IMPERIAL

NEURAL INFORMATION PROCESSING SYSTEMS

MatchLab

OpenDlign: Enhancing Open-World 3D Learning with Depth-Aligned Images

Ye Mao, Junpeng Jing, Krystian Mikolajczyk

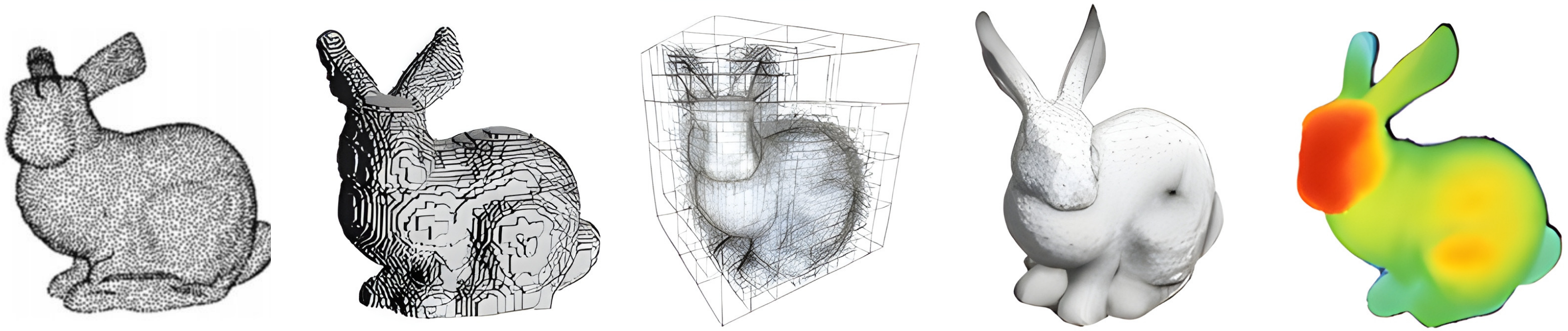
https://yebulabula.github.io/OpenDlign/

IMPERIAL

MatchLab

What is open-world 3D learning?

- 3D learning: Learn representations/features from 3D data (e.g., point cloud, depth map)

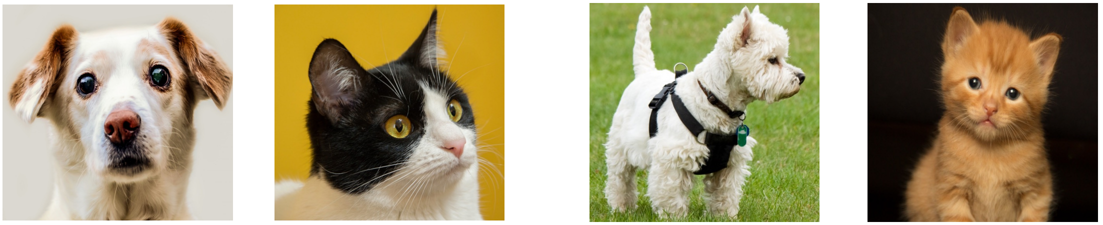


Point Cloud Voxel Octree Mesh Depth Map

3D Representations → Classification / Detection / Segmentation

- Open-World: Generalize beyond 'seen' categories in pretraining dataset.

Train Data Test Data



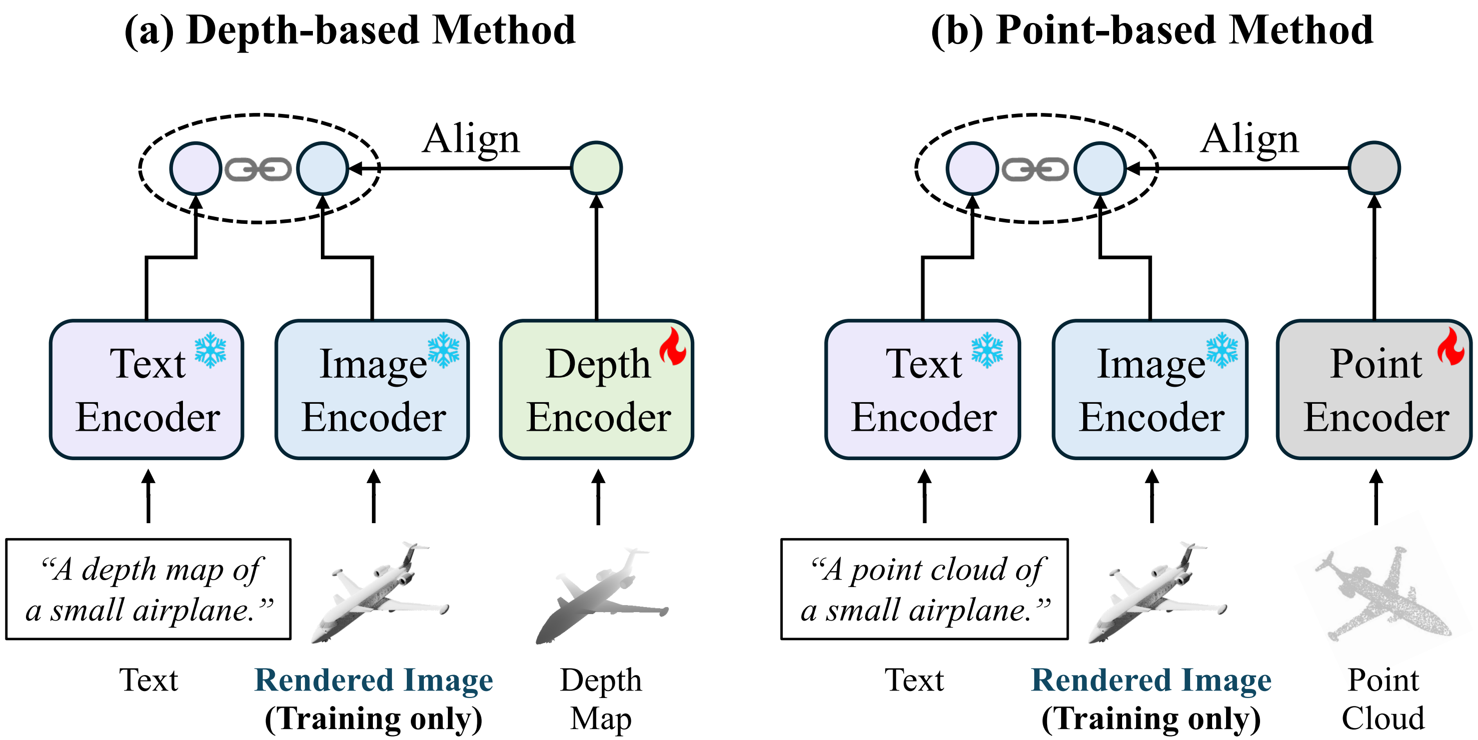
closed-world model

Train Data Test Data

open-world model

IMPERIAL

Previous methods: Depth-based vs Point-based open-world 3D models.



Rendered image:

- Multi-view, rendered from CAD models, used for training only.

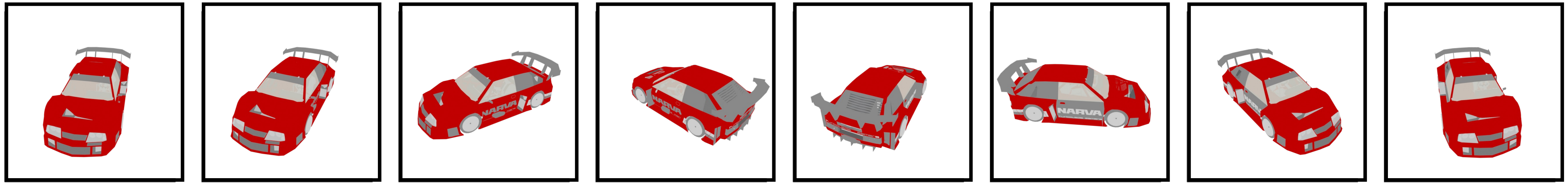
Image and text modalities are pre-aligned by CLIP.

Feature alignment of 3D data, CAD-rendered images, and text is a standard practice.

IMPERIAL

CLIP's secret of success: Pretraining on an extremely diverse set of image-text pairs.

- 3D dataset is scarce ($\leq$ 1M, synthetic). CLIP Pretraining data: LAION5B, DFN5B (5B).
- CAD-rendered images: unrealistic, simplistic texture, limited visual diversity.



Example of multi-view CAD-rendered images.

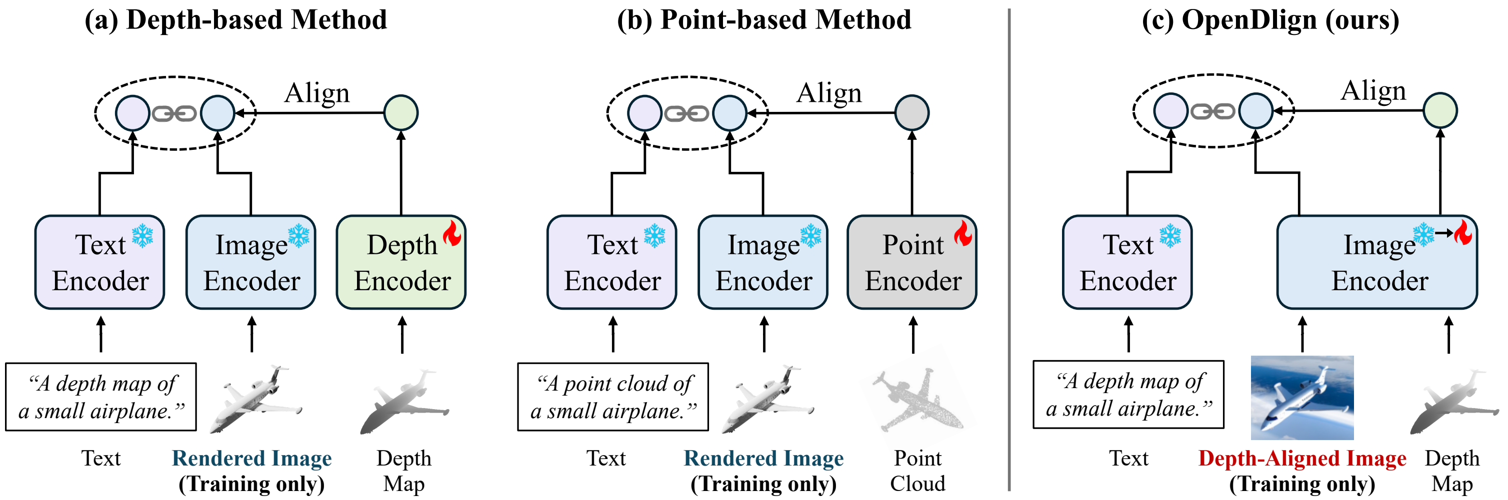
Directly training a 3D version of CLIP from scratch is impractical.

IMPERIAL

OpenDlign: a novel open-world 3D model.

- Align 3D data with visually diverse *depth-aligned images* instead of CAD renderings.

- **Fine-tunes** CLIP to maximally leverage its extensive knowledge for 3D learning, avoiding the need to **train a separate encoder from scratch**.



IMPERIAL

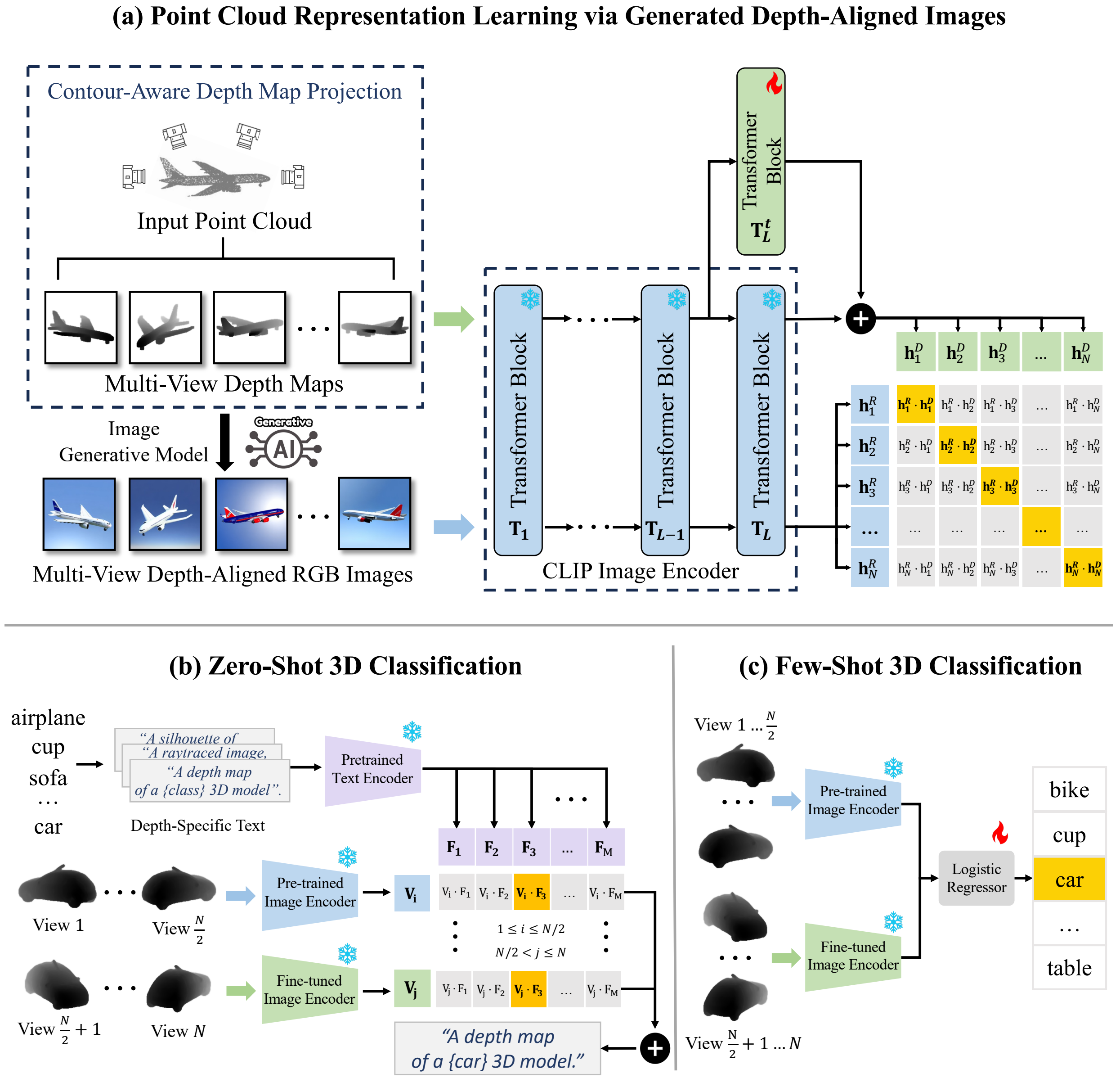
1. Project point cloud into multi-view depth maps with clear contours.

2. Use projected depth maps to generate depth-aligned images using ControlNet.

3. Contrastive learning between features from depth maps and depth-aligned images *(6M param)*.

4. Multi-view logit aggregation, avoid catastrophic forgetting.

5. Depth-specific text prompts for 3D zero-shot classification.



IMPERIAL

MatchLab

Experiment: zero-shot classification

- OpenDlign outperforms the leading baseline by 8.0% on ModelNet40, 16.4% on OmniObject3D.
- Using depth-aligned images consistently enhances the performance of other SOTA models

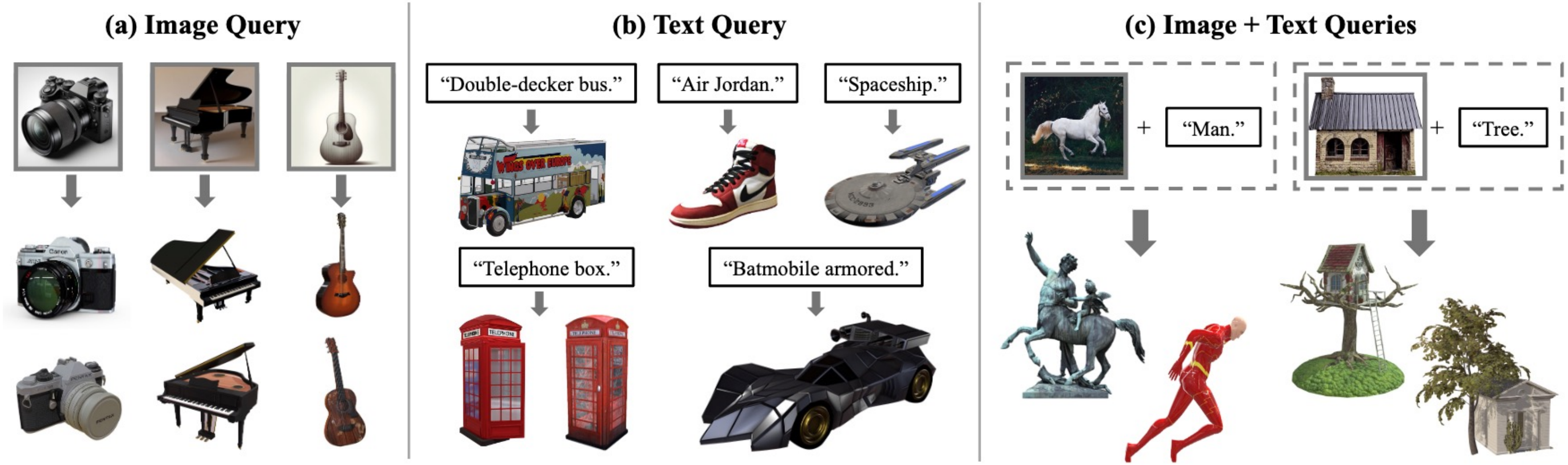
Table 1: Zero-shot classification results on ModelNet40 [50], ScanObjectNN [51] and OmniObject3D [52]. The best-performing results are presented in bold, while the second-best results are underlined. Our models are highlighted in blue.

Training Source	3D Open-World Methods	CLIP Variant	ModelNet40 [50] Top1	ModelNet40 [50] Top3	ModelNet40 [50] Top5	ScanObjectNN [51] Top1	ScanObjectNN [51] Top3	ScanObjectNN [51] Top5	OmniObject3D [52] Top1	OmniObject3D [52] Top3	OmniObject3D [52] Top5
2D inferences No Training	PointCLIP [16]	ResNet-50	19.3	28.6	34.8	10.5	20.8	30.6	0.3	1.0	1.8
	PointCLIP V2 [15]	ViT-B-16	63.6	77.9	85.0	42.2	63.3	74.5	3.9	9.6	14.4
ShapeNet	CLIP2Point [17]	ViT-B-32	49.5	71.3	81.2	25.5	44.6	59.4	1.4	3.7	7.1
	ULIP-PointBERT [19]	SLIP [54]	60.4	79.0	84.4	51.5	71.1	80.2	8.4	15.2	19.7
	OpenShape-SparseConv [20]	ViT-bigG-14	72.9	87.2	93.0	52.7	72.7	83.6	13.7	24.2	30.0
	OpenShape-PointBERT [20]	ViT-bigG-14	70.3	86.9	91.3	51.3	69.4	78.4	13.0	23.3	29.4
	TAMM-SparseConv [23]	ViT-bigG-14	74.6	88.2	94.0	57.9	75.3	83.1	-	-	-
	TAMM-PointBERT [23]	ViT-bigG-14	73.1	88.5	91.9	54.8	74.5	83.3	14.9	26.2	33.4
	OpenShape-SparseConv (+dlign)	ViT-bigG-14	74.9	89.5	94.1	56.3	75.2	**85.4**	15.0	26.1	32.8
	OpenShape-PointBERT (+dlign)	ViT-bigG-14	73.7	87.1	91.3	52.7	72.4	82.6	13.4	23.7	29.9
	TAMM-PointBERT (+dlign)	ViT-bigG-14	73.7	89.1	92.2	57.3	73.6	82.3	15.8	27.4	33.0
	OpenDlign-B32	ViT-B-32	68.4	86.4	92.6	46.7	72.0	83.0	17.3	29.2	36.3
	OpenDlign-B16	ViT-B-16	74.2	90.5	95.4	49.3	74.0	84.4	23.2	37.5	44.3
	OpenDlign-L	ViT-L-14	77.8	93.1	96.4	52.1	74.6	82.8	27.5	41.3	47.8
	OpenDlign	**ViT-H-14**	**82.6**	**96.2**	**98.4**	**59.5**	**76.8**	83.7	**31.3**	**46.7**	**53.2**
Ensemble	OpenShape-SparseConv [20]	ViT-bigG-14	83.4	95.6	97.8	56.7	78.9	88.6	33.7	49.3	57.4
	OpenShape-PointBERT [20]	ViT-bigG-14	84.4	96.5	98.0	52.2	79.7	88.7	34.0	49.7	57.9
	TAMM-PointBERT [23]	ViT-bigG-14	85.0	96.6	98.1	55.7	80.7	88.9	37.1	53.5	61.8
	TAMM-SparseConv [23]	ViT-bigG-14	85.4	96.4	98.1	58.5	81.3	89.5	-	-	-
	OpenShape-SparseConv (+dlign)	ViT-bigG-14	85.0	96.1	97.9	56.2	78.5	87.8	34.1	50.5	58.5
	OpenShape-PointBERT (+dlign)	ViT-bigG-14	85.4	96.6	**98.2**	51.1	77.4	88.2	35.6	50.4	57.9
	TAMM-PointBERT (+dlign)	**ViT-bigG-14**	**86.2**	**96.6**	97.5	**60.5**	**82.5**	**90.4**	**37.5**	**54.9**	**62.1**

IMPERIAL

MatchLab

Cross-modal retrieval



3D Object Detection

Table 3: Zero-shot 3D object detection results on ScanNet V2 [55].

	Method	Mean	Cabinet	Bed	Chair	Sofa	Table	Door	Window	Counter	Desk	Sink	Bathtub
AP_{25}	PointCLIP [16]	6.00	3.99	4.82	45.16	4.82	7.36	4.62	2.19	1.02	4.00	13.40	6.46
	PointCLIP V2 [15]	18.97	19.32	20.98	61.89	15.55	23.78	13.22	17.42	12.43	21.43	14.54	16.77
	OpenDlign (ours)	**50.72**	**38.91**	**67.27**	**86.33**	**72.01**	**58.72**	**44.58**	**32.07**	**50.49**	**62.04**	**51.98**	**64.29**
AP_{50}	PointCLIP [16]	4.76	1.67	4.33	39.53	3.65	5.97	2.61	0.52	0.42	2.45	5.27	1.31
	PointCLIP V2 [15]	11.53	10.43	13.54	41.23	6.60	15.21	6.23	11.35	6.23	10.84	11.43	10.14
	OpenDlign (ours)	**37.97**	**17.04**	**66.68**	**73.92**	**54.96**	**50.03**	**24.73**	**12.84**	**20.44**	**41.64**	**34.17**	**64.29**