Supplementary Materials to ”Descrip3D: Enhancing Large Language Model-based 3D Scene Understanding with Object-Level Text Descriptions”

Jintang Xue [1], Ganning Zhao [1], Jie-En Yao [1], Hong-En Chen [1], Yue Hu [1], Meida Chen [1],
Suya You [2], C.-C. Jay Kuo [1]
University of Southern California, Los Angeles, California, USA[1]
DEVCOM Army Research Laboratory, Los Angeles, California, USA[2]

Abstract

This supplementary file provides additional details on the Descrip3D framework, including qualitative examples of generated object-level relational descriptions, ablation experiments on object labels overlay in multi-view inputs, additional comparisons across multiple benchmarks, and the prompt template we use. These results further validate the effectiveness of our dual-level integration strategy and highlight the importance of explicitly modeling inter-object relationships through natural language.

1. Examples of Generated Descriptions

We present qualitative examples of our object-level textual descriptions in Fig. 1. Starting from detected object proposals and their corresponding multi-view images, we overlay the object names at the center of the projection areas in the image, as illustrated in the upper left part of the example. We then generate relational descriptions using a vision-language model. Key objects, typically those centrally positioned in the scene, are selected as query anchors. For each key object, we prompt the model to describe its spatial relationships with all other detected objects, resulting in detailed, contextually grounded descriptions. The prompt used is: *"Describe clearly and briefly the relationships between the <Key Object>in the scene and nearby objects (<Other Object 1>, <Other Object 2>, ..., <Other Object n>). Do not describe objects you cannot see."* For example, the objects in the image are a desk, two curtains, a window, a cabinet, and a table. There are two curtains, but only the one on the right is considered a key object because the other is positioned at the edge of the image. The chosen curtain is described as covering the window and situated near the table, the cabinet, and the desk. These relational descriptions offer interpretable summaries of local neighborhoods and equip downstream models with structured scene understanding for improved reasoning.

2. Ablation Study on Object Labels in Description Generation

To examine the impact of explicitly overlaying object category names during relational description generation, we conduct an ablation study comparing two variants: one where multi-view images include projected object labels (ours), and one without. As shown in Tab. 1, incorporating object labels consistently improves performance across all five benchmarks. The improvement is particularly notable in Scan2Cap and SQA3D, where more precise object references in the descriptions likely benefit caption generation and question answering. These results confirm that providing explicit category labels helps the vision-language model better ground each object and generate more informative relational descriptions.

3. Ablation Study on Prompt Design for Object-Level Descriptions

To assess how different prompt formulations influence the quality of generated object-level relational descriptions and downstream 3D scene understanding, we compare two designs: a default prompt (Prompt A) that emphasizes relational conciseness, and a spatially grounded prompt (Prompt B) that encourages explicit spatial terms and appearance details.

Fig. 1 presents qualitative examples generated using Prompt A (Default). This prompt directs the vision-language model (LLaVA-1.5) to describe relational context in a concise, human-like way, without placing heavy emphasis on precise spatial markers. As shown, descriptions tend to mention object co-occurrence and proximity in natural, readable sentences. For example, "The curtain is covering the window, and it is also close to a table, a cabinet, and a desk." While this phrasing lacks precise positional anchoring, it aligns with how humans intuitively describe contextual relevance.

In contrast, Fig. 2 illustrates the results of Prompt B

All Objects: ["desk", "curtain", "window", "cabinet", "curtain", "table"]
Key Objects: ["curtain", "cabinet", "table", "desk", "window"]

Object-level Descriptions:
Curtain: "There is a curtain in the room, which is located near a window. The curtain is covering the window, and it is also close to a table, a cabinet, and a desk. The table is placed in front of the cabinet, and the desk is situated next to the table."
Cabinet: "There is a cabinet with a window and a curtain nearby. The cabinet is located next to a desk, and a table is placed in front of it. A chair is also present in the scene, positioned in front of the desk. The arrangement of these objects suggests a functional and comfortable workspace, with the cabinet acting as a storage area and the desk serving as a work surface."
Table: "There is a table with a window, desk, cabinet, and curtain nearby. The table is located in the middle of the scene, and the window is positioned above it. The desk is situated to the left of the table, and the cabinet is located to the right of the table. The curtain is hanging in front of the window, providing privacy and possibly controlling the amount of light entering the room."
Desk: "There is a desk with a tablet on it, with a table, a cabinet, a window, and a curtain. The desk is located near the table, which is placed next to the cabinet. The window is situated above the desk, and the curtain is hanging in front of it. The desk is also close to the cabinet, creating a functional workspace with easy access to various items and resources."
Window: "There is a window located next to a table, desk, cabinet, and curtain. The window is positioned above the table, which is placed on the floor. The desk is situated in front of the window, and the cabinet is located near the desk. The curtain is hanging in front of the window, providing privacy and controlling the amount of light entering the room. The presence of these objects suggests that the room is a living space, possibly a bedroom or a home office, where the window serves as a source of natural light and a connection to the outdoors."

Figure 1. Qualitative examples of object-level relational descriptions generated using **Prompt A (Default)** with LLaVA-1.5
. The upper left part displays the image with object names, and the lower left shows the 3D bounding boxes. The right side contains relational descriptions emphasizing general proximity and object co-occurrence without enforcing strict spatial language.

Multi-view Image Input	**ScanRefer** Acc@0.5	**Multi3DRefer** F1@0.5	**Scan2Cap** C@0.5	**ScanQA** CIDEr	**SQA3D** EM
Without Object Labels	51.5	54.8	75.6	93.5	54.6
With Object Labels (Ours)	**51.8**	**55.1**	**77.2**	**93.7**	**55.7**

Table 1. Ablation study on the effect of overlaying object category labels in multi-view images during relational description generation. **Adding object labels leads to consistent performance improvements across all benchmarks**, demonstrating their importance in guiding the vision-language model toward accurate grounding.

(Spatially Focused). This prompt explicitly encourages the use of geometric relations ("on the left," "in front of," "behind") and appearance details ("white," "rectangular"), resulting in descriptions that are shorter but more spatially grounded. For instance, "The desk is located in a corner of the room… the window is above the desk… the cabinet is in front of the desk," offers clearer positional context but less nuanced interpretation of function or co-usage.

To better understand how different prompts influence downstream model performance, Tab. 2 provides the full text of each design, and Tab. 3 summarizes quantitative results across five 3D vision-language tasks. While Prompt B encourages explicit spatial expressions (e.g., "to the left of," "in front of") and produces shorter sentences, it tends to focus narrowly on positional details, omitting functional or contextual cues. This results in descriptions that are more rigid but less informative overall. In contrast, Prompt A (Ours) generates a richer relational context with broader object co-occurrence and usage clues. As shown, Prompt A consistently outperforms Prompt B, suggesting that general, semantically rich descriptions better support multimodal reasoning than strictly spatial ones.

4. Ablation Study on Description Generator Choice

We further analyze the effect of the model used to generate object-level descriptions. While our main experi-

All Objects: ["desk", "curtain", "window", "cabinet", "curtain", "table"]
Key Objects: ["curtain", "cabinet", "table", "desk", "window"]

Object-level Descriptions:

Curtain: "The curtain is white and hanging on the wall, covering a window. It is positioned next to a cabinet and a desk. The desk has a table in front of it, and a chair is located in front of the desk. There is also a book on the desk, and a cup is placed on the table."

Cabinet: "The cabinet is a small, white, rectangular piece of furniture. It is positioned next to a window, under a curtain, and in front of a desk. The cabinet is also located on the left side of the room."

Table: "The table is a small, white, rectangular piece of furniture. It is positioned in front of a window, with a cabinet and a desk nearby. The table is located on the left side of the image, with the cabinet and desk on the right side. The window is situated behind the table, providing natural light to the room."

Desk: "The desk is a small, white, rectangular table with a black chair positioned in front of it. The desk is located in a corner of the room, with a window and a cabinet nearby. The cabinet is on the right side of the desk, and the window is above the desk. The desk is also situated under a curtain, which is hanging above it. The desk appears to be a functional workspace, with the necessary furniture and accessories for daily use."

Window: "The window is located on the left side of the room, positioned behind a curtain. It is also situated next to a cabinet and a desk. The window is rectangular in shape and has a white frame."

Figure 2. Qualitative examples of object-level relational descriptions generated using **Prompt B (Spatially Focused)** with LLaVA-1.5. Compared to Prompt A, these descriptions include more explicit spatial terms (e.g., "on the left," "behind") and visual attributes, resulting in shorter but more positionally grounded sentences.

Prompt Version	Prompt Text
Prompt A (Default)	`Describe clearly and briefly the relationships between the <key_object> in the scene and nearby objects (<other_obj1>, <other_obj2>, ...). Do not describe objects you cannot see. Do not describe green labels.`
Prompt B (Spatially Focused)	`Based on the image, describe both the appearance and spatial relationships of the <key_object> in relation to nearby visible objects (<other_obj1>, <other_obj2>, ...). Include visual details like color, shape, size, or texture of the <key_object>, and explain precisely how it is positioned relative to nearby visible objects (<other_obj1>, <other_obj2>, ...) using terms such as 'on the left', 'next to', 'under', 'in front of', 'behind', or 'on top of'. Only refer to what is clearly visible. Do not mention green text labels or objects not shown in the image."`

Table 2. Comparison of prompt designs used for generating object-level relational descriptions with LLaVA-1.5. Prompt A is our default, concise formulation emphasizing relational grounding. Prompt B explicitly encourages spatial terms (e.g., "left," "in front of") and detailed appearance cues.

ments adopt Vicuna-7B as the trainable backbone (to ensure fair comparison with prior methods such as Chat-Scene and 3DGraphLLM), the relational text can, in principle, be generated by any frozen captioner. Tab. 4 compares two options: generating descriptions with Vicuna-7B versus with LLaVA-1.5. Results show that replacing Vicuna with LLaVA-1.5 as the description generator improves downstream performance, especially on language-intensive tasks

Prompt Design	ScanRefer Acc@0.5	Multi3DRefer F1@0.5	Scan2Cap C@0.5	ScanQA CIDEr	SQA3D EM
Prompt B	51.4	**55.1**	74.1	92.3	55.2
Prompt A (Ours)	**51.8**	**55.1**	**77.2**	**93.7**	**55.7**

Table 3. Downstream performance using different prompts for generating object-level descriptions. Prompt B emphasizes spatial precision, while Prompt A (ours) encourages concise, general relational reasoning. Despite lacking explicit directional terms, Prompt A outperforms or matches Prompt B, suggesting that overly specific spatial descriptions may omit broader contextual signals useful for multimodal understanding.

(e.g., +2.0 CIDEr on Scan2Cap, +2.1 CIDEr on ScanQA). This suggests that our framework is flexible with respect to the choice of description generator, and benefits from relation-dense captions produced by stronger multimodal models.

5. Additional Qualitative Results and Failure Cases

Additional Qualitative Examples To further demonstrate the strengths of Descrip3D, we present additional qualitative comparisons of both question answering and object grounding tasks in Fig. 3. In the QA task Fig. 3a, Descrip3D produces accurate answers in cases where Chat-Scene fails due to limited spatial awareness or insufficient contextual cues. For example, in the first question, while Chat-Scene incorrectly places the "single seat sofa" behind the brown chair, Descrip3D correctly identifies it as "in the corner of the room," grounded by relational language. Similarly, Descrip3D succeeds in localizing queried objects such as the laptop and chairs based on complex object-to-object references, demonstrating its enhanced relational understanding. In the grounding task Fig. 3b, Descrip3D resolves ambiguous references more reliably. For example, given a query like "the black couch next to a tall shelf and a fan," Descrip3D identifies the correct object using spatial and contextual signals provided by the object descriptions. These results emphasize Descrip3D's ability to perform robust reasoning in cluttered indoor environments where visual and geometric cues alone may be insufficient.

Failure Case Analysis Despite its improved performance, Descrip3D is not immune to errors. Fig. 4 illustrates several representative failure cases in both QA (Fig. 4a) and grounding (Fig. 4b). In question answering, a common failure mode arises in counting tasks where performance is limited by upstream 3D detection accuracy. For instance, when the detector undercounts chairs around a table, Descrip3D cannot recover the correct answer solely through textual reasoning. Additionally, discrepancies between query phrasing (e.g., "square table") and detector-generated object names (e.g., "coffee table") introduce challenges in aligning language inputs with the available descriptions. In grounding, failures often occur when key attributes mentioned in the query (e.g., color or material) are missing or omitted in the generated object descriptions. As shown in Fig. 4b, the system fails to ground "the black cotton pillow" because the corresponding object description lacks explicit mention of its color, resulting in ambiguity during matching. These cases highlight limitations in both object detection accuracy and object description completeness, pointing to future directions for improving the robustness and coverage of relational grounding in 3D scene understanding.

6. Additional Quantitative Results

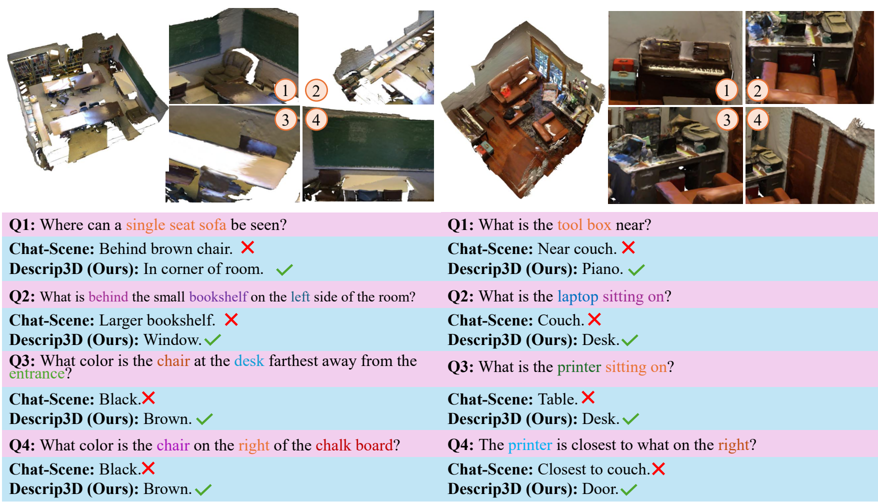
We evaluate our method using the standard metrics established in the original papers for each 3D scene-language dataset. To thoroughly assess the effectiveness of our approach, we perform extensive comparisons against a diverse set of baselines across multiple benchmarks. To complement the main results, we report additional evaluation metrics on the same datasets (ScanRefer, Multi3DRefer, and ScanQA) used in the main paper. The results, summarized in Tab. 6 (ScanRefer), Tab. 7 (Multi3DRefer), and Tab. 8 (ScanQA), show our method consistently outperforms prior approaches across grounding and question answering tasks. On ScanRefer, Descrip3D achieves the highest overall accuracy. On Multi3DRefer, it leads in almost all grounding settings, with the best overall F1 scores. On ScanQA, it outperforms baselines in nearly all language metrics, including ROUGE-L, METEOR, and CIDEr. These results confirm the effectiveness of incorporating object-level textual descriptions through dual-level integration for 3D vision-language tasks.

7. Prompt Template

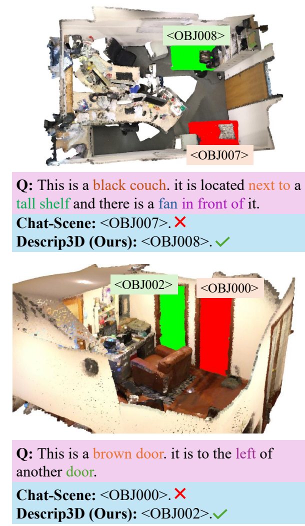
We adopt the same dialogue-style prompt format as Chat-Scene [10], consisting of a system message, a user instruction, and the corresponding assistant response. The system message sets the interaction context and introduces the object-level representation of the scene. Specifically, the scene is serialized as a flat sequence of object identifiers and

VLM	ScanRefer Acc@0.5	Multi3DRefer F1@0.5	Scan2Cap C@0.5	ScanQA CIDEr	SQA3D EM
Vicuna-7B	**51.8**	55.0	75.2	91.6	55.6
LLaVa-1.5 (Ours)	**51.8**	**55.1**	**77.2**	**93.7**	**55.7**

Table 4. Impact of the choice of description generator. We compare object-level relational text produced by Vicuna-7B and by LLaVA-1.5 (used as frozen captioners). Using LLaVA-1.5 yields stronger downstream results.



(a) Additional qualitative comparison of question answering. Descrip3D correctly answers more challenging questions by leveraging precise spatial relations and context-aware relational cues from its textual descriptions.



(b) Additional qualitative comparison of grounding. Descrip3D provides more accurate object localization by resolving ambiguous object references using context-enhanced descriptions, such as ”left of the fan” or “next to another door.”

Figure 3. Additional Qualitative comparison of 3D scene understanding tasks. **Descrip3D outperforms Chat-Scene, especially in cases involving complex spatial grounding or multi-object reasoning**, due to its use of a dual-level integrated relational textual descriptions that enhance contextual understanding.

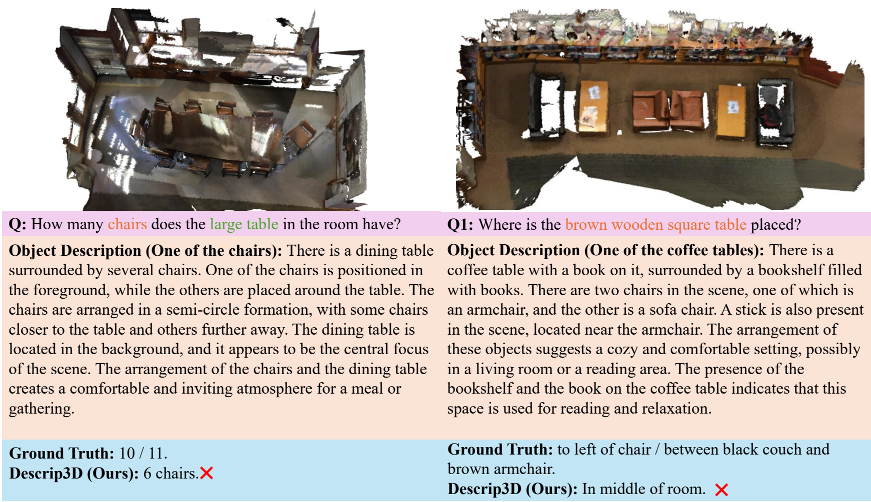
features: [<OBJ001>$\mathbf{F}_1$ <OBJ002>$\mathbf{F}_2$ ... <OBJn>$\mathbf{F}_n$], where $\mathbf{F}_i$ represents the feature embedding of the ith object. Each object identifier uniquely refers to a detected object in the scene. Users interact with the system by referencing these identifiers directly, and the assistant generates responses based on the identifiers. Tab. 5 provides an example of this prompt format.

8. IoU Variance and Distribution

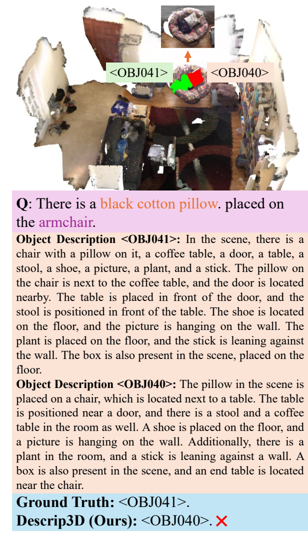
To provide a more fine-grained view of grounding robustness beyond thresholded metrics (Acc@0.25, Acc@0.50), we report IoU distributions and variance. Fig. 5 shows per-sample IoU histograms for ScanRefer and Multi3DRefer, both following the typical bimodal pattern with concentrations near IoU = 0 (failure) and IoU = 1 (successful grounding). Compared to Chat-Scene, Descrip3D consistently reduces the density of near-zero IoUs and increases the high-

System: A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user’s questions. The conversation centers around an indoor scene: [<OBJ001>$\mathbf{F}_1$ <OBJ002>$\mathbf{F}_2$...<OBJn>$\mathbf{F}_n$].
User: What is the ID of the object that matches the description ”this is a brown chair. It is to the right of another chair near the end of the table.”? [Generated description (may be noisy)] <OBJ002>: there is a <OBJ002>positioned next to a <OBJ003>...
Assistant: <OBJ002>.

Table 5. Prompt template used during training and evaluation.

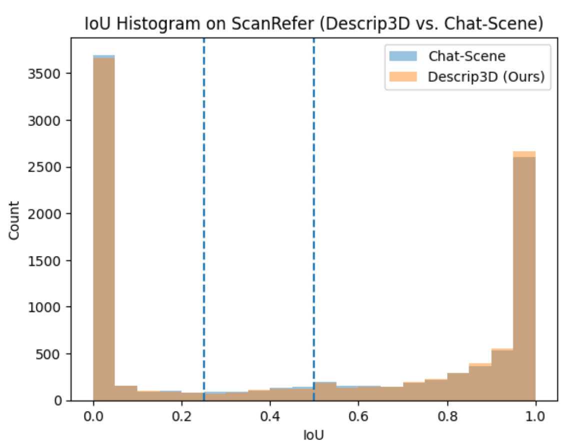


(a) Additional qualitative comparison of question answering. Descrip3D fails in cases where the 3D detector misses objects or the query uses object names not aligned with detector output, limiting the effectiveness of textual reasoning.

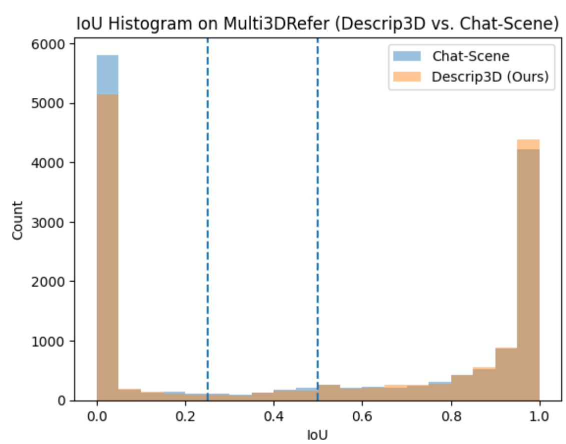


(b) Additional qualitative comparison of grounding. Failures in grounding are often caused by incomplete object descriptions that omit key attributes (e.g., "black cotton pillow"), which are essential for accurate reference resolution.

Figure 4. Failure cases of 3D scene understanding tasks. While Descrip3D improves contextual reasoning, failure can still occur due to missing or ambiguous descriptions (e.g., color not mentioned) or mismatches between detection names and query expressions, especially in counting or spatial referencing.



(a)



(b)

Figure 5. IoU distribution comparison between Chat-Scene and Descrip3D on (a) ScanRefer and (b) Multi3DRefer. The histograms show the per-sample IoU distributions, with dashed vertical lines indicating the 0.25 and 0.5 thresholds commonly used for visual grounding. Compared to Chat-Scene, Descrip3D consistently reduces the concentration of low-IoU cases and increases the density of high-IoU predictions, demonstrating more robust performance across the full distribution.

IoU mass. The measured variance is stable across datasets: ScanRefer variance ≈ 0.19 (std ≈ 0.44) and Multi3DRefer variance ≈ 0.20 (std ≈ 0.44). These results demonstrate that improvements are not driven by a handful of outlier cases but reflect consistent robustness across samples.

References

[1] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–

Method	Venue	Unique		Multiple		Overall	
		Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
ScanRefer [3]	ECCV20	76.33	53.51	32.73	21.11	41.19	27.40
TGNN [12]	AAAI21	68.61	56.80	29.84	23.18	37.37	29.70
SAT [21]	ICCV21	73.21	50.83	37.64	25.16	44.54	30.14
InstanceRefer [22]	ICCV21	75.72	64.66	29.41	22.99	38.40	31.08
3DVG-Transformer [24]	ICCV21	81.93	60.64	39.30	28.42	47.57	34.67
MVT [13]	CVPR22	77.67	66.45	31.92	25.26	40.80	33.26
3D-SPS [16]	CVPR22	84.12	66.72	40.32	29.82	48.82	36.98
ViL3DRel [6]	NeurIPS22	81.58	68.62	40.30	30.71	47.94	37.73
3DJCG [2]	CVPR22	83.47	64.34	41.39	30.82	49.56	37.33
D3Net [4]	ECCV22	–	72.04	–	30.05	–	37.87
BUTD-DETR [14]	ECCV22	84.2	66.3	46.6	35.1	52.2	39.8
HAM [5]	ArXiv22	79.24	67.86	41.46	34.03	48.79	40.60
3DRP-Net [18]	EMNLP23	83.13	67.74	42.14	31.95	50.10	38.90
3D-VLP [15]	CVPR23	84.23	64.61	43.51	33.41	51.41	39.46
EDA [20]	CVPR23	85.76	68.57	**49.13**	37.64	54.59	42.26
M3DRef-CLIP [23]	ICCV23	85.3	77.2	43.8	36.8	51.9	44.7
3D-VisTA [25]	ICCV23	81.6	75.1	43.7	39.1	50.6	45.8
ConcreteNet [17]	ECCV24	86.40	82.05	42.41	38.39	50.61	46.53
DORa [19]	ArXiv24	–	–	–	–	52.80	44.80
Chat-Scene [10]	NeurIPS24	89.59	82.49	47.78	42.90	55.52	50.23
Descrip3D (Ours)	–	**90.79**	**83.23**	**49.62**	**44.72**	**57.24**	**51.84**

Table 6. Performance comparison on the validation set of ScanRefer [3].

Method	Venue	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
3DVG-Trans+ [24]	ICCV21	87.1	45.8	–	–	16.7	–	26.5	–	25.5	–
D3Net (Grounding) [4]	ECCV22	81.6	32.5	–	–	23.3	–	35.0	–	32.2	–
3DJCG (Grounding) [2]	CVPR22	**94.1**	66.9	–	–	16.7	–	26.2	–	26.6	–
M3DRef-CLIP [23]	ICCV23	81.8	39.4	53.5	47.8	34.6	30.6	43.6	37.9	42.8	38.4
Chat-Scene [10]	NeurIPS24	90.3	62.6	82.9	**75.9**	49.1	44.5	45.7	41.1	57.1	52.4
Descrip3D (Ours)	–	92.0	**70.4**	**83.1**	**75.9**	**51.4**	**47.4**	**49.2**	**45.2**	**59.4**	**55.1**

Table 7. Performance comparison on the validation set of Multi3DRefer [23].

19139, 2022. 8

[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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16464–16473, 2022. 7

[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*, pages 202–221. Springer, 2020. 7

[4] Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X. Chang. D3net: A speaker-listener architecture for semi-supervised dense captioning and visual grounding in rgb-d scans, 2021. 7

[5] Jiaming Chen, Weixin Luo, Xiaolin Wei, Lin Ma, and Wei Zhang. Ham: Hierarchical attention model with high performance for 3d visual grounding. *arXiv preprint arXiv:2210.12513*, 2(3), 2022. 7

[6] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. *Advances in neural information processing systems*, 35:20522–20535, 2022. 7

[7] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26428–26438, 2024. 8

[8] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024. 8

[9] 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. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023. 8

Method	Venue	EM@1	B-1	B-2	B-3	B-4	ROUGE-L	METEOR	CIDEr
ScanQA [1]	CVPR22	21.05	30.24	20.40	15.11	10.08	33.33	13.14	64.86
3D-VLP [15]	CVPR22	21.65	30.53	21.33	16.67	11.15	34.51	13.53	66.97
3D-LLM [9]	NeurIPS23	20.5	39.3	25.2	18.4	12.0	35.7	14.5	69.4
LL3DA [7]	CVPR24	–	–	–	–	13.53	37.31	15.88	76.79
LEO [11]	ICML24	–	–	–	–	11.5	39.3	16.2	80.0
Scene-LLM [8]	WACV25	**27.2**	43.6	26.8	19.1	12.0	40.0	16.6	80.0
Chat-Scene [10]	NeurIPS24	21.62	43.20	29.06	20.57	14.31	41.56	18.00	87.70
Descrip3D (Ours)	–	22.67	**44.36**	**30.51**	**22.08**	**15.70**	**43.01**	**19.06**	**93.71**

Table 8. Performance comparison on the validation set of ScanQA [1].

[10] Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. *Advances in Neural Information Processing Systems*, 37: 113991–114017, 2024. 4, 7, 8

[11] 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. *arXiv preprint arXiv:2311.12871*, 2023. 8

[12] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3d instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1610–1618, 2021. 7

[13] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15524–15533, 2022. 7

[14] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *European Conference on Computer Vision*, pages 417–433. Springer, 2022. 7

[15] Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. Context-aware alignment and mutual masking for 3d-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10984–10994, 2023. 7, 8

[16] Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si Liu. 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16454–16463, 2022. 7

[17] Ozan Unal, Christos Sakaridis, Suman Saha, Fisher Yu, and Luc Van Gool. Three ways to improve verbo-visual fusion for dense 3d visual grounding. *arXiv preprint arXiv:2309.04561*, 2:15, 2023. 7

[18] Zehan Wang, Haifeng Huang, Yang Zhao, Linjun Li, Xize Cheng, Yichen Zhu, Aoxiong Yin, and Zhou Zhao. 3drp-net: 3d relative position-aware network for 3d visual grounding. *arXiv preprint arXiv:2307.13363*, 2023. 7

[19] Tung-Yu Wu, Sheng-Yu Huang, and Yu-Chiang Frank Wang. Dora: 3d visual grounding with order-aware referring. *CoRR*, 2024. 7

[20] Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19231–19242, 2023. 7

[21] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. Sat: 2d semantics assisted training for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1856–1866, 2021. 7

[22] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Sheng Wang, Zhen Li, and Shuguang Cui. 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*, pages 1791–1800, 2021. 7

[23] Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15225–15236, 2023. 7

[24] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2928–2937, 2021. 7

[25] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2911–2921, 2023. 7