, toilet
B12/N7	wall, floor, cabinet, sofa, door, window, counter, desk, curtain, refrigerator, showercurtain, toilet	bed, chair, table, bookshelf, picture, sink, bathtub
B10/N9	wall, floor, cabinet, bed, chair, sofa, table, door, window, curtain	bookshelf, picture, counter, desk, refrigerator, showercurtain, toilet, sink, bathtub

Table S1. Category partitions for open-vocabulary semantic segmentation on ScanNet.

Partition	Base Categories	Novel Categories
B13/N4	cabinet, bed, chair, table, door, window, picture, counter, curtain, refrigerator, showercurtain, sink, bathtub	sofa, bookshelf, desk, toilet
B10/N7	cabinet, sofa, door, window, counter, desk, curtain, refrigerator, showercurtain, toilet	bed, chair, table, bookshelf, picture, sink, bathtub
B8/N9	cabinet, bed, chair, sofa, table, door, window, curtain	bookshelf, picture, counter, desk, refrigerator, showercurtain, toilet, sink, bathtub

Table S2. Category partitions for open-vocabulary instance segmentation on ScanNet.

the average number of points in an instance mask for each class, which assists proposal grouping, are removed to avoid leakage of novel class information. We empirically show that those modifications cause little degradation of fully-supervised performance by 1.1% mAP_{50}, as demonstrated in Table S4. Note that we train the model from scratch rather than fine-tuning a supervised pretrained model, as SoftGroup does, to prevent leakage of novel classes during training. Additionally, we use a smaller hidden dimension size for the UNet backbone. Consequently, our reproduced performance differs from that in the original paper.

S1.3. Baseline Setups

As mentioned in Sec. 4.1 of the main paper, we follow LSeg [9] to implement LSeg-3D as a baseline with UNet [6, 4] backbone, vision-language adapter implemented by MLP and the CLIP [13] ViT-B/16 text encoder. For the other two 3D zero-shot methods, 3DGenZ [11] and 3DTZSL [3], we reproduce them with the same network and CLIP text embedding for fair comparisons. Specifically, for 3DGenZ [11], instead of training on samples that only contain base classes, we train it on the whole training dataset with points belonging to novel classes ignored during optimization. Besides, we remove calibrated stacking that aims to alleviate bias towards seen classes since it brings extremely minor performance gains in our implementations. As for 3DTZSL [3] designed for object classification, we extend it to segmentation via learning with triplet loss on the point level instead of the sample level. We implement its projection net with 2 fully-connected layers and the Tanh activation function, the same as its paper claimed.

S1.4. Hyper-Parameter Configurations

We train 19,216 iterations on ScanNet and 4,080 iterations on S3DIS for semantic segmentation. For instance segmentation, we train 24,020 iterations on ScanNet and 9,160 iterations on S3DIS. The learning rate is initialized as 0.004 with cosine decay. We adopt the AdamW [10] optimizer and run all experiments with 32 batch size on 8 NVIDIA V100 or NVIDIA A100.

For entity-level captions, we filter out some $\langle \hat{\mathbf{p}}^{\mathbf{e}}, \mathbf{t}^{e} \rangle$ pairs to guarantee the point set $\hat{\mathbf{p}}^{\mathbf{e}}$ is small enough containing only a few entities. Specifically, we set the minimal points γ as 100 and the ratio that controls the maximum number of points δ as 0.3. As for the caption loss, we set α_1, α_2 and α_3 as 0, 0.05 and 0.05 for scene-level $\mathcal{L}^{s}_{\text{cap}}$, view-level $\mathcal{L}^{v}_{\text{cap}}$ and entity-level loss $\mathcal{L}^{e}_{\text{cap}}$ for ScanNet, respectively. For S3DIS, we set α_1, α_2, and α_3 as 0, 0.08, and 0.02 separately.

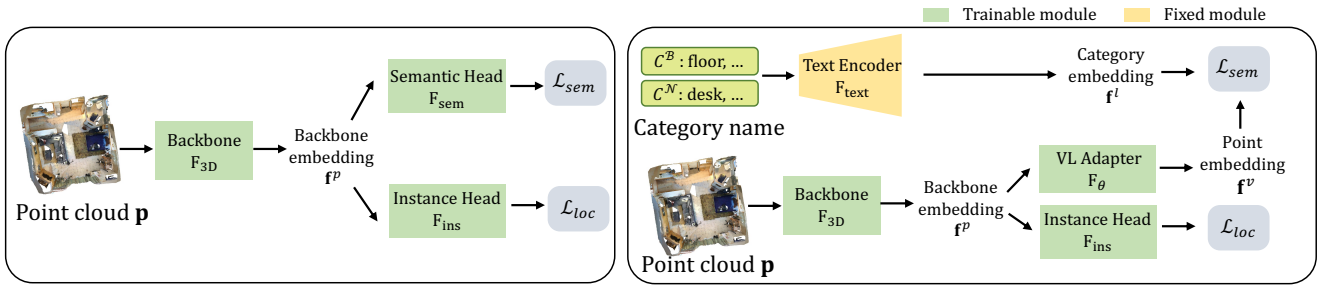
S1.5. Usage of Images

For ScanNet, we use a 25,000-frame subset* from ScanNet images for captioning. For S3DIS, as each scene contains a widely varying number of images, we subsample its images to caption at most 50 images per scene. It is worth noting that some S3DIS scenes lack corresponding images; we consequently cannot provide language supervision for those scenes without images during training.

*https://kaldir.vc.in.tum.de/scannet_benchmark/documentation

Partition	Base Categories	Novel Categories
B8/N4	ceiling, floor, wall, beam, column, door, chair, board	window, table, sofa, bookcase
B6/N6	ceiling, wall, beam, column, chair, bookcase	floor, window, door, table, sofa, board

Table S3. Category partitions for open-vocabulary semantic and instance segmentation on S3DIS.



(a) Close-set framework (b) Open-vocabulary framework

Figure S1. Comparison between close-set scene understanding framework and open-vocabulary scene understanding framework.

Components			mAP_{50}
per-class seg head and score head	cls head	class statistics	
✓	✓	✓	61.8
	✓	✓	62.0
		✓	61.1
			60.7

Table S4. Fully-supervised instance segmentation results of different SoftGroup variants upon ScanNet in terms of mAP_{50}.

S2. Analysis of PointCLIP for Scene Understanding

In recent years, 2D open-vocabulary understanding [7, 14, 16, 9] achieves unprecedented success driven by transferable vision-language models such as CLIP [13] trained on large-scale image-caption pairs. Inspired by that success, PointCLIP [8] has made the first attempt to transfer the knowledge of CLIP into the 3D domain for zero-shot and few-shot object classification tasks. PointCLIP projects 3D point clouds into 2D multi-view depth maps and leverages CLIP to process multi-view depth images to obtain predictions. Finally, the predictions are assembled into 3D predictions. Though some progress has been made in object-level understanding, our experimental results show that PointCLIP is not suitable for scene-level understanding tasks with poor performance and heavy inference overheads.

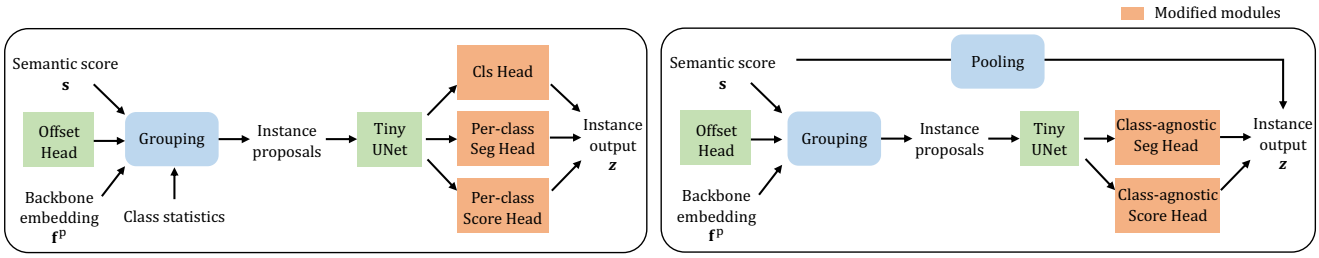
Task-specific modifications. To extend PointCLIP for 3D scene understanding, we make the following modifications. First, we follow the state-of-the-art 2D open-vocabulary semantic segmentation method MaskCLIP [17] to modify the attentive pooling layer of CLIP's vision encoder for obtaining pixel-wise dense predictions. Second, instead of using self-rendered images, we utilize collected depth images captured by depth sensors since they are realistic with more accurate depth values. We also explore utilizing collected RGB images to avoid modal gaps caused by using depth images. Finally, to assemble multi-view 2D results into 3D, other than voting to get object-wise predictions, we back-project all multi-view image predictions into 3D space via 3D geometry and assign predictions to each point of 3D scenes by searching nearest neighbors in back-projected 3D point clouds.

Results. As shown in Table S5, with depth images as input,

Input	2D mIoU	3D mIoU	latency (ms)
depth images	02.2	01.7	1667
RGB images	17.8	17.2	1667

Table S5. Results of zero-shot 3D semantic segmentation using PointCLIP on ScanNet.

the modified PointCLIP obtains only 2.2% mIoU on 2D semantic segmentation with 5,436 validation samples of ScanNet. The assembled 3D prediction only attains 1.7% mIoU on 312 samples, which is very close to random guesses. When alternated to use RGB images as input, the performance lifts to 17.8% mIoU on 2D and 17.2% mIoU on 3D, demonstrating that using RGB images can avoid annoying modal gaps. However, the performance is still moderate, which suggests this projection-based stream of work is sub-optimal for tackling 3D scene understanding tasks. Though further fine-tuning on seen categories might benefit model performance, this line of research has a key limita-



(a) Close-set instance head

(a) Open-vocabulary instance head

Figure S2. Comparison between close-set instance head and open-vocabulary instance head.

tion: by projecting 3D data to 2D, it suffers from information loss and makes the model unable to directly learn from information-rich 3D data.

In addition, to assess the model efficiency, we use latency to measure the execution speed of model inference on a single GeForce RTX 2080Ti. As shown in Table S5, PointCLIP takes an average of 1667ms to process images of one 3D scene, which is rather costly, not to mention the post-processing time for back-projection and results ensemble. Instead, our 3D network only costs 83ms to process one 3D sample, which is 20 times more efficient than PointCLIP.

In sum, the poor zero-shot performance, information loss from projection, and heavy computation costs render this line of work not suitable for 3D scene understanding and prevent us from exploring further on this stream of work.

S3. Additional Experimental Results

S3.1. Re-partition Experiments

Splits	hIoU / mIoU$^{\mathcal{B}}$ / mIoU$^{\mathcal{N}}$	
	LSeg-3D [9]	Ours
random-sample 1	00.0 / 61.7 / 00.0	**65.3 / 68.3 / 62.4**
random-sample 2	00.0 / 48.5 / 00.0	**53.1 / 70.1 / 42.7**
random-sample 3	00.3 / 66.1 / 00.2	**60.9 / 69.2 / 54.5**
frequency-sample	00.0 / 68.7 / 00.0	**62.6 / 69.0 / 57.3**

Table S6. Results of re-sampled base and novel categories.

To ensure the reliability of results, we randomly re-sample base and novel categories three times and sample it based on class frequency for the B15/N4 ScanNet semantic segmentation task. As shown in Table S6, our method consistently exceeds LSeg-3D baseline among four different splits by a large margin of $53.1\% \sim 65.3\%$ hIoU, which reveals the robustness of our methods in handling different novel classes.

S3.2. Per-class Results

We present per-category performances of our open-vocabulary 3D scene understanding framework on semantic and instance segmentation. As shown in Table S7 and Table S8, novel classes generally perform worse than base classes without annotation supervision. With the space of novel categories enlarged (*e.g.* from B15/N4 to B12/N7 partition), the performance on novel classes degrades (*e.g.* 'bookshelf' obtains 7.4% mIoU drop from B15/N4 to B12/N7 partition on semantic segmentation) due to the insufficient seen-category data to tune the model.

S3.3. Error Bar

Here, to show the robustness of our experimental results, we repeat the experiments on open-vocabulary semantic and instance segmentation three times and report their average along with standard deviation. As shown in Table S9 and Table S10, the results on base classes are slightly more stable than novel classes with lower standard deviations, which demonstrates the higher confidence uncertainty of novel class predictions. Besides, results on ScanNet are more stable than S3DIS, which indicates that the sample size and diversity contribute a lot to the performance stability.

S3.4. Equipping Fully-Supervised Model with Point-Caption Supervision.

As demonstrated in Table S11, fully-supervised models equipped with caption supervision loss perform similarly to those without it, as they already have access to annotations for all categories. In this scenario, our language supervision neither hinders nor enhances fully-supervised performance, validating our fairness in using the fully-supervised model for comparison in the main paper.

S3.5. Combination of Caption Supervisions.

The combination of three captions, including the scene-level caption, can result in a 0.6% increase in hIoU, as shown in Table S12. However, finding such a right balance between these captions requires sophisticated loss trade-off

Task	Partition	wall	floor	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	desk	curtain	fridge	shower c.	toilet	sink	bathtub
Sem.	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
Inst.	B13/N4	–	–	50.5	77.0	82.9	43.4	75.4	49.0	46.0	43.7	46.5	33.7	23.2	54.1	49.6	56.0	97.8	47.5	85.8
	B10/N7	–	–	53.7	62.7	11.2	70.5	27.2	47.7	45.7	30.0	01.5	39.9	40.8	50.6	68.6	84.6	92.9	24.6	00.0
	B8/N9	–	–	45.1	77.4	82.2	84.2	74.2	48.9	51.0	30.0	00.5	02.1	16.8	44.9	28.3	35.1	94.3	16.6	00.0

Table S7. Per-class results of 3D open-vocabulary scene understanding on ScanNet. Performance on novel class are marked in blue.

Task	Partition	ceiling	floor	wall	beam	column	window	door	table	chair	sofa	bookcase	board
Sem.	B8/N4	93.9	97.8	82.9	00.0	17.2	15.6	53.7	35.8	86.3	05.3	37.3	43.3
	B6/N6	93.7	79.1	80.1	00.1	28.5	24.1	08.4	37.6	87.0	54.0	24.0	06.9
Inst.	B8/N4	89.5	100.0	50.8	00.0	35.3	36.2	60.5	00.1	84.6	01.9	00.8	59.4
	B6/N6	89.5	60.2	17.9	00.0	41.5	10.2	02.1	00.6	86.2	45.1	00.1	02.2

Table S8. Per-class results of 3D open-vocabulary scene understanding on S3DIS.

techniques that are not universally applicable across different datasets. Therefore, the scene-level caption is not used in our paper for the sake of generalization. Further studies on effectively combining caption supervisions would be a future investigation.

S4. Caption Examples

In this section, we present examples of image-caption pairs obtained by vision-language (VL) foundation models and examples of hierarchical associated point-caption pairs. As illustrated in Fig. S3, image captions describe main entities of images along with room types (*e.g.* kitchen), texture (*e.g.* leather), color (*e.g.* green) or spatial relationships (*e.g.* on top of), conveying rich semantic clues with large vocabulary size. Moreover, uncommon classes such as 'buddha statue' are also correctly detected, reflecting the generalizability of existing VL foundation models and semantic comprehensiveness of generated captions.

With obtained image-caption pairs, we are capable to associate 3D points and captions hierarchically leveraging geometric constraints between 3D point clouds and multi-view images. As shown in Fig. S4 (a), the scene-level caption describes each area/room (*e.g.* kitchen, living room) in the whole scene with abundant vocabulary, providing semantic-rich language supervision. View-level caption in Fig. S4 (b) focuses on single view frustums of the 3D point cloud, capturing more local details with elaborate text descriptions, which enables the model to learn region-wise vision-semantic relationships. Additionally, as shown in Fig. S4 (c), the entity-level caption covers only a few entities in small 3D point sets with concrete words, providing more fine-grained supervisions to learn object-level understanding and localization.

S5. Qualitative Results

Here, we provide some qualitative results on open-vocabulary semantic segmentation and instance segmentation as illustrated in Fig. S5. Compared to the LSeg-3D baseline that always confuses unseen classes as seen classes, our framework successfully recognizes novel categories with accurate semantic masks, which shows our point-caption association injects rich semantic concepts into the 3D network. Additionally, the instance prediction masks of our framework are also accurate, while the LSeg-3D baseline misses novel objects or predicts incomplete object masks. It reflects the strong generalized localization ability of our framework.

S6. Limitation and Open Problems

Although our language-driven open-vocabulary 3D scene understanding framework introduces rich semantic concepts for learning adequate visual-semantic relationships, it still suffers from limitations in the following aspects. First comes the calibration problem that the model tends to produce over-confident predictions on base classes, which lies in both semantic and instance segmentation tasks. For semantic segmentation, though the binary head is developed to calibrate semantic scores for in-domain

Round	ScanNet									S3DIS					
	B15/N4			B12/N7			B10/N9			B8/N4			B6/N6		
	hIoU	mIoU$^{\mathcal{B}}$	mIoU$^{\mathcal{N}}$	hIoU	mIoU$^{\mathcal{B}}$	mIoU$^{\mathcal{N}}$	hIoU	mIoU$^{\mathcal{B}}$	mIoU$^{\mathcal{N}}$	hIoU	mIoU$^{\mathcal{B}}$	mIoU$^{\mathcal{N}}$	hIoU	mIoU$^{\mathcal{B}}$	mIoU$^{\mathcal{N}}$
1	66.3	68.4	64.2	54.3	69.5	44.6	52.8	76.2	40.6	33.2	58.2	23.3	39.4	57.2	30.0
2	65.2	68.6	62.2	54.8	69.7	45.2	53.3	75.6	40.9	37.0	59.5	26.9	39.5	55.1	30.8
3	64.5	67.8	60.8	59.7	69.2	48.0	53.2	76.6	40.8	33.7	59.4	23.5	36.5	54.3	27.5
Average	65.3	68.3	62.4	55.3	69.5	45.9	53.1	76.2	40.8	34.6	59.0	24.5	38.5	55.5	29.4
Std	00.9	00.4	01.7	01.3	00.2	01.8	00.3	00.5	00.2	02.1	00.7	02.0	01.7	01.5	01.7

Table S9. Repeat results for open-vocabulary 3D semantic segmentation on ScanNet and S3DIS in terms of hIoU, mIoU$^{\mathcal{B}}$ and mIoU$^{\mathcal{N}}$.

Round	ScanNet									S3DIS					
	B13/N4			B10/N7			B8/N9			B8/N4			B6/N6		
	hAP$_{50}$	mAP$^{\mathcal{B}}_{50}$	mAP$^{\mathcal{N}}_{50}$	hAP$_{50}$	mAP$^{\mathcal{B}}_{50}$	mAP$^{\mathcal{N}}_{50}$	hAP$_{50}$	mAP$^{\mathcal{B}}_{50}$	mAP$^{\mathcal{N}}_{50}$	hAP$_{50}$	mAP$^{\mathcal{B}}_{50}$	mAP$^{\mathcal{N}}_{50}$	hAP$_{50}$	mAP$^{\mathcal{B}}_{50}$	mAP$^{\mathcal{N}}_{50}$
1	54.9	58.1	52.0	33.1	52.5	24.1	34.5	62.1	23.9	19.3	59.2	11.5	10.9	49.2	06.1
2	56.7	57.9	55.5	28.4	55.1	19.1	37.5	63.8	26.5	9.2	57.4	05.0	19.8	46.7	12.6
3	55.0	59.5	51.1	32.1	56.3	22.5	35.7	63.5	24.8	16.8	60.0	09.7	17.4	44.9	10.8
Average	55.5	58.5	52.9	31.2	54.6	21.9	35.9	63.1	25.1	15.0	59.0	08.6	16.0	46.9	09.8
Std	01.0	00.9	02.3	02.5	01.9	02.6	01.5	00.9	01.3	04.3	01.1	02.7	04.6	02.2	03.4

Table S10. Repeat results for open-vocabulary 3D instance segmentation on ScanNet and S3DIS in terms of hAP$_{50}$, mAP$^{\mathcal{B}}_{50}$ and mAP$^{\mathcal{N}}_{50}$.

Method	mIoU	mIoU$^{\mathcal{B}}$ / mIoU$^{\mathcal{N}}$		
		B15/N4	B12/N7	B10/N9
Fully-Sup.	70.62	68.4 / 79.1	70.0 / 71.8	75.8 / 64.9
Fully-Sup. + Caption	70.82	68.7 / 78.9	70.3 / 71.7	76.7 / 64.6

Table S11. Fully-supervised results equipped with point-caption supervision.

α_1(scene)	α_2(view)	α_3(entity)	hIoU / mIoU$^{\mathcal{B}}$ / mIoU$^{\mathcal{N}}$
0.000	0.050	0.050	65.3 / 68.3 / 62.4
0.033	0.033	0.033	64.6 / **69.0** / 60.8
0.010	0.045	0.045	**65.9** / 68.2 / **63.8**

Table S12. Ablation for caption loss weights on ScanNet B15/N4.

open-vocabulary scene understanding, it fails to rectify predictions for out-of-domain transfer tasks. Trained on the dataset-specific base/novel partition, the binary head is hard to generalize to other datasets with data distribution shifts, which encourages us to design more transferable score calibration modules in the future. As for the instance segmentation task, though we largely address the localization problem for novel classes through fine-grained point-caption pairs, the calibration problem also exists in the proposal grouping process, where objects of novel classes cannot group well and probably obtain incomplete instance masks. We also leave it as a challenge that needs to be resolved further.

The second problem is that S3DIS achieves slightly worse open-vocabulary performance than ScanNet, largely due to its limited sample size and diversity, as well as much fewer point-caption associations. Inspired by our zero-shot transfer results, we believe it is an appealing alternative to pre-train on a large dataset with rich semantic information and then fine-tune it on the small-scale dataset, which we leave for future study.

References

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

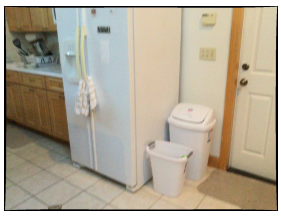
[2] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1534–1543, 2016.

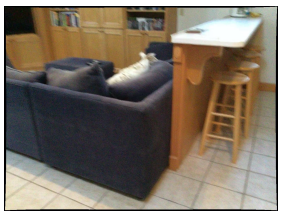
[3] Ali Cheraghian, Shafin Rahman, Dylan Campbell, and Lars Petersson. Transductive zero-shot learning for 3d point cloud classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 923–933, 2020.

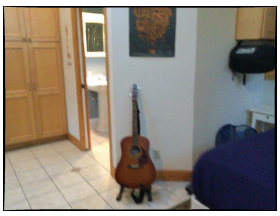
[4] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019.

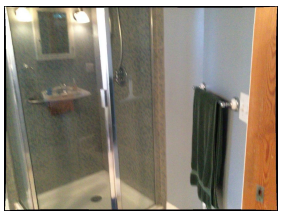
[5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

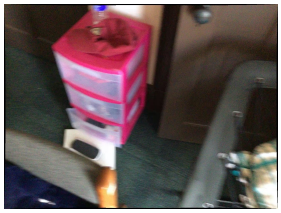
[6] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the*

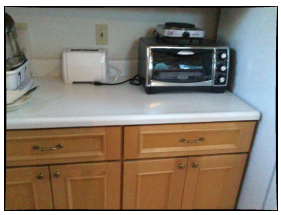

a kitchen with a refrigerator and a trash can


a living room with a couch and a bar


a guitar sitting on the floor in a room

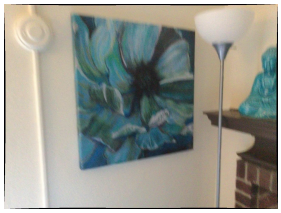

a bathroom with a shower and a green towel

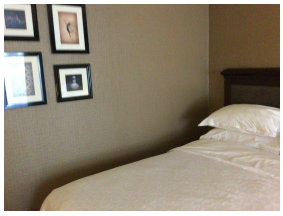

a pink plastic container with a bunch of boxes on the floor

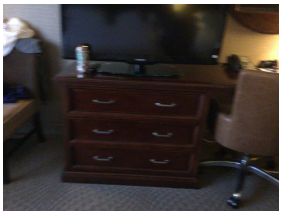

a toaster oven sitting on top of a kitchen counter


three leather chairs and a stool in a living room


the back of a computer screen on a table


a painting of a flower next to a lamp and a buddha statue


a bedroom with a bed and pictures on the wall


a dresser with drawers and a tv on top of it

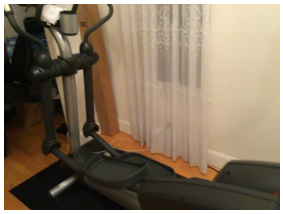

a treadmill in the corner of a room

Figure S3. Examples of image-caption pairs by image-captioning model ViT-GPT2 [1].

IEEE conference on computer vision and pattern recognition, pages 9224–9232, 2018.

[7] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[8] Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson WH Lau, Wanli Ouyang, and Wangmeng Zuo. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. *arXiv preprint arXiv:2210.01055*, 2022.

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

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[11] Björn Michele, Alexandre Boulch, Gilles Puy, Maxime Bucher, and Renaud Marlet. Generative zero-shot learning for semantic segmentation of 3d point clouds. In *2021 International Conference on 3D Vision (3DV)*, pages 992–1002. IEEE, 2021.

[12] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017.

[13] 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.

[14] Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. In *36th Conference on Neural Information Processing Systems (NIPS)*, 2022.

[15] Thang Vu, Kookhoi Kim, Tung M. Luu, Xuan Thanh Nguyen, and Chang D. Yoo. Softgroup for 3d instance segmentation on 3d point clouds. In *CVPR*, 2022.

[16] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. *arXiv preprint arXiv:2112.14757*, 2021.

[17] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Com-*

Video shows a person sitting on a couch with their feet on a rug. A guitar is sitting in a room next to a bed. A toaster oven is sitting on top of a kitchen counter. A bike is parked in a living room with a tiled floor.

A living room is clean and ready for the flooring to be installed. A bed with a gold blanket and a laptop on top of it. A bag of clothes sitting on a chair in a living room. A treadmill in the corner of a room. an exercise bike in a room with a white curtain.

(a) scene-level caption

a kitchen with a refrigerator and a trash can

a bedroom with a bed and pictures on the wall

a dresser with drawers and a tv on top of it

a toaster oven sitting on top of a kitchen counter

(b) view-level caption

table couch living

chair couch

hotel lamp bed

tv

(c) entity-level caption

Figure S4. Examples of hierarchical point-caption pairs from ScanNet [5]

.

puter Vision (ECCV), 2022.

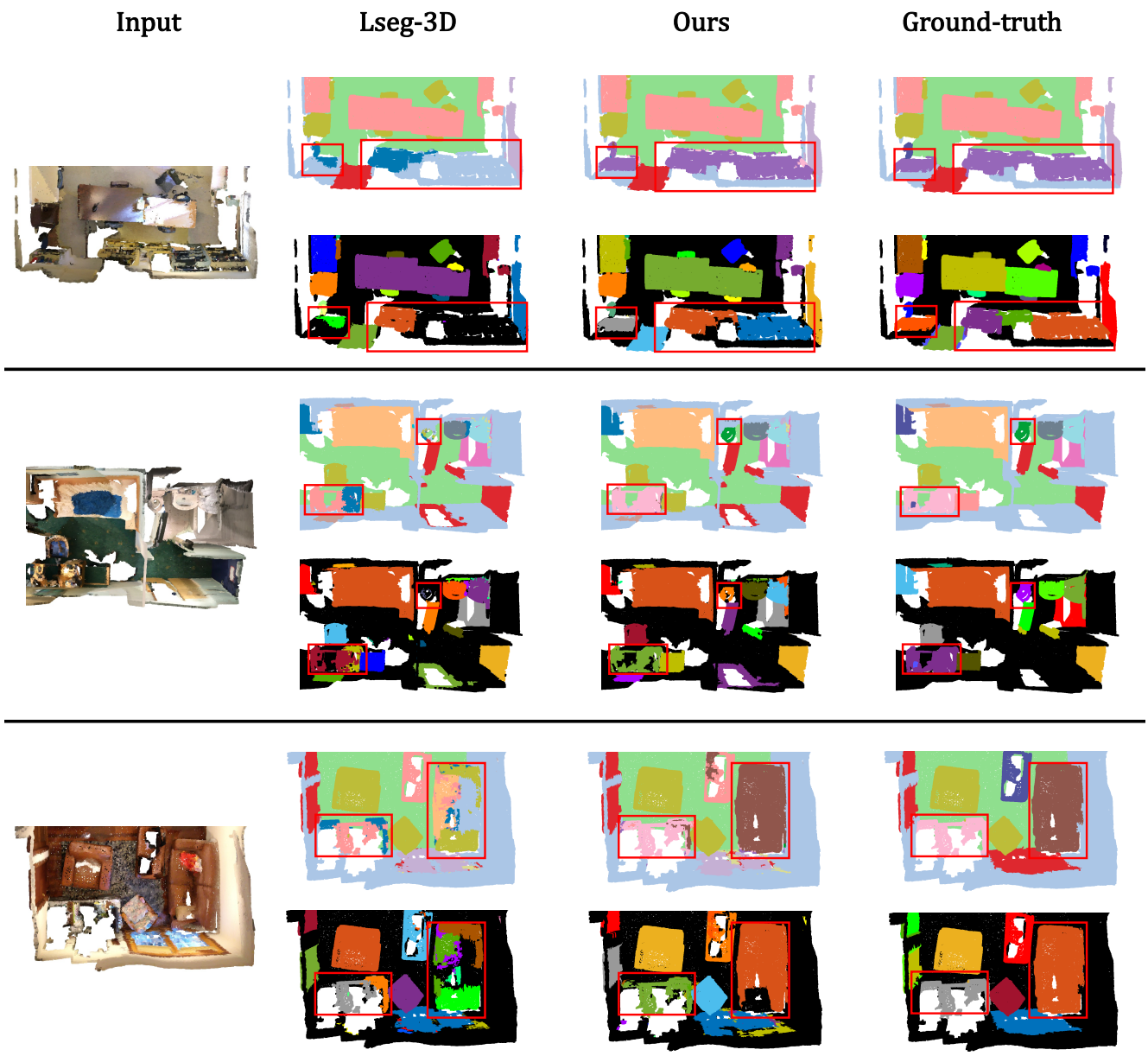


Figure S5. Qualitative results of open-vocabulary semantic segmentation and instance segmentation. In each example, the first row illustrates the semantic masks and the second row shows the instance masks. Novel classes are highlighted in red bounding boxes.