2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

EMBOSR: Embodied Spatial Reasoning for Enhanced Situated Question Answering in 3D Scenes

Yu Hao[1*], Fan Yang[1*] , Nicholas Fang[2] and Yu-Shen Liu[3]

***Abstract*— 3D Embodied Spatial Reasoning, emphasizing an agent's interaction with its surroundings for spatial information inference, is adeptly facilitated by the process of Situated Question Answering in 3D Scenes (SQA3D). SQA3D requires an agent to comprehend its position and orientation within a 3D scene based on a textual situation and then utilize this understanding to answer questions about the surrounding environment in that context. Previous methods in this field face substantial challenges, including a dependency on constant retraining on limited datasets, which leads to poor performance in unseen scenarios, limited expandability, and inadequate generalization. To address these challenges, we present a new embodied spatial reasoning paradigm for enhanced SQA3D, fusing the capabilities of foundation models with the chain of thought methodology. This approach is designed to elevate adaptability and scalability in a wide array of 3D environments. A new aspect of our model is the integration of a chain of thought reasoning process, which significantly augments the model's capability for spatial reasoning and complex query handling in diverse 3D environments. In our structured experiments, we compare our approach against other methods with varying architectures, demonstrating its efficacy in multiple tasks including SQA3D and 3D captioning. We also assess the informativeness contained in the generated answers for complex queries. Ablation studies further delineate the individual contributions of our method to its overall performance. The results consistently affirm our proposed method's effectiveness and efficiency.**

I. INTRODUCTION

3D Embodied Spatial Reasoning, which focuses on an agent's ability to interact with and interpret spatial information within its surroundings, is effectively addressed through the process of Situated Question Answering in 3D Scenes (SQA3D) [1], [2]. It entails the need for an agent to accurately comprehend its location and orientation within a three-dimensional scene, as described by a textual situation. This understanding is critical for the agent to effectively respond to questions related to the environment from the current situation [3]. The task of SQA3D not only demonstrates the agent's ability to interpret and interact with spatial information but also highlights the importance of integrating situational awareness into the systems for realistic applications [4].

The field of SQA3D has seen considerable development through a diverse range of methodologies and approaches.

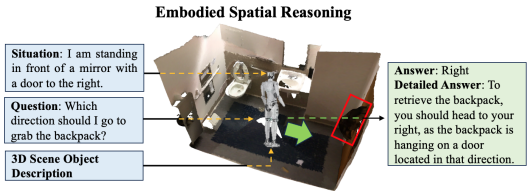


Fig. 1. Illustration of Embodied Spatial Reasoning.

Among the most significant contributions are models such as ScanQA [5] and ScanRefer [6], which effectively utilizes 3D scan data to enhance reasoning about spatial environments. These models represent a step forward in understanding and interpreting complex 3D scenes, allowing for more accurate responses to queries based on an agent's position and orientation within a given space. Additionally, other approaches like Clip-BERT [7] and MCAN [3] have been instrumental in advancing the field. These models utilize egocentric videos and bird's eye view (BEV) images to enhance the understanding of spatial layouts and object relationships in 3D environments, offering a deeper insight into agent interaction and perception of their surroundings.

Previous methods in SQA3D generally depend on 3D scene encoders and text encoders that require extensive offline training on specific, often limited-scale datasets. This reliance on predefined datasets poses a significant challenge: the models struggle to effectively respond to complex questions in unseen environments. The scale of the data used for training is usually not extensive enough to encompass the wide variety of real-world scenarios, leading to a gap in the models' adaptability and understanding. Consequently, these models often exhibit limited spatial reasoning capabilities when confronted with novel or complex situations and questions outside their training scope. This observation has motivated us to employ language foundation models [8] due to their powerful generalization abilities and enhanced reasoning capabilities in diverse and dynamic 3D environments without necessitating additional training, offering a significant advantage in addressing the limitations of previous approaches [9].

In this paper, as depicted in Figure 2, we introduce a novel approach in Situated Question Answering in 3D Scenes (SQA3D), combining the strengths of the language foundation model with the chain of thought methodology. Our model's first phase employs a vision module to create detailed descriptions of individual objects within a 3D scene. Subsequently, we leverage a scene graph approach to identify and understand the spatial relationships among these objects, thus forming a scene graph description that describes the direct spatial relationship in neighbouring objects. Then, the

[1]Yu Hao and Fan Yang are affiliated with NYU Tandon and also with the Embodied AI and Robotics (AIR) Lab at NYU Abu Dhabi. Yu Hao is the corresponding author: yuhao@nyu.edu

[2]Nicholas Fang is affiliated with Cranleigh Abu Dhabi in the UAE. It should be noted that the work he has conducted was during his research internship at NYU Abu Dhabi's AIR Lab.

[3]Yu-Shen Liu is affiliated with School of Software at Tsinghua University

*These authors contributed equally to this work.

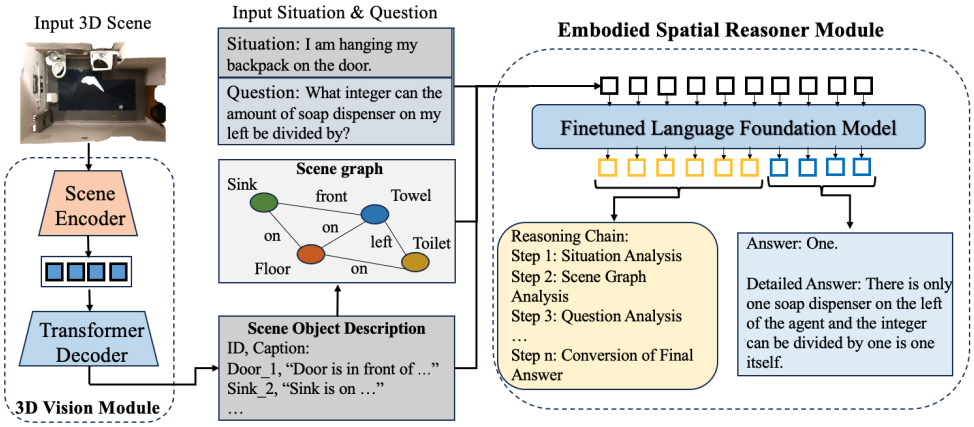


Fig. 2. Flowchart illustrating our proposed EMBOSR.

textual scene object description, scene graph description, along with situations and questions is then fed into the Embodied Spatial Reasoner (EMBOSR) module. The EMBOSR's integration of a chain of thought reasoning process with a language foundation model substantially boosts its ability to understand and analyze textual situations and questions in diverse 3D environments. This enhancement is essential for handling complex queries and scenarios, allowing the EMBOSR to effectively process textual descriptions of 3D scenes and generate answers that are both precise and rich in information. Our key contributions are summarized as follows: 1) We propose a new paradigm that employs a language foundation model for embodied spatial reasoning, as depicted in Figure 1, offering enhanced reasoning capabilities and better adaptability to the diverse and dynamic nature of 3D environments for SQA3D.

2) We introduce a new integration of chain of thought reasoning and scene graph analysis, which significantly improves the model's ability to comprehend and analyze the spatial relationships between objects, as represented in textual format. This advancement also extends to a better understanding and analysis of textual situations and questions, deepening the model's contextual awareness.

3) We conduct extensive experiments across multiple tasks and demonstrate the effectiveness of our model. Our approach not only shows proficiency in handling different 3D vision-language tasks but also produces answers that are richer in information and contextually relevant, highlighting the model's advanced reasoning and analytical capabilities.

II. Methods

A. Problem Setup

The 3D Situated Question Answering task is designed to generate an accurate answer vector $A \in \mathbb{R}^{d_a}$ from an input 3D point cloud scene P, a textual situation description S, and a textual question Q. The 3D point cloud P consists of a set of points in a three-dimensional space, each point represented by its coordinates and attributes as $P \in \mathbb{R}^{N\times(3+c)}$, where N is the number of points and c the number of additional attributes like color. The textual situation S and question Q are encoded into fixed-size vectors, $S \in \mathbb{R}^{d_s}$ and $Q \in \mathbb{R}^{d_q}$, providing the necessary context and inquiry relating to the scene.

We define the SQA problem with a function f that maps the inputs P,S,Q to the answer A [1]:

$$f : (\mathbb{R}^{N\times(3+c)}, \mathbb{R}^{d_s}, \mathbb{R}^{d_q}) \rightarrow \mathbb{R}^{d_a} \tag{1}$$

The goal is to develop a model that can approximate the function f to generate an answer **a** that is not only accurate but also contextually relevant to the given 3D scene P, situation S, and question Q.

In the following subsections, we will detail the architecture of our model. This includes a 3D vision module tasked with describing scene objects, followed by the use of a scene graph to analyze spatial relationships between neighboring objects. Subsequently, we integrate a chain of thought methodology and employ a fine-tuned language foundation model. This model is responsible for generating a reasoning chain and producing the final answer based on analyzed data.

B. 3D Vision Module

The 3D vision module is to describe 3D point cloud data as depicted on the left of Figure 2. Its primary objective is to generate text-based descriptions of the objects within the 3D scene. These descriptions include both spatial and semantic information about the objects, such as their location and categories.

Given an input point cloud $P \in \mathbb{R}^{N\times(3+c)}$, where each point is specified by its spatial coordinates and c additional features, the module employs a scene encoder to convert P into a set of scene tokens T, following the approach in [10]:

$$T = \text{SceneEncoder}(P) \tag{2}$$

These scene tokens T are then processed by a transformer decoder. The decoder's task is to produce the final scene object descriptions C, encompassing extensive details about objects' shapes, sizes, positions, and identities, thereby enabling a comprehensive understanding of the 3D scene [10]:

$$C = \text{TransformerDecoder}(T) \tag{3}$$

The resultant C represents the predicted geometric spatial and semantic information of each object in the scene, conveyed in a textual format. For example, the module might generate a scene description for an object like, "*This is a white door. It is to the left of the sink.*" This output offers

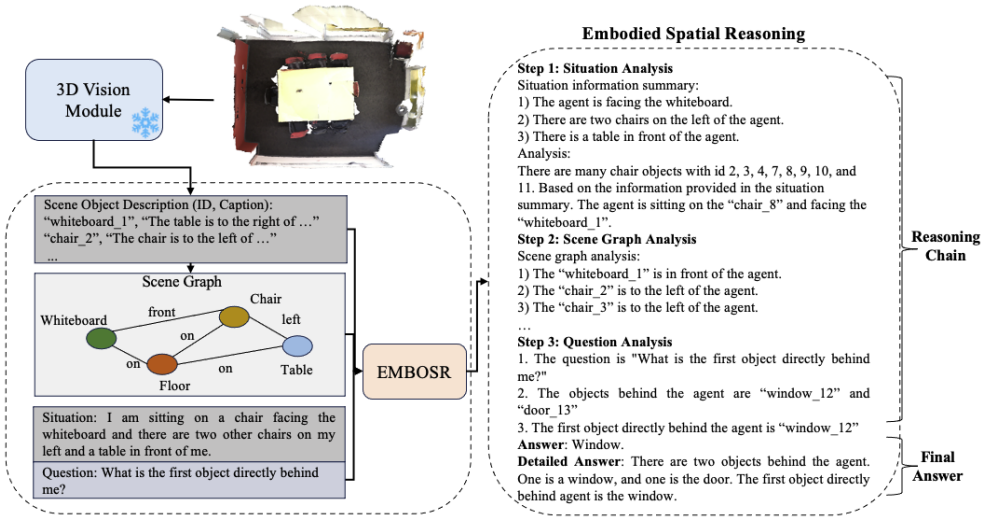


Fig. 3. Example of Embodied Spatial Reasoning.

a precise and detailed portrayal of the 3D environment, emphasizing the positions, dimensions, and identifying characteristics of the objects within.

C. Scene Graph Analysis

We've identified that beyond individual object descriptions C obtained from the 3D vision module, the relationships between objects are vital for a comprehensive understanding of 3D scenes. Therefore, as illustrated on the middle of Figure 2, integrating scene graph analysis [4] into our reasoning framework is essential, as it enhances the language foundation model's ability to discern spatial and semantic inter-object relationships. This understanding is fundamental for generating answers both coherent and contextually relevant in 3D Situated Question Answering scenarios.

Firstly, we define the scene graph as $G = (V, E)$, where V represents the nodes and E the edges. The nodes in V correspond to objects detected in the scene, each identified by the 3D vision module. The edges in E signify the semantic relationships between these objects, capturing spatial and contextual connections such as 'on', 'behind', and 'left'. We define $\langle v_i, e_{ij}, v_j \rangle$ delineates the spatial relationship e_{ij} between two neighbouring objects v_i and v_j. For every such relationship, we generate descriptive sentences following the template: "*The [object v_i] is [relation e_{ij}] to [neighbour v_j]*". In this context, a neighbor refers to other objects within a certain range of a given object, establishing a spatial relationship. This method provides a textual representation of scene graph, systematically mapping the network of object relationships. These descriptions are then incorporated into the reasoning process of our proposed EMBOSR, resulting in accurate and context-aware answers.

D. Spatial Reasoning for SQA3D

In this section, as illustrated on the right of Figure 2, we present our novel approach to spatial reasoning in the context of SQA3D. Our method initiates by utilizing embodied spatial reasoning information, which includes the scene object description, the scene graph description, the situation, and the posed question. This comprehensive analysis of these elements enables our system to derive well-informed and accurate answers. We observed that a singular, one-step reasoning process often yields answers that lack the desired level of detail and precision. To address this, we advocate for a multi-step reasoning approach that refines the results. This concept is rooted in the principles of the chain of thought reasoning [11], as applied in language foundation models. Our proposed scheme utilizes the chain of thought approach to sequentially construct a detailed reasoning process as shown in Figure 3. This method allows for a step-by-step analysis, leading to progressively refined answers. Below, we detail the specifics of our approach.

Given the scene object descriptions C, the scene graph description D, the situation S, and the question Q, we utilize a fine-tuned language foundation model that employs a probabilistic framework p_θ to sequentially generate the reasoning chain R with a length of N and the final answer A with a length of M [12]:

$$p(R,A|C,D,S,Q) = \prod_{i=1}^{N} p_\theta(R_i|C,D,S,Q,R_{<i}) \cdot \prod_{j=1}^{M} p_\theta(A_j|C,D,S,Q,R_{\leq N},A_{<j}) \tag{4}$$

Each step in the reasoning chain R_i is formulated considering the scene object descriptions C, graph scene description D, the situation S, the question Q, and all preceding reasoning steps $R_{<i}$. Upon constructing the full reasoning chain R, the model evaluates the probability of the final answer A, taking into account the entire reasoning process and the initial inputs. This multi-step reasoning approach provides an detailed intermediary step, elucidating the model's decision-making pathway and culminating in the generation of a well-defined answer. We demonstrate an example of our embodied spatial reasnoning process in Figure 3.

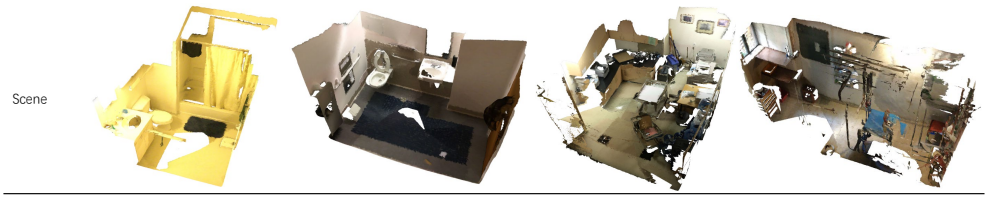
E. Fine-tuning of Language Foundation Model

General language foundation models, when confronted with spatial reasoning tasks, often do not perform as optimally as desired. This is particularly evident in scenarios involving complex spatial relationships and properties within

Models	Source	Format	What	Is	How	Can	Which	Others	Avg.
Blind test		SQ→A	26.75	63.34	43.44	69.53	37.89	43.41	43.65
ScanQA (w/o s)	3D scan	VQ→A	28.58	65.03	47.31	66.27	43.87	42.88	45.27
ScanQA	3D scan	VSQ→A	31.64	63.80	46.02	69.53	43.87	45.34	46.58
ScanQA + aux. task	3D scan	VSQ→AL	33.48	66.10	42.37	69.53	43.02	46.40	47.20
MCAN	BEV	VSQ→A	28.86	59.66	44.09	68.34	40.74	40.46	43.42
ClipBERT	Ego. video	VSQ→A	30.24	60.12	38.71	63.31	42.45	42.71	43.31
Unified QALarge	ScanRefer	VSQ→A	33.01	50.43	31.91	56.51	45.17	41.11	41.00
Unified QALarge	ReferIt3D	VSQ→A	27.58	47.99	34.05	59.47	40.91	39.77	38.71
GPT-3	ScanRefer	VSQ→A	39.67	45.99	40.47	45.56	36.08	38.42	41.00
GPT-3	ReferIt3D	VSQ→A	28.90	46.42	28.05	40.24	30.11	36.07	34.57
EMBOSR	ScanRefer	VSQ→A	**42.62**	**72.22**	**53.57**	**72.73**	**45.83**	**56.67**	**55.22**

TABLE I

QUANTITATIVE RESULTS ON SQA3D DATASET. THE 'SOURCE' COLUMN (SECOND COLUMN) SPECIFIES THE INPUT SOURCES OF THE 3D SCENE. THE 'FORMAT' COLUMN (THIRD COLUMN) INDICATES THE TYPES OF INPUTS AND OUTPUTS, WHERE 'V' REPRESENTS 3D VISUAL INPUT, 'S' FOR THE SITUATION, 'Q' FOR THE QUESTION, 'A' FOR THE ANSWER, AND 'L' DENOTES LOCATION LIKE POSITION AND ORIENTATION, WHICH WERE UTILIZED BY SQA3D [1] AS ADDITIONAL LOSS PARAMETERS DURING THEIR TRAINING PHASE.



Situation	I am using the towel and there is a bathroom cabinet on my right.	I am emptying litter from the trash can with the backpack in my six o'clock direction.	I am facing and using the photo copier to make copies of my biology report.	I am hanging the towel with a bucket by my left foot.
Question	Is the larger towel hanging in front of or behind me?	What color is the toilet to my right?	Is the number of cabinets behind me odd or even?	Which direction should I go if I want to take the rack?
SQA3D	The larger towel is hanging behind you.	The color of the toilet to your right is white.	Based on the given information, it is not possible to determine whether the number of cabinets behind you is odd or even.	You should go to the left if you want to take the rack.
Ours	The larger towel is hanging behind you.	The color of the toilet to my right is white.	The number of cabinets behind me is odd.	You should go to the right direction.

Fig. 4. Qualitative Results of Random Selected Examples on the SQA3D dataset. Green text indicate correct answers, while red text are wrong answers.

3D environments. Their lack of experience in spatial reasoning tasks means that their answers can be less specific and comprehensive. This observation has prompted us to explore fine-tuning techniques to better models for handling the intricacies of domain-specific spatial reasoning challenges.

For the specific task of fine-tuning the language foundation model for SQA3D, we employed a paired dataset $\{(C_i, D_i, S_i, Q_i, R_i, A_i)\}_{i=1}^{n}$, where each tuple contains the input data of 3D scene description (C), scene graph description (D), situations (S) and questions (Q), along with the corresponding reasoning chains (R) and final answers (A) and n is the total number of pair data. This dataset has a diverse set of scenarios to challenge the model's reasoning capabilities.

The adaptation of the language foundation model was achieved through the Proximal Policy Optimization (PPO) algorithm [13], which iteratively fine-tunes the model's policy to better predict the reasoning chains and final answers. The PPO objective function for our task is given by [13]:

$$\mathscr{L}_{\text{PPO}} = \min\left(\frac{\pi_{\text{New}}(R,A|C,D,S,Q)}{\pi_{\text{Old}}(R,A|C,D,S,Q)} A^{\text{adv}}, \text{clip}\left(\frac{\pi_{\text{New}}(R,A|C,D,S,Q)}{\pi_{\text{Old}}(R,A|C,D,S,Q)}, 1-\varepsilon, 1+\varepsilon\right) A^{\text{adv}}\right) \tag{5}$$

where π_{New} denotes the policy of the finetuned LLM, π_{Old} represents the policy before fine-tuning, and A^{adv} is the advantage function that calculates the relative benefit of the chosen action. The clip function bounds the policy update within the interval $[1-\varepsilon, 1+\varepsilon]$ to prevent destabilizing the learning process with large updates.

Through this process, the language foundation model is fine-tuned to enhance its capacity for generating reasoning chains and final answers that are pertinent to SQA3D tasks.

III. EXPERIMENT

A. Implementation Details

Our Embodied Spatial Reasoner (EMBOSR) comprises two integral modules. The vision module we used to generate 3D scene object description is Vote2Cap-DETR [10] which is pre-trained on the ScanRefer [6] dataset. The second module of EMBOSR comprises two distinct components: the fine-tuning of a language foundation model and the chain of thought reasoning. For the fine-tuning aspect, we selected the fine-tuned GPT-3.5 model [14]. To tailor the GPT-3.5 model to understand and analyze complex scene object descriptions, sence graph descriptions, situations and answers, we prepared a dataset to fine-tune the GPT-3.5 model.

As for the situated question answering in 3D scenes, we focused on the dataset of SQA3D [1], which consists of 650 3D indoor point cloud scenes sourced from ScanNet [15]. It follows a methodology that begins by establishing a situation, and subsequently formulates relevant problems under that situation. The test set of SQA3D consists of 67 different scenes with 3519 different question-answer pairs [1]. To evaluate SQA3D Task, we use the accuracy by comparing the generated answers with the ground truth answers.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGH-L
Scene1	33.33	24.62	19.40	15.39	31.74	46.55
Scene2	49.35	37.03	28.23	23.21	38.59	64.36
Scene3	50.00	34.21	25.57	19.19	33.02	57.95

TABLE II

QUANTITATIVE RESULTS ON 3D CAPTIONING OF 3 SCENES.

B. Comparative Experiments on SQA3D

Experimental Setting: We conducted comparisons with several methods on the SQA3D task. In line with the evaluation of SQA3D [1], our experiments compare three categories of baseline models. These categories are distinguished by their input sources, as detailed in the second column of Table I. For the model utilizing 3D scan as input source, we focus on the ScanQA [5]. There are three distinct settings of ScanQA we examine: the first, "*w/o s*," only takes the question as input and excludes the situation. The second adheres to the default setting of ScanQA which takes both situation and question as input. The third, "*aux. task*," incorporates both position and rotation as additional losses in the model during the training process. For the model utilizing images and video as input source, we analyze MCAN [3] and ClipBERT [7]. MCAN employs Bird's Eye View (BEV) images to generate responses, while ClipBERT uses video inputs to derive its answers. In the realm of models based on language foundation models, we explore Unified QA [16] and GPT-3 [17]. Both these models utilize scene object descriptions extracted from ScanRefer [6] and ReferIt3D [18]. Following the model setting based on language foundation models as in SQA3D, our EMBOSR also extracts 3D scene object descriptions from ScanRefer, enabling it to engage with and respond to queries in the context of the 3D environment.

We also set question of six types to verify models' effectiveness.

Results: Our evaluation demonstrates both qualitative and quantitative enhancements achieved by our model. As illustrated in Figure 4, we present four case studies highlighting the qualitative results of our proposed Embodied Spatial Reasoning. These examples showcase the accuracy and detail of our model's responses. For instance, in second scenario, our model advising, "*You should go to the **right** direction.*" while SQA3D generate wrong answer "*You should go to the **left** if you want to take the rack.*", which demonstrates our model's better spatial analysis capabilities in providing more precise and context-aware answers.

We also conduct statistical quantitative analysis to evaluate our model's performance across various types of situated questions in Table I. As we can see from the table, our model demonstrates superior performance across all six types of situated questions. This quantitative assessment underscores our model's overall effectiveness in understanding and responding to complex 3D situated queries.

C. Real-world Experiment

We observed that our approach to scene object descriptions, while detailed, was limited to simulated environments. Therefore, we are intrigued to explore whether our model could answer questions from real environment, providing a more complete understanding of the 3D environment.

Experimental Setting: Our experimental apparatus employs the Unitree B1, an industrial-grade quadrupedal robot. The evaluation comprises three scenarios, including laboratory, bedroom, and living room. These three scenarios are scanned into point clouds and inputted into our model. To validate the reliability of our system in real environments, we generate 5 pairs of scenarios and questions for each scenario. These questions involve various scenarios and topics, effectively assessing our system. This comprehensive evaluation aimed to elucidate the system's capability to generalize across different real-world scenarios and its adaptability to environmental changes. As for evaluation metrics, we follow 3D-LLM [3] and use the BLEU [19], ROUGE-L [20], METEOR [21] metrics to evaluate our results.

Results: To better illustrate the evaluation outcomes, we first present some of the questions in Figure 5, which depicts the positions of the quadruped robot and the corresponding Q&A logs. This demonstrates the ability of our system to operate across various scenarios and inquiries. Our quantitative results of real experiment are listed in Table II, where we utilize metrics to showcase the data from experiments in three different scenarios. These results highlight how our system can provide meaningful answers in diverse environments, further substantiating the efficacy of the system.

D. Ablation Study

In this section, we conduct an ablation study focusing on two key components of our proposed EMBOSR: the chain of thought reasoning and the fine-tuned language foundation model. Our investigation encompasses three different model configurations: GPT-3.5 without chain of thought reasoning (w/o CoT), GPT-3.5 with chain of thought reasoning (w/ CoT), and Finetuned GPT-3.5 with chain of thought reasoning (w/ CoT). In the first configuration, GPT-3.5 (w/o CoT), the model directly generates answers without employing the chain of thought reasoning. The second configuration, GPT-3.5 (w/ CoT), enhances this process by including a chain of thought reasoning chain that encompasses situation analysis, scene graph analysis, and question analysis. The third configuration takes a step further by employing a version of GPT-3.5 that has been fine-tuned with our custom dataset, as elaborated in section III-A. Our evaluation covers six different question types and uses average accuracy as our evaluation metric.

The experimental results are detailed in Table III. The results of GPT-3.5 (w/ CoT) outperform GPT3.5 (w/o CoT), underscoring the effectiveness of the chain of thought reasoning enhancing comprehension and answer accuracy. Furthermore, Finetuned GPT3.5 (w/ CoT) demonstrates superior performance over GPT3.5 (w/ CoT). This improvement highlights the effectiveness of the finetuning process in customizing the GPT-3.5 model for the specific needs of

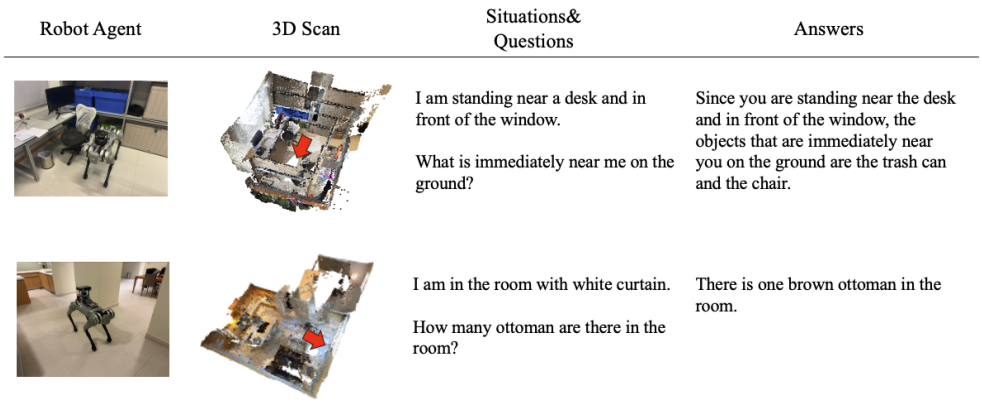


Fig. 5. Qualitative Results of random selected examples on the real dataset. Red arrow indicate the location and orientation of robot agent.

Model	What	Is	How	Can	Which	Others	Avg.
GPT-3.5 (w/o CoT)	27.87	43.24	21.43	45.45	33.33	27.59	32.34
GPT3.5 (w/ CoT)	29.51	62.16	26.92	50.00	40.91	37.93	40.20
Fine-tuned GPT3.5 (w/ CoT)	**42.62**	**72.22**	**53.57**	**72.73**	**45.83**	**56.67**	**55.22**

TABLE III

ABLATION STUDY ON CHAIN OF THOUGHT REASONING.

the SQA3D task, resulting in more precise and contextually appropriate answers.

We propose a new embodied spatial reasoning framework that integrates the robust capabilities of foundational models with the chain of thought reasoning. This combination aims to improve the model's adaptability and scalability across diverse 3D settings. Our model's distinct feature is its use of a chain of thought reasoning process, enhancing its ability to perform spatial reasoning and address complex questions within varied 3D environments. We conducted experiments on the SQA3D dataset to verify the effectiveness of our proposed method. Additionally, we designed a real-world experiment to assess the model's performance in a robotic agent setting. We also evaluate the depth of information in the answers to complex questions and conduct ablation studies to outline our method's specific contributions to its overall performance. Our findings consistently validate the efficiency and effectiveness of the proposed method.

REFERENCES

[1] X. Ma, S. Yong, Z. Zheng, Q. Li, Y. Liang, S.-C. Zhu, and S. Huang, "Sqa3d: Situated question answering in 3d scenes," *arXiv preprint arXiv:2210.07474*, 2022.

[2] J. Duan, S. Yu, H. L. Tan, H. Zhu, and C. Tan, "A survey of embodied ai: From simulators to research tasks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 2, pp. 230–244, 2022.

[3] Y. Hong, H. Zhen, P. Chen, S. Zheng, Y. Du, Z. Chen, and C. Gan, "3d-llm: Injecting the 3d world into large language models," *arXiv preprint arXiv:2307.12981*, 2023.

[4] Z. Zhu, X. Ma, Y. Chen, Z. Deng, S. Huang, and Q. Li, "3d-vista: Pre-trained transformer for 3d vision and text alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2911–2921.

[5] D. Azuma, T. Miyanishi, S. Kurita, and M. Kawanabe, "Scanqa: 3d question answering for spatial scene understanding," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 19 129–19 139.

[6] D. Z. Chen, A. X. Chang, and M. Nießner, "Scanrefer: 3d object localization in rgb-d scans using natural language," in *European conference on computer vision*. Springer, 2020, pp. 202–221.

[7] Y. Arase and J. Tsujii, "Monolingual phrase alignment on parse forests," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1–11.

[8] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[9] OpenAI, "Gpt-4 technical report," 2023.

[10] S. Chen, H. Zhu, M. Li, X. Chen, P. Guo, Y. Lei, G. Yu, T. Li, and T. Chen, "Vote2cap-detr++: Decoupling localization and describing for end-to-end 3d dense captioning," *arXiv preprint arXiv:2309.02999*, 2023.

[11] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[12] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola, "Multimodal chain-of-thought reasoning in language models," *arXiv preprint arXiv:2302.00923*, 2023.

[13] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[14] L. Floridi and M. Chiriatti, "Gpt-3: Its nature, scope, limits, and consequences," *Minds and Machines*, vol. 30, pp. 681–694, 2020.

[15] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5828–5839.

[16] D. Khashabi, S. Min, T. Khot, A. Sabharwal, O. Tafjord, P. Clark, and H. Hajishirzi, "Unifiedqa: Crossing format boundaries with a single qa system," *arXiv preprint arXiv:2005.00700*, 2020.

[17] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[18] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. Guibas, "Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 422–440.

[19] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[20] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[21] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.