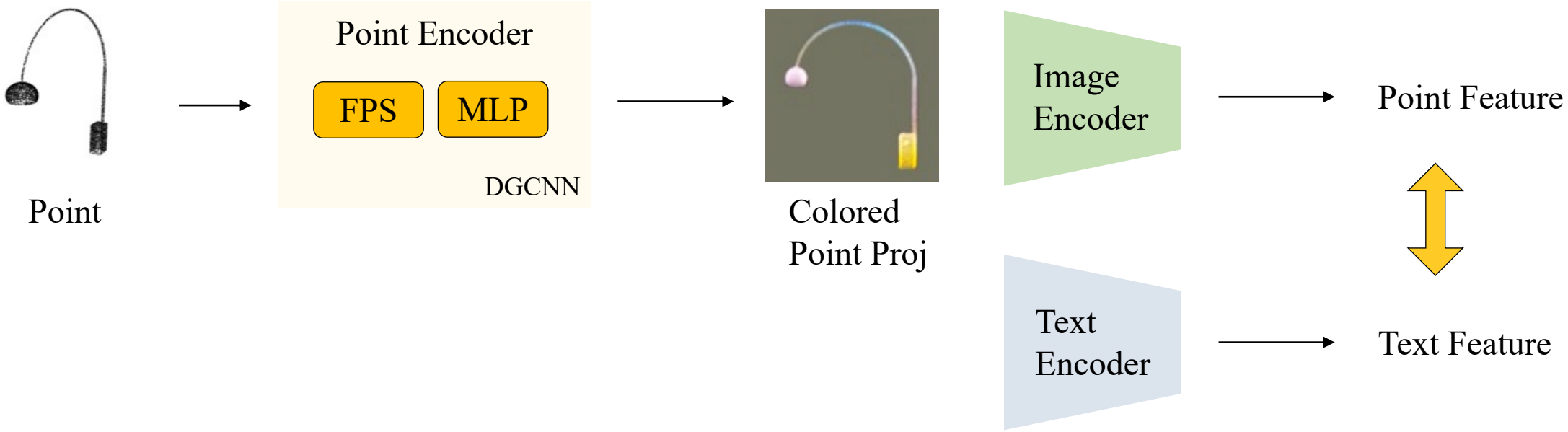
Report on

Bridging 3D-2D Understanding

2023/04/10

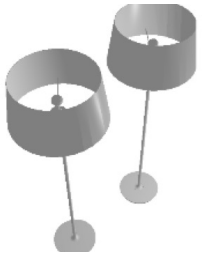
Our Experiments



Results

	Base										Novel								
	airplane	bathtub	bed	bench	bookshelf	bottle	bowl	car	chair	cone	cup	curtain	desk	door	dresser	flower pot	glass box	guitar	keyboard
Acc/ 1e	100	12	59	0	0	97	95	98	95	95	70	60	0	0	0	0	1	84	25
Acc/ 100e	100	46	99	35	85	96	100	99	96	95	75	90	1.16	55	0	45	16	86	5
Diff	0	34	40	35	85	-1	5	1	1	0	5	30	1.16	55	0	45	15	2	-20

Our Experiments



Image

Point

Point Encoder

FPS MLP

DGCNN

Image-Point Guiding

Point Patches

1. Projection

$$F_p = \sigma(F_{C_1}, F_{C_2}, \ldots, F_{C_N})$$

2. Attention

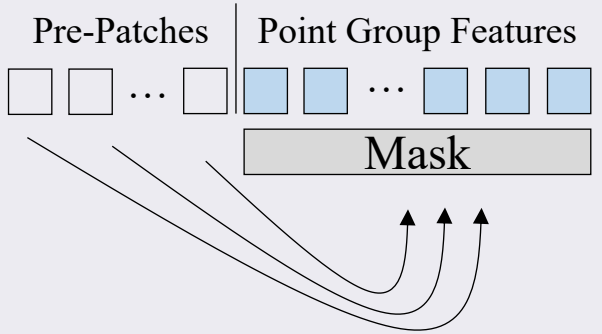


Image Encoder

Point Feature

Text Encoder

Text Feature

Results

- Novel class remains 0!

Why?

- Frozen CLIP Model is An Efficient Point Cloud Backbone
- CLIP goes 3D: Leveraging Prompt Tuning for Language Grounded 3D Recognition

Frozen CLIP Model is An Efficient Point Cloud Backbone

Point

Point Encoder

FPS MLP

PointBert

Point Patches

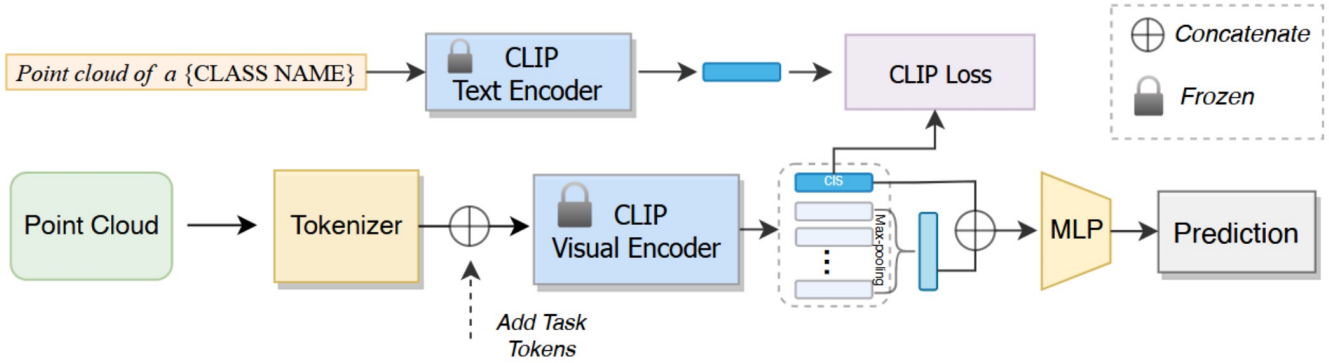
Task Token

Image Encoder

Point Feature

Text Encoder

Text Feature

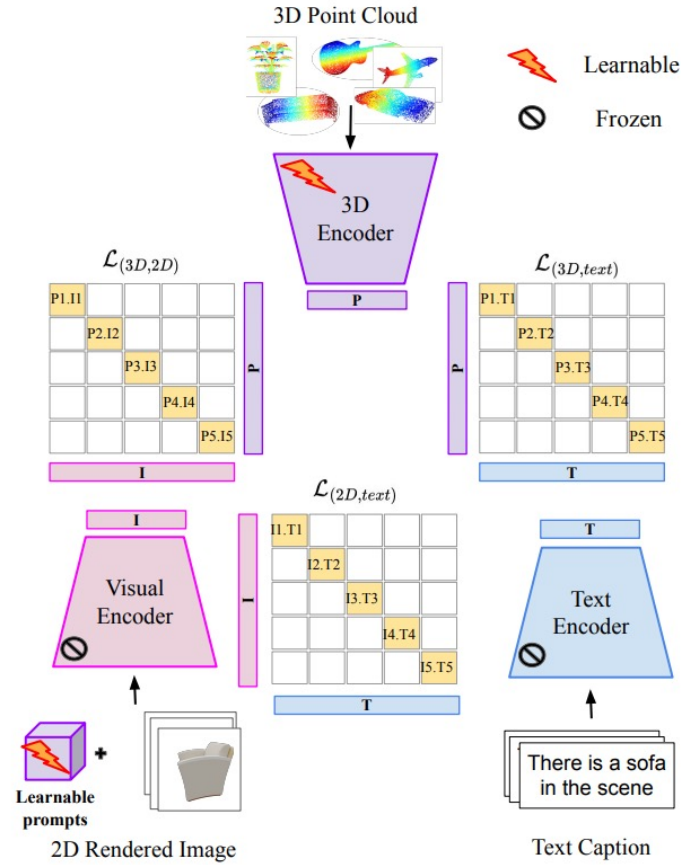


Img-Point Feature Loss: L2

Point-Text Loss: Cross Entropy

Point-Text Loss: contrastive loss function

CLIP goes 3D: Leveraging Prompt Tuning for Language Grounded 3D Recognition



$$\mathcal{L}_{(3D,2D)} = - \sum_{(f^{3D}, f^{2D}) \in B} (\log \text{NCE}(f^{3D}, f^{2D}) + \log \text{NCE}(f^{2D}, f^{3D}))$$

$$\mathcal{L}_{(3D,text)} = - \sum_{(f^{3D}, f^{text}) \in B} (\log \text{NCE}(f^{3D}, f^{text}) + \log \text{NCE}(f^{text}, f^{3D}))$$

Results

Method	3D Pretrain	OA
PointBERT [35]	✓	93.8
MaskPoint [21]	✓	**93.8**
PointNet [22]	✗	89.2
DGCNN [31]	✗	92.2
P2P [32]	✗	92.7
Simple3D-Former [30]	✗	92.0
Ours + w/o CLIP frozen	✗	92.3
Ours	✗	**92.9**

Table 4. Classification on ModelNet40. Our method achieves best accuracy in the methods using 2D pretrained models.

Tuning Method	**5-w,10-s**	**5-w,20-s**	**10-w,10-s**	**10-w,20-s**	**30-w,10-s**
PointBert	94.6 ± 3.1	96.3 ± 2.7	91.0 ± 5.4	92.7 ± 5.1	81.4 ± 2.4
MaskPoint	$\underline{95.0} \pm 3.7$	$\underline{97.2} \pm \underline{1.7}$	$\mathbf{91.4 \pm 4.0}$	$93.4 \pm \mathbf{3.5}$	80.7 ± 4.9
Ours	$\mathbf{95.1 \pm 2.7}$	$\mathbf{97.3 \pm 1.6}$	$\underline{91.1} \pm \underline{4.2}$	$\mathbf{93.5} \pm \underline{3.8}$	$\mathbf{81.7 \pm 0.7}$

Table 3. Few-shot learning accuracy of 3D pretrained methods and frozen CLIP model on ModelNet40.

Method	Overall accuracy	
	ModelNet40	ScanObjectNN
Pointnet [44]	89.2	68.0
Pointnet++ [45]	90.5	77.9
PointCNN [27]	92.2	78.5
DGCNN [59]	92.9	78.1
Point-BERT [69]	93.2	83.07
Point-MAE [42]	93.8	85.18
PointTransformer [74]	91.62 ± 0.29	75.56 ± 0.24
PointTransformer [74] + CG3D	92.93 ± 0.06	80.95 ± 0.54
PointMLP [34]	92.61 ± 0.13	84.08 ± 0.55
PointMLP [34] + CG3D	93.35 ± 0.18	85.78 ± 0.75

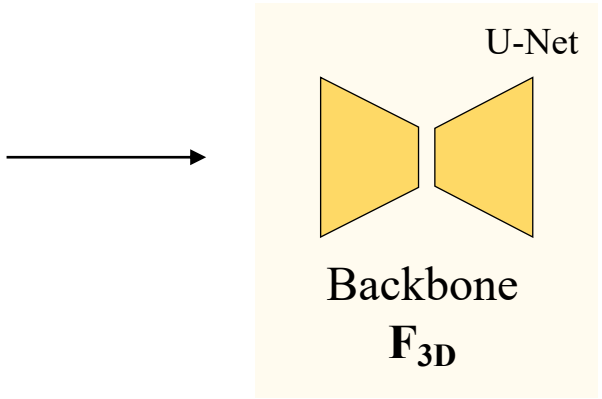
Table 2: Comparison of fine-tuning performance of CG3D with initial weights on ModelNet40 and ScanObjectNN (hardest variation: PB-T50-RS) against previous methods.

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

Closed-set framework

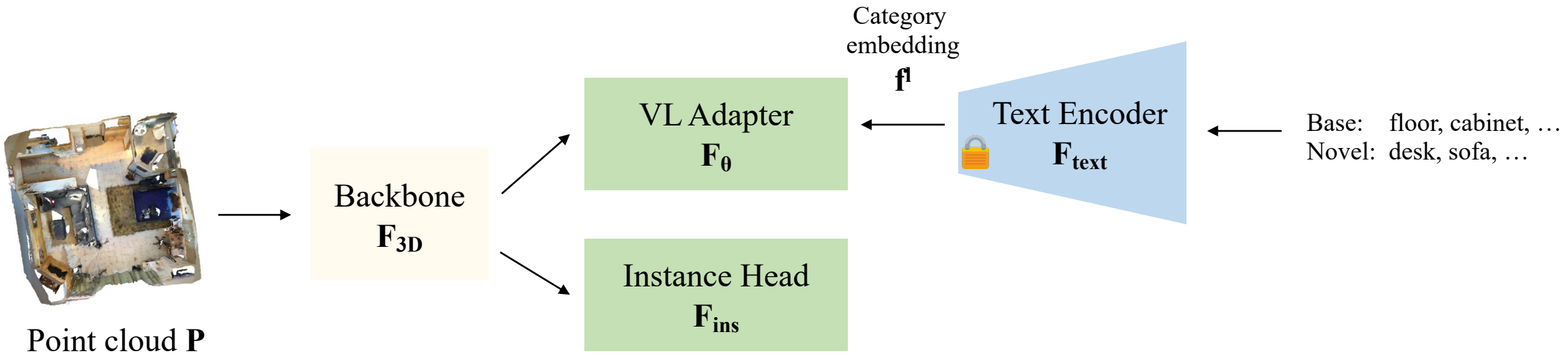
Point cloud **P**



Semantic Head
$\mathbf{F_{sem}}$

Instance Head
$\mathbf{F_{ins}}$

Open-vocabulary framework



Open-vocabulary framework

Point cloud **P**

Backbone
$\mathbf{F_{3D}}$

Binary Head
$\mathbf{F_b}$

VL Adapter
$\mathbf{F_\theta}$

Instance Head
$\mathbf{F_{ins}}$

Caption embedding
$\mathbf{f^t}$

Category embedding
$\mathbf{f^l}$

Text Encoder
$\mathbf{F_{text}}$

Image-bridged Point-Language Association Module

Caption: a living room with …

Base: floor, cabinet, …
Novel: desk, sofa, …

Binary Head

$$s^b = \mathrm{F_b}(f^p), \quad \mathcal{L}_{bi} = \mathrm{BCELoss}(s^b, y^b)$$

$$s = s_B \cdot (1 - s^b) + s_N \cdot s^b$$

Image-bridged Point-Language Association

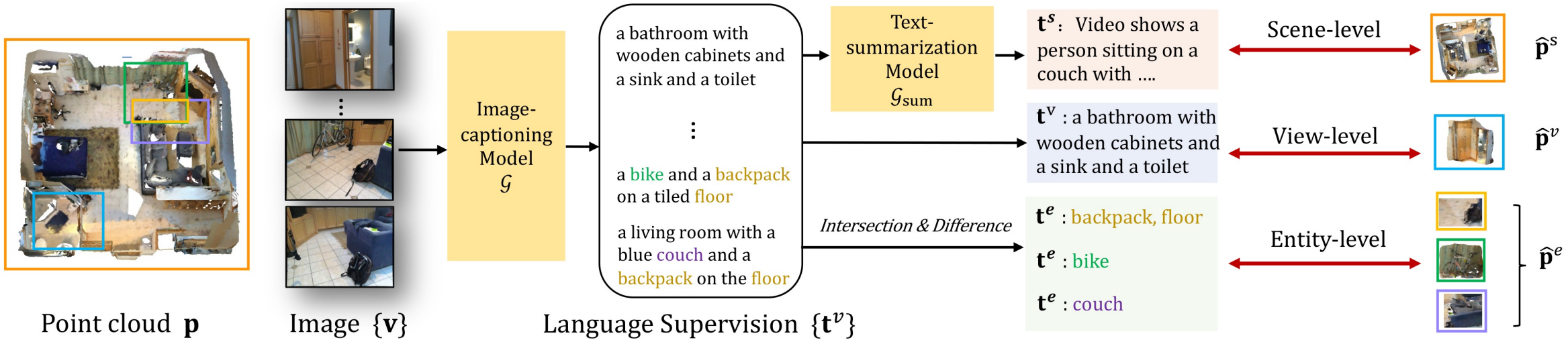


Image-bridged Point-Language Association

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

Feature:

$$f^t = F_{text}(t), \quad f^{\hat{p}} = Pool(\hat{p}, f^v)$$

Contrastive loss:

$$\mathcal{L}_{cap} = -\frac{1}{n_t}\sum_{i=1}^{n_t} \log \frac{\exp\left(\mathbf{f_i^{\hat{p}}} \cdot \mathbf{f_i^t}/\tau\right)}{\sum_{j=1}^{n_t} \exp\left(\mathbf{f_i^{\hat{p}}} \cdot \mathbf{f_j^t}/\tau\right)}$$

Overall loss:

$$\mathcal{L}_{cap}^{all} = \alpha_1 * \mathcal{L}_{cap}^{s} + \alpha_2 * \mathcal{L}_{cap}^{v} + \alpha_3 * \mathcal{L}_{cap}^{e}$$

$$\mathcal{L} = \mathcal{L}_{sem} + \mathcal{L}_{loc} + \mathcal{L}_{cap}^{all} + \mathcal{L}_{bi}$$

Results

Method	$\mathcal{C}^N$ prior	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$^{\mathcal{N}}$
LSeg-3D [26]	×	00.0	64.4	00.0	00.9	55.7	00.1	01.8	68.4	00.9	00.1	49.0	00.1	00.0	30.1	00.0
3DGenZ [28]	✓	20.6	56.0	12.6	19.8	35.5	13.3	12.0	63.6	06.6	08.8	50.3	04.8	09.4	20.3	06.1
3DTZSL [5]	✓	10.5	36.7	06.1	03.8	36.6	02.0	07.8	55.5	04.2	08.4	43.1	04.7	03.5	28.2	01.9
PLA (w/o Cap.)	×	39.7	**68.3**	28.0	24.5	**70.0**	14.8	25.7	75.6	15.5	13.0	58.0	07.4	12.2	54.5	06.8
PLA	×	**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**
PLA (w/ self-train)	✓	70.3	68.9	71.7	61.1	70.4	54.0	59.2	76.9	48.2	36.1	59.7	26.0	46.7	58.9	38.7
Fully-Sup.	✓	73.3	68.4	79.1	70.6	70.0	71.8	69.9	75.8	64.9	67.5	61.4	75.0	65.4	59.9	72.0

Table 2. Results for open-vocabulary 3D semantic segmentation on ScanNet and S3DIS in terms of hIoU, mIoU$^{\mathcal{B}}$ and mIoU$^{\mathcal{N}}$. $\mathcal{C}^N$ prior denotes whether novel category names $\mathcal{C}^N$ need to be known during training. PLA (w/o Cap.) denotes training without point-caption pairs as supervision. Best open-vocabulary results are highlighted in **bold**.

Method	$\mathcal{C}^N$ prior	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$^{\mathcal{N}}_{50}$
LSeg-3D [26]	×	05.1	57.9	02.6	02.0	50.7	01.0	02.4	59.4	01.2	00.5	58.3	00.3	01.1	41.4	00.5
PLA (w/o Cap.)	×	21.0	**59.6**	12.6	11.1	**56.2**	06.2	15.9	**63.2**	09.1	01.8	**59.3**	00.9	01.3	**49.2**	01.2
PLA	×	**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**
PLA (w/ self-train)	✓	58.6	58.0	59.2	41.4	56.9	32.6	42.1	61.1	32.1	26.7	60.3	17.2	23.4	45.6	15.8
Fully-Sup.	✓	64.5	59.4	70.5	62.5	57.6	62.0	62.0	65.1	62.0	57.6	60.8	54.6	57.4	50.0	67.5

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

Ablation Study & Supplementary

Components				hIoU / mIoU$^{\mathcal{B}}$ /mIoU$^{\mathcal{N}}$	hAP$_{50}$ / mAP$^{\mathcal{B}}_{50}$ / mAP$^{\mathcal{N}}_{50}$
Binary	Cap^{s}	Cap^{v}	Cap^{e}		
				00.0 / 64.4 / 00.0	05.1 / 57.9 / 02.6
✓				39.8 / **68.5** / 28.1	21.0 / **59.6** / 12.8
✓	✓			54.6 / 67.9 / 45.7	52.8 / 57.8 / 36.6
✓		✓		61.3 / **68.5** / 55.5	55.9 / 58.9 / 53.3
✓			✓	63.6 / 67.8 / 60.0	**56.6** / 59.0 / **54.4**
✓	✓	✓		61.9 / 68.1 / 56.8	54.9 / 59.5 / 51.0
✓		✓	✓	**65.3** / 68.3 / **62.4**	55.5 / 58.5 / 52.9
✓	✓	✓	✓	64.6 / 69.0 / 60.8	54.5 / 58.2 / 51.4

Table 5. Component analysis on ScanNet. Binary denotes binary head calibration. Cap^{s}, Cap^{v} and Cap^{e} denotes scene-level, view-level and entity-level caption supervision, respectively.

Caption Composition	hIoU / mIoU$^{\mathcal{B}}$ / mIoU$^{\mathcal{N}}$
(a) keep only entities	**65.7 / 69.0 / 62.7**
(b) keep only label names	57.6 / 68.5 / 49.6
(c) ground-truth label names	64.8 / 68.1 / 61.9
(d) full caption	65.3 / 68.3 / 62.4

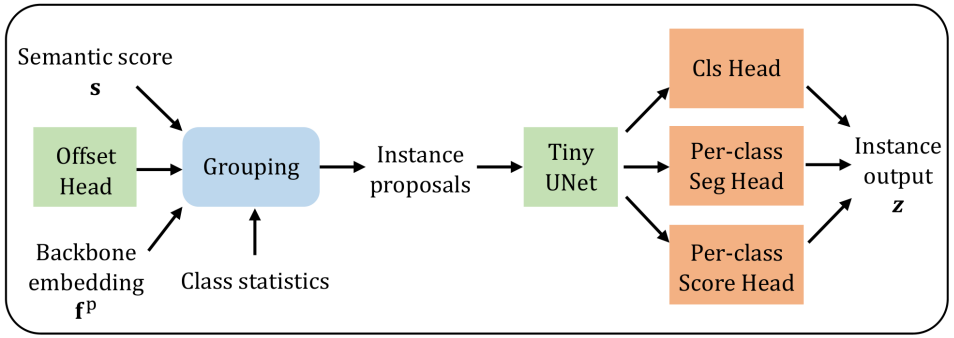
Table 6. Ablation of caption composition.

S1.4. Hyper-Parameter Configurations

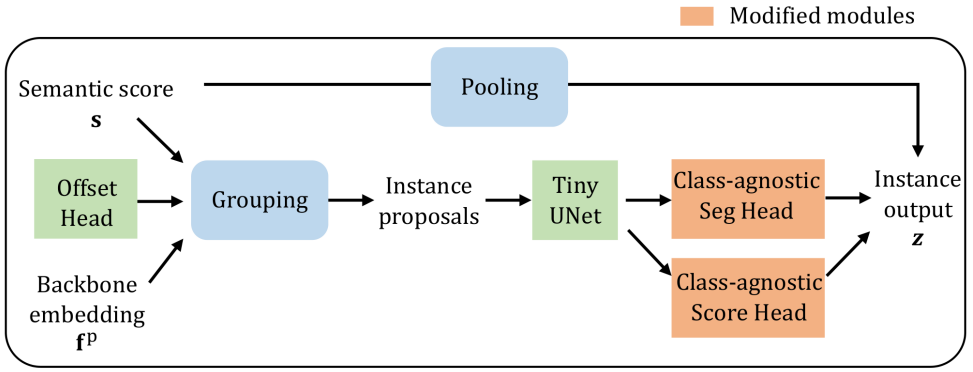
We train 19,216 iterations on ScanNet and 4,080 iterations on S3DIS for semantic segmentation. For instance segmentation, we train 24,020 iterations on ScanNet and 9,160 iterations on S3DIS. The learning rate is initialized as 0.004 with cosine decay. We adopt the AdamW [27] optimizer and run all experiments 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.

Ablation Study & Supplementary



(a) Close-set instance head



(a) Open-vocabulary instance head

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 S15. 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

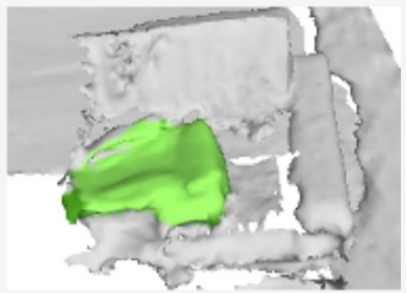
Table S16. Per-class results of 3D open-vocabulary scene understanding on S3DIS.

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

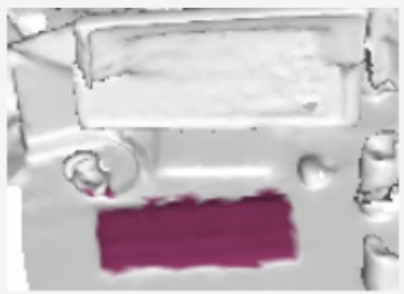
Jihan Yang[1] Runyu Ding[1] Zhe Wang[2] Xiaojuan Qi[1]
[1]The University of Hong Kong [2]SenseTime Research

https://jihanyang.github.io/projects/RegionPLC

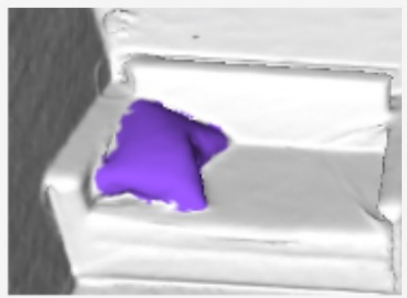
(a) Novel Catgory Segmentation



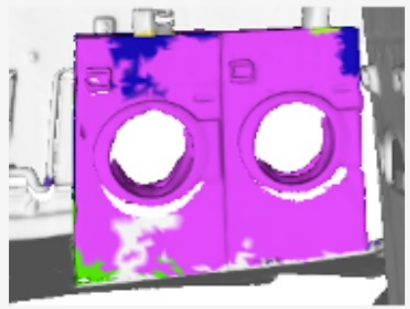
Backpack



Keyboard



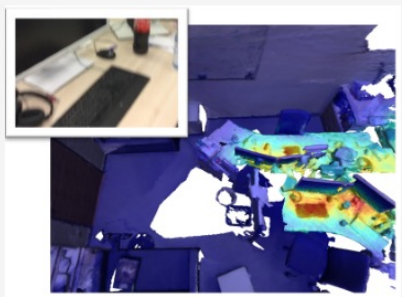
Pillow



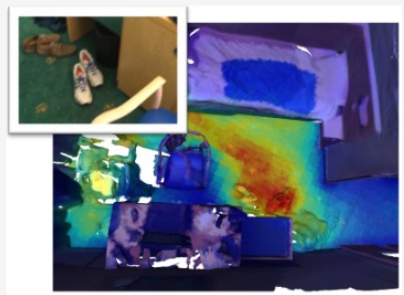
Washing Machine

(b) Novel Catgory Heatmap

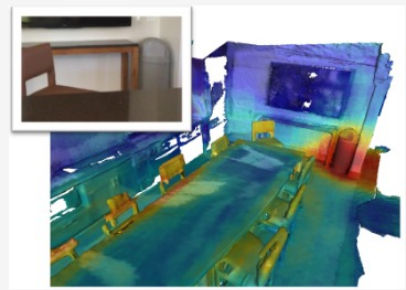
paper



Keyboard

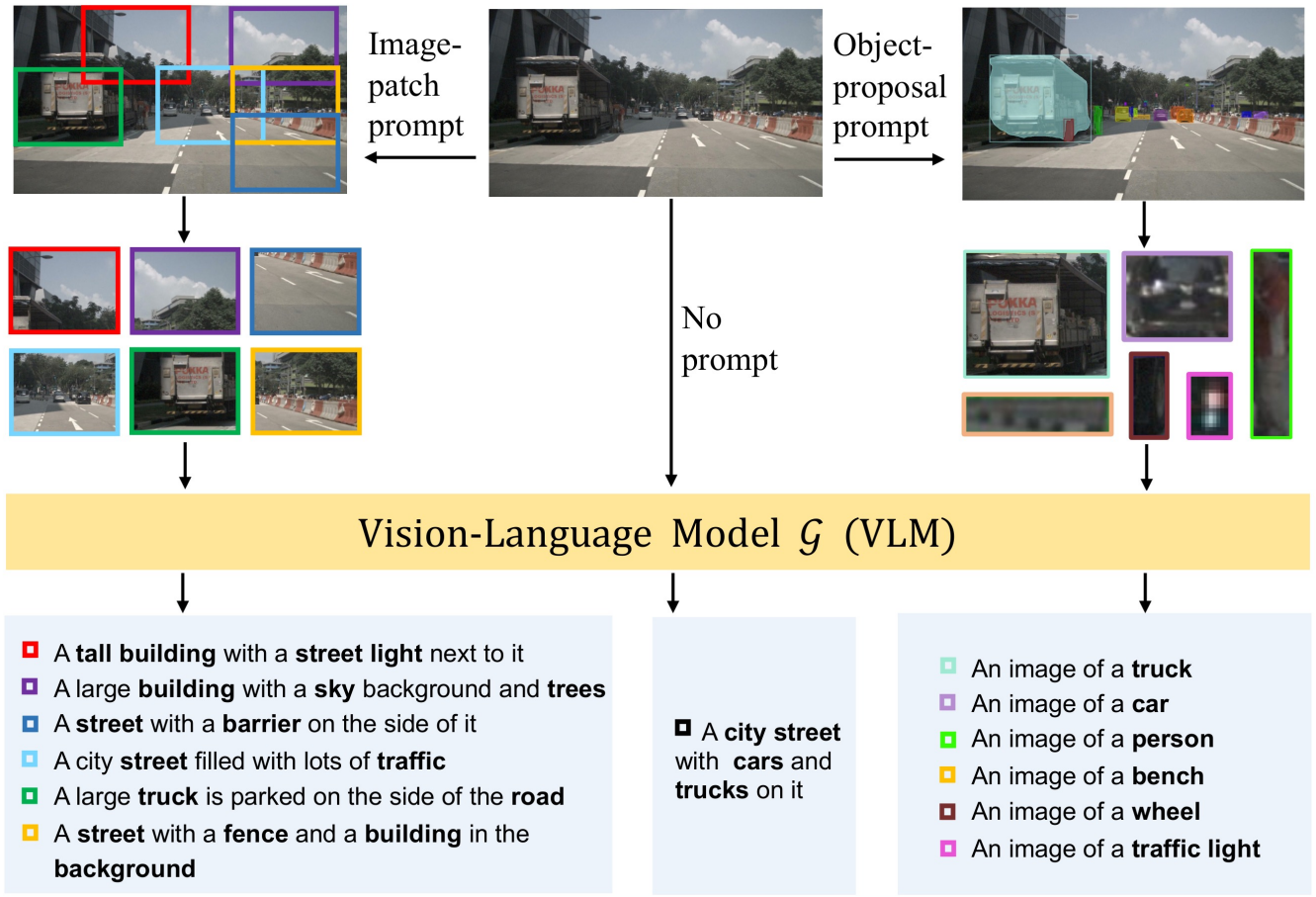


shoes



Trash Can

Motivation



Metrics	$\mathbf{t}^v$ [8]	$\mathbf{t}^e$ [8]	$\mathbf{t}^{obj}$	$\mathbf{t}^{sw}$
# $\mathbf{t}$ Per Scene	16	5	212	142
# Unique Concept	450	270	806	1,496
# Points Per $\mathbf{t}$	24,294	3,933	942	2,479
% Points with $\mathbf{t}$	73%	11%	44%	71%

Table 1. Comparisons among different point-language pairing manners. We use ScanNet frames 25K here.[1]

Point-Discriminative Loss

- **Contrastive loss:**

$$\mathcal{L}_{\text{cap}} = -\frac{1}{\mathrm{n_t}} \sum_{\mathrm{i}=1}^{\mathrm{n_t}} \log \frac{\exp\left(\mathbf{f_i^{\hat{r}}}\right.}{\sum_{j=1}^{n_t} \exp(}$$

$$\hat{z} = f^{\hat{p}} \cdot F^t, \quad \hat{s} = \sigma(\hat{z}, \tau), \quad \mathcal{L}_c = -y^t \cdot \ln \hat{s}$$

- **Pooling after**

$$z = f^p \cdot F^t, \quad s = \sigma(z), \quad \mathcal{L}_{pdc} = -y^t \cdot \mathrm{Pool}(\hat{p}, \ln s)$$

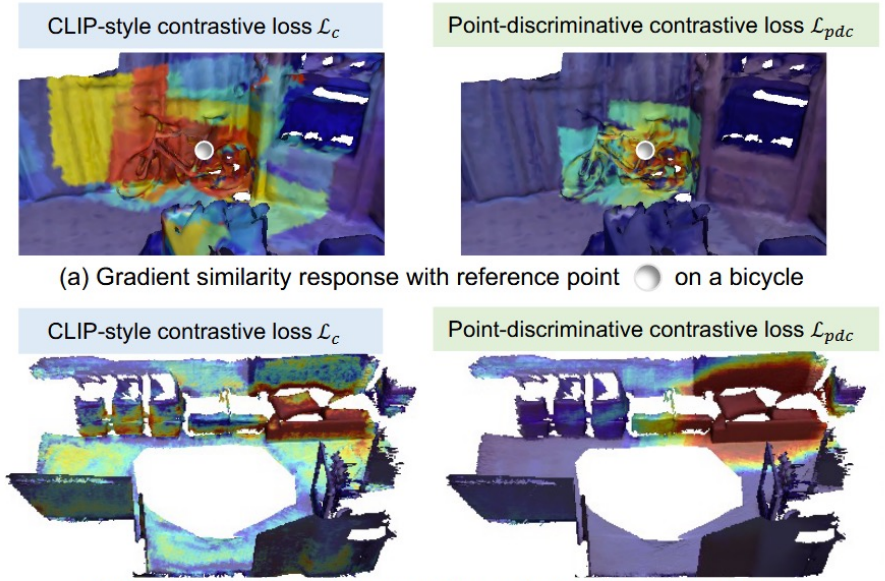


Figure 3. Visualizations of the similarity between gradients and the reference point and between point features and the text embedding of "sofa". Colors closer to red indicate higher similarity.

Results

Method	ScanNet [7]			nuScenes [3]		ScanNet200 [30]	
	B15/N4	B12/N7	B10/N9	B12/N3	B10/N5	B170/N30	B150/N50
3DGenZ [25]	20.6 / 56.0 / 12.6	19.8 / 35.5 / 13.3	12.0 / 63.6 / 06.6	01.6 / 53.3 / 00.8	01.9 / 44.6 / 01.0	02.6 / 15.8 / 01.4	03.3 / 14.1 / 01.9
3DTZSL [5]	10.5 / 36.7 / 06.1	03.8 / 36.6 / 02.0	07.8 / 55.5 / 04.2	01.2 / 21.0 / 00.6	06.4 / 17.1 / 03.9	00.9 / 04.0 / 00.5	00.7 / 03.8 / 00.4
LSeg-3D [8]	00.0 / 64.4 / 00.0	00.9 / 55.7 / 00.1	01.8 / 68.4 / 00.9	00.6 / 74.4 / 00.3	0.00 / 71.5 / 0.00	01.5 / **21.1** / 00.8	03.0 / 20.6 / 01.6
PLA w/o $\mathbf{t}$ [8]	39.7 / 68.3 / 28.0	24.5 / **70.0** / 14.8	25.7 / 75.6 / 15.5	25.5 / **75.8** / 15.4	10.7 / 76.0 / 05.7	07.5 / 21.0 / 04.6	06.4 / 21.0 / 03.8
PLA [8]	65.3 / 68.3 / 62.4	55.3 / 69.5 / 45.9	53.1 / 76.2 / 40.8	47.7 / 73.4 / 35.4	24.3 / 73.1 / 14.5	11.4 / 20.9 / 07.8	10.1 / 20.9 / 06.6
RegionPLC	**69.9 / 68.4 / 71.5**	**65.1** / 69.6 / **61.1**	**58.8 / 76.6 / 47.7**	**62.0 / 75.8 / 52.4**	**36.6 / 76.7 / 24.1**	**15.1** / 20.9 / **11.8**	**12.5 / 21.8 / 08.8**
Fully-Sup.	73.3 / 68.4 / 79.1	70.6 / 70.0 / 71.8	69.9 / 75.8 / 64.9	73.7 / 76.6 / 71.1	74.8 / 76.8 / 72.8	20.9 / 21.7 / 20.1	20.6 / 22.0 / 19.4

Table 2. Results for open-world 3D semantic segmentation on ScanNet and nuScenes in terms of hIoU / mIoU$^{\mathcal{B}}$ / mIoU$^{\mathcal{N}}$. PLA w/o $\mathbf{t}$ denotes training without language supervision in [8]. Best results are presented in **bold**.

Method	ScanNet		
	B13/N4	B10/N7	B8/N9
LSeg-3D [23]	05.1 / 57.9 / 02.6	02.0 / 50.7 / 01.0	02.4 / 59.4 / 01.2
PLA (w/o $\mathbf{t}$) [8]	21.0 / **59.6** / 12.6	11.1 / **56.2** / 06.2	15.9 / **63.2** / 09.1
PLA [8]	55.5 / 58.5 / 52.9	31.2 / 54.6 / 21.9	35.9 / 63.1 / 25.1
RegionPLC	**56.6** / 58.2 / **55.0**	**40.7** / 54.7 / **32.3**	**42.3** / 61.6 / **32.2**
Fully-Sup.	64.5 / 59.4 / 70.5	62.5 / 57.6 / 62.0	62.0 / 65.1 / 62.0

Table 4. Results for open-world 3D instance segmentation on ScanNet in terms of hAP$_{50}$ / mAP$^{\mathcal{B}}_{50}$ / mAP$^{\mathcal{N}}_{50}$.

OpenScene-3D‡ [27]	PLA [8]	RegionPLC ($\mathbf{t}^{\text{sw}}$ only)	Fully-Sup.
5.9 (10.2)	1.8 (3.1)	**6.5** (**15.9**)	23.9 (32.9)

Table 6. Annotation-free open-world semantic segmentation on ScanNet200 [30] in terms of mIoU† (mAcc†).

Method	Images Infer	Task-specific Model	mIoU†	mAcc†	Train Hours	Extra Storage	Latency
PointCLIP-Seg‡ [46]	✓	✓	2.1	5.5	-	-	1.7 s
MaskCLIP‡ [47]	✓	✓	23.1	40.9	-	-	1.7 s
OpenScene-2D‡ [27]	✓	✓	**58.0**	**71.0**	-	-	106.1 s
OpenScene-3D‡ [27]		✓	57.9	70.7	24.8 h	342 G	0.08 s
OpenScene-3D + RegionPLC ($\mathbf{t}^{\text{sw}}$ only)		✓	**60.0**	**75.8**	25.5 h	344 G	0.08 s
PLA‡ [8]			17.7	33.5	11.6 h	1.8 G	0.08 s
RegionPLC ($\mathbf{t}^{\text{sw}}$ only)			**43.8**	**65.6**	12.7 h	1.7 G	0.08 s
Fully-Sup.	-	-	75.9	84.8	11.0 h	-	0.08 s

Table 5. Annotation-free open-world 3D semantic segmentation on ScanNet. ‡ indicates results are reproduced by us.

Ablation Study

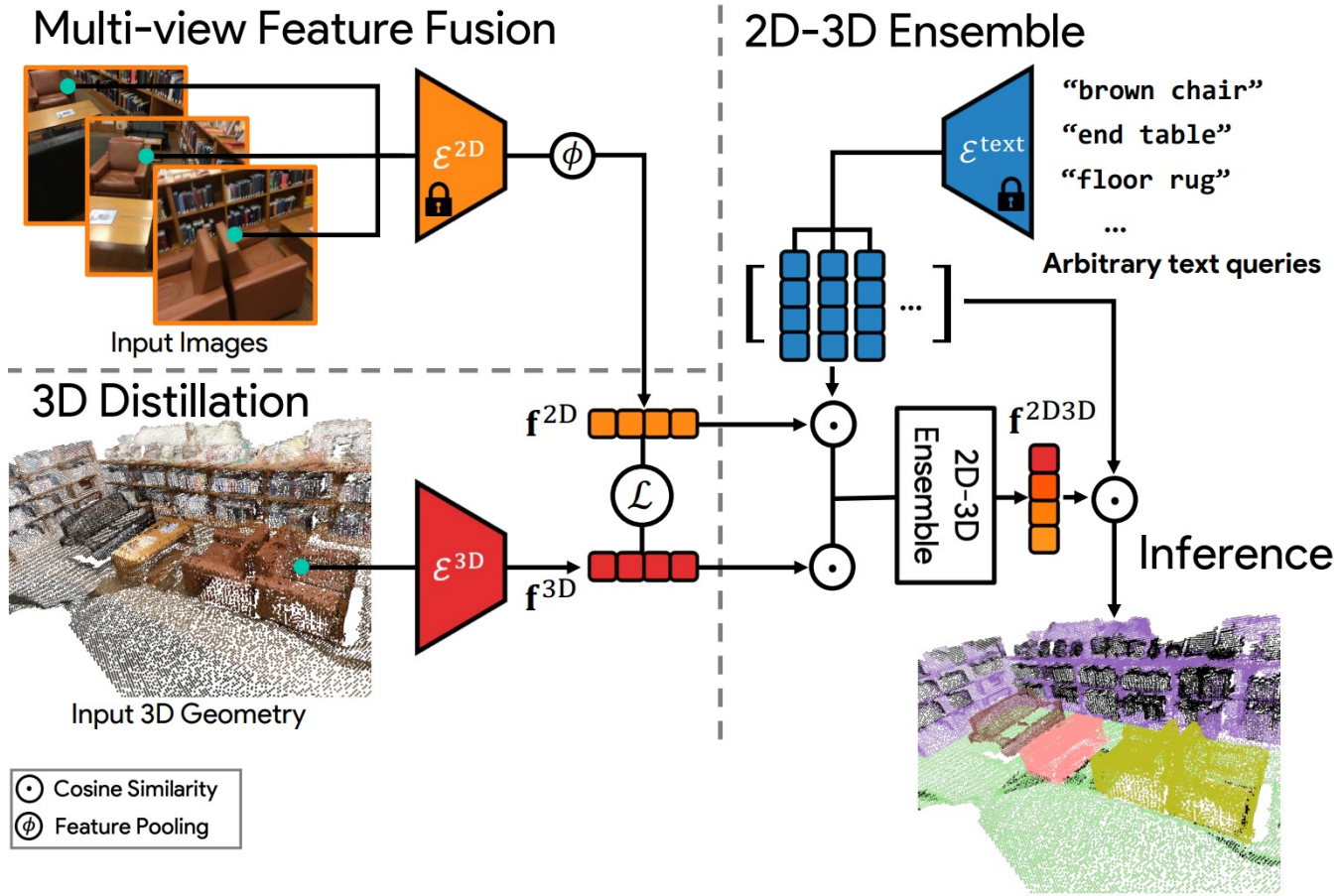
Components			ScanNet	ScanNet200	ScanNet
$\mathbf{t}^{v+e}$	$\mathbf{t}^{r}$	$\mathcal{L}_{\text{pdc}}$	B15/N4	B170/N30	B0/N19
			39.7 / 68.3 / 28.0	07.5 / 21.0 / 04.6	00.3 (05.3)
✓			65.3 / 68.3 / 62.4	11.4 / 20.9 / 07.8	17.7 (33.5)
	✓		66.3 / **68.7** / 64.1	12.8 / 21.1 / 09.2	30.8 (54.6)
✓		✓	67.5 / 68.2 / 66.8	12.1 / **21.3** / 08.5	23.7 (46.9)
	✓	✓	**69.9** / 68.4 / **71.5**	**15.1** / 20.9 / **11.8**	**43.8** (**65.6**)

Table 7. Component analysis on ScanNet and ScanNet200. $\mathbf{t}^{v+e}$, $\mathbf{t}^{r}$ and $\mathcal{L}_{\text{pdc}}$ denotes the combination of view and entity language supervision [8], our region-level language supervision and our point-discrimination contrastive loss, respectively.

OpenScene: 3D Scene Understanding with Open Vocabularies

Songyou Peng[1,2,3] Kyle Genova[1] Chiyu "Max" Jiang[4] Andrea Tagliasacchi[1,5]
Marc Pollefeys[2] Thomas Funkhouser[1]

[1] Google Research [2] ETH Zurich [3] MPI for Intelligent Systems, Tübingen [4] Waymo LLC [5] Simon Fraser University



2D backbone: LSeg / OpenSeg

3D backbone: MinkowskiNet18A

$$L = 1 - \cos(\mathbf{F}^{2D}, \mathbf{F}^{3D})$$