FaST-3D: Integrating Fast and Slow Thinking for 3D Visual Question Answering

Yuhang Liu[1, 2], **Boyi Sun**[1, 2], **Yuzheng Hu**[3], **Jing Yang**[1], **Yutong Wang**[1, 2], **Fei-Yue Wang**[1*]

[1]Institute of Automation, Chinese Academy of Sciences
[2]Zhongke JingYu Sensing Technology Co., Ltd
[3]Department of Computer Science, University of Illinois at Urbana-Champaign
{liuyuhang2021, sunboyi2024, yangjing2020, yutong.wang, feiyue.wang}@ia.ac.cn, yh46@illinois.edu

Abstract

3D Visual Question Answering (VQA) is one of the most challenging tasks in the 3D Vision-Language (3D-VL) domain, as it requires not only precise interpretation of 3D environments but also effective reasoning over natural language questions. While existing approaches have conducted preliminary explorations in this area, they still suffer from significant issues such as limited robustness and an oversimplified treatment of VQA's inherently open-ended nature. To address these issues, we propose a novel FaST-3D agent based on dual process theory, which integrates two complementary reasoning systems: a fast-thinking system for rapid visual reasoning using representational memory and a slow-thinking system for detailed logical reasoning based on abstract memory. The fast system utilizes a Multimodal Large Language Model (MLLM) to quickly process visual input, enhanced by an adaptive image retriever and a confidence reflection module. In cases of low confidence, the slow system will be activated to invoke a well-designed toolset for step-by-step reasoning. Extensive experiments on ScanQA and OpenEQA datasets demonstrate that FaST-3D achieves SOTA performance across all metrics in zero-shot settings, particularly in the open-ended LLM-Match score. It effectively enhances model robustness across different scenarios and offers substantial practical value in embodied intelligence.

Introduction

In recent years, 3D scene understanding has garnered widespread attention across various fields, particularly in autonomous driving and robotics (Mao et al. 2023; Song et al. 2024). Early research primarily focused on fundamental tasks such as 3D object detection (Liu et al. 2023; Chen et al. 2023) and semantic segmentation (Li et al. 2023; Ando et al. 2023), achieving remarkable progress. With the rapid development of large language models (LLMs) (Naveed et al. 2023), more complex 3D vision-language (3D-VL) tasks have emerged, among which 3D visual question answering (3D VQA) stands out as one of the most challenging (Ma et al. 2024). This task requires agents to accurately comprehend 3D environments and reason over natural language queries to generate appropriate responses, underscoring its significant practical value in embodied robotics.

*Corresponding Author.

Recent studies have initiated preliminary explorations into 3D VQA and achieved promising results. The methods can be broadly divided into three main categories. The first and most common strategy involves training task-specific expert models for 3D VQA (Azuma et al. 2022; Parelli et al. 2023; Delitzas et al. 2023; Mo and Liu 2024). However, these methods exhibit notable shortcomings: 1) they lack robustness—even minor rephrasing of a question can cause drastic performance drops (Deng et al. 2024); and 2) they reduce the QA task to a classification problem with a fixed answer set, which may not be practical in real-world applications. The second approach adopts a pretrain-finetune strategy to enhance VQA accuracy (Hong et al. 2023; Li et al. 2024; Jin et al. 2023), where the encoder is pretrained on large-scale point cloud-text pairs before being finetuned on the 3D VQA task. However, this method still suffers from poor robustness due to the scarcity of high-quality 3D data pairs. The third approach leverages LLMs to enable fully zero-shot 3D VQA (Singh, Pavlakos, and Stamoulis 2024; Majumdar et al. 2024) by converting 3D scenes into textual descriptions as context. Unfortunately, this conversion process often sacrifices crucial spatial information, resulting in significant accuracy degradation.

To address these limitations, we introduce *dual process theory* (Frankish 2010) to offer a new perspective on 3D spatial cognition. Dual process theory posits that human cognition is driven by two complementary systems: *System 1* for fast thinking, and *System 2* for slow thinking (Kahneman 2011). *System 1* enables rapid information processing through representational memory, which is highly context-dependent and encompasses intuitive environmental perceptions, such as visual and auditory cues. In contrast, *System 2* relies on abstract memory to infer deeper spatial relationships through logical reasoning. Abstract memory stores structured rules and knowledge systems, granting it strong generalization capabilities. By integrating the intuitive responses of *System 1* with the analytical reasoning of *System 2*, the dual process mechanism establishes a robust foundation for comprehensive spatial understanding.

Inspired by these insights, we propose a novel FaST-3D agent for zero-shot 3D VQA. As illustrated in Figure 1, FaST-3D features two complementary thinking systems. The fast-thinking system processes first-person visual representations in a single forward pass of a MLLM to generate

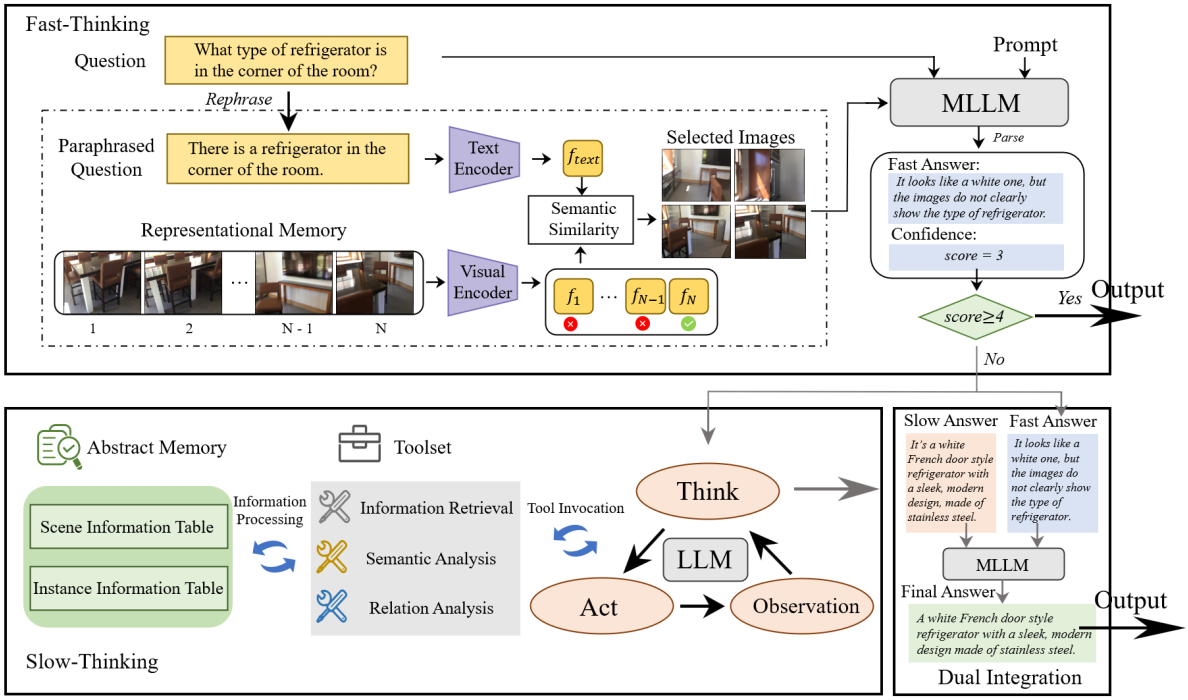

Figure 1: The overall framework of FaST-3D Agent.

immediate responses. To boost data retrieval efficiency, we incorporate an adaptive image retriever inspired by selective recall mechanisms. Additionally, a confidence reflection module enables the MLLM to self-assess its generated answers—if the confidence score is high, the agent outputs the answer directly; otherwise, it triggers the slow-thinking system for further reasoning. In this slow-thinking system, the 3D scene is preprocessed into two structured tables: a scene information table and an instance information table. We leverage an off-the-shelf LLM as the reasoning unit, decomposing tasks into multiple sub-steps and autonomously invoking well-designed tools to interact with these tables. Finally, a dual integration module fuses the outputs from both systems to produce the final answer. We conduct extensive experiments on the ScanQA (Azuma et al. 2022) and OpenEQA (Majumdar et al. 2024) datasets, and the results demonstrate that the proposed FaST-3D agent achieves outstanding performance across all evaluation metrics in zero-shot settings. In particular, given the open-ended nature of 3D VQA, we focus on the LLM-Match score (Majumdar et al. 2024), where FaST-3D achieves SOTA performance with significant improvements.

Our main contributions are summarized as follows:

- We propose a unified memory mechanism for 3D scene understanding, which includes representational memory for storing ego-centric visual images and abstract memory for maintaining structured information tables.
- We develop a novel FaST-3D agent for zero-shot 3D VQA, featuring tightly coupled fast-thinking and slow-thinking systems.
- Extensive experiments on ScanQA and OpenEQA datasets demonstrate that FaST-3D excels in all metrics, especially in the open-ended LLM-Match score.

Related Work

3D VQA

3D VQA is a challenging task in 3D scene understanding, requiring models to comprehend spatial environments and reason effectively to answer user queries. ScanQA (Azuma et al. 2022) stands as the first large-scale 3D VQA dataset derived from ScanNet (Dai et al. 2017). Unfortunately, it simplifies the VQA task by formulating it as a classification problem, where the goal is to select the correct answer from a closed set of options. Considering the open-vocabulary nature of 3D VQA, (Majumdar et al. 2024) introduces a OpenEQA dataset, which uses the LLM-Match score to simulate human preferences in evaluation.

Currently, most studies are conducted on ScanQA, following its classification-based assumptions for 3D VQA. These works primarily focus on refining end-to-end model architectures to enhance EM@1 accuracy. For example, (Delitzas et al. 2023) leverages the rich representations of CLIP to better align 3D scene features, 2D image embeddings, and text descriptions. Similarly, (Mo and Liu 2024) introduces a novel Twin-Transformer module to enable more efficient cross-modal fusion. Additionally, some studies adopt a pretrain-finetune strategy to enhance model performance by injecting prior knowledge from large-scale 3D data pairs. (Huang et al. 2023) proposes a two-stage pretraining framework that aligns point cloud and text em-

beddings at both object-level and scene-level granularity. In contrast, (Li et al. 2024) challenges the complexity of multi-stage pipelines by advocating for end-to-end finetuning with a frozen 3D encoder. Despite these advancements, most existing approaches remain limited by the scale of available 3D data pairs. With the rapid progress of LLMs, fully zero-shot 3D VQA has emerged as a promising direction (Singh, Pavlakos, and Stamoulis 2024; Liu et al. 2024a). They convert 3D scenes into textual descriptions, such as scene graphs or frame captions, and provide them as context for LLMs to answer user queries. While LLM-based methods exhibit strong generalization, the loss of spatial information during conversion significantly impacts accuracy, leaving ample room for improvement in future research.

Multimodal Agents

Recently, LLM-based agents have gained increasing attention in both academia and industry (Xi et al. 2023). Unlike standalone LLMs, these agents operate as autonomous systems with planning capabilities: they can decompose complex tasks into substeps, execute them incrementally, and invoke external tools as needed. This allows them to extend the generalization power of LLMs to more complex visual tasks, demonstrating strong practical potential.

Several pioneering works have applied LLM-based agents to image and video understanding. VisProg (Gupta and Kembhavi 2023) and Vipergpt (Surís, Menon, and Vondrick 2023) use GPT-3 as a planner to break down visual reasoning tasks into modular programs executed via external tools, achieving promising results. In video understanding, multimodal agents help reduce training costs and address the inherent challenges of long-form video analysis. For example, (Pan et al. 2023) introduces a text-based memory module to enhance video representations through contextual retrieval, (Fan et al. 2025) extracts temporal and object memories from long-form videos and designs a toolset for memory querying, while (Wang et al. 2025) incorporates self-reflection to iteratively improve observational accuracy and reasoning performance. In terms of 3D vision, early efforts have explored LLM-based agents for 3D visual grounding. LLM-Grounder (Yang et al. 2024) is the first agent work for zero-shot 3D visual grounding. It first identifies object categories from the query and then utilizes a *Target Finder* tool to filter candidates. (Yuan et al. 2024) extends visual programming to 3D visual grounding, leveraging an LLM planner to generate structured execution scripts. Despite these advances, no prior research has investigated the application of multimodal agents in 3D VQA. To bridge this gap, this paper presents the first study of multimodal agents for the more challenging 3D VQA task.

FaST-3D Agent

Overview

Figure 1 presents an overview of FaST-3D agent designed for zero-shot 3D VQA. It employs a unified memory mechanism, comprising both representational memory and abstract memory, to store and process 3D scene information. When a user query is received, the fast-thinking system is first activated to generate an intuitive answer based on visual information. Depending on the confidence level of this initial response, the agent determines whether to invoke the slow-thinking system for step-by-step logical reasoning. Finally, a dual integration module combines the outputs from both systems to produce the final answer.

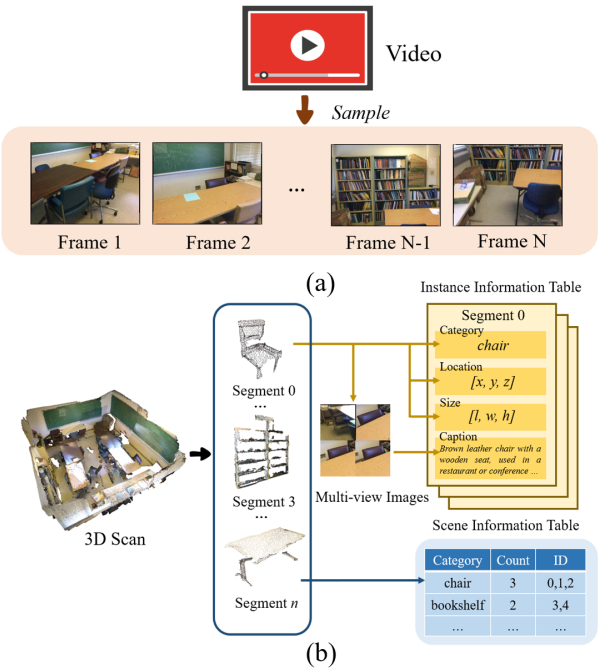

Figure 2: The construction process of (a). representational memory and (b). abstract memory.

Memory Mechanisms

Representational Memory As shown in Figure 2(a), representational memory captures the most intuitive visual information of 3D scenes, serving as the foundation for rapid reasoning. It consists of a sequence of images $I = \{I_1, I_2, \ldots, I_N\}$, uniformly sampled at 1-second intervals from the ego-centric video stream V. This memory not only preserves fine-grained visual details, such as object texture and lighting conditions, but also accurately reflects the temporal dynamics of motion.

Abstract Memory Abstract memory encodes structured 3D scene knowledge in standardized relational tables. As illustrated in Figure 2(b), we first apply the Mask3D model (Schult et al. 2023) to perform instance segmentation on the 3D scan or directly use ground-truth (GT) annotations. Based on the segmentation results, we design two queryable tables as abstract memory: a scene information table and an instance information table. Scene information table provides a global summary of the scene, allowing for accurate instance identification according to category names. It is structured as a relational database with three fields: category name, object count, and corresponding segment IDs.

Instance information table records fine-grained attributes for each segment, indexed by a unique ID. Each entry records the following essential properties, including: category, size, location, and a brief textual caption. Specifically, we adopt the algorithm proposed in (Jia et al. 2025) to generate high-quality object descriptions. For further details on segment processing in abstract memory, refer to App. A.

Fast-Thinking System

When the user poses a query q, the FaST-3D agent first activates its fast-thinking system, which utilizes an MLLM to generate an initial response based on representational memory. To further enhance reasoning efficiency and accuracy, we introduce two key components in the fast system: an adaptive image retriever and a self-assessment module.

Adaptive Image Retriever In cognitive science, the theory of selective attention posits that humans prioritize processing information most relevant to the current task or context, thereby reducing cognitive load and improving efficiency (Driver 2001). Inspired by this, we propose a novel adaptive image retriever that filters irrelevant visual cues and focuses on images most relevant to the query q. To improve semantic clarity, we first use an LLM to transform q into a declarative statement q_{dec}. Then, we use the CLIP model (Radford et al. 2021) to compute the semantic similarity between q_{dec} and each image in representational memory I, selecting the top 4 images with the highest similarity scores to construct the image set S. Notably, both the original question q and its declaration q_{dec} can be used in the retrieval process, with further analysis provided in App. E. The rephrasing prompt is detailed in App. B.1.

Self-Assessment Module The fast-thinking system is designed to deliver rapid responses via visual reasoning, while its reliability remains a critical concern. To address this issue, we introduce a self-assessment module which prompts the MLLM to generate a confidence score alongside each answer during inference. This score, ranging from *1* (very uncertain) to *5* (highly confident), reflects the agent's self-evaluation of the response's reliability and serves as a decision criterion for invoking the slow-thinking system. If the confidence score is *4* or above, the response is deemed reliable and returned to the user directly. Otherwise, the agent activates the slow-thinking system to perform step-by-step reasoning using structured knowledge from abstract memory. Details on the prompts used in the fast-thinking system can be found in App. B.2.

Slow-Thinking System

The slow-thinking system is designed to handle complex spatial reasoning and logical inference in 3D scenes. It breaks down challenging queries into manageable subtasks and incrementally processes structured data by invoking external tools. Rather than adopting a large toolset, we streamline the system with five specialized tools, grouped into three functional categories: *Information Retrieval (IR)*, *Semantic Analysis (SA)*, and *Relation Analysis (RA)*. Table 1 defines each tool in detail. This minimalist design avoids redundancy, and ensures that each tool is used with precision and intent, thereby improving overall reasoning efficiency.

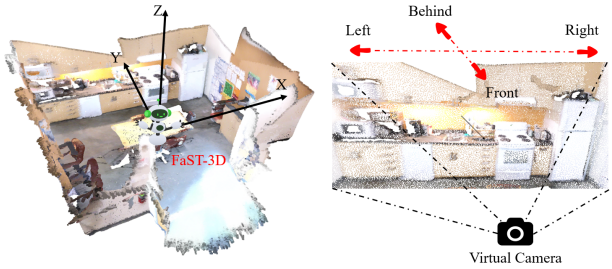

Figure 3: Egocentric 2D projection method to determine view-dependent relations.

Category	Tool	Definition
IR	*SIR*	Return the corresponding item from scene information table for a given category c.
	IIR	Return instance information for each object in the list x.
SA	*SF*	Return objects in the list x that semantically match the query q.
RA	*VDR*	Return objects in the list x that satisfy a given view-dependent relation r_d with the anchor a
	VIR	Return objects in the list x that satisfy a given view-independent relation r_i with the anchor a

Table 1: Tool list in the slow-thinking system: *Scene Information Retrieval (SIR)*, *Instance Information Retrieval (IIR)*, *Semantic Filter (SF)*, *View Dependent Relations (VDR)*, and *View Independent Relations (VIR)*.

The *IR* category consists of two tools: *SIR* and *IIR*, both designed to query structured tables in abstract memory. *SIR* retrieves entries based on a given object category, while *IIR* fetches complete instance information for a list of object IDs. To ensure robustness, *IIR* automatically skips invalid IDs, enabling smooth and uninterrupted queries.

In the *SA* category, we introduce the *SF* tool, which determines whether an object matches a given open-vocabulary description. Specifically, it iterates over each instance in the ID list, extracting its 'category' and 'caption' attributes. Then we integrate this information with the query to create a complete prompt, which is fed into *Llama-3.1-8B-Instruct* for binary semantic judgment. The response is parsed to produce a final decision of either 'yes' or 'no'. More details and prompt templates are provided in App. B.3.

For the *RA* category, we develop two specialized tools to compute spatial relations between targets and anchors in 3D space. *VDR* assesses four egocentric spatial relations—'front', 'back', 'left', and 'right'—which dynamically vary with the agent's viewpoint. To tackle this challenge, we adopt an egocentric 2D projection method (Yuan et al.

2024), as illustrated in Figure 3. The agent is assumed to be positioned at the center of the room, denoted as P_c, and is equipped with a virtual camera. The virtual camera rotates to face the anchor position P_a and projects all target objects onto its 2D image plane for calculation. The projection process is formulated as follows:

$$R, T = VT(P_c, P_a, up) \tag{1}$$

$$(u, v, w)^T = I \cdot (R \mid T) \cdot P_t \tag{2}$$

VT is a view transformation function that calculates the rotation matrix R and translation matrix T (Vince and Vince 2010), with $up = (0, 0, 1)^T$ as a unit vector along the z-axis. P_t represents the position of the target object, I is the intrinsic matrix, and (u, v, w) correspond to the final projections on the image plane. The u-coordinate determines the left-right positioning: a smaller u indicates the object is to the left. The w-coordinate differentiates between front and back: a lower w suggests the object is in front.

In contrast, the *VIR* tool evaluates three view-independent spatial relations: 'high', 'low', and 'distance'. The 'high' and 'low' relations are determined by comparing the vertical positions of the target and anchor objects, while 'distance' is computed by measuring the Euclidean distance between each target object and the anchor. The resulting distance list is then provided to the FaST-3D agent to support subsequent reasoning steps.

Dual Integration

FaST-3D agent leverages two complementary thinking systems to generate individual answers. To integrate the strengths of both systems, we design a straightforward dual integration module. It aggregates the answers from both systems along with the original query and in-context examples, forming a comprehensive prompt. The prompt is then fed into an LLM to generate the final response. The prompt template is provided in App. B.4.

Experiments

Datasets

ScanQA (Azuma et al. 2022) is a large-scale 3D VQA dataset comprising 41,000 questions across 800 indoor scenes from ScanNet (Dai et al. 2017). We follow its standard data split and evaluate on the validation set.

OpenEQA (Majumdar et al. 2024) is the first open-vocabulary benchmark for 3D Embodied Question Answering (EQA). Our experiments focus on its EM-EQA subset, which contains 1,636 high-quality questions across 180 real-world environments sourced from ScanNet and HM3D (Ramakrishnan et al. 2021).

Experimental Setup

Implementation Details We develop two versions of FaST-3D agent for experiments: *Mask3D* and *GT*, which primarily differ in their segmentation stage. The *Mask3D* version constructs structured tables using instance masks predicted by a Mask3D model (Schult et al. 2023) trained on ScanNet200, while the *GT* version directly uses GT masks. Since GT masks are missing for part of the ScanNet test split in EM-EQA, we exclude the affected QA pairs and conduct experiments on the remaining 1,510 samples.

In all experiments, FaST-3D consistently uses *GPT-4o-2024-08-06* as the reasoning engine. We use the CLIP model to select the top 4 semantically relevant images in adaptive image retriever, ensuring SOTA performance in 3D VQA across all configurations. The confidence score threshold is set to 4 in the self-assessment module.

Evaluation Metrics For ScanQA, we adopt standard metrics from previous work, including EM@1, BLEU scores (Papineni et al. 2002), ROUGE-L (Lin 2004), METEOR (Banerjee and Lavie 2005), and CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015). However, these metrics prioritize lexical precision in a fixed vocabulary, ignoring the open-vocabulary nature of VQA tasks. To address this limitation, we introduce the LLM-Match score (Majumdar et al. 2024), which utilizes *GPT-4* to simulate human evaluation by assessing the semantic relevance between predicted and GT answers. Additional details on the LLM-Match metric are provided in App. C. For OpenEQA, we rely solely on the LLM-Match score for evaluation.

Comparative Methods We compare the FaST-3D agent with several existing methods on ScanQA, including expert models (Azuma et al. 2022; Mo and Liu 2024), pre-trained models (Jin et al. 2023; Li et al. 2024; Huang et al. 2023; Hong et al. 2023), and LLM-based approaches (Singh, Pavlakos, and Stamoulis 2024). Among LLM-based methods, *blind GPT-4* processes only the query without additional context, while *GPT-4 w/ SGC (Scene-Graph Captions)* enhances reasoning by incorporating scene descriptions generated by *GPT-4V* from 3D mesh views.

For OpenEQA, we select advanced LLM-based baselines for comparison. *GPT-4 w/ FC (Frame Captions)* samples 50 frames from episodic memory and use *LLaVa-1.5* (Liu et al. 2024b) to generate captions as context. *GPT-4 w/ SGC* constructs an object-centric scene graph from visual images, providing textual descriptions for each object. Specifically, we explore two scene graph construction methods: CG (ConceptGraph) (Gu et al. 2024) and SVM (Sparse Voxel Map). *GPT-4V* directly processes 50 sampled frames using MLLM for visual reasoning. See App. D for more details on comparative methods.

Quantitative Results

***ScanQA* Results** Table 2 presents the performance of all methods on the ScanQA validation set. Both versions of the FaST-3D agent outperform previous zero-shot approaches across all evaluation metrics. In particular, FaST-3D *(GT)* achieves SOTA performance, showing a slight improvement over *Mask3D*. This gain is attributed to the more accurate instance information stored in abstract memory. In terms of conventional EM@1, the expert model *BridgeQA* maintains the highest accuracy of 27.0%, while FaST-3D *(GT)* demonstrates a 0.8% improvement in the zero-shot setting, narrowing the gap with supervised models. Moreover, FaST-3D exhibits substantial improvements across other closed-set metrics, illustrating the potential of zero-shot methods to rival

Method	Type	EM@1	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	LLM-Match
Supervised Methods										
VoteNet + MCAN (Azuma et al. 2022)	E	17.3	28.1	16.7	10.8	6.2	11.4	29.8	54.7	-
ScanRefer + MCAN (Azuma et al. 2022)	E	18.6	26.9	16.6	11.6	7.9	11.5	30.0	55.4	-
ScanQA (Azuma et al. 2022)	E	20.3	29.5	19.8	14.7	9.6	12.6	32.4	61.7	-
BridgeQA (Mo and Liu 2024)	E	27.0	-	-	-	-	-	-	-	-
3D-VLP (Jin et al. 2023)	P	21.7	30.5	21.3	16.7	11.2	13.5	34.5	67.0	-
3DMIT (Li et al. 2024)	P	13.0	27.6	-	-	5.2	10.7	26.2	48.0	-
Chat-3D v2 (Huang et al. 2023)	P	21.1	38.4	-	-	7.3	16.1	40.1	77.1	-
3D-LLM (Hong et al. 2023)	P	20.5	39.3	25.2	18.4	12.0	14.5	35.7	69.4	-
Zero-shot Methods										
Blind GPT-4 (Singh, Pavlakos, and Stamoulis 2024)	L	14.6	28.7	13.6	6.8	3.8	13.5	30.9	53.6	37.6
GPT-4 w/ Vocab-agnostic *SGC* (Singh, Pavlakos, and Stamoulis 2024)	L	10.1	16.3	7.0	3.2	1.0	8.8	20.0	34.2	25.7
GPT-4 w/ Vocab-grounded *SGC* (Singh, Pavlakos, and Stamoulis 2024)	L	18.0	24.5	10.7	4.6	1.6	14.2	33.4	58.3	42.0
FaST-3D *(Mask3D)*	A	18.5	33.2	17.1	8.6	4.6	15.9	33.6	69.2	48.5
FaST-3D *(GT)*	A	**18.8**	**34.6**	**18.7**	**9.5**	**5.4**	**16.3**	**34.8**	**72.1**	**50.3**

Table 2: Comparisons on the ScanQA validation set (E: Expert model, P: Pretrained model, L: LLM-based method, A: Agentic method). **Bold** denotes the best performance of zero-shot methods, and underline indicates the overall best result.

supervised approaches in 3D VQA. For instance, FaST-3D *(GT)* achieves a notable 10.1% gain in BLEU-1 and reaches a peak 16.3% in METEOR. Additionally, FaST-3D *(GT)* attains SOTA performance with a 50.3% LLM-Match score, reflecting an 8.3% increase and improved alignment with human preferences.

***OpenEQA* Results** Table 3 summarizes the performance of all methods on the OpenEQA dataset. Among the baseline methods, *GPT-4V* achieves the highest accuracy of 49.6% on LLM-Match using 50 images as visual context. In contrast, our FaST-3D agent attains a significantly higher accuracy of 58.5% using only 4 images as context, marking an 8.9% improvement while also substantially reducing token costs during reasoning. Furthermore, FaST-3D exhibits exceptional generalization capability, with accuracy gains of 12.7% on ScanNet and 2.4% on HM3D, respectively. Figure 4 presents a category-level performance analysis on the EM-EQA benchmark, where FaST-3D outperforms all competing methods across seven categories. Notably, it achieves a 15.5% improvement in the 'spatial understanding' category, underscoring its superior ability to handle complex spatial relationships and geometric reasoning.

Method	EM-EQA		
	ScanNet	HM3D	ALL
Blind GPT-4	32.5	35.5	33.5
GPT-4 w/ LLaVA-1.5 *FC*	45.4	40.0	43.6
GPT-4 w/ CG *SGC*	37.8	34.0	36.5
GPT-4 w/ SVM *SGC*	40.9	35.0	38.9
GPT-4V	51.3	46.6	49.6
FaST-3D* *(Mask3D)*	62.5	47.8	57.1
FaST-3D* *(GT)*	**64.0**	**49.0**	**58.5**

Table 3: Comparisons on the EM-EQA set of OpenEQA. FaST-3D* results are based on a subset of 1,510 samples.

Ablation Studies

To control API costs, we randomly sample three subsets from the ScanQA validation set for ablation analysis, each containing 500 questions.

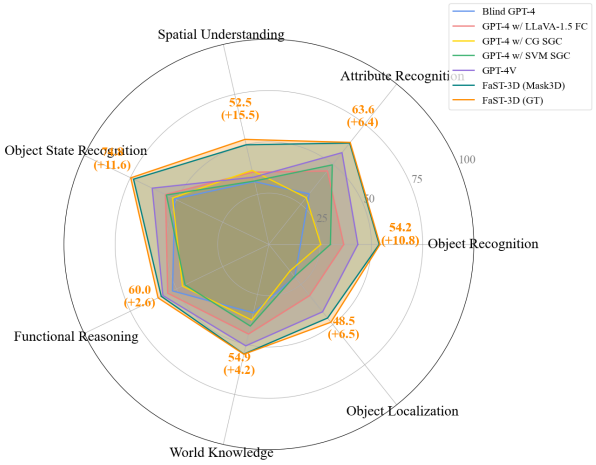

Figure 4: Category-level performance on EM-EQA.

Fast System	Slow System	LLM-Match	Token
✓		$45.7_{\pm 1.7}$	$1043.8_{\pm 25.7}$
	✓	$25.5_{\pm 1.4}$	$4088.7_{\pm 79.8}$
✓	✓	$50.2_{\pm 1.3}$	$1259.7_{\pm 45.9}$

Table 4: Ablation study on the thinking system.

Effectiveness of Thinking Systems We evaluate the individual contributions of two thinking systems in FaST-3D, with results recorded in Table 4. When both systems are enabled, the VQA accuracy reaches 50.2%, with an average token cost of 1259.7. In contrast, using only the fast system reduces accuracy to 45.7%, while relying solely on the slow system causes a sharp drop to 25.5%. These findings indicate the dominant role of visual information in 3D spatial understanding, with structured table data providing essential complementary support.

Choice of Image Sampling and Quantity We investigate the effect of image sampling strategies and quantity on agent performance. As shown in Figure 5(a), with a fixed image count, the adaptive strategy achieves higher accuracy by extracting more semantically relevant images. FaST-3D

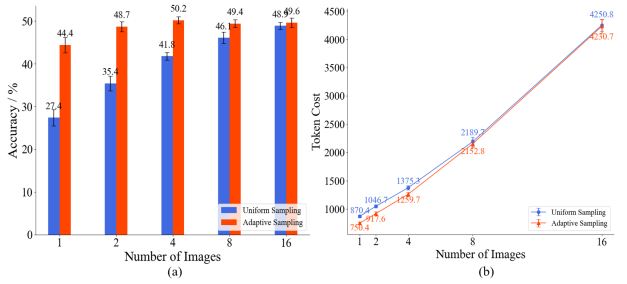

Figure 5: Ablation study on image sampling strategy and quantity: (a). VQA accuracy, (b). Token cost.

reaches its peak accuracy of 50.2% with 4 images. However, further increasing the number of images slightly reduces accuracy, indicating that redundant visual information can impede spatial understanding. In addition, the token analysis in Figure 5(b) shows that the uniform sampling strategy requires more tokens than the adaptive strategy under identical conditions. This discrepancy arises from the lower reliability of fast answers, which triggers more frequent use of the slow system, thereby increasing token usage.

SIR & IIR	*SF*	*VDR*	*VIR*	LLM-Match
✓		✓	✓	$47.2_{\pm 1.1}$
✓	✓		✓	$48.6_{\pm 1.4}$
✓	✓	✓		$49.1_{\pm 0.9}$
✓	✓	✓	✓	$50.2_{\pm 1.3}$

Table 5: Ablation study on the toolkit.

Ablation for Toolkit We also perform ablation experiments to evaluate the contribution of each tool in the toolkit. Since *SIR* and *IIR* are essential for information retrieval, we retain them in all settings. As illustrated in Table 5, the removal of any tool results in a decrease in VQA accuracy. Notably, eliminating *SF* results in a notable 3.0% drop, emphasizing its critical role in the slow-thinking system. These results demonstrate that the combined use of all tools is necessary to achieve optimal 3D VQA performance.

We also provide ablation results on the image retriever, confidence score, MLLM engines, and paraphrasing operations in App. E.

Qualitative Analysis

To provide a more intuitive illustration of FaST-3D's performance, we present two inference examples in Figure 6. Figure 6(a) depicts a case where the fast-thinking system successfully retrieves the most semantically relevant image and outputs a high-confidence answer directly output as the final response. In contrast, Figure 6(b) shows a failure of the fast system to extract the geometric attributes of 'lamp', thereby triggering the slow system for spatial reasoning. It first retrieves the 'lamp' from abstract memory, identifies the reference object 'curtain' through associative reasoning, and then calculates their spatial relationship to determine the

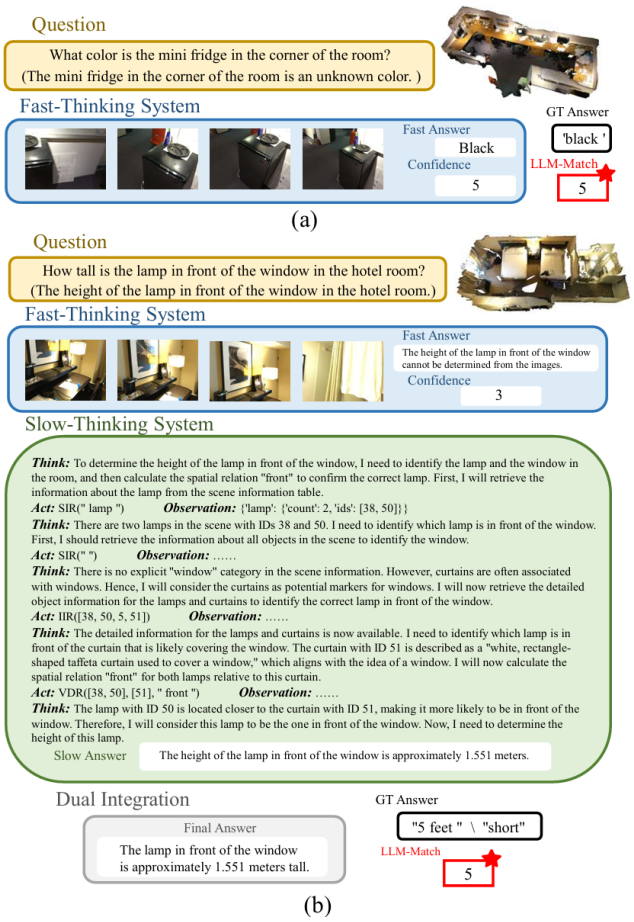

Figure 6: Qualitative results of FaST-3D's reasoning process: (a) fast-thinking system only, (b) both systems are activated.

correct instance and infer its height. Additional reasoning examples are provided in App. F. These cases highlight how the dual-system design of FaST-3D leverages complementary strengths to strike an optimal balance between VQA accuracy and reasoning efficiency.

Conclusions

In this paper, we present FaST-3D, a novel agent-based approach for zero-shot 3D VQA with broad implications for embodied intelligence. FaST-3D integrates two complementary reasoning systems: a fast-thinking module for rapid visual inference and a slow-thinking module for step-by-step logical reasoning. It achieves exceptional performance in zero-shot settings and demonstrates strong generalization across diverse scenarios. In future work, we will explore the deployment of FaST-3D agent on resource-constrained edge devices to evaluate its efficiency and scalability in real-world applications.

Acknowledgments

This work is supported by the Beijing Natural Science Foundation (L245025).

References

Ando, A.; Gidaris, S.; Bursuc, A.; Puy, G.; Boulch, A.; and Marlet, R. 2023. RangeViT: Towards Vision Transformers for 3D Semantic Segmentation in Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5240–5250.

Azuma, D.; Miyanishi, T.; Kurita, S.; and Kawanabe, M. 2022. ScanQA: 3D Question Answering for Spatial Scene Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19129–19139.

Banerjee, S.; and Lavie, A. 2005. 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*, 65–72.

Chen, Y.; Liu, J.; Zhang, X.; Qi, X.; and Jia, J. 2023. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21674–21683.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Niessner, M. 2017. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Delitzas, A.; Parelli, M.; Hars, N.; Vlassis, G.; Anagnostidis, S.; Bachmann, G.; and Hofmann, T. 2023. Multi-clip: Contrastive vision-language pre-training for question answering tasks in 3d scenes. *arXiv preprint arXiv:2306.02329*.

Deng, W.; Yang, J.; Ding, R.; Liu, J.; Li, Y.; Qi, X.; and Ngai, E. 2024. Can 3D Vision-Language Models Truly Understand Natural Language? *arXiv preprint arXiv:2403.14760*.

Driver, J. 2001. A selective review of selective attention research from the past century. *British journal of psychology*, 92(1): 53–78.

Fan, Y.; Ma, X.; Wu, R.; Du, Y.; Li, J.; Gao, Z.; and Li, Q. 2025. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, 75–92. Springer.

Frankish, K. 2010. Dual-process and dual-system theories of reasoning. *Philosophy Compass*, 5(10): 914–926.

Gu, Q.; Kuwajerwala, A.; Morin, S.; Jatavallabhula, K. M.; Sen, B.; Agarwal, A.; Rivera, C.; Paul, W.; Ellis, K.; Chellappa, R.; Gan, C.; de Melo, C. M.; Tenenbaum, J. B.; Torralba, A.; Shkurti, F.; and Paull, L. 2024. ConceptGraphs: Open-Vocabulary 3D Scene Graphs for Perception and Planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 5021–5028.

Gupta, T.; and Kembhavi, A. 2023. Visual Programming: Compositional Visual Reasoning Without Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14953–14962.

Hong, Y.; Zhen, H.; Chen, P.; Zheng, S.; Du, Y.; Chen, Z.; and Gan, C. 2023. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36: 20482–20494.

Huang, H.; Wang, Z.; Huang, R.; Liu, L.; Cheng, X.; Zhao, Y.; Jin, T.; and Zhao, Z. 2023. Chat-3d v2: Bridging 3d scene and large language models with object identifiers. *arXiv preprint arXiv:2312.08168*.

Jia, B.; Chen, Y.; Yu, H.; Wang, Y.; Niu, X.; Liu, T.; Li, Q.; and Huang, S. 2025. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In *European Conference on Computer Vision*, 289–310. Springer.

Jin, Z.; Hayat, M.; Yang, Y.; Guo, Y.; and Lei, Y. 2023. Context-Aware Alignment and Mutual Masking for 3D-Language Pre-Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10984–10994.

Kahneman, D. 2011. Thinking, fast and slow. *Farrar, Straus and Giroux*.

Li, J.; Dai, H.; Han, H.; and Ding, Y. 2023. MSeg3D: Multi-Modal 3D Semantic Segmentation for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21694–21704.

Li, Z.; Zhang, C.; Wang, X.; Ren, R.; Xu, Y.; Ma, R.; Liu, X.; and Wei, R. 2024. 3dmit: 3d multi-modal instruction tuning for scene understanding. In *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 1–5. IEEE.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, B.; Dong, Y.; Wang, Y.; Rao, Y.; Tang, Y.; Ma, W.-C.; and Krishna, R. 2024a. Coarse correspondence elicit 3d spacetime understanding in multimodal language model. *arXiv preprint arXiv:2408.00754*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D. L.; and Han, S. 2023. BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2774–2781.

Ma, X.; Bhalgat, Y.; Smart, B.; Chen, S.; Li, X.; Ding, J.; Gu, J.; Chen, D. Z.; Peng, S.; Bian, J.-W.; et al. 2024. When LLMs step into the 3D World: A Survey and Meta-Analysis of 3D Tasks via Multi-modal Large Language Models. *arXiv preprint arXiv:2405.10255*.

Majumdar, A.; Ajay, A.; Zhang, X.; Putta, P.; Yenamandra, S.; Henaff, M.; Silwal, S.; Mcvay, P.; Maksymets, O.; Arnaud, S.; Yadav, K.; Li, Q.; Newman, B.; Sharma, M.; Berges, V.; Zhang, S.; Agrawal, P.; Bisk, Y.; Batra, D.; Kalakrishnan, M.; Meier, F.; Paxton, C.; Sax, A.; and Rajeswaran, A. 2024. OpenEQA: Embodied Question Answering in the Era of Foundation Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16488–16498.

Mao, J.; Shi, S.; Wang, X.; and Li, H. 2023. 3D object detection for autonomous driving: A comprehensive survey. *International Journal of Computer Vision*, 131(8): 1909–1963.

Mo, W.; and Liu, Y. 2024. Bridging the Gap between 2D and 3D Visual Question Answering: A Fusion Approach for 3D VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4261–4268.

Naveed, H.; Khan, A. U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; and Mian, A. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Pan, J.; Lin, Z.; Ge, Y.; Zhu, X.; Zhang, R.; Wang, Y.; Qiao, Y.; and Li, H. 2023. Retrieving-to-Answer: Zero-Shot Video Question Answering with Frozen Large Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 272–283.

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

Parelli, M.; Delitzas, A.; Hars, N.; Vlassis, G.; Anagnostidis, S.; Bachmann, G.; and Hofmann, T. 2023. CLIP-Guided Vision-Language Pre-Training for Question Answering in 3D Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 5607–5612.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramakrishnan, S. K.; Gokaslan, A.; Wijmans, E.; Maksymets, O.; Clegg, A.; Turner, J.; Undersander, E.; Galuba, W.; Westbury, A.; Chang, A. X.; et al. 2021. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*.

Schult, J.; Engelmann, F.; Hermans, A.; Litany, O.; Tang, S.; and Leibe, B. 2023. Mask3D: Mask Transformer for 3D Semantic Instance Segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 8216–8223.

Singh, S.; Pavlakos, G.; and Stamoulis, D. 2024. Evaluating Zero-Shot GPT-4V Performance on 3D Visual Question Answering Benchmarks. *arXiv preprint arXiv:2405.18831*.

Song, Z.; Liu, L.; Jia, F.; Luo, Y.; Jia, C.; Zhang, G.; Yang, L.; and Wang, L. 2024. Robustness-Aware 3D Object Detection in Autonomous Driving: A Review and Outlook. *IEEE Transactions on Intelligent Transportation Systems*, 25(11): 15407–15436.

Surís, D.; Menon, S.; and Vondrick, C. 2023. ViperGPT: Visual Inference via Python Execution for Reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 11888–11898.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Vince, J.; and Vince, J. A. 2010. *Mathematics for computer graphics*, volume 5. Springer.

Wang, X.; Zhang, Y.; Zohar, O.; and Yeung-Levy, S. 2025. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, 58–76. Springer.

Xi, Z.; Chen, W.; Guo, X.; He, W.; Ding, Y.; Hong, B.; Zhang, M.; Wang, J.; Jin, S.; Zhou, E.; et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Yang, J.; Chen, X.; Qian, S.; Madaan, N.; Iyengar, M.; Fouhey, D. F.; and Chai, J. 2024. LLM-Grounder: Open-Vocabulary 3D Visual Grounding with Large Language Model as an Agent. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 7694–7701.

Yuan, Z.; Ren, J.; Feng, C.-M.; Zhao, H.; Cui, S.; and Li, Z. 2024. Visual Programming for Zero-shot Open-Vocabulary 3D Visual Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20623–20633.