

---

# MLLMs Need 3D-Aware Representation Supervision for Scene Understanding (*Supplementary Material*)

---

Anonymous Author(s)

Affiliation

Address

email

## 1 Detailed Comparison

In this section, we provide a detailed comparison with other methods using all metrics across 5 benchmarks.

**Scanrefer.** Tab. 1 shows that our method 3DRS achieves the best overall performance on the ScanRefer validation set, especially in the challenging “Multiple” scenario where precise target discrimination is required. These results demonstrate that 3DRS effectively leverages multi-view images for robust spatial understanding and accurate object localization.

Table 1: Performance comparison on the validation set of ScanRefer [3]. “Unique” and “Multiple” depends on whether there are other objects of the same class as the target object.

Method	Unique		Multiple		Overall	
	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
ScanRefer [3]	76.3	53.5	32.7	21.1	41.2	27.4
MVT [13]	77.7	66.4	31.9	25.3	40.8	33.3
3DVG-Transformer [19]	81.9	60.6	39.3	28.4	47.6	34.7
ViL3DRel [5]	81.6	68.6	40.3	30.7	47.9	37.7
3DJCG [2]	83.5	64.3	41.4	30.8	49.6	37.3
D3Net [4]	–	72.0	–	30.1	–	37.9
M3DRef-CLIP [17]	85.3	77.2	43.8	36.8	51.9	44.7
3D-VisTA [23]	81.6	75.1	43.7	39.1	50.6	45.8
3D-LLM (Flamingo) [10]	–	–	–	–	21.2	–
3D-LLM (BLIP2-flant5) [10]	–	–	–	–	30.3	–
Grounded 3D-LLM [7]	–	–	–	–	47.9	44.1
PQ3D [24]	86.7	78.3	51.5	46.2	57.0	51.2
ChatScene [11]	89.6	82.5	47.8	42.9	55.5	50.2
LLaVA-3D [22]	–	–	–	–	54.1	42.2
Video 3D-LLM [20]	88.0	78.3	50.9	45.3	58.1	51.7
<b>3DRS (Ours)</b>	87.4	77.9	57.0	50.8	62.9	56.1

**Multi3DRefer.** In Tab. 2, 3DRS achieves the best overall results on the Multi3DRefer validation set, with top F1 scores in both standard and challenging scenarios. Our method consistently outperforms previous approaches, especially in the difficult zero-target and distractor settings, demonstrating superior robustness and spatial understanding.

**ScanQA.** In Tab. 3, 3DRS achieves the best performance on the ScanQA validation set across almost all metrics, including EM, BLEU scores, METEOR, and CIDEr, demonstrating its strong effectiveness for 3D question answering.

Table 2: Performance comparison on the validation set of Multi3DRefer [17]. ZT: zero-target, ST: single-target, MT: multi-target, D: distractor.

Method	ZT w/o D	ZT w/ D	ST w/o D		ST w/ D		MT		ALL	
	F1	F1	F1@0.25	F1@0.5	F1@0.25	F1@0.5	F1@0.25	F1@0.5	F1@0.25	F1@0.5
M3DRef-CLIP [17]	81.8	39.4	53.5	47.8	34.6	30.6	43.6	37.9	42.8	38.4
D3Net [4]	81.6	32.5	—	38.6	—	23.3	—	35.0	—	32.2
3DJCG [2]	94.1	66.9	—	26.0	—	16.7	—	26.2	—	26.6
Grounded 3D-LLM [7]	—	—	—	—	—	—	—	—	45.2	40.6
PQ3D [24]	85.4	57.7	—	68.5	—	43.6	—	40.9	—	50.1
ChatScene [11]	90.3	62.6	82.9	75.9	49.1	44.5	45.7	41.1	57.1	52.4
Video 3D-LLM [20]	94.7	78.5	82.6	73.4	52.1	47.2	40.8	35.7	58.0	52.7
<b>3DRS (Ours)</b>	95.6	79.4	79.6	71.4	57.0	51.3	43.0	37.8	60.4	54.9

Table 3: Performance comparison on the validation set of ScanQA [1]. EM indicates exact match accuracy, and B-1, B-2, B-3, B-4 denote BLEU-1, -2, -3, -4, respectively.

Method	EM	B-1	B-2	B-3	B-4	ROUGE-L	METEOR	CIDEr
ScanQA [1]	21.05	30.24	20.40	15.11	10.08	33.33	13.14	64.86
3D-VisTA [23]	22.40	—	—	—	10.40	35.70	13.90	69.60
Oryx-34B [14]	—	38.00	24.60	—	—	37.30	15.00	72.30
LLaVA-Video-7B [18]	—	39.71	26.57	9.33	3.09	44.62	17.72	88.70
3D-LLM (Flamingo) [10]	20.40	30.30	17.80	12.00	7.20	32.30	12.20	59.20
3D-LLM (BLIP2-flant5) [10]	20.50	39.30	25.20	18.40	12.00	35.70	14.50	69.40
Chat-3D [16]	—	29.10	—	—	6.40	28.50	11.90	53.20
NaviLLM [21]	23.00	—	—	—	12.50	38.40	15.40	75.90
LL3DA [6]	—	—	—	—	13.53	37.31	15.88	76.79
Scene-LLM [9]	27.20	43.60	26.80	19.10	12.00	40.00	16.60	80.00
LEO [12]	—	—	—	—	11.50	39.30	16.20	80.00
Grounded 3D-LLM [7]	—	—	—	—	13.40	—	—	72.70
ChatScene [11]	21.62	43.20	29.06	20.57	14.31	41.56	18.00	87.70
LLaVA-3D [22]	27.00	—	—	—	14.50	50.10	20.70	91.70
Video 3D-LLM [20]	30.10	47.05	31.70	22.83	16.17	49.02	19.84	102.06
<b>3DRS (Ours)</b>	30.30	48.37	32.67	23.79	17.22	49.82	20.47	104.78

15 **SQA3D**. In Tab. 4, 3DRS achieves the highest scores on the SQA3D test set, outperforming all previ-  
16 ous approaches on almost every question type as well as in the overall average, which demonstrates  
17 its superior capability for 3D question answering across diverse scenarios.

18 **Scan2cap**. In Tab. 5, 3DRS achieves the best performance on the Scan2Cap validation set in terms  
19 of CIDEr (C), and remains highly competitive on other metrics such as BLEU-4, METEOR, and  
20 ROUGE-L, demonstrating strong overall effectiveness for 3D captioning.

Table 4: Performance comparison on the test set of SQA3D [15].

Method	Test set						Avg.
	What	Is	How	Can	Which	Others	
SQA3D [15]	31.60	63.80	46.00	69.50	43.90	45.30	46.60
3D-VisTA [2]	34.80	63.30	45.40	69.80	47.20	48.10	48.50
LLaVA-Video[18]	42.70	56.30	47.50	55.30	50.10	47.20	48.50
Scene-LLM [9]	40.90	69.10	45.00	70.80	47.20	52.30	54.20
LEO [12]	—	—	—	—	—	—	50.00
ChatScene [11]	45.40	67.00	52.00	69.50	49.90	55.00	54.60
LLaVA-3D [22]	—	—	—	—	—	—	55.60
Video 3D-LLM [20]	51.10	72.40	55.50	69.80	51.30	56.00	58.60
<b>3DRS (Ours)</b>	54.40	75.20	57.00	72.20	49.90	59.00	60.60

Table 5: Performance comparison on the validation set of Scan2Cap [8].

Method	@0.5			
	C	B-4	M	R
Scan2Cap [8]	39.08	23.32	21.97	44.48
3DJCG [2]	49.48	31.03	24.22	50.80
D3Net [4]	62.64	35.68	25.72	53.90
3D-VisTA [23]	66.90	34.00	27.10	54.30
LL3DA [6]	65.19	36.79	25.97	55.06
LEO [12]	68.40	36.90	27.70	57.80
ChatScene [11]	77.19	36.34	28.01	58.12
LLaVA-3D [22]	79.21	41.12	30.21	63.41
Video 3D-LLM [20]	83.77	42.43	28.87	62.34
<b>3DRS (Ours)</b>	86.11	41.63	28.97	62.29

## 2 Qualitative Results

22 Figs. 1 and 2 provide a visual summary of how our method performs on three challenging 3D  
23 scene understanding tasks. These tasks include identifying objects based on language, generating  
24 descriptions for specific regions, and answering spatial questions about the scene.

In the visual grounding examples at the top, the model is challenged to find the correct object in a complex 3D environment based on a textual description. The comparison highlights three bounding boxes for each case: blue for the ground truth, red for the baseline, and green for our result. Our predictions consistently align with the intended targets, showing our model’s ability to accurately interpret spatial and semantic cues from language.

The object captioning section in the middle presents how each model describes a highlighted object or area. For each instance, the ground truth, baseline output, and our generated caption are shown, along with their respective CIDEr scores. Our model’s captions are both more precise and more faithful to the scene’s content, as reflected in the higher evaluation scores.

At the bottom, the question answering task demonstrates the model’s reasoning abilities within a 3D environment. The figures show the posed question, the correct answer, the baseline’s response, and our model’s answer. Even for questions that require counting or locating objects, our approach tends to provide accurate answers, often supported by clear visual evidence in the scene.

Altogether, these qualitative results illustrate that our approach delivers more reliable scene understanding across a variety of tasks, outperforming the baseline in both accuracy and descriptive quality.

## References

- [1] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *CVPR*, 2022. License: Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.
- [2] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *CVPR*, 2022.
- [3] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, 2020. License: Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.
- [4] Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X. Chang. D<sup>3</sup>net: A unified speaker-listener architecture for 3d dense captioning and visual grounding. In *ECCV*, volume 13692, 2022.
- [5] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. In *NeurIPS*, 2022.
- [6] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. LL3DA: visual interactive instruction tuning for omni-3d understanding, reasoning, and planning. In *CVPR*, 2024.
- [7] Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Ruiyuan Lyu, Runsen Xu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens. *Arxiv e-prints*, 2024.
- [8] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *CVPR*, 2021. License: Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.
- [9] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *Arxiv e-prints*, 2024.
- [10] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. In *NeurIPS*, 2023.
- [11] Haifeng Huang, Zehan Wang, Rongjie Huang, Luping Liu, Xize Cheng, Yang Zhao, Tao Jin, and Zhou Zhao. Chat-3d v2: Bridging 3d scene and large language models with object identifiers. *Arxiv e-prints*, 2023.
- [12] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *ICML*, 2024.
- [13] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *CVPR*, 2022.

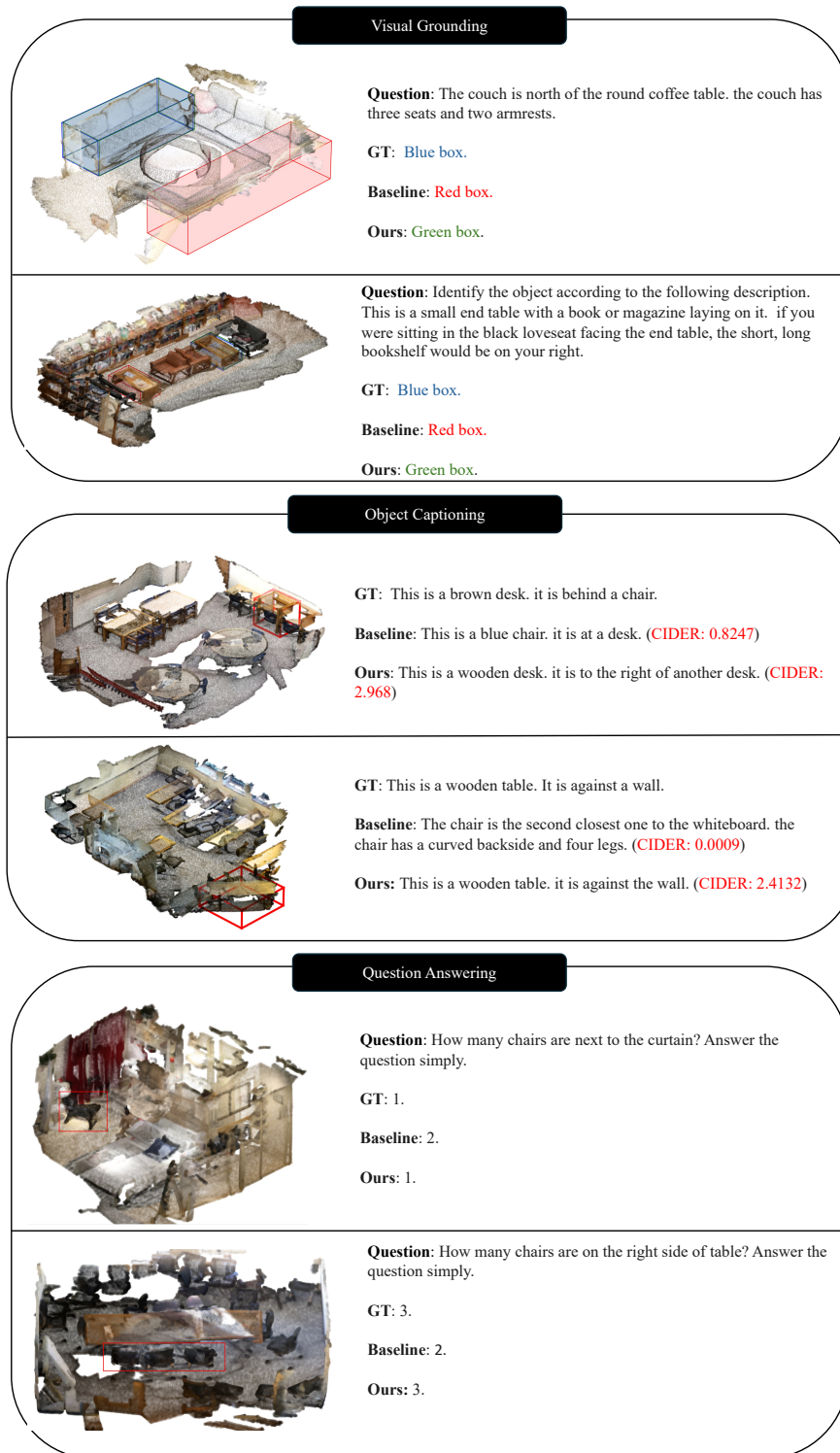


Figure 1: **Visualization of Results Across Different Tasks.** (a) Visual Grounding: The predicted bounding box closely aligns with the ground truth. (b) Object Captioning: Our method generates accurate captions for each referred object. (c) Question Answering: The model provides precise answers, where we use the red rectangles to indicate the visual cues utilized for each response. Best viewed when zoomed in.

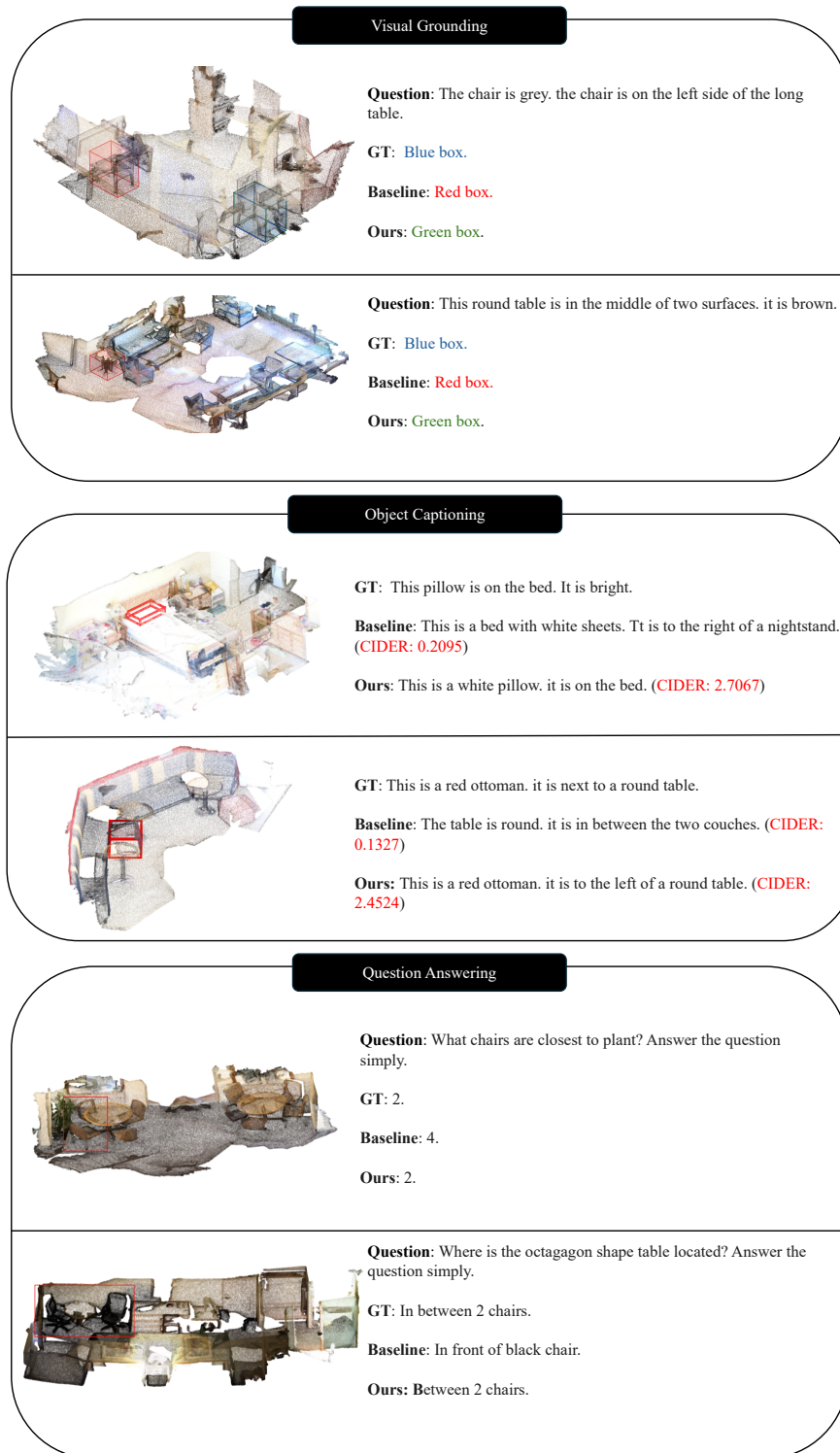


Figure 2: **Visualization of Results Across Different Tasks.** (a) Visual Grounding: The predicted bounding box closely aligns with the ground truth. (b) Object Captioning: Our method generates accurate captions for each referred object. (c) Question Answering: The model provides precise answers, where we use the red rectangles to indicate the visual cues utilized for each response. Best viewed when zoomed in.

- 73 [14] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx MLLM: on-demand  
74 spatial-temporal understanding at arbitrary resolution. *Arxiv e-prints*, 2024.
- 75 [15] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang.  
76 SQA3D: situated question answering in 3d scenes. In *ICLR*, 2023. License: CC-BY-4.0.
- 77 [16] Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3d: Data-efficiently tuning  
78 large language model for universal dialogue of 3d scenes. *Arxiv e-prints*, 2023.
- 79 [17] Yiming Zhang, ZeMing Gong, and Angel X. Chang. Multi3drefer: Grounding text description to multiple  
80 3d objects. In *ICCV*, 2023. License: MIT.
- 81 [18] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction  
82 tuning with synthetic data, 2024.
- 83 [19] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual  
84 grounding on point clouds. In *ICCV*, 2021.
- 85 [20] Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation  
86 for 3d scene understanding. *Arxiv e-prints*, 2024.
- 87 [21] Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model  
88 for embodied navigation. In *CVPR*, 2024.
- 89 [22] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet  
90 effective pathway to empowering llms with 3d-awareness. *Arxiv e-prints*, 2024.
- 91 [23] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained  
92 transformer for 3d vision and text alignment. In *ICCV*, 2023.
- 93 [24] Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan  
94 Huang, and Qing Li. Unifying 3d vision-language understanding via promptable queries. In *ECCV*, 2024.