Reason3D: Searching and Reasoning 3D Segmentation via Large Language Model

Supplementary Material

We provide additional details on our dataset in Section 1. Section 2 elaborates on the implementation and presents further results. Sections 3 to 5 offer comprehensive discussions on potential failure cases, limitations, and the broader impact of our method.

1. Reason3D Dataset

1.1. Dataset Annotation

Each 3D scene in the Reason3D dataset consists of a textual query and a binary segmentation mask to identify the target objects. As mentioned in the paper, we utilize ScanNetV2 [8, 19] and Matterport3D [4] as our data sources. Considering single room space as one scene, we first extract the instance object annotation and the room type information from these two datasets as the tags of 3D scenes. After that, we utilize these tags as parts of the text prompt to incorporate with GPT-4. The illustration of the prompt construction process is shown in Table 1, and some samples utilized for prompting are shown in Table 2.

1.2. Dataset Statistics

The collected Reason3D dataset incorporates the ScanNetV2 [8] and Matterport3D [4] datasets. We adhere to their official training and validation splits for data annotation. Specifically, the Matterport3D dataset provides 934 training samples and 837 validation samples. Meanwhile, the ScanNetV2 dataset contributes 405 training samples and 308 validation samples.

1.3. 3D Hierarchical Searching Extension

The 3D hierarchical searching task extends the reasoning-based 3D segmentation task by incorporating a specified target room type where the queried object should be located. As detailed in the main paper, we utilize only a subset of the Matterport3D dataset, chosen for its diversity in room types. In this task, we include the template "In <ROOM_TYPE>" to specify the target room, where <ROOM_TYPE> represents the various room categories defined in the Matterport3D dataset. For experiments involving different numbers of rooms, we expand the target room's space by including its neighboring rooms, adjusting this space according to the specified number of rooms.

2. Experiments

2.1. Implementation Details

Our models are primarily executed on two NVIDIA RTX A6000 GPUs, using a batch size of 16 during training and 1 during inference. The AdamW optimizer is employed with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a weight decay of 0.05. Additionally, we implement a linear warm-up strategy for the learning rate during the initial 1,000 steps, gradually increasing it from 10^{-8} to 10^{-4}, followed by a cosine decay schedule. All experiments are conducted using the PyTorch framework.

2.2. Used Datasets

In addition to the reason3D dataset, our model utilizes the following datasets for training:

ScanNet [8, 19], a comprehensive 3D indoor dataset, covers diverse environments including apartments and various room types. The dataset is structured into 1201 training scenes, 312 validation scenes, and 100 testing scenes.

Matterpor3D [4] dataset is a large-scale, real-world dataset comprising 90 houses. Each house is divided into various regions. We do not fine-tune the scene encoder on the Matterport3D dataset.

ScanRefer [5], a dataset annotated using ScanNet for 3D express referring segmentation tasks, including 36,665 natural language descriptions related to 7,875 objects across 562 scenes for training, and 9,508 descriptions of 2,068 objects from 141 ScanNet scenes for evaluation.

Nr3D [1], another 3D referring segmentation dataset derived from ScanNet, comprises 32,919 language descriptions associated with 4,664 objects from 511 scenes for training purposes. We further employ this dataset to train for 3D express referring segmentation tasks.

ScanQA [2] is a dataset for the 3D question answering task based on ScanNet, consisting of 25,563 question-answer pairs on 562 scenes for training and 4,675 question-answer pairs on 71 scenes for validation.

2.3. Training

Following the approach in [20, 22], we first train a sparse 3D UNet [11] as our backbone and compute superpoint features using pre-computed superpoints [10, 15]. Once the backbone is fully trained, its weights are frozen. For each target task, we then pre-train the entire network using data from the other tasks, followed by task-specific fine-tuning.

```
messages = [ {"role":"system", "content": You are an AI visual assistant, and you are seeing a 3D scene. What
you see is provided with several words to represent objects with tag <objects>, describing the scene you are looking at, and
also the room type <type> to describe the type of the scene. Design a question <question> that can be answered confidently
with the <answer> from one of the provided objects in <objects>. Please do not ask any <question> that cannot be answered
confidently. Each question should have one clear answer that is most relevant, without ambiguity or multiple possible answers
in the list of description words. Please include complex questions relevant to the scene's content, such as inquiries into the
background knowledge of objects or discussions about events related to these objects. Avoid questions about uncertain or
unclear details. The question should be natural.}
]
for sample in few_shot_samples do
    messages.append({"role":"user", "content":sample['context']})
    messages.append({"role":"assistant", "content":sample['response']} )
messages.append({"role":"user", "content":'\n'.join(query)})
```

Table 1. The illustration of the prompt construction process for generating 3D reasoning dataset with ChatGPT / GPT-4.

'context': **<room>** game room **<objects>** 'armchair', 'ceiling', 'door', 'doorframe', 'fireplace', 'floor', 'pool table', 'post', 'rail', 'stair', 'stool', 'tv', 'wall', 'window'
'response': **<question>** In a game room, what object in the scene could be used for playing a competitive and strategic game involving balls and cues? **<answer>** pool table

'context': **<room>** living room **<objects>** 'armchair', 'bedroom', 'bookshelf', 'lamp', 'bureau', 'carpet', 'ceiling fan', 'chair', 'computer', 'computer desk', 'couch', 'table', 'drawer', 'dresser'
'response': **<question>** It's very hot outside. After coming back home, what appliance would you turn on to help cool down the temperature? **<answer>** ceiling fan

'context': **<room>** game room **<objects>** 'table', 'door', 'cabinet', 'desk', 'office chair', 'picture', 'lamp', 'bathtub', 'bag', 'trash can', 'mirror', 'radiator'
'response': **<question>** When staying at a hotel, what part of the room in the scene can provide additional lighting for reading or working while in bed? **<answer>** lamp

'context': **<room>** game room **<objects>** 'floor', 'door', 'cabinet', 'shelf', 'desk', 'office chair', 'window', 'monitor', 'book', 'box', 'keyboard', 'trash can', 'file cabinet', 'fan', 'telephone', 'cup', 'paper towel roll', 'windowsill', 'clock' , 'headphones'
'response': **<question>** If someone wanted to check the time after getting ready in the morning, what object in this scene would they most likely use? **<answer>** clock

Table 2. The few shot samples used for ChatGPT prompting.

2.4. Instruction Template

In this section, we present the instructions and outputs used for task-specific templates. Following previous works [7, 23], we utilize a "human:" identifier to initiate the instruction, followed by an "assistant:" identifier for the LLM-generated response. We use <scene> to represent the token corresponding to the point cloud scene. The tokens <SEG> and <LOC> are used for segmentation and location, respectively, as part of the prompting process for generating segmentation results, as described in the main paper. Below are examples for various tasks, where {description} refers to the target object's description, {question} represents a query based on the given scene, and {answer} is the corresponding response.

3D Reasoning Segmentation: "Human: <scene> can you segment the object in the scene with the following descriptions: description? Assistant: Sure, it's <SEG>."

3D Hierarchical Searching. "Human: <scene> can you segment the object based on the description: description? Please segment the target room first, then output the object mask. Assistant: Sure, the room is <LOC>, and the object

Method	Venue	B-4	METEOR	ROUHE-L	CIDER
VoteNet+MCAN	-	6.2	11.4	29.8	54.7
ScanRefer+MCAN	-	7.9	11.5	30	55.4
ScanQA [2]	CVPR 2022	10.1	13.1	33.3	64.9
3D-VLP [14]	CVPR 2023	11.2	13.5	34.5	67.0
3D-LLM [12]	NeurIPS 2023	12.0	14.5	35.7	69.4
Reason3D (Ours)	-	**12.1**	**15.1**	**37.4**	**73.5**

Table 3. **3D question answering results** on ScanQA validation dataset. The first two results are from [2]. B-4 denotes BLEU-4. Our model achieves better results than all baseline models.

Method	Venue	Acc@0.25	Acc@0.50
ScanRefer [5]	ECCV 2020	38.97	26.10
InstanceRefer [24]	ICCV 2021	40.23	32.93
3D-SPS [17]	CVPR 2022	47.65	36.43
ViL3DRel [6]	NeurIPS 2022	47.94	37.73
3D-LLM [12]	NeurIPS 2023	30.3	-
TGNN [13]	AAAI 2021	37.37	29.70
3D-STMN [22]	AAAI 2024	46.8	36.6
Reason3D (Ours)	-	**49.60**	**41.10**

Table 4. **3D visual grounding results** on ScanRefer validation dataset. Our approach does not use 3D box annotation for training.

is <SEG>."

3D Express Referring Segmentation. "Human: <scene> please segment the object from the given scene: description. Assistant: It's <SEG>."

3D Question Answering. "Human: <scene> please answer the question based on the given scene: question and output the related segmentation mask. Assistant: answer <SEG>."

2.5. 3D Question Answering Results

Evaluation Metrics. For the QA task, the evaluation metrics include BLEU-4 [18], ROUGE-L [16], METEOR [3], and CIDEr [21] to ensure robust answer matching. These metrics evaluate the precision, fluency, and semantic accuracy of the responses.

Results. In addition to excelling in 3D reasoning and referring tasks, our approach also performs well in 3D question answering tasks. We present our results on the ScanQA validation set in Table 3, where we observed a significant improvement in evaluation metrics over both baseline methods and the recent LLM-based method, 3D-LLM [12]. Our approach not only answers questions accurately but also visualizes the related segmentation masks to further demonstrate the effectiveness.

2.6. 3D Visual Grounding Results.

Evaluation Metrics. Similar to the 3D expressive referring task, the evaluation metrics used are Acc@0.25 and Acc@0.5, indicating the percentage of correctly predicted bounding boxes with an IoU greater than 0.25 or 0.5, respectively, compared to the ground truth.

Results. Although our primary focus is on 3D segmentation, our method also effectively predicts 3D bounding boxes as supplementary outputs, facilitating comparison with 3D visual grounding methods. To generate a 3D bounding box for a referred object, we first apply DBSCAN [9] to eliminate noisy points, and then calculate the minimum and maximum XYZ coordinates from the points within the segmentation mask to form the 3D box. As demonstrated in Table 4, our approach not only outperforms recent 3D visual grounding methods but also significantly surpasses LLM-based methods, such as 3D-LLM [12], which struggle to integrate textual and numerical data to accurately localize objects in 3D space.

2.7. More Visualization Results.

3D Reasoning Segmentation. We provide more qualitative examples for the 3D reasoning segmentation task and the predictions by our Reason3D in Figures 1 and 2.

3D Referring Segmentation. We show the visualization results of the 3D referring segmentation task compared with 3D-STMN [22] in Figure 3. We observe that our approach can have correct predictions when the scenes contain multiple similar objects or when the query sentence is long, which proves the effectiveness of our approach.

3. Failure Case.

In Figure 4, we present representative failure cases as follows: (a) If the question involves querying a small object in the scene, our model may fail to generate the correct prediction. (b) The presence of similar objects in the scene may lead to false positive predictions by our model. (c) Similar structures in the point cloud, such as mirrors or sensor-induced fragments, can mislead our model. (d) Complex world knowledge required by the question may hinder our model's ability to generate accurate mask predictions.

4. Limitations

While our model introduces a novel approach to 3D reasoning segmentation, it does have limitations that open promising avenues for future research. For example, our system currently struggles with large-scale scenes—such as identifying an object within a 30-room house in the Matterport dataset. Future work could refine the hierarchical decoder or integrate adaptive multi-scale processing tech-

Figure 1. **Visualization Results for 3D Reasoning Segmentation Tasks.** The purple regions highlight the predicted segmentation masks generated by our model. Best viewed with zoom in.

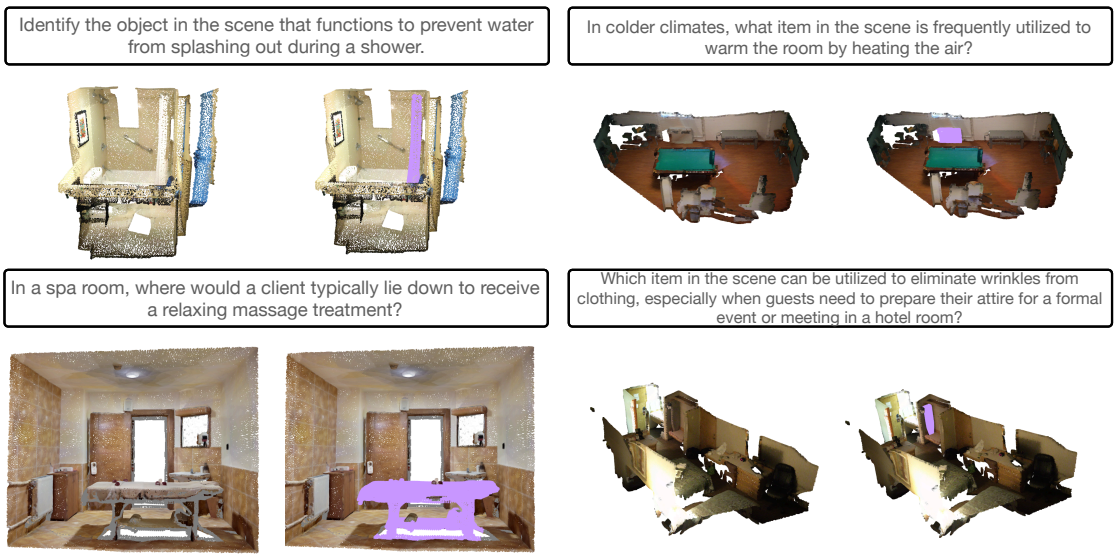


Figure 2. **More Visualization Results for 3D Reasoning Segmentation Tasks.** The purple regions highlight the predicted segmentation masks generated by our model. Best viewed with zoom in.

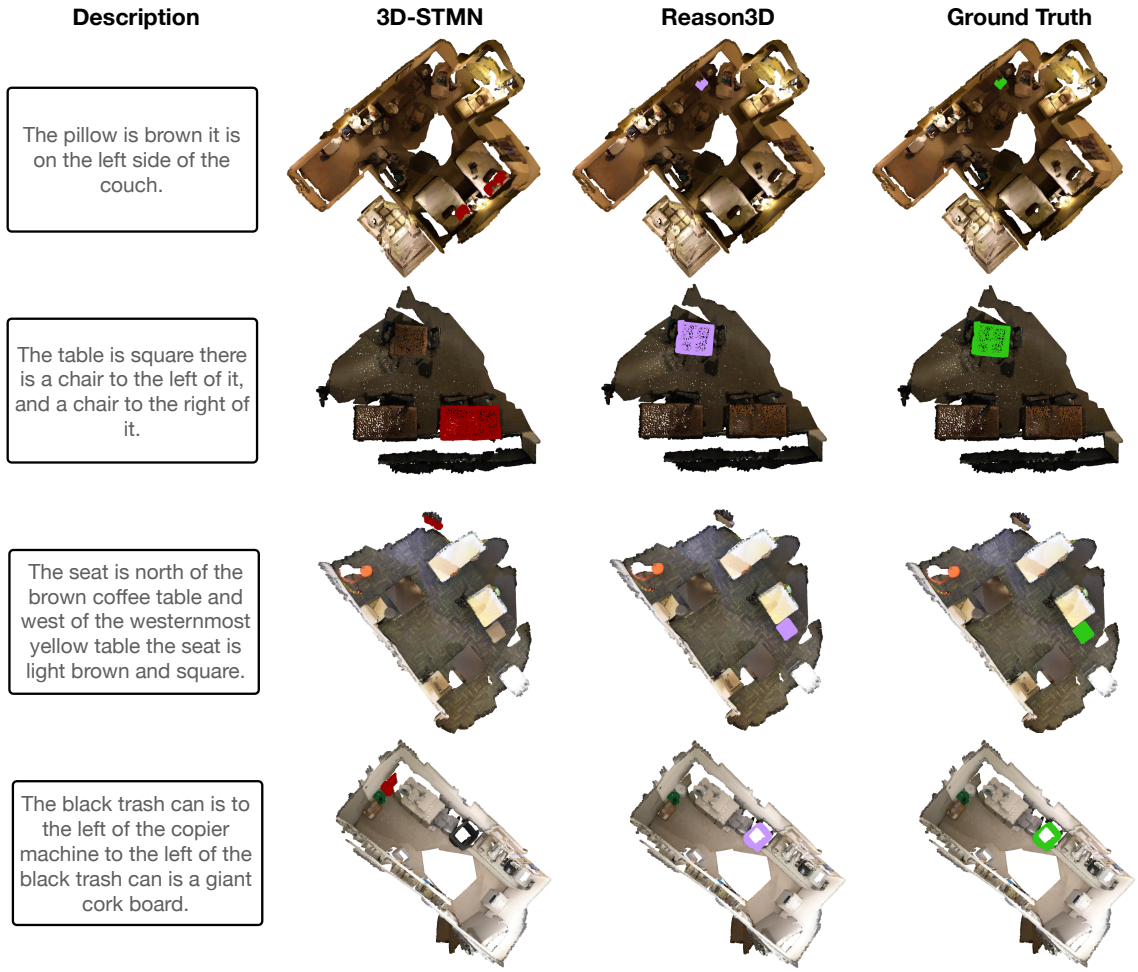
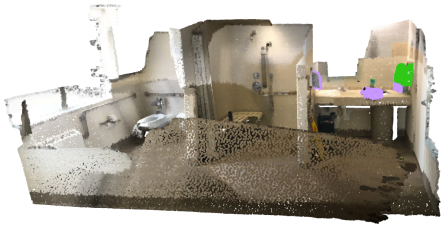


Figure 3. **Visualization Results for 3D Referring Segmentation Tasks.** The purple regions denotes the predicted segmentation masks from our Reason3D. The red and green means the predictions from 3D-STMN and ground truth, respectively. Best viewed with zoom in.

(a.) Small queried object

Safety is crucial in public spaces like spas. What object in the scene is designed to detect and suppress fires to ensure the well-being of everyone in the spa?



(b.) Similar objects in the scene.

In a public restroom, if someone wishes to dry their hands after washing, which item in the scene can they use for this task?

(c.) Similar structures of the point cloud.

What object in the scene would be most useful for checking one's appearance before leaving the bathroom?

(d.) Requiring complicate world knowledge.

In the bathroom, there may be decorative elements on the walls that hold symbolic significance. What object in the scene could be a religious symbol often associated with Christianity?

Figure 4. **Failure cases.** (a.) Small queried objects. (b.) Similar object in the scene. (c.) Similar structures of the point cloud. (d.) The question requires complicated world knowledge. The purple regions denote the predicted segmentation masks from our Reason3D, and the green means the ground truth. Best viewed with zoom in.

niques to better manage complex spatial extents. Additionally, the model does not yet handle scenarios with false premises—such as querying for an object that may not be present—suggesting that incorporating uncertainty estimation or robust error-detection mechanisms could help validate query assumptions before processing. Moreover, since our implementation is designed primarily for single-object queries, its performance on multi-object or multi-category tasks remains untested. This limitation points to the need for developing enriched query representations and joint optimization strategies that can simultaneously manage multiple objects. Addressing these challenges would significantly enhance the robustness and versatility of our LLM-based 3D reasoning framework.

5. Broader Impact

Reason3D is designed to segment objects in 3D space based on language inputs. Compared to traditional 3D segmentation algorithms, Reason3D models have a lower barrier to customization, enabling users to identify objects using natural language. However, this increased accessibility also raises the potential for misuse. Furthermore, the datasets and pre-trained models used in Reason3D may carry inherent biases, which could influence the model's performance.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *ECCV*, 2020. 1

[2] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *CVPR*, 2022. 1, 3

[3] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshop*, 2005. 3

[4] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *3DV*, 2017. 1

[5] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, 2020. 1, 3

[6] Shizhe Chen, Makarand Tapaswi, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. In *NeurIPS*, 2022. 3

[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 *CVPR*, 2024. 2

[8] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 1

[9] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996. 3

[10] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 2004. 1

[11] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, 2018. 1

[12] 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. 3

[13] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3d instance segmentation. In *AAAI*, 2021. 3

[14] Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. Context-aware alignment and mutual masking for 3d-language pre-training. In *CVPR*, 2023. 3

[15] Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *CVPR*, 2018. 1

[16] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 2004. 3

[17] 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. *arXiv preprint arXiv:2204.06272*, 2022. 3

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 3

[19] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *ECCV*, 2022. 1

[20] Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint transformer for 3d scene instance segmentation. *arXiv preprint arXiv:2211.15766*, 2022. 1

[21] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 3

[22] Changli Wu, Yiwei Ma, Qi Chen, Haowei Wang, Gen Luo, Jiayi Ji, and Xiaoshuai Sun. 3d-stmn: Dependency-driven superpoint-text matching network for end-to-end 3d referring expression segmentation. In *AAAI*, 2024. 1, 3

[23] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16911*, 2023. 2

[24] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Zhen Li, and Shuguang Cui. Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In *ICCV*, 2021. 3