Supplementary of Uni3DL: A Unified Model for 3D Vision-Language Understanding

Xiang Li[1,*], Jian Ding[1,*], Zhaoyang Chen[2,†], Mohamed Elhoseiny[1]

[1] King Abdullah University of Science and Technology
{xiang.li.1,jian.ding,mohamed.elhoseiny}@kaust.edu.sa
[2] École Polytechnique
zhaoyang.chen@polytechnique.edu

1 Experimental Settings

1.1 Pretraining

As described in the main paper, we use the ScanNet (v2), ScanRefer, and Cap3D Objaverse datasets for joint pretraining. For the Cap3D Objaverse caption dataset, we only include objects whose captions contain any object name from the ScanNet, S3DIS, or ModelNet categories. Our Uni3DL model is pretrained for 10,000 steps. We set the initial learning rate to 1e-4 and reduce it by 0.1 after 50% and 70% of the total training steps. A linear warmup is applied for the first 10 iterations.

During pertaining, we set voxel size to 0.02m for 3D scans (e.g., S3DIS) and 0.01 for normalized 3D shapes (e.g., Cap3D Objaveerse), with a batch size of 8 for 3D scans and 12 for 3D-text pairs. For the S3DIS/ScanNet datasets, we randomly crop $5m \times 5m \times 5m$ blocks from each scene, ensuring a minimum of 25,000 points per scene. Input scenes are augmented by random flips along the X and Y axes, and rotations along the X, Y, and Z axes. Color augmentations, including jittering, brightness, and contrast adjustments, are also applied.

1.2 Finetuning

Finetuning for 3D semantic/instance Segmentation. Our Uni3DL model is finetuned for 25 epochs with an initial learning rate of 2e-5, which is multiplied by 0.1 after 50% and 70% of the total finetuning steps. For ScanNet segmentation, we finetune our Uni3DL model for 30 epochs on ScanNet with the same learning rate strategy as in S3DIS segmentation.
Finetuning for Grounded Segmentation. We finetune our Uni3DL model on grounded segmentation for 20 epochs with an initial learning rate of 1e-5, decaying it by 0.1 after reaching 50% and 70% of the total training steps.
Finetuning for 3D Captioning. Our Uni3DL model is finetuned for 30 epochs on the Cap3D Objaverse dataset. The learning rate starts at 1e-4 and is reduced by 0.2 after 50% and 70% of the training steps.

* Equal contribution
† This work was done when Zhaoyang Chen was an intern at KAUST.

Finetuning for Text-3D Cross-Modal Retrieval. We finetune the Uni3DL model for 30 epochs on the Text2Shape retrieval task, following a similar learning rate strategy as in 3D Captioning. For data augmentation, we apply random scaling to the training shapes, using a scale factor uniformly sampled from the range [0.8, 1.2]. Additionally, we randomly rotate the shapes along the z-axis, selecting rotation angles within the range $[-\pi/2, \pi/2]$.

2 More quantative results

2.1 Zero-Shot 3D Classification

We use our Uni3DL model fine-tuned on the Cap3D Objaverse dataset to evaluate zero-shot 3D classification performance on ModelNet40 and ModelNet10 datasets. ModelNet40 includes 40 different categories with 12, 311 CAD models, while ModelNet10, a smaller subset, consists of 10 categories with 4, 899 models. We use the same validation set as [7] for performance evaluation.

Table 1 summarizes the performance on ModelNet10 and ModelNet40 test datasets. From this Table, we can see that our method achieves competitive performance on both datasets, with a top-1 classification accuracy of 70.4% on ModelNet10 and 57.0% on ModelNet40. Moreover, our Uni3DL achieves the best top-5 classification accuracy on ModelNet40 dataset. *It should be noted that all compared methods rely on projecting 3D data to multiview 2D images and use a pretrained CLIP for image-text alignment; while our method does not require view projection.*

Method	Input	Pretraing dataset	Pretrained FM	ModelNet10	ModelNet40	
				top-1	top-1	top-5
PointCLIP [10]	MV Images	ShapeNet	Yes (CLIP)	30.2	23.8	-
CLIP2Point [6]	MV Images	ShapeNet	Yes (CLIP)	66.6	49.4	-
PointCLIP V2 [13]	MV Images	ShapeNet	Yes (CLIP+GPT3)	**73.1**	<u>64.2</u>	-
ULIP [8]	MV Images	ShapeNet	Yes (CLIP)	-	60.4	<u>84.0</u>
ULIP [8]	MV Images	Cap3D Objaverse	Yes (CLIP)	-	**67.2**	83.1
Ours	Point Cloud	Cap3D Objaverse	No	<u>70.4</u>	57.0	**88.8**

Table 1: Zero-shot 3D shape classification performance on ModelNet10 and ModelNet40 datasets. We show input types, pretrained datasets, and foundation model (FM) requirements for detailed comparison. Our method does not require projected multi-view images as inputs and does not require pretrained foundation models. The results highlighted in **bold** and <u>underline</u> denote the best and second-best performance, respectively.

2.2 Grounded Localization

In the main paper, we report the performance of our Uni3DL model for grounded *segmentation*. Previous methods have also explored the grounded *localization* task. To produce grounded object location, we directly use grounded object

Model	Single Stage	Detector	Overall	
			Acc@0.25	Acc@0.5
ScanRefer [2]	✗	VoteNet	39.0	26.1
InstanceRefer [9]	✗	PointGroup	38.2	31.4
3DVG-Transformer [12]	✗	VoteNet	45.9	34.5
3DJCG [1]	✗	VoteNet	**47.6**	36.1
D3Net [3]	✗	PointGroup	-	35.6
UniT3D [4]	✗	PointGroup	-	36.5
M3DRef [11]	✗	PointGroup	-	**40.4**
TGNN [5]	✓	N/A	37.4	29.7
Uni3DL (Ours)	✓	N/A	**37.8**	**33.7**

Table 2: Comparative analysis of grounded localization performance on the ScanRefer [2] dataset. We report the ratios of correctly predicted bounding boxes with IoU thresholds of 0.25 and 0.5. We report the performance of all comparing methods with only 3D point clouds as inputs.

masks to calculate their bounding boxes. Table 2 summarizes the performance of Uni3DL and previous state-of-the-art methods for grounded localization. Note that all compared methods, except for TGNN [5], employ a dual-stage process, where a 3D object detector identifies potential bounding box candidates, followed by a disambiguation module employed to fuse visual and textural features and determine the precise target bounding box. *In contrast, our Uni3DL model is a single-stage model, without using second-stage object-text fusion modules.* Specifically, our Uni3DL model achieves better performance than another single-stage model TGNN [5] which also generates bounding boxes from object segmentation masks. It should be noted that in the *grounded localization* task, akin to TGNN, our model tends to underperform, likely due to our basic approach in generating bounding boxes from masks. Minor boundary inaccuracies in segmentation masks minimally impact segmentation IOU, but can significantly alter bounding box locations. As our focus is primarily on *grounded segmentation*, minimizing bounding-box loss is not a priority.

3 More qualitative results

3.1 3D Captioning

We show more qualitative results of 3D captioning on the Cap3D objaverse dataset in Fig. 1. As shown in the figure, our Uni3DL can generate text descriptions well aligned with ground truth captions.

3.2 Grounded Segmentation

Fig. 2 presents additional grounded segmentation results on the ScanRefer dataset. As illustrated, our Uni3DL model accurately predicts the grounded masks corresponding to each referring sentence.

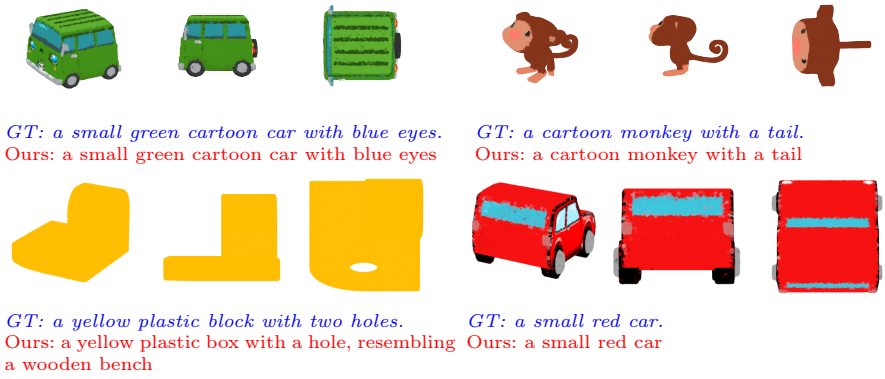

Fig. 1: 3D captioning results on Cap3D Objaverse dataset.

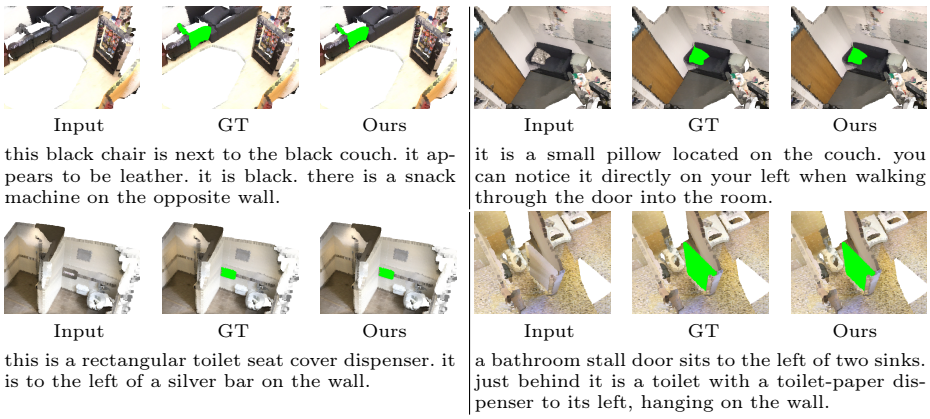

Fig. 2: Results of grounded segmentation on ScanRefer dataset.

3.3 Text-3D cross-modal retrieval

We show text-to-3D and 3D-to-text retrieval results in Fig. 3 and Fig. 4 respectively. From the two figures, our Uni3DL model learns satisfying text-3D feature alignments and produces satisfying cross-modal retrieval results.

4 Limitation and Future Work

In this study, we introduced Uni3DL, a novel unified model for 3D vision-language understanding, operating directly on raw point clouds. This approach differs from conventional 3D vision-language models that predominantly rely on projected multi-view images. While these projection-based methods are limited by their handling of geometric information, their integration with powerful 2D

	top1	top2	top3	top4	top5
a round table with differnt type of look and is good	 **0.91(GT)**	 0.90	 0.90	 0.90	 0.88
it is an oblong table with distressed wooden top and six spindle shaped legs.	 **0.91(GT)**	 0.86	 0.86	 0.83	 0.83
a red sofa that is sitting on a black carpet. the sofa is round and ovalular.	 0.90	 **0.90(GT)**	 0.87	 0.86	 0.82
a unique design brown wooden table with white color at top is great for outdoor	 0.94	 0.87	 **0.83(GT)**	 0.82	 0.81

Fig. 3: Text-to-Shape Retrieval results on Text2Shape dataset, For each query sentence, we show the top-5 ranked shape, the scores of ground truth shape are marked in red.

pretrained foundation models, such as CLIP, has yielded promising results. To leverage the benefits of both point-based and projection-based techniques, our future work will focus on a hybrid approach. This strategy aims to simultaneously learn joint 2D and 3D features, integrating insights from 2D foundation models. This advancement is expected to significantly enhance the sophistication and accuracy of 3D vision-language understanding in upcoming versions of our model.

Query Shape	Retrieval Results
	**1. it is an oblong table with distressed wooden top and six spindle shaped legs. (Prob: 0.91, GT)** 2. elliptical table with brown wooden top and grey straight legs (Prob: 0.88) 3. a brown oblong wooden topped table with four grey supporting legs (Prob: 0.86) 4. oval table with shape oval , 4 legs and high qualit wood from alaska that will make you happy (Prob: 0.86) 5. this is a dining table that is oval with the insert, but could collapse down to a circle table. it has 4 legs. (Prob: 0.85)
	1. a grey rectangular shaped wooden table with four short legs. (Prob: 0.91) **2. grey colored, wooden table. four short solid legs with rectangular top. (Prob: 0.89, GT)** 3. a grey rectangular short table with four short grey legs. (Prob: 0.89) 4. a white colored rectangular table which has rectangular top painted in white and has four short legs colored in black. (Prob: 0.89) **5. a low and long grey table with four legs. (Prob: 0.89, GT)**
	**1. a white conference table with legs (Prob: 0.92, GT)** 2. a table with a white colored oval type top and four grey colored plate type legs (Prob: 0.88) 3. simple white table. lunch room table. 4 legs. metal legs. formica top. wide. (Prob: 0.87) 4. an ash colored oval shaped steel coffee table which has skinny rectangular shaped long four legs. (Prob: 0.87) 5. outdoor table, wooden, gray, oval shape, with four legs. (Prob: 0.86)
	**1. red colour plastic chair with u shape iron legs and chair was looking variety (Prob: 0.90, GT)** 2. a red chair with curved back legs. probably made of plastic. (Prob: 0.89) 3. a basket backed, red seated high bar stool with thin metal legs (Prob: 0.88) **4. red high back chair made of plastic. four legs are made of metal. (Prob: 0.88, GT)** **5. this is a red molded chair with back and no arms. the chair has 4 metal/plastic legs. (Prob: 0.87, GT)**

Fig. 4: Shape-to-Text Retrieval results on Text2Shape dataset, For each query shape, we show the top-5 ranked sentences, the ground truth sentences are marked in red.

References

1. Cai, D., Zhao, L., Zhang, J., Sheng, L., Xu, D.: 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16464–16473 (2022)
2. Chen, D.Z., Chang, A.X., Nießner, M.: Scanrefer: 3d object localization in rgb-d scans using natural language. In: European conference on computer vision. pp. 202–221. Springer (2020)
3. Chen, D.Z., Wu, Q., Nießner, M., Chang, A.X.: D 3 net: A unified speaker-listener architecture for 3d dense captioning and visual grounding. In: European Conference on Computer Vision. pp. 487–505. Springer (2022)
4. Chen, Z., Hu, R., Chen, X., Nießner, M., Chang, A.X.: Unit3d: A unified transformer for 3d dense captioning and visual grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18109–18119 (2023)
5. Huang, P.H., Lee, H.H., Chen, H.T., Liu, T.L.: Text-guided graph neural networks for referring 3d instance segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1610–1618 (2021)
6. Huang, T., Dong, B., Yang, Y., Huang, X., Lau, R.W., Ouyang, W., Zuo, W.: Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22157–22167 (2023)
7. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1912–1920 (2015)
8. Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., Savarese, S.: Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1179–1189 (2023)
9. Yuan, Z., Yan, X., Liao, Y., Zhang, R., Wang, S., Li, Z., Cui, S.: Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1791–1800 (2021)
10. Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H.: Pointclip: Point cloud understanding by clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8552–8562 (2022)
11. Zhang, Y., Gong, Z., Chang, A.X.: Multi3drefer: Grounding text description to multiple 3d objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15225–15236 (2023)
12. Zhao, L., Cai, D., Sheng, L., Xu, D.: 3dvg-transformer: Relation modeling for visual grounding on point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2928–2937 (2021)
13. Zhu, X., Zhang, R., He, B., Guo, Z., Zeng, Z., Qin, Z., Zhang, S., Gao, P.: Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2639–2650 (2023)