Agentic Collaborative Cognition for Zero-Shot 3D Understanding

Wenxin Wang[1⋆], Bo Zhang[1⋆], Feng Chen[2†], Zixuan Wang[1], Wen Li[3], Changsheng Li[4], and Yinjie Lei[1‡]

[1] Sichuan University, Chengdu, China
[2] Adelaide University, Adelaide, Australia
[3] University of Electronic Science and Technology of China, Chengdu, China
[4] Beijing Institute of Technology, Beijing, China
2024222050082@stu.scu.edu.cn, yinjie@scu.edu.cn

Abstract. Recent advancements have explored agentic zero-shot 3D understanding by reformulating it as video keyframe understanding with Multimodal Large Language Models (MLLMs). However, existing methods face an intrinsic bottleneck due to the finite observation perspectives inherent in videos and the implicit perception of 3D scenes. In this paper, we propose a collaborative multi-agent framework that assigns a Planning Agent to handle high-level viewpoint planning and supplement novel perspectives, and a Perception Agent to explicitly summarize the 3D scene into a structured holistic cognitive map. Specifically, Planning Agent first analyzes this cognitive map to determine query-relevant viewpoints and supplements missing critical perspectives to ensure comprehensive observation. Subsequently, Perception Agent documents object-level attributes from these views by assigning consistent instance identifiers across viewpoints, thereby integrating fragmented observations into the holistic cognitive map. In parallel, it provides feedback to filter out mismatched candidate objects and guide subsequent viewpoint planning. Through this closed-loop iterative process, two agents collaboratively figure out candidates until Perception Agent determines that sufficient information has been captured to complete the task. Extensive experiments demonstrate that our method achieves state-of-the-art performance on 6 benchmarks, with improvements of 11.1% Acc@0.5 on ScanRefer, 14.6 BLEU-1 on 3D-assisted dialog, and 2.1 EM on SQA3D. **Project Page**: https://zhangbo135.github.io/agentic-collaborative-cognition/

Keywords: Zero-Shot 3D Understanding · Collaborative Cognition · Comprehensive Scene Observation

1 Introduction

3D scene understanding [21, 23, 44, 57] focuses on interpreting spatial layouts, semantic attributes, and inter-object relationships in complex environments, enabling critical applications from robotics [43, 46] to augmented reality [3, 28].

⋆ Equal contribution. † Project lead. ‡ Corresponding author.

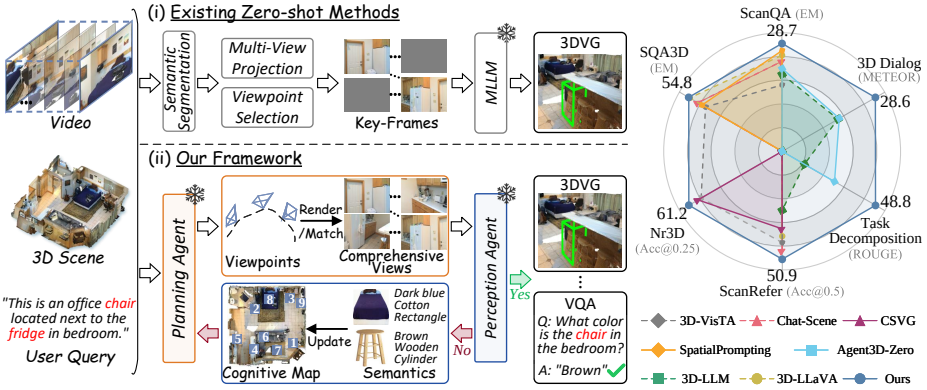


Fig. 1: Comparison of existing methods with our method. Left: Existing zero-shot methods [22,26] typically rely on video keyframe selection for 3D visual grounding. Our method actively plans viewpoints to obtain comprehensive views and explicitly summarizes the 3D scene into a structured holistic cognitive map from these views for diverse 3D tasks. Right: Our method achieves substantial improvements over existing zero-shot and most fully supervised methods across 6 benchmarks.

Previous studies mainly rely on fully supervised paradigms to achieve 3D visual grounding [8, 60], 3D dense captioning [9, 16], 3D question answering [4, 29], *etc.* Despite their effectiveness, the heavy reliance on large-scale 3D-text paired datasets limits scalability and adaptability to diverse real-world environments [51, 55].

Zero-shot 3D understanding [39, 52, 56] has emerged to reduce the requirement for manually annotated datasets [11,35]. As shown in the top-left of Fig. 1, previous methods [22, 26] typically leverage the general world knowledge of Multimodal Large Language Models (MLLMs) [19, 42] to interpret keyframes extracted from scene videos (an inherent component of scanned 3D datasets) through object-centric keyframe selection, which is enabled by key proposal selection and multi-view projection. However, relying solely on video keyframe selection inevitably has a performance bottleneck because: (1) *Finite observation perspectives.* During 3D scene scanning, videos are typically captured from restricted viewpoints close to the scene surfaces to ensure reconstruction quality. However, they provide only finite frames with fixed viewpoints, which often lack query-relevant critical perspectives; (2) *Implicit spatial perception.* Existing methods [22, 39] typically leverage MLLMs to implicitly infer the spatial layout from keyframes. Although capable of capturing inter-object relationships under similar viewpoints [34, 41, 60], they struggle to maintain unified spatial cognition when facing significant viewpoint variations [47, 50].

In this paper, we propose a collaborative multi-agent framework for zero-shot 3D scene understanding. As illustrated in the bottom-left of Fig. 1, our method consists of two specialized agents: a Planning Agent that performs flexible planning of query-relevant viewpoints and supplements missing critical perspectives,

and a Perception Agent that explicitly summarizes the 3D environment into a structured holistic cognitive map. Specifically, Planning Agent first analyzes this cognitive map to retrieve critical candidate objects and determine query-relevant viewpoints to further figure out these objects. For missing critical perspectives, it supplements rendered views to ensure comprehensive observation for Perception Agent. Subsequently, Perception Agent assigns unique instance identifiers to all objects in each view to ensure cross-view consistency, thereby accurately documenting object appearance and physical attributes, and integrating originally fragmented observations into the cognitive map. Additionally, it provides feedback that filters out mismatched objects to narrow the scope of candidates and guides the next round of viewpoint planning. This iterative collaboration allows the two agents to progressively refine 3D scene cognition until Perception Agent determines that sufficient information has been captured.

Empirically, as illustrated in the right of Fig. 1, our method surpasses previous state-of-the-art zero-shot and most fully supervised methods across 6 benchmarks, achieving substantial improvements of 11.1% Acc@0.5 on ScanRefer [8], 14.6 BLEU-1 on 3D-assisted dialog [15], and 2.1 EM on SQA3D [29]. In brief, our main contributions can be summarized as follows:

- We propose a collaborative multi-agent framework that reformulates zero-shot 3D understanding as an interactive planning-perception process.
- We design a Planning Agent to plan query-relevant viewpoints for comprehensive observation, and a Perception Agent to update the holistic cognitive map, enabling the MLLM to directly observe and understand 3D scenes.
- Extensive experiments on 6 benchmarks validate the state-of-the-art performance and generality of our method across diverse 3D understanding tasks.

2 Related Work

Zero-shot 3D Scene Understanding. Zero-shot 3D scene understanding has been explored as a promising alternative to fully supervised paradigms. Early works [14, 49, 55] typically reformat the 3D scene into textual descriptions for LLMs. For example, LLM-Grounder [49] uses the LLM to decompose queries into semantic constituents and employs a grounding tool to identify the target. CSVG [52] constructs structured scene graphs to model inter-object spatial relationships. However, these methods rely solely on text and struggle to capture rich visual details and fine-grained cues. Recent studies [26, 39, 48] leverage MLLMs to understand the 3D scene from 2D videos. SeqVLM [26] develops a proposal-guided projection strategy that performs object-centric keyframe selection to localize the target object. SPAZER [22] proposes a coarse-to-fine progressive reasoning mechanism to enable 3D-2D joint decision-making. Despite their effectiveness, they remain limited by the finite observation perspectives and fail to explicitly model the entire 3D scene.

3D Scene Documenting. Effective 3D scene documenting is a key approach to enhance understanding capabilities, evolving from geometric maps to rich semantic structures [32, 36]. SceneGraphFusion [45] proposes a real-time method

that incrementally constructs a globally consistent representation to document the 3D scene. More recently, Clio [30] leverages LLMs to construct language-based open-vocabulary 3D scene graphs. MTU3D [64] uses FastSAM [59] and DINO [33] to implicitly construct a dynamic spatial memory bank. Embodied VideoAgent [12] analyzes egocentric video and embodied sensory inputs to build a persistent scene memory. However, these representations often focus on categorical labels and coarse scene-level state, lack explicit multi-dimensional attributes, and typically require several tools to query the memory, whereas our cognitive map provides a flexible and fine-grained representation and further serves as a shared state for inter-agent communication.

MLLM-Based Multi-Agent Collaboration. With the rapid advancement of MLLMs [5,19], recent research has increasingly explored a wide range of applications that focus on MLLM-based multi-agent collaboration for multimodal reasoning. This allows multiple agents to work together, which enables the collective intelligence to efficiently solve complex and multi-step tasks via explicit role specialization and communication. For example, leveraging MLLMs for role-playing can accomplish multiple challenging tasks, such as embodied navigation [38], social simulation [27,40], video understanding [62], and video generation [53]. This paradigm is particularly relevant to 3D scene understanding, which naturally involves two heterogeneous sub-tasks: viewpoint acquisition and perceptual verification. Relying on a single agent inevitably leads to functional entanglement, causing insufficient verification and potential mislocalization of the target object. Therefore, we adopt a collaborative multi-agent framework that explicitly decouples viewpoint planning from perceptual verification.

3 Method

Given a query, the goal of 3D scene understanding is to either localize the referring target object or answer questions about object attributes and spatial layout. In this work, we propose a general collaborative multi-agent framework that progressively refines the cognition of the 3D scene through iterative interactions between planning and perception for diverse 3D understanding tasks. Here are two key benefits of our method: (1) Rather than relying on static object-centric keyframe selection from the video [22, 26], we actively identify critical objects, capture their relationships with surrounding objects, and allow targeted observations of uncertain objects, ensuring that query-relevant information is fully captured; (2) The holistic cognitive map serves as an explicit and structured documentation of the 3D scene, shared among agents. By integrating spatial and semantic information, it facilitates consistent understanding and effective collaboration between agents.

3.1 Overview

The overall framework is illustrated in Fig. 2. Given a textual query Q, our method incorporates a Planning Agent and a Perception Agent to collaboratively capture query-relevant scene information. Based on this, we first initialize

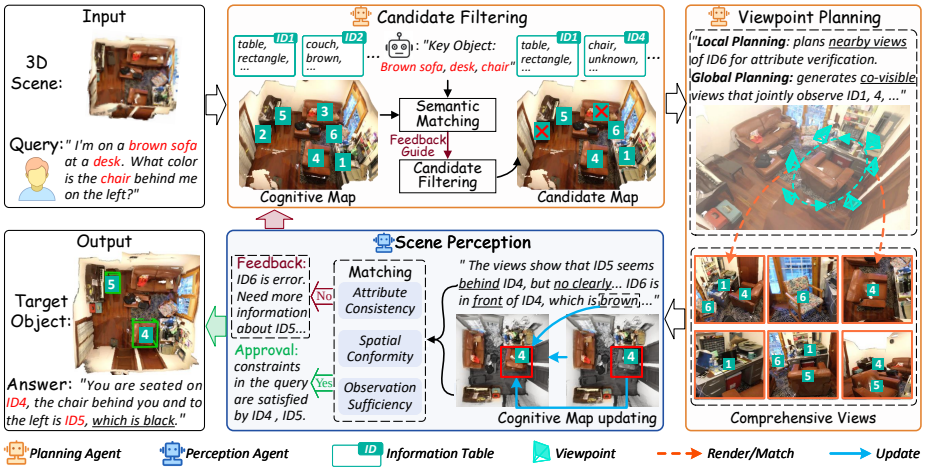


Fig. 2: Overview of our method. Given a 3D scene and query, our method achieves zero-shot 3D understanding through collaboration of planning and perception agents. We first construct a holistic cognitive map that contains an information table and a BEV annotated with instance identifiers. Planning Agent conducts candidate filtering to retrieve query-relevant objects, plans viewpoints and maps them to obtain comprehensive views. Perception Agent then perceives the scene from these views, and updates this map with documented object attributes. It matches current observations with query, and provides feedback to refine candidates or generates answer.

a holistic cognitive map $\mathcal{M} = \{\mathcal{T}, I_B\}$ that explicitly summarizes the scene. $\mathcal{T} = \{\sigma_i\}_{i=1}^{N}$ is an object information table that stores object-level information $\sigma_i = \{b_i, l_i, a_i\}$, where b_i and l_i represent the 3D bounding box and category label extracted by a 3D detection framework [37], and a_i contains attributes of each object. At initialization, a_i is empty and will be progressively updated by Perception Agent. $\mathcal{T}$ can be converted into a text format compatible with the MLLM input, providing an accurate spatial and semantic description of the 3D scene. I_B is a bird's-eye view (BEV), a 2D projection of the scene, annotated with the instance identifier i of each object.

We first parse the textual query Q to identify the key object set $\mathcal{O} = \{o_j\}_{j=1}^{J}$ of all objects involved in Q, where J denotes the number of key objects. During the t-th interaction round, Planning Agent first filters out irrelevant and erroneous candidate objects from $\mathcal{M}$ to obtain the candidate map $\mathcal{C}^t$, and then plans a set of viewpoints $\mathcal{V}^t$ to further figure out candidates. These viewpoints $\mathcal{V}^t$ are then matched with real images or supplemented with rendered images to obtain comprehensive observations $\mathcal{I}^t$. For each image in $\mathcal{I}^t$, we annotate all candidate objects with their unique identifiers, serving as visual prompts to ensure identity consistency across different viewpoints. Based on the annotated images, Perception Agent performs fine-grained perception, including object recognition, attribute extraction, and spatial relationships reasoning. The extracted attributes of each object are integrated into $\mathcal{T}$ of $\mathcal{M}$ to refine the description of the 3D

scene. Subsequently, Perception Agent provides feedback $\mathcal{F}^{t+1}$, which indicates erroneous candidate objects and guides the next round of viewpoint planning of Planning Agent. This iterative process continues until Perception Agent determines that attribute consistency, spatial conformity, and observation sufficiency have been achieved to generate the final answer.

3.2 Planning Agent

To overcome the limitations of previous methods [22, 26] that rely on finite perspectives in the video, we introduce a Planning Agent that supplements missing critical perspectives. As shown in Fig. 2, Planning Agent first leverages a candidate filtering strategy to retrieve candidate objects. Subsequently, it analyzes the candidate map to actively figure out these objects through flexible viewpoint planning, which enables broader and more diverse viewpoint coverage. These planned viewpoints are then mapped to obtain comprehensive views.

Candidate Filtering. We propose a retrieval-augmented candidate filtering strategy to progressively filter out irrelevant and erroneous candidate objects. For each key object $o_j \in \mathcal{O}$ in the query, we represent its semantics as $o_j = (l_j^q, a_j^q)$, consisting of the target category and attributes. We define a matching function $\text{Match}(\sigma_i, o_j)$ to determine whether a scene object $\sigma_i \in \mathcal{T}$ matches o_j:

$$\text{Match}(\sigma_i, o_j) = \begin{cases} \mathbb{1}\big[l_i = l_j^q\big], & a_i = \varnothing \ \vee \ a_j^q = \varnothing, \\ \mathbb{1}\big[l_i = l_j^q\big] \cdot \mathbb{1}\big[a_i = a_j^q\big], & \text{otherwise}, \end{cases} \tag{1}$$

where $\mathbb{1}\big[\,\cdot\,\big]$ denotes the indicator function and $\varnothing$ denotes that the attribute is undefined or missing. By applying this matching rule, we effectively filter out mismatched scene objects from the holistic cognitive map, thereby obtaining the initial candidate map:

$$\mathcal{C}^0 = \{\sigma_i \in \mathcal{T} \mid \exists\, o_j \in \mathcal{O},\ \text{Match}(\sigma_i, o_j) = 1\}\,. \tag{2}$$

Unlike previous methods [22, 26] that rely solely on object categories for retrieval, we leverage the fine-grained semantics provided by the holistic cognitive map to additionally incorporate object attributes into this matching process.

During the interaction, Planning Agent leverages feedback from Perception Agent, which identifies erroneous candidates that conflict with the query in terms of attributes or spatial relationships (as elaborated in the following Sec. 3.3), to progressively refine $\mathcal{C}^0$ and focus on critical objects for subsequent viewpoint planning. After the t-th interaction round, $\mathcal{C}^0$ is refined to $\mathcal{C}^t$.

Viewpoint Planning. To further figure out candidate objects, Planning Agent plans a set of viewpoints based on the candidate map. Specifically, by jointly considering the query and the fine-grained guidance provided by the perceptual feedback, Planning Agent plans viewpoints close to the target objects to observe and verify their attributes, and viewpoints that simultaneously observe multiple objects to directly capture spatial relationships. These viewpoints constitute a set $\mathcal{V}^t = \{V_k^t\}_{k=1}^K$, where K denotes the number of planned viewpoints. Each

viewpoint V_k^t is defined by its position and orientation on the 2D BEV plane, and is subsequently converted into a 3D camera pose represented by a rotation matrix R_k^t and a translation vector T_k^t.

Then each planned viewpoint is mapped to a real camera image or a rendered image: real images capture rich visual details but are limited by fixed camera viewpoints, whereas rendered images allow flexible observations at the cost of geometric and textural inaccuracies. Specifically, we design a viewpoint matching strategy that determines whether each planned viewpoint can be reliably matched with any real camera image. If a match is found, the corresponding real camera image is used; otherwise, a rendered image is synthesized to supplement the missing perspective and provide additional visual evidence.

Inspired by [39], we compute the pose distance $D_{k,m}^t$ for each planned viewpoint V_k^t with T_k^t and R_k^t to each real image I_m in the set $\mathcal{I} = \{I_m\}_{m=1}^M$, where M denotes the number of real images:

$$D_{k,m}^t = \frac{1}{\theta} \cdot \arccos\left(\frac{\mathrm{Tr}((R_k^t)^\top R_m) - 1}{2}\right) + \frac{1}{d} \cdot \left\|T_k^t - T_m\right\|_2, \tag{3}$$

where R_m and T_m denote the rotation matrix and translation vector of the m-th real camera image, $\mathrm{Tr}(\cdot)$ denotes the matrix trace operator, θ and d are normalization factors corresponding to the maximum rotation and translation difference. The corresponding image I_k^t is given by:

$$I_k^t = \begin{cases} I_{\hat{m}}, & \text{if } D_{k,\hat{m}}^t < \tau_D, \\ \mathrm{Render}(\mathcal{P}, R_k^t, T_k^t), & \text{otherwise,} \end{cases} \tag{4}$$

where $\hat{m} = \underset{m=1,\ldots,M}{\arg\min}\ D_{k,m}^t$ and τ_D is the viewpoint matching threshold. Finally, these matched and rendered 2D images are aggregated into a comprehensive view set $\mathcal{I}^t = \{I_k^t\}_{k=1}^K$ for Perception Agent.

3.3 Perception Agent

The initialized holistic cognitive map contains only object layout information and lacks fine-grained attributes, such as color, shape, or material, which are crucial for precise scene understanding. To this end, as shown in Fig. 2, we introduce a Perception Agent that continuously documents object attributes, figures out candidate objects, and generates the final answer.

Scene Perception. We annotate each candidate object with its unique instance identifier for each image in $\mathcal{I}^t$, thereby ensuring cross-view consistent understanding. Subsequently, Perception Agent interprets $\mathcal{I}^t$ and $\mathcal{C}^t$ to figure out the candidate objects and extract the appearance attributes and physical attributes of each object. Considering the domain gap between real and rendered images, Perception Agent is prompted to prioritize extracting fine-grained semantic attributes from real images, while using rendered images primarily to infer inter-object relationships and spatial layout. These extracted attributes are dynamically integrated into $\mathcal{T}$, progressively updating and enriching the holistic

cognitive map. Importantly, each attribute is assigned a confidence score, and attributes with higher confidence will overwrite previous ones, thereby identifying and correcting errors inferred from previous perceptions.

Perception Agent precisely matches the current observations with respect to the query using three distinct criteria: (1) Attribute conformity. Check whether the attributes of the candidates match the query; (2) Spatial consistency. Verify whether the spatial relations of the candidates satisfy the query constraints; (3) Observation sufficiency. Determine whether the current observations provide sufficient visual evidence to resolve the query. If any condition fails, Perception Agent provides feedback $\mathcal{F}^{t+1}$, which either identifies mismatched candidate objects in the case of attribute conflict or spatial conflict, or provides structured guidance for subsequent planning when observations are insufficient. Leveraging this guidance, Planning Agent actively seeks new viewpoints to resolve spatial ambiguities or capture fine-grained attribute evidence, thereby focusing on critical objects and avoiding exhaustive observations of the entire scene in an indiscriminate manner.

Through this closed-loop iterative process, the two agents collaboratively figure out the complex 3D environment until Perception Agent confirms that the above conditions are satisfied to generate the final answer.

4 Experiments

4.1 Experimental Settings

Benchmarks. To evaluate the comprehensive performance effectiveness of our method, we conduct extensive comparative experiments on 6 benchmarks covering diverse 3D scene understanding tasks: 3D visual grounding on ScanRefer [8] and Nr3D [1], situation estimation on SQA3D [29], 3D visual question answering on ScanQA [4] and SQA3D [29], 3D-assisted dialog and task decomposition on 3D-LLM held-in dataset [15].

Evaluation Metrics. For 3D visual grounding, we evaluate grounding accuracy using Acc@0.25 and Acc@0.5 on ScanRefer [8], and Acc@0.25 on Nr3D [1]. For situation estimation, we evaluate localization accuracy using Acc@0.5m and Acc@1.0m, and orientation accuracy using Acc@15° and Acc@30°. For 3D question answering, we report exact match accuracy (EM) and the refined exact match protocol (EM-R) proposed by LEO [17] on SQA3D [29]. Similarly, we report CIDEr, BLEU-4, METEOR, ROUGE, and EM on ScanQA [4]. For 3D-assisted dialog and task decomposition, we adopt BLEU-1, BLEU-4, METEOR, and ROUGE on 3D-LLM held-in dataset [15].

Implementation Details. We adopt Qwen2.5-VL-72B [6] as Planning Agent and GPT-4o (gpt-4o-2024-08-06 [19]) as Perception Agent. The rendered images are generated at a resolution of 512×512 pixels. We set the viewpoint matching threshold to $\tau_D = 0.8$. The number of planned viewpoints per interaction round is fixed at $K = 8$. The default number of Perception Agent is set to 1. The number of interaction rounds is limited to a maximum of 6. Following SeeGround [25]

Table 1: Evaluation of 3D question answering on SQA3D [29] test set and ScanQA [4] validation set. Gray font means the refined exact match, EM-R.

Method	SQA3D							ScanQA				
	What	Is	How	Can	Which	Other	Overall	CIDEr	BLEU-4	METEOR	ROUGE	EM
Supervision: Fully Supervised												
ScanQA [4]	28.6	65.0	47.3	66.3	43.9	42.9	45.3	60.2	10.8	12.6	31.1	20.9
SQA3D [29]	33.5	66.1	42.4	69.5	43.0	46.4	47.2	-	-	-	-	-
3D-VLP [21]	-	-	-	-	-	-	-	67.0	-	13.5	34.5	21.7
3D-LLM [15]	36.5	65.6	47.2	68.8	48.0	46.3	48.1	69.4	12.0	14.5	35.7	20.5
LEO [17]	-	-	-	-	-	-	50.0 (52.4)	101.4	13.2	20.0	49.2	24.5
SIG3D [31]	35.6	67.2	48.5	71.4	49.1	45.8	52.6	68.8	12.4	13.4	35.9	-
Chat-Scene [16]	45.4	67.0	52.0	69.5	49.9	55.0	54.6	87.7	14.3	18.0	41.6	21.6
LSceneLLM [61]	-	-	-	-	-	-	-	88.2	-	18.0	40.8	-
Scene-LLM [13]	40.9	69.1	45.0	70.8	47.2	52.3	54.2	80.0	12.0	16.6	40.0	27.2
Video-3D LLM [60]	51.1	72.4	55.5	69.8	51.3	56.0	58.6	102.1	16.3	19.8	49.0	30.1
3DRS [18]	54.4	75.2	57.0	72.2	49.9	59.0	60.6	104.8	17.2	20.5	49.8	30.3
Ross3D [41]	56.0	79.8	60.6	70.4	55.3	60.1	63.0 (65.7)	107.0	17.9	20.9	50.7	30.8
Supervision: Zero-shot												
Agent3D-Zero [56]	-	-	-	-	-	-	-	71.8	4.4	16.0	37.0	17.5
SpatialPrompting [39]	48.7	64.3	**53.6**	58.9	33.6	55.3	52.7	87.7	10.9	16.9	43.4	27.3
Ours	**50.2**	**65.0**	46.7	**67.8**	**46.4**	**55.7**	**54.8** (57.1)	**91.1**	**12.3**	**17.4**	**44.5**	**28.7**

Table 2: Evaluation of situation estimation on SQA3D [29] test set. "Separate" indicates models trained exclusively for situation estimation. "*" denotes experimental results reported in [54].

Method	Localization		Orientation	
	Acc@0.5m	Acc@1.0m	Acc@15°	Acc@30°
Supervision: Fully Supervised				
SQA3D [29]	9.5	29.6	8.7	16.5
SQA3D (separate)	10.3	31.4	17.1	22.8
3D-VisTA [63]	11.7	34.5	16.9	24.2
SIG3D* [31]	16.8	35.2	23.4	26.3
3DSA-LLM [54]	17.4	36.9	24.1	28.5
Supervision: Zero-shot				
Ours	**20.7**	**48.9**	22.8	**32.2**

Table 3: Evaluation of 3D-assisted Dialog and task decomposition on 3D-LLM Held-In Dataset [15]. B-1, B-4, M, and R denote BLEU-1, BLEU-4, METEOR, and ROUGE, respectively.

Method	3D-assisted Dialog				Task Decomposition			
	B-1	B-4	M	R	B-1	B-4	M	R
Supervision: Fully Supervised								
flant5 [10]	27.4	8.7	9.5	27.5	25.5	6.0	13.9	28.4
flamingo [2]	30.6	9.1	10.4	27.9	33.1	7.3	16.1	33.2
BLIP2-flant5 [24]	32.4	9.5	11.0	29.5	33.1	6.9	15.5	34.0
3D-LLM [15]	39.0	16.6	18.9	39.3	33.9	7.4	15.9	37.8
Supervision: Zero-shot								
Agent3D-Zero [56]	32.8	9.8	19.3	39.3	42.0	15.5	**22.9**	45.1
Ours	**47.4**	**24.0**	**28.6**	**47.3**	**43.8**	**24.9**	21.0	**48.8**

and SPAZER [22], we exclude the top 0.3m of the scene to better account for closed room setup.

4.2 Quantitative Results

3D Question Answering. As illustrated in Tab. 1, the quantitative comparison shows that our method outperforms existing zero-shot methods in most evaluation metrics. Compared with SpatialPrompting [39], it achieves improvements of 2.1% EM on SQA3D [29] and 3.4 CIDEr on ScanQA [4]. Moreover, our method already exceeds most fully supervised approaches, such as LEO [17], Chat-Scene [16], and Scene-LLM [13]. These results demonstrate that our approach attains a better understanding of spatial layouts and fine-grained object semantics by explicitly documenting the 3D scene.

Table 4: Evaluation of 3DVG on ScanRefer [8] and Nr3D [1] validation sets. In ScanRefer, queries are annotated as "Unique" (single target) or "Multiple" (with same-class distractors). In Nr3D, queries are categorized as "Easy" (one distractor) or "Hard" (multiple distractors), and as "View-Dependent" or "View-Independent", depending on whether grounding requires specific viewpoints. *0.25* and *0.5* denote Acc@0.25 and Acc@0.5, respectively. We also follow VLM-Grounder [48] to adopt the same *"250 queries"* setting for a fair comparison.

Method	ScanRefer						Nr3D				
	Unique		Multiple		Overall		Easy	Hard	Dep.	Indep.	Overall
	0.25	*0.5*	*0.25*	*0.5*	*0.25*	*0.5*	*0.25*	*0.25*	*0.25*	*0.25*	*0.25*
Supervision: Fully Supervised											
ScanRefer [8]	67.6	46.2	32.1	21.3	39.0	26.1	-	-	-	-	-
ReferIt3DNet [1]	-	-	-	-	-	-	43.6	27.9	32.5	37.1	35.6
3DVG-T [58]	77.2	58.5	38.4	28.7	45.9	34.5	48.5	34.8	34.8	43.7	40.8
BUTD-DETR [20]	84.2	66.3	46.6	35.1	52.2	39.8	60.7	48.4	46.0	58.0	54.6
3D-VisTA [63]	81.6	75.1	43.7	39.1	50.6	45.8	72.1	56.7	61.5	65.1	64.2
MiKASA [7]	-	-	-	-	-	-	69.7	59.4	65.4	64.0	64.4
Video-3D LLM [60]	88.0	78.3	50.9	45.3	58.1	51.7	-	-	-	-	-
3DRS [18]	87.4	77.9	57.0	50.8	62.9	56.1	-	-	-	-	-
Ross3D [41]	87.2	77.4	54.8	48.9	61.1	54.4	-	-	-	-	-
Supervision: Zero-shot (250 queries)											
VLM-Grounder [48]	66.0	29.8	48.3	33.5	51.6	32.8	55.2	39.5	45.8	49.4	48.0
SeqVLM [26]	77.3	72.7	47.8	41.3	55.6	49.6	58.1	47.4	51.0	54.5	53.2
SPAZER [22]	80.9	72.3	51.7	43.4	57.2	48.8	68.0	58.8	59.9	66.2	63.8
Ours	**83.0**	**74.5**	**52.2**	**44.3**	**58.0**	**50.0**	**69.9**	**60.5**	**61.5**	**68.2**	**65.6**
Supervision: Zero-shot (full)											
LLM-Grounder [49]	-	-	-	-	17.1	5.3	-	-	-	-	-
ZSVG3D [55]	63.8	58.4	27.7	24.6	36.4	32.7	46.5	31.7	36.8	40.0	39.0
SeeGround [25]	75.7	68.9	34.0	30.0	44.1	39.4	54.5	38.3	42.3	48.2	46.1
CSVG [52]	68.8	61.2	38.4	27.3	49.6	39.8	**67.1**	51.3	53.0	62.5	59.2
Ours	**81.2**	**72.9**	**50.8**	**43.9**	**58.1**	**50.9**	65.7	**57.0**	**56.7**	**63.7**	**61.2**

Situation Estimation. In Tab. 2, the results on SQA3D [29] demonstrate that our method achieves the best localization performance over existing methods, with 20.7% Acc@0.5m and 48.9% Acc@1.0m. Owing to its flexible viewpoint planning, our method can effectively map situation descriptions to accurate spatial locations and orientations, thereby enhancing spatial understanding.

3D-assisted Dialog. The results in Tab. 3 illustrate that our method establishes a new state-of-the-art zero-shot performance on 3D-LLM Held-In dataset [15], surpassing fully supervised method 3D-LLM [15] and zero-shot method Agent3D-Zero [56] by 8.4 and 14.6 BLEU-1, respectively.

Task Decomposition. As demonstrated in Tab. 3, our method outperforms existing zero-shot methods in most evaluation metrics. Compared with zero-shot method Agent3D-Zero [56], our method achieves improvements of 1.8 BLEU-1, 9.4 BLEU-4, and 3.7 ROUGE on the 3D-LLM Held-In dataset [15].

3D Visual Grounding. As shown in Tab. 4, compared with the previous state-of-the-art zero-shot method CSVG [52], our method achieves substantial improvements of 8.5% Acc@0.25 and 11.1% Acc@0.5 on ScanRefer [8], and 2.0% Acc@0.25 on Nr3D [1]. Moreover, even on the subset of 250 queries proposed by the VLM-Grounder [48], our method still achieves advanced performance.

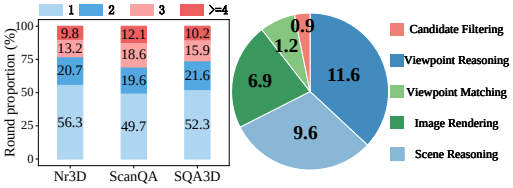


Fig. 3: Left: Distribution of interaction rounds. Right: Inference time of each step.

Table 5: Inference efficiency comparison with zero-shot methods. "*" denotes our reproduced results.

Method	Tokens	Time (s)	Acc.
VLM-Grounder	-	50.3	48.0
SeqVLM*	7.5k	42.3	53.2
Ours	**4.3k**	**30.2**	**65.6**

Table 6: Ablation on different components.

Method	Nr3D	SQA3D	ScanQA
Ours (Full)	**56.7**	**53.2**	**26.1**
w/o Holistic Cognitive Map	47.8	49.3	18.6
w/o Filtering and Feedback	49.3	50.4	20.9
w/o Rendered Image	52.7	50.6	21.0
w/o Real Image	53.9	51.7	21.4

Table 7: Performance comparison with different baseline.

Method	Nr3D	SQA3D	ScanQA
Random	13.8	12.6	6.4
Co-Visible Template	43.1	42.0	13.6
Ours	**56.7**	**53.2**	**26.1**

Notably, it also achieves competitive performance on the challenging "Multiple" subset, indicating its ability to efficiently disambiguate among multiple similar objects through unified spatial cognition.

Interaction Rounds and Inference Efficiency. As shown in the left of Fig. 3, we investigate the distribution of interaction rounds across different benchmarks. We observe that our multi-agent collaborative framework efficiently resolves approximately 50% of the queries in a single round, while only a small fraction requires four or more iterations. Meanwhile, Nr3D [1] exhibits fewer interaction rounds compared to other benchmarks. This is because Nr3D primarily involves selecting and localizing the target object, whereas ScanQA [4] and SQA3D [29] additionally require understanding more complex object semantics, spatial relationships, and intricate layouts. In the right of Fig. 3, we break down the inference time of each step within a single round, which in total takes 30.2s. The reasoning processes of the two agents take up a significant portion of the total time. For the multi-round case, the total inference time increases linearly with the number of interaction rounds. Moreover, Tab. 5 shows that our method significantly reduces the token cost from 7.5k to 4.3k and achieves higher efficiency from 41.3s to 30.2s, while still achieving strong performance on Nr3D [1]. This is because our method leverages the cognitive map to filter out a large number of erroneous candidates and plans viewpoints around the critical target, which avoids the need to sample and filter out all video frames.

4.3 Ablation Study

In this section, we adopt Acc@0.25 as the default evaluation metric for Nr3D [1], and EM for SQA3D [29] and ScanQA [4]. We employ Qwen2.5-VL-72B [6] as both Planning Agent and Perception Agent.

Effect of Holistic Cognitive Map. As illustrated in Tab. 6, removing the object information table and retaining only the BEV leads to notable performance drops of 8.9% Acc@0.25 on Nr3D [1] and 3.9% EM on SQA3D [29]. We find that without the holistic cognitive map to refine the cognition of 3D scenes, this variant is forced to repeatedly process previously observed objects, leading to redundant planning and reduced efficiency.

Effect of Filtering and Feedback. We disable both candidate filtering and feedback, which jointly sustain the collaborative process between the two agents. As demonstrated in Tab. 6, this variant results in a notable performance degradation of 7.4% Acc@0.25 on Nr3D [1]. Without candidate filtering and feedback, mismatched objects remain in the candidate map, thereby preventing Planning Agent from focusing on critical objects.

Effect of Image Sources. Our method balances visual fidelity and perspective coverage by prioritizing real images and using rendered images only when necessary. As shown in Tab. 6, without Rendered Images results in a performance drop of 5.1% EM on ScanQA [4]. This performance gap occurs because, without rendered images, this variant can only perceive the 3D scene from limited and fixed viewpoints, thereby missing critical perspectives that are crucial for spatial localization. In contrast, w/o Real Images leads to a performance drop of 4.7% EM on ScanQA [4]. This variant limits the ability of our method to capture fine-grained object semantics, such as material and texture details. These results highlight the complementary roles of rendered and real images in supporting a comprehensive understanding of the 3D scene.

Influence of MLLMs. The MLLM serves as the "brain" of our method, effectively parsing the query and performing spatial relationship reasoning. To further study how the choice of the MLLM affects overall performance, we perform a comparative analysis of several models under the same evaluation setting, including open-source backbones (Qwen2.5-VL-72B [6], Qwen2.5-VL-7B [6], and Qwen2-VL-72B [42]) and a closed-source alternative (GPT-4o [19]). As demonstrated in Tab. 8, we observe that GPT-4o exhibits stronger 3D scene understanding capabilities than open-source models. Moreover, our method still outperforms most zero-shot methods even when using smaller backbones, indicating that its effectiveness stems from explicit scene cognition, candidate filtering, and comprehensive observation, rather than relying solely on more advanced MLLMs.

Table 8: Ablation study of the MLLM on Nr3D [1]

Method	Overall
Qwen2.5-VL-7B	49.3
Qwen2-VL-72B	54.2
Qwen2.5-VL-72B	56.7
GPT-4o	**61.2**

Effect of Viewpoint Matching Threshold τ_D. As demonstrated in the left of Fig. 4, we investigate the effect of viewpoint matching threshold τ_D across different datasets. As τ_D increases, performance steadily improves, peaking at 0.75/0.8 as more real images are provided to capture fine-grained semantics. However, as τ_D further increases, performance degrades as fewer rendered views lead to missing critical perspectives for spatial localization.

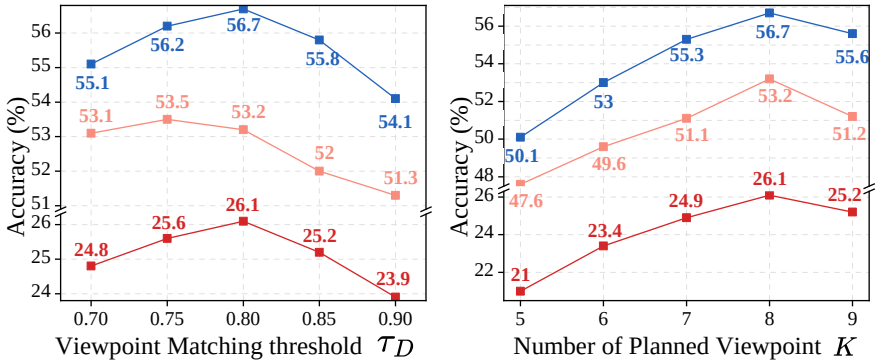


Fig. 4: Left: Influence of the viewpoint matching threshold τ_D. Right: Influence of the number of planned viewpoints K.

Effect of Number of Planned Viewpoints K. As shown in the right of Fig. 4, we analyze how varying the number of planned viewpoints K across three benchmarks. As K increases, the accuracy steadily improves, reaching its maximum at $K = 8$. This indicates that providing an appropriate number of viewpoints supplies Perception Agent with more comprehensive scene information, allowing a more reliable observation of the target object. However, when K increases from 8 to 9, the accuracy begins to drop. This is because an excessive number of planned viewpoints increases the burden on Planning Agent, resulting in lower-quality viewpoint planning. More importantly, the increased overlap among views introduces redundant information that hampers the inference of Perception Agent.

4.4 Discussion

Effectiveness of Viewpoint Planning. As mentioned in Sec. 3.2, our method actively plans observation perspectives to capture scene information. To analyze the effectiveness of such design, we further compare the following two baselines: (1) Random, which randomly plans viewpoints based on the BEV [56]. (2) Co-Visible Template, which generates fixed near and far viewpoints around candidate objects to ensure broad scene coverage. As demonstrated in Tab. 7, Random performs poorly because it may fail to reliably capture the target object. Co-Visible Template performs better than Random by increasing scene coverage. In contrast, our method leverages the cognitive map to directly observe the 3D scene and actively capture more informative viewpoints, leading to performance improvements of 42.9% and 13.6% Acc@0.25 over Random and Co-Visible Template on Nr3D [1], respectively. These results reveal an important principle: the key to 3D understanding lies in capturing decisive visual evidence from critical perspectives, rather than increasing coverage alone.

More Perception Agents across Scene Scales. As demonstrated in the left of Fig. 5, we analyze the effect of varying the number of Perception Agents across

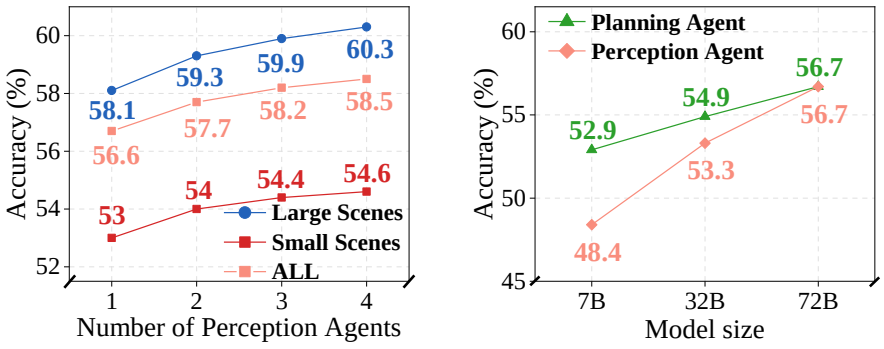


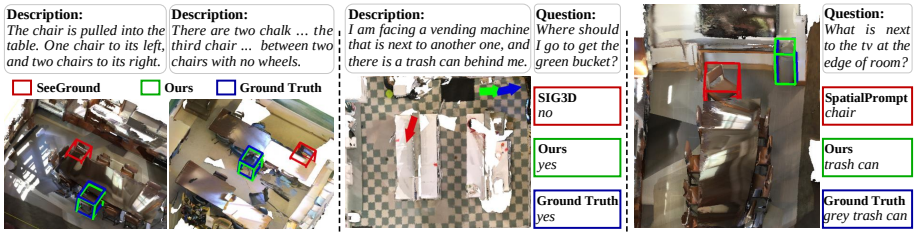
Fig. 5: Left: Influence of the number of Perception Agents on Nr3D [1]. Right: Influence of the model size of Planning Agent and Perception Agent on Nr3D [1].

different scene scales, with each Perception Agent tasked with perceiving a different region in parallel. For scene scale analysis, we divide ScanNet [11] into small scenes (average area 16.32m^2) and large scenes (average area 40.85m^2). We find that as the number of Perception Agents increases, the performance continues to improve, with larger improvements observed in large scenes compared to small ones. However, the rate of improvement gradually slows down. This is because the relatively small scene scale of ScanNet [11] limits the advantage of our multi-agent collaborative scene cognition, which is better suited for larger-scale 3D environments.

Better Input or Better Perception? Planning Agent provides query-relevant viewpoint inputs, and Perception Agent is tasked to perform scene perception. This decoupled design motivates an investigation centered around the key question: *Which matters more, better input or better perception?* As shown in the right of Fig. 5, we further conduct two symmetric experiments by keeping one agent at 72B while varying the other agent's model size. We find that better input is a promising direction that provides reliable support for subsequent perception tasks, while better perception ability further enhances the understanding of the 3D scene and leads to greater performance improvement. Moreover, the effectiveness of both improvements is largely influenced by model size, and continued advancements in MLLMs may further unlock its potential.

4.5 Qualitative Results

Fig. 6 illustrates the comprehensive qualitative results on ScanRefer [8], SQA3D [29], and ScanQA [4]. In Fig. 6a, multiple visually identical objects (chairs) appear in the 3D environment. Our method can accurately and effectively understand the spatial relationships (*e.g.*, "to its right", "between") to localize the target object, but SeeGround [25] merely relies on limited single-view localization, which severely limits its ability to resolve confusing spatial ambiguity. In Fig. 6b, the query includes egocentric situations (*e.g.*, "facing a vending machine", "behind



(a) 3D visual grounding **(b)** Situated understanding **(c)** Visual question answering

Fig. 6: Qualitative comparison of our method with previous approaches across various 3D understanding tasks on ScanRefer [8], SQA3D [29], and ScanQA [4], respectively.

me") and fine-grained semantics (*e.g.*, "green bucket"). Our explicit cognitive map enables the model to directly observe the 3D scene and understand egocentric spatial locations, and it further verify the orientation through flexible viewpoint observations. In contrast, SIG3D relies on fixed anchors and struggles to interpret complex spatial location descriptions. In Fig. 6c, SpatialPrompting [39] fails to capture the spatial layout (*e.g.*, "edge of room") because its static object-centric keyframe selection inevitably ignores critical relations among objects and overall scene layout.

5 Conclusion

In this paper, we propose a collaborative multi-agent framework for zero-shot 3D scene understanding. Unlike previous methods that rely on static object-centric keyframe selection from the scene video, we integrate a Planning Agent for flexible viewpoint planning and critical perspective completion, and a Perception Agent for explicit environment perception, enabling more comprehensive cognition of query-relevant scene information. Extensive experiments demonstrate that our approach achieves promising performance across a variety of 3D tasks, highlighting its potential as a foundation for future multi-agent systems in 3D scene understanding.

Limitations and Future Work. We follow previous works and adopt a pre-trained 3D detection model to obtain 3D bounding boxes for all scene objects, which makes the overall performance related to their raw detection quality. In future work, we will try to integrate a high-fidelity online scene reconstruction model to automatically refine detection errors and reduce this performance dependency.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (U23B2013, U2441242 and 62276176). This work was also partly supported by the SICHUAN Provincial Natural Science Foundation (No. 2024NSFJQ0023).

References

1. Achlioptas, P., Abdelreheem, A., Xia, F., Elhoseiny, M., Guibas, L.: ReferIt3D: Neural listeners for fine-grained 3D object identification in real-world scenes. In: European Conference on Computer Vision (ECCV). pp. 422–440 (2020)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems (NeurIPS) **35**, 23716–23736 (2022)
3. Arena, F., Collotta, M., Pau, G., Termine, F.: An overview of augmented reality. Computers **11**(2), 28 (2022)
4. Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: Scanqa: 3d question answering for spatial scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19129–19139 (2022)
5. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 **1**(2), 3 (2023)
6. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
7. Chang, C.P., Wang, S., Pagani, A., Stricker, D.: MiKASA: Multi-key-anchor & scene-aware transformer for 3D visual grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14131–14140 (2024)
8. Chen, D.Z., Chang, A.X., Nießner, M.: ScanRefer: 3D object localization in RGB-D scans using natural language. In: European Conference on Computer Vision (ECCV). pp. 202–221 (2020)
9. Chen, Z., Gholami, A., Nießner, M., Chang, A.X.: Scan2cap: Context-aware dense captioning in rgb-d scans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3193–3203 (2021)
10. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. Journal of Machine Learning Research (JMLR) **25**(70), 1–53 (2024)
11. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5828–5839 (2017)
12. Fan, Y., Ma, X., Su, R., Guo, J., Wu, R., Chen, X., Li, Q.: Embodied videoagent: Persistent memory from egocentric videos and embodied sensors enables dynamic scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6342–6352 (2025)
13. Fu, R., Liu, J., Chen, X., Nie, Y., Xiong, W.: Scene-llm: Extending language model for 3d visual reasoning. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2195–2206. IEEE (2025)
14. Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K.M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., et al.: Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 5021–5028. IEEE (2024)

15. Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., Gan, C.: 3d-llm: Injecting the 3d world into large language models. Advances in Neural Information Processing Systems (NeurIPS) **36**, 20482–20494 (2023)
16. Huang, H., Chen, Y., Wang, Z., Huang, R., Xu, R., Wang, T., Liu, L., Cheng, X., Zhao, Y., Pang, J., et al.: Chat-scene: Bridging 3d scene and large language models with object identifiers. Advances in Neural Information Processing Systems (NeurIPS) **37**, 113991–114017 (2024)
17. Huang, J., Yong, S., Ma, X., Linghu, X., Li, P., Wang, Y., Li, Q., Zhu, S.C., Jia, B., Huang, S.: An embodied generalist agent in 3D world. In: Proceedings of the International Conference on Machine Learning (ICML) (2024)
18. Huang, X., Wu, J., Xie, Q., Han, K.: MLLMs Need 3D-Aware Representation Supervision for Scene Understanding. arXiv preprint arXiv:2506.01946 (2025)
19. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
20. Jain, A., Gkanatsios, N., Mediratta, I., Fragkiadaki, K.: Bottom up top down detection transformers for language grounding in images and point clouds. In: European Conference on Computer Vision (ECCV). pp. 417–433. Springer (2022)
21. Jin, Z., Hayat, M., Yang, Y., Guo, Y., Lei, Y.: 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). pp. 10984–10994 (2023)
22. Jin, Z., Tu, R.C., Liao, J., Sun, W., Luo, X., Liu, S., Tao, D.: SPAZER: Spatial-Semantic progressive reasoning agent for zero-shot 3d visual grounding. arXiv preprint arXiv:2506.21924 (2025)
23. Ke, F., Cai, Z., Li, B., Chen, L., Lin, B., Wang, W., Haghighi, P.D., Haffari, G., Rezatofighi, H.: VIEW2SPACE: Studying multi-view visual reasoning from sparse observations. arXiv preprint arXiv:2603.16506 (2026)
24. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 19730–19742. PMLR (2023)
25. Li, R., Li, S., Kong, L., Yang, X., Liang, J.: SeeGround: See and ground for Zero-Shot Open-Vocabulary 3D visual grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
26. Lin, J., Bian, S., Zhu, Y., Tan, W., Zhang, Y., Xie, Y., Qu, Y.: SeqVLM: Proposal-Guided Multi-View sequences reasoning via VLM for Zero-Shot 3D visual grounding. In: Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM). pp. 3094–3103 (2025)
27. Liu, Y., Liu, W., Gu, X., Rui, Y., He, X., Zhang, Y.: Lmagent: A large-scale multimodal agents society for multi-user simulation. arXiv preprint arXiv:2412.09237 (2024)
28. Liu, Y., Kong, L., Cen, J., Chen, R., Zhang, W., Pan, L., Chen, K., Liu, Z.: Segment any point cloud sequences by distilling vision foundation models. Advances in Neural Information Processing Systems (NeurIPS) **36**, 37193–37229 (2023)
29. Ma, X., Yong, S., Zheng, Z., Li, Q., Liang, Y., Zhu, S.C., Huang, S.: SQA3D: Situated question answering in 3D scenes. In: International Conference on Learning Representations (ICLR) (2023)
30. Maggio, D., Chang, Y., Hughes, N., Trang, M., Griffith, D., Dougherty, C., Cristofalo, E., Schmid, L., Carlone, L.: Clio: Real-time task-driven open-set 3d scene graphs. IEEE Robotics and Automation Letters **9**(10), 8921–8928 (2024)

31. Man, Y., Gui, L.Y., Wang, Y.X.: Situational awareness matters in 3d vision language reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13678–13688 (2024)
32. Mascaro, R., Chli, M.: Scene representations for robotic spatial perception. Annual Review of Control, Robotics, and Autonomous Systems **8** (2024)
33. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
34. Qi, Z., Zhang, Z., Fang, Y., Wang, J., Zhao, H.: Gpt4scene: Understand 3d scenes from videos with vision-language models. arXiv preprint arXiv:2501.01428 (2025)
35. Ramakrishnan, S.K., Gokaslan, A., Wijmans, E., Maksymets, O., Clegg, A., Turner, J., Undersander, E., Galuba, W., Westbury, A., Chang, A.X., et al.: Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. arXiv preprint arXiv:2109.08238 (2021)
36. Rosinol, A., Gupta, A., Abate, M., Shi, J., Carlone, L.: 3d dynamic scene graphs: Actionable spatial perception with places, objects, and humans. arXiv preprint arXiv:2002.06289 (2020)
37. Schult, J., Engelmann, F., Hermans, A., Litany, O., Tang, S., Leibe, B.: Mask3D: Mask transformer for 3D semantic instance segmentation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 8216–8223. IEEE (2023)
38. Shen, Z., Luo, H., Chen, K., Lv, F., Li, T.: Enhancing multi-robot semantic navigation through multimodal chain-of-thought score collaboration. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 14664–14672 (2025)
39. Taguchi, S., Deguchi, H., Hamazaki, T., Sakai, H.: SpatialPrompting: Keyframe-driven Zero-Shot spatial reasoning with Off-the-Shelf Multimodal Large Language Models. arXiv preprint arXiv:2505.04911 (2025)
40. Vera, A., Sanchez, K., Hinojosa, C., Hamid, H.B., Kim, D., Ghanem, B.: Multimodal safety evaluation in generative agent social simulations. arXiv preprint arXiv:2510.07709 (2025)
41. Wang, H., Zhao, Y., Wang, T., Fan, H., Zhang, X., Zhang, Z.: Ross3d: Reconstructive visual instruction tuning with 3d-awareness. arXiv preprint arXiv:2504.01901 (2025)
42. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
43. Wang, X., Huang, Q., Celikyilmaz, A., Gao, J., Shen, D., Wang, Y.F., Wang, W.Y., Zhang, L.: Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6629–6638 (2019)
44. Wu, D., Liu, F., Hung, Y.H., Duan, Y.: Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. arXiv preprint arXiv:2505.23747 (2025)
45. Wu, S.C., Wald, J., Tateno, K., Navab, N., Tombari, F.: Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7515–7525 (2021)
46. Xia, F., Zamir, A.R., He, Z., Sax, A., Malik, J., Savarese, S.: Gibson env: Real-world perception for embodied agents. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9068–9079 (2018)
47. Xie, Q., Liang, Z., Zeng, L.: DSM: Building a diverse semantic map for 3D visual grounding. arXiv preprint arXiv:2504.08307 (2025)

48. Xu, R., Huang, Z., Wang, T., Chen, Y., Pang, J., Lin, D.: VLM-Grounder: A VLM agent for zero-shot 3D visual grounding. In: Conference on Robot Learning (CoRL) (2024)
49. Yang, J., Chen, X., Qian, S., Madaan, N., Iyengar, M., Fouhey, D.F., Chai, J.: LLM-Grounder: Open-vocabulary 3D visual grounding with large language model as an agent. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 7694–7701. IEEE (2024)
50. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10632–10643 (2025)
51. Yang, Y., Hayat, M., Jin, Z., Zhu, H., Lei, Y.: Zero-shot point cloud segmentation by semantic-visual aware synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11586–11596 (2023)
52. Yuan, Q., Li, K., Zhang, J.: Solving zero-shot 3d visual grounding as constraint satisfaction problems. arXiv preprint arXiv:2411.14594 (2024)
53. Yuan, Z., Liu, Y., Cao, Y., Sun, W., Jia, H., Chen, R., Li, Z., Lin, B., Yuan, L., He, L., et al.: Mora: Enabling generalist video generation via a multi-agent framework. arXiv preprint arXiv:2403.13248 (2024)
54. Yuan, Z., Peng, Y., Ren, J., Liao, Y., Han, Y., Feng, C.M., Zhao, H., Li, G., Cui, S., Li, Z.: Empowering large language models with 3d situation awareness. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19435–19445 (2025)
55. Yuan, Z., Ren, J., Feng, C.M., Zhao, H., Cui, S., Li, Z.: Visual programming for zero-shot open-vocabulary 3D visual grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20623–20633 (2024)
56. Zhang, S., Huang, D., Deng, J., Tang, S., Ouyang, W., He, T., Zhang, Y.: Agent3d-zero: An agent for zero-shot 3d understanding. In: European Conference on Computer Vision (ECCV). pp. 186–202. Springer (2024)
57. Zhang, Y., Luo, H., Lei, Y.: Towards CLIP-driven language-free 3D visual grounding via 2D-3D relational enhancement and consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13063–13072 (2024)
58. Zhao, L., Cai, D., Sheng, L., Xu, D.: 3DVG-Transformer: Relation modeling for visual grounding on point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2928–2937 (2021)
59. Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., Wang, J.: Fast segment anything. arXiv preprint arXiv:2306.12156 (2023)
60. Zheng, D., Huang, S., Wang, L.: Video-3d llm: Learning position-aware video representation for 3d scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8995–9006 (2025)
61. Zhi, H., Chen, P., Li, J., Ma, S., Sun, X., Xiang, T., Lei, Y., Tan, M., Gan, C.: Lscenellm: Enhancing large 3d scene understanding using adaptive visual preferences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3761–3771 (2025)
62. Zhou, Y., He, Y., Su, Y., Han, S., Jang, J., Bertasius, G., Bansal, M., Yao, H.: ReAgent-V: A Reward-Driven Multi-Agent framework for video understanding. arXiv preprint arXiv:2506.01300 (2025)

63. Zhu, Z., Ma, X., Chen, Y., Deng, Z., Huang, S., Li, Q.: 3d-vista: Pre-trained transformer for 3d vision and text alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2911–2921 (2023)
64. Zhu, Z., Wang, X., Li, Y., Zhang, Z., Ma, X., Chen, Y., Jia, B., Liang, W., Yu, Q., Deng, Z., et al.: Move to understand a 3d scene: Bridging visual grounding and exploration for efficient and versatile embodied navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8120–8132 (2025)