Agent3D-Zero: An Agent for Zero-shot 3D Understanding

Sha Zhang[1,2], Di Huang[3], Jiajun Deng[4]✉, Shixiang Tang[5], Wanli Ouyang[2,5], Tong He[2]✉, and Yanyong Zhang[1,6] ✉

[1] University of Science and Technology of China, Hefei, China
[2] Shanghai AI lab, Shanghai, China
[3] The University of Sydney, Sydney, Australia
[4] The University of Adelaide, Adelaide, Australia
[5] The Chinese University of Hong Kong, Hong Kong, China
[6] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China

Abstract. The ability to understand and reason the 3D real world is a crucial milestone towards artificial general intelligence. The current common practice is to finetune Large Language Models (LLMs) with 3D data and texts to enable 3D understanding. Despite their effectiveness, these approaches are inherently limited by the scale and diversity of the available 3D data. Alternatively, in this work, we introduce **Agent3D-Zero**, an innovative 3D-aware agent framework addressing the 3D scene understanding in a zero-shot manner. The essence of our approach centers on reconceptualizing the challenge of 3D scene perception as a process of understanding and synthesizing insights from multiple images, inspired by how our human beings attempt to understand 3D scenes. By consolidating this idea, we propose a novel way to make use of a Large Visual Language Model (VLM) via *actively* selecting and analyzing a series of viewpoints for 3D understanding. Specifically, given an input 3D scene, **Agent3D-Zero** first processes a bird's-eye view image with custom-designed visual prompts, then iteratively chooses the next viewpoints to observe and summarize the underlying knowledge. A distinctive advantage of **Agent3D-Zero** is the introduction of novel visual prompts, which significantly unleash the VLMs' ability to identify the most informative viewpoints and thus facilitate observing 3D scenes. Extensive experiments demonstrate the effectiveness of the proposed framework in understanding diverse and previously unseen 3D environments. project page

Keywords: 3D Scene Understanding · Agent · Multi-Modal Large Language Model

1 Introduction

Understanding three-dimensional (3D) scenes [36] is a fundamental task in computer vision, especially vital for robotics [5], autonomous driving [17,21,53], and

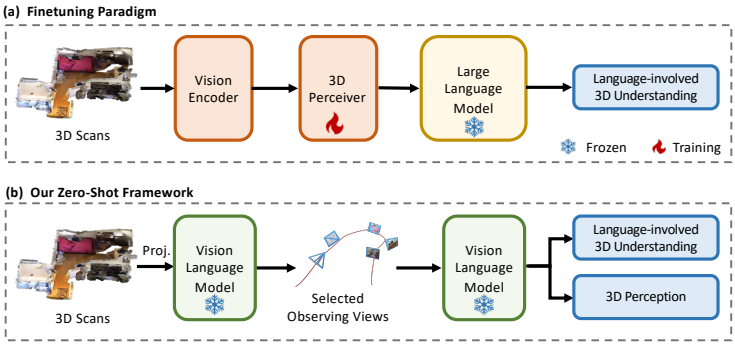


Fig. 1: Illustration of (a) finetuning-based paradigm and (b) our proposed zero-shot paradigm. The finetuning-based paradigm exploits an external 3D perceiver, and finetunes it with a frozen LLM. On the contrary, our proposed zero-shot paradigm is simple and efficient, directly utilizing the VLM to actively select and interpret multiple observing views for zero-shot 3D task resolution.

augmented reality applications [26]. The emergence of an intelligent assistant who can fully comprehend the 3D world and find her way seamlessly in a given space is a crucial milestone on the path to artificial general intelligence.

Recently, the convergence of visual perception with Large Language Models (LLMs) [1,34] represents a significant leap forward, showcasing remarkable proficiency in a variety of 2D understanding tasks. Building on this advancement, extending these capabilities to the 3D realm seems a natural progression. This involves integrating 3D data into LLMs or Large Vision Language Models (VLMs) through fine-tuning [22,32] and enabling the models to process 3D data formats directly. By leveraging their extensive prior knowledge, these LLMs/VLMs are poised to significantly enhance 3D understanding in open-world scenarios [3].

However, collecting a large quantity of 3D data is a rather challenging task, which requires specialized equipment like depth cameras [23] or LiDAR sensors [28], along with sophisticated reconstruction algorithms [14,48]. Annotating the 3D data with textual descriptions can be much more difficult and labor-intensive compared to the 2D counterpart. Moreover, the diversity of publicly available 3D data is severely limited, with existing datasets often confined to CAD models [24,47], indoor environments [15], and autonomous driving scenarios [6,19]. Such issues motivate us to rethink the potential solutions toward 3D understanding with large foundation models.

Formally, in this work, we explore an alternative approach for developing a 3D-aware intelligent assistant. Contrary to the common practice of fine-tuning LLMs/VLMs on 3D data and text pairs, we introduce **Agent3D-Zero**, an agent framework for VLMs addressing the 3D scene understanding in a zero-shot manner. Our approach draws inspiration from the human cognitive ability to comprehend the real world without the need for explicit 3D reconstruction, but observe the 3D scenario from multiviews. Humans, for example, can intuitively under-

stand the spatial relationships among objects through visual perception, even without precise measurements of distance.

By consolidating this idea, **Agent3D-Zero** utilizes multiple images from diverse viewpoints, enabling VLMs, such as GPT-4V [37], to perform robust reasoning and achieve a reliable understanding of spatial relationships between objects. This approach allows **Agent3D-Zero** to perceive the 3D world using the extensive knowledge embedded in pre-trained VLMs, thereby achieving zero-shot scene understanding.

Furthermore, **Agent3D-Zero** is designed to actively select subsequent viewpoints for observing and reasoning about spatial information across multiple images. While introducing multiple viewpoints enriches scene comprehension, it simultaneously imposes substantial memory and processing demands on current VLMs. To address this, we propose a novel visual prompting technique termed *Set-of-Line Prompting (SoLP)*. By employing bird's-eye view images, we delineate the scene's boundaries and superimpose a Cartesian coordinate system equipped with uniform grid lines and directional markers. This straightforward yet effective strategy significantly enhances the VLM's capability to understand 3D spatial concepts.

We demonstrate the effectiveness of **Agent3D-Zero** in various 3D reasoning and perception tasks. In the 3D Question Answering task, our approach surpasses all related works in the ScanQA dataset, even without the use of annotated data for training or fine-tuning. Additionally, as a zero-shot method, **Agent3D-Zero** gets advantages over the previous fine-tuning method in task decomposition and 3D-assisted dialog tasks. Our contributions can be summarized as follows:

- We pioneer the design of **Agent3D-Zero**, a 3D-aware Agent for scene understanding, which excels in zero-shot learning using only images, thereby eliminating the dependence on explicit 3D data structures such as point clouds or meshes.
- We develop a comprehensive framework for proactive perception in agents, enabling the VLM to identify location and direction through inherent reasoning capabilities.
- **Agent3D-Zero** demonstrates exceptional performance across a range of tasks and multitasking scenarios, outperforming existing methodologies in experiments involving the ScanQA dataset, 3D-assisted dialogue, and zero-shot 3D segmentation.

2 Related Work

LLMs have shown remarkable potential in advancing artificial intelligence, demonstrating their effectiveness in various assessments [12,13,38,52]. Initially limited to processing textual information, there has been a significant shift towards Multi-modal Large Language Models (MLLMs) to overcome these constraints by integrating multi-modal inputs. This discussion outlines the transition from 2D to 3D scene understanding within MLLMs, providing a foundation for exploring advancements in multimodal learning.

2.1 MLLM for 2D Scene Understanding

The application of MLLMs for 2D scene understanding [1,9–11,20,27,27,30,31,34,37,40] forms a critical foundation for our method, which utilizes VLMs to comprehend the 3D world in a zero-shot manner. As a precursor to our approach, insights into the operational mechanisms of VLMs offer valuable lessons. To establish a foundational model capable of interpreting 2D vision, researchers have pursued two distinct paradigms: training from scratch using internet-scale text-image pairs, exemplified by models such as CLIP [40] by OpenAI, BLIP series [30,31] by Facebook AI, and ViLT [27] by Google; and fine-tuning existing LLMs with additional 2D vision data, as seen in models like GPT-4V [37] and LLava [34]. The former approach fosters a deeper, more intrinsic understanding of visual-textual relationships, whereas the latter is more cost-effective and often yields superior general performance. Leveraging these insights from LLMs and VLMs, the research community has extensively explored various applications and achievements of VLMs in 2D scene understanding, including image captioning [29–31], visual question answering [2,43], semantic segmentation [42,49], and zero-shot classification tasks [35,46].

2.2 MLLM for 3D Scene Understanding

Extending LLMs to 3D scene understanding marks a significant shift beyond traditional 2D analyses, embracing the complexity of our three-dimensional environment. Initial efforts focused on recognizing single 3D objects [46]. Meanwhile, Models like Chat-3D [45] and SpatialVLM [7] enhance spatial understanding in a single image through innovative techniques. Recent advancements [7,22,25,32,45] have adapted LLMs and VLMs for more comprehensive 3D spatial comprehension. 3D-LLM [22] pioneers the incorporation of 3D worlds into LLMs by training perceivers to interpret reconstructed 3D features. Similarly, 3DMIT [32] focuses on refining the integration of 3D spatial data into LLMs through dedicated scene and object projectors. However, these approaches are limited by the unavailability of 3D scene data, restricting the scope of 3D understanding compared to the extensive resources available for 2D analysis. Our method diverges by using VLMs to perform 3D tasks in a zero-shot manner, getting rid of the labor-intensive 3D data collection and giving an elegant receipt of interpreting and understanding complicated 3D scenes with actively selected 2D inputs.

3 Method

3.1 Overview

Agent3D-Zero presents a novel agent framework that utilizes VLMs for 3D scene understanding in a zero-shot manner. An overview is illustrated in Figure 2. The process initiates with a bird's-eye view (BEV) image I_b derived from a given 3D mesh M. The selection of I_b serves a critical purpose: it provides a comprehensive layout of the scene, which is instrumental in planning the camera viewpoints that

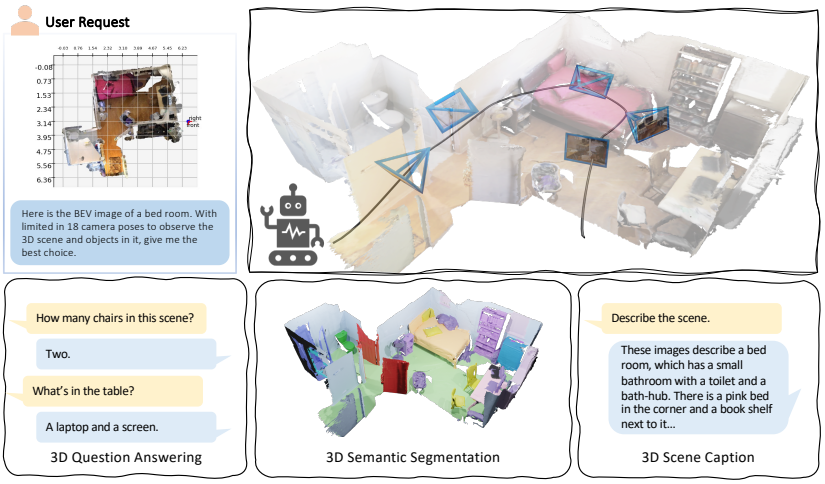


Fig. 2: System Overview of Agent3D-Zero. The upper segment illustrates our viewpoints-selection progress. We initiate the process by overlaying grid lines and tick marks on the BEV images, constituting the prompt along with a scene type description. This prompt guides the VLM to retrieve camera poses for images observing the 3D scene. The lower section demonstrates the versatility of Agent3D-Zero, showcasing its proficiency in addressing various 3D reasoning and perception tasks through strategic prompting and tool utilization.

are crucial for a thorough scene analysis. Given I_b and the camera intrinsic metric K, we aim to strategically plan N camera viewpoints. These are defined by the extrinsic matrix $\mathcal{T} = \{(R, t)_i | i \in [1, N]\}$, with R denoting a 3x3 rotation matrix for the camera's orientation and t a translation vector [x, y, z] for its position. The x and y are predicted by the VLM, while z is set as the average height. For the rotation matrix, we set pitch and roll angles as 0, and discretize azimuth to four directions $[0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}]$. This is formalized as:

$$\mathcal{T} = \mathbf{VLM}(I_b, P_b), \tag{1}$$

where P represents the textual prompt guiding the VLM. The structured prompt communicates the goal and specific requirements, such as:

> Given a bird's-eye view of a scene, please provide N pictures to comprehensively understand the scene...
> Could you suggest camera positions and orientations for each shot?

Upon determining the camera extrinsic matrix $\mathcal{T}$, new images $\mathcal{I} = \{I_i | i \in [1, N]\}$ are rendered from the 3D mesh:

$$I_i = \pi(R_i, t_i, K, M), \tag{2}$$

where π represents the rendering process. Subsequently, VLMs analyze images from these selected viewpoints, synthesizing insights for a coherent scene inter-

pretation. This empowers the agent with a detailed understanding of the scene, enabling it to tackle diverse scene-understanding tasks. With a task-specific prompt P_t, Agent3D-Zero can query the VLM to address specific questions:

$$A_t = \mathbf{VLM}(\mathcal{I}, P_t), \tag{3}$$

where A symbolizes the answer to the question.

In practice, instead of directly asking the model to output N camera positions, we make the VLM iteratively output N' each time, enabling the VLM to select the viewpoint with previous experiences.

Regrettably, when relying on raw BEV image inputs, the VLM struggles to generate meaningful viewpoints due to inherent limitations in distance measurement. To address this challenge, we propose a visual prompt, the Set-of-Line Prompting (SoLP) technique based on BEV images. Subsequent section 3.2 elaborates on the concept and application of SoLP. This is followed by a detailed exploration of Agent3D-Zero's capabilities in zero-shot 3D scene understanding, focusing on its reasoning (section 3.3) and perception (section 3.4) skills.

3.2 Set-of-Line Prompting

Recognizing the challenge of precise location determination faced by VLMs, we introduce SoLP as an innovative solution to improve the VLMs' understanding of mathematical and spatial concepts. Traditional prompt engineering focuses on textural inputs for LLMs. However, our approach aims to adopt similar methodologies for visual inputs, enhancing the VLMs' planning and localization capabilities. Given that VLMs primarily produce textual outputs, the auxiliary visual prompts added to input images must be both interpretable by the models and describable in textual form.

Drawing inspiration from how humans interpret maps, where longitude and latitude are used to specify locations on a spherical surface and the Cartesian coordinate system represents points in a plane, we devised SoLP. This method involves superimposing grid lines and tick marks onto a BEV image, transforming the original image I_b into a prompted image I_b^p. Consequently, the process described in Equation 1 and 3 is refined as shown in Equation 4 and 5:

$$\mathcal{T} = \mathbf{VLM}(I_b^p, P_b^p), \tag{4}$$

$$A_t = \mathbf{VLM}(\mathcal{I}, P_t^p), \tag{5}$$

where P_t^p represents the text prompt paired with the proposed SoLP.

SoLP not only aids in enhancing the VLM's comprehension of the scene's geometric aspects but also enables the generation of more precise camera poses. While directly determining $\mathcal{T}$ poses a significant challenge for VLMs, the introduction of SoLP facilitates this process by incorporating additional output format controls. This allows for the specification of camera positions as grid points within the image (e.g., (0, 0)) and orientations from a predefined set of

directions ['left', 'right', 'front', 'back'], simplifying the VLM's task. Compared with P_t, P_t^p has additional format control sentences:

> % Output format control
> The position can be present as the grid point in the picture, like (0, 0). The orientations can be chosen from ['left', 'right', 'front', 'back'].

Upon acquiring the camera poses $\mathcal{T}$, we can render the corresponding images $\mathcal{I} = \{I_i | i \in [1, N]\}$ within the 3D scene, setting the stage for subsequent analyses and evaluations.

3.3 Exploration of 3D Reasoning Capabilities

Upon processing the scene images, Agent3D-Zero empowers VLMs to assimilate comprehensive information about the scene, thereby equipping the model to tackle a wide array of downstream tasks, including question answering (QA), caption generation, and dialogues. As indicated by the framework outlined in Equation 3, Agent3D-Zero adapts to diverse tasks by employing specific task-oriented prompts P_t. An illustrative prompt P_t is provided below to demonstrate how Agent3D-Zero can guide VLMs in understanding a 3D scene through images from varied viewpoints and subsequently respond to a series of queries:

> % Task prompt
> Understand a 3D scene without direct access to the point clouds but only images from different viewpoints. Later, I'll ask you a series of questions about the scene, and I'd like your responses one-by-one with correspondence number, in the order the questions are presented. Please keep each response short and clear.
> Examples: questions: [1. How many chairs are around the table? 2. what's the color of the table? 3. Where is the beige wooden working table placed? 4. What is in the corner of the bath?].
> Answers: [1. 3 2. Brown 3. right of tall cabinet 4. shower]

Similar to a 3D-LLM [22], Agent3D-Zero boasts the capability to execute various downstream tasks using a singular model framework. This is a departure from most prior approaches, which typically specialize in a single aspect of 3D reasoning. This highlights the general applicability and adaptability of Agent3D-Zero, showcasing its potential to serve as a versatile tool for 3D scene analysis and understanding.

3.4 Exploration of 3D Perception Capabilities

Agent3D-Zero distinguishes itself not only through its proficiency in language-centric 3D tasks but also by adeptly handling conventional perception tasks, such as 3D semantic segmentation. Although Agent3D-Zero does not possess inherent perception abilities, it functions as an agent that effectively utilizes a variety of

vision tools to enhance its recognition capabilities. We define the tool function as f, the per-view perception results as Res_i. This process can be formalized as:

$$Res_i = \mathbf{VLM}(I_i, P_f, f), \tag{6}$$

Taking the 3D semantic segmentation task as an illustrative example, our approach unfolds in two primary steps. The first step involves performing 2D semantic segmentation on each image selected for analysis. Second, with the aid of depth information, the 2D segmented results are back-projected into 3D point clouds to construct a comprehensive 3D semantic map.

Inspired by the Set-of-Mark (SoM) method [50], Agent3D-Zero utilizes a Segment Anything Model (SAM) to segment the imagery into distinct regions initially without semantic labels. This segmentation facilitates the arrangement of the regions within each image, which, in turn, aids the VLM in assigning accurate semantic labels to each region. By annotating every selected image with semantic labels and then employing backprojection and concatenation techniques, Agent3D-Zero successfully accomplishes 3D semantic segmentation.

4 Experiment

4.1 Datasets and Metrics

We conduct experiments on the ScanQA dataset [3] and Scannet v2 dataset [15] to evaluate the performance of 3DQA and semantic segmentation, respectively. Moreover, we follow the practice of 3D-LLM [22] to separate the scene dataset to evaluate the other 3D reasoning tasks.

ScanQA [3]: This dataset comprises over 40,000 human-annotated question-and-answer pairs, with each pair grounded within the objects of 800 indoor 3D scenes from the ScanNet dataset. Our experiments are conducted exclusively on the evaluation and test subsets of this dataset.

ScanNet v2 [15]: ScanNet v2 is an extensive collection of 1,513 indoor scenes, meticulously annotated and equipped with multi-view RGB-D images alongside reconstructed meshes. For the task of semantic segmentation, this dataset offers 20 distinct classes of annotated 3D objects.

3D-LLM held-in dataset [22]: To compare our method with the most related work (i.e. 3D-LLM [22]), we follow it to evaluate the performance in 3D-assisted dialogue and task decomposition. This dataset contains 300k 3D-language pairs conducted by the data-generation pipeline in 3D-LLM, which is based on Scannet [15], Habitat-Matterport [41], and Objaverse [16].

Evaulation metrics: We employ a suite of metrics to quantitatively evaluate the performance on language-related 3D reasoning tasks. These include BLEU [39], ROUGE-L [33], METEOR [4], and CIDEr [44], which collectively assess the quality of generated textual responses. METEOR, ROUGE-L, and CIDEr are designed to capture meaning and semantic coherence, often rewarding answers that are thematically consistent and informative, even if they diverge in exact wording. On the other hand, BLEU and EM metrics focus heavily on

Table 1: Performance comparison on the ScanQA validation set. 'Two-stage' means the models use explicit object representations. 'Fine-tune' means extra training. Our proposed Agent3D-Zero is training-free.

		B-1	B-4	METEOR	ROUGE-L	CIDEr	EM
Two-stage	VoteNet+MCAN	28.0	6.2	11.4	29.8	54.7	17.3
	ScanRefer+MCAN	26.9	7.9	11.5	30	55.4	18.6
	ScanQA	30.2	10.1	13.1	33.3	64.9	**21.0**
Fine-tune	flamingo-SingleImage	23.8	8.5	10.7	29.6	52	16.9
	flamingo-MultiView	25.6	8.4	11.3	31.1	55	18.8
	BLIP2-flant5-SingleImage	28.6	5.1	10.6	25.8	42.6	13.3
	BLIP2-flant5-MultiView	29.7	5.9	11.3	26.6	45.7	13.6
	3D-LLM (flamingo)	30.3	7.2	12.2	32.3	59.2	20.4
	3D-LLM (BLIP2-opt)	35.9	9.4	13.8	34.0	63.8	19.3
	3D-LLM (BLIP2-flant5)	**39.3**	**12.0**	14.5	35.7	69.4	20.5
Zero-Shot	LLaVA-SingleImage	7.1	0.3	10.5	12.3	5.7	0.0
	Agent3D-Zero (random)	16.4	2.1	12.2	26.9	40.0	4.9
	Agent3D-Zero (selected)	28.6	4.4	**16.0**	**37.0**	**71.8**	17.5

Table 2: Performance comparison on the ScanQA test set. B-1, B-4 denote BLEU-1, BLEU-4 respectively. Our model outperforms all related models and the baseline model for evaluation metrics METEOR, ROUGE-L, and CIDEr.

		B-1	B-4	METEOR	ROUGE-L	CIDEr	EM
Two-stage	SingleImage+MCAN	16.5	0.0	8.4	21.5	38.6	15.8
	VoteNet+MCAN*	29.5	6.0	12.0	30.9	58.2	19.7
	ScanRefer+MCAN*	27.9	7.5	11.9	30.7	57.4	20.6
	ScanQA*	31.6	**12.0**	13.5	34.3	67.3	**23.5**
Fine-tune	3D-LLM (flamingo)	32.6	8.4	13.5	34.8	65.6	23.2
	3D-LLM (BLIP2-opt)	37.3	10.7	14.3	34.5	67.1	19.1
	3D-LLM (BLIP2-flant5)	**38.3**	11.6	14.9	35.3	69.6	19.1
Zero-Shot	Agent3D-Zero	31.4	5.1	**16.9**	**39.3**	**77.5**	21.3

n-gram precision, rewarding exact matches between the predicted and reference sequences. For 3D Semantic Segmentation, Mean IoU (MIoU) serves as our primary metric, offering a comprehensive measure of segmentation accuracy.

4.2 Compared Methods

To evaluate the effectiveness of Agent3D-Zero, we compare it with existing methods across several dimensions: Two-stage, Finetune, and Zero-Shot types.

Two-stage methods: These approaches, including combinations like VoteNet [18]+MCAN [51], ScanRefer [8]+MCAN [51], and ScanQA [3], first recognize objects in the 3D scene, then integrate language information to address Scan Question Answering tasks.

Finetune methods: Following the fine-tuning paradigm [22], this category includes models like flamingo [1] and BLIP2-flant5 [13,30]. They train a perceiver to adapt LLM/VLMs to comprehend reconstructed 3D representations through a three-step process: encoding in 2D/3D, fine-tuning perceivers to align features

with VLMs, and producing language-based answers. The inputs for these perceivers vary, being single-image, multi-view, or reconstructed 3D features.
Zero-Shot methods: These methods, such as LLaVA [34] and our Agent3D-Zero, utilize VLMs to understand 3D scenes without any additional training. In all experiments, we utilize GPT4-V [37] as the VLM in Agent3D-Zero.

4.3 Zero-shot Performance

ScanQA. In the 3D-QA task, models must utilize visual data from comprehensive RGB-D indoor scans to answer textual queries about the 3D scene. Unlike conventional 2D-QA, models face challenges in spatial understanding and object identification from textual descriptions in 3D contexts.

Evaluation results (Table 1) reveal that, despite its zero-shot operation, Agent3D-Zero demonstrates competitive performance against other methods. Notably, with iteratively selected image inputs, Agent3D-Zero outperforms previous benchmarks with scores of 16.0 vs 14.5 in METEOR, 37.0 vs 35.7 in ROUGE-L, and 71.8 vs 69.4 in CIDEr. However, it does not lead in EM or BLEU metrics, which prioritize n-gram precision. This discrepancy highlights the limitation of these metrics in evaluating the nuanced understanding and flexibility of zero-shot methods. Metrics like METEOR, ROUGE-L, and CIDEr are better suited for appreciating diverse expressions and capturing the essence of responses, as they do not solely reward exact matches.

Furthermore, Table 2 highlights Agent3D-Zero's superior performance in most metrics compared to previous methods (77.5 in CIDEr, 39.3 in ROUGE-L, and 16.8 in METEOR), indicating its robust capability in 3D QA.
3D-assisted dialog. 3D-assisted dialog systems incorporate spatial awareness to enhance conversational interactions. According to results (Table 3), Agent3D-Zero matches or exceeds former methods in key metrics like ROUGE-L and METEOR, reaffirming its capability in nuanced conversational contexts. BLEU metrics are discussed in the previous ScanQA analysis.
3D scene caption. 3D captioning requires a holistic scene understanding, extending beyond mere image analysis to comprehend spatial arrangements and object interactions. As the validation set for 3d scene captioning in the 3D-LLM held-in dataset only focuses on single 3d objects, we randomly select 51 scenes from the train set in the 3D-LLM held-in dataset to evaluate Agent3D-Zero's 3d scene caption ability. The results are presented in Table 3. **Agent3D-Zero**'s performance is comparable to the baseline method, indicating its potential to generate descriptive captions of scenes.
Task decomposition. Task decomposition involves breaking down a task into subtasks based on the 3D scene's spatial relationships. **Agent3D-Zero** shows competitive performance in zero-shot style (Table 3), underscoring its ability to leverage spatial information effectively.
3D semantic segmentation. 3D semantic segmentation, a cornerstone for understanding environments, involves assigning semantic labels to each 3D point in a scene. Our investigation employs VLMs for zero-shot segmentation, showcasing the potential of VLMs to navigate this task without extensive labeled data.

Table 3: Performance comparison on the Held-In Dataset, which is introduced in 3D-LLM [22]. Our zero-shot method outperforms related methods.

Tasks	Models	BLEU-1	BLEU-4	METEOR	ROUGE-L
3D-assisted Dialog	flant5	27.4	8.7	9.5	27.5
	flamingo-SingleImage	29.4	9.4	10.0	26.8
	flamingo-MultiView	30.6	9.1	10.4	27.9
	BLIP2-flant5-SingleImage	28.4	9.1	10.2	27.4
	BLIP2-flant5-MultiView	32.4	9.5	11.0	29.5
	3D-LLM (flamingo)	35.0	10.6	16.0	34.2
	3D-LLM (BLIP2-opt)	**39.6**	16.2	18.4	38.6
	3D-LLM (BLIP2-flant5)	39.0	**16.6**	18.9	**39.3**
	Agent3D-Zero (random)	26.9	7.1	17.2	30.9
	Agent3D-Zero (selected)	32.8	9.8	**19.3**	**39.3**
Task Decomposition	flant5	25.5	6.0	13.9	28.4
	flamingo-SingleImage	31.4	7.1	15.6	30.6
	flamingo-MultiView	33.1	7.3	16.1	33.2
	BLIP2-flant5-SingleImage	32.2	6.9	15.0	31.0
	BLIP2-flant5-MultiView	33.1	6.9	15.5	34.0
	3D-LLM (flamingo)	32.9	6.4	16.0	33.5
	3D-LLM (BLIP2-opt)	34.1	7.6	16.5	35.4
	3D-LLM (BLIP2-flant5)	33.9	7.4	15.9	37.8
	Agent3D-Zero (random)	33.8	6.7	16.7	36.6
	Agent3D-Zero (selected)	**42.0**	**15.5**	**22.9**	**45.1**
3D Captioning	Agent3D-Zero (random)	26.1	1.0	13.9	14.3
	Agent3D-Zero (selected)	**29.5**	**7.2**	**15.9**	**16.1**

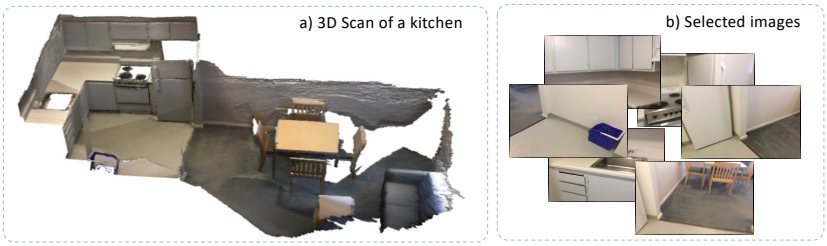
Our approach yielded a mIoU of 5.6 using randomly selected images. By iteratively refining viewpoint selection, we improved the mIoU to 8.7, highlighting the importance of strategic view selection in enhancing zero-shot segmentation performance.

While our zero-shot method does not yet rival the accuracy of traditional 3D semantic segmentation techniques, it illustrates the untapped potential of leveraging VLMs for complex 3D perception challenges. This work not only demonstrates the adaptability of language models to perception tasks but also opens new research avenues in integrating linguistic and visual understanding.

4.4 Qualitative Results

This section showcases two qualitative examples that highlight the efficacy and versatility of the proposed Agent3D-Zero in addressing complex scenarios.

Qualitative result for 3D reasoning. Our first case study focuses on ScanQA, as illustrated in Figure 3. This example vividly demonstrates Agent3D-Zero's ability to accurately identify and describe objects and their relationships within 3D environments. Through intelligent scene analysis based on selected images, Agent3D-Zero excels in synthesizing information from multiple objects to provide precise 3D scene captions, highlighting its advanced 3D reasoning capabilities. Additionally, Agent3D-Zero can decompose tasks in specific environments with minimal supplementary information. Further details and results are available in the supplementary materials.



3D Scene Caption: The scene is a compact and utilitarian kitchen space. The kitchen is equipped with essential appliances including a white refrigerator with a top freezer compartment, and a white electric range oven with coil burners and control knobs at the front. There is a corner kitchen counter that provides a modest amount of workspace, flanked on one side by the range and on the other by a stainless steel sink with a faucet. Above the counter, there are white cabinets offering storage for kitchen essentials. One cabinet has a built-in microwave nook above the range. The adjacent wall has a set of drawers and additional under-counter storage. Flooring transitions are visible from carpet to a more wipe-clean surface. Overall, the kitchen is modest but seems well-organized.

Task Decomposition: prepare a meal in the kitchen

1. go to the kitchen area
2. open the kitchen cabinets
3. take out ingredients and utensils from the kitchen cabinets
4. go to the refrigerator
5. open the refrigerator
6. take out items from the refrigerator
7. go to the kitchen counter
8. wash vegetables and prepare them on counter
9. go to the sink
10. wash hands in the sink
11. go to the stove
12. turn on the stove
13. cook the meal on the stove
14. turn off the stove
15. go to the table

Fig. 3: Visualization of 3D Scene Caption and Task Decomposition of Agent3D-Zero. The top part presents the raw 3D scan and some of the images selected from different viewpoints. We show examples of 3D Scene Caption and Task Decomposition at the bottom.

Qualitative result for real-world navigation. A compelling application of Agent3D-Zero is its utility in real-world navigation tasks. In this case study, we explore a scenario involving the navigation towards a printer located in a typical office setting. Uniquely, for this experiment, we forgo the initial BEV image input, challenging Agent3D-Zero to iteratively engage with its surroundings to determine optimal viewpoints for progression. Although the office overview (depicted in the upper part of Figure 4) is not directly fed into Agent3D-Zero, it serves as a context for understanding the complexity of the navigation task, with the printer situated in the southeast direction amidst numerous obstacles.

Agent3D-Zero adeptly navigates this environment by making informed decisions at each juncture, incorporating historical data to look around and select the most promising path forward. This process exemplifies the system's capacity to not only *circumnavigate obstacles* efficiently but also to successfully identify and reach the target destination, all while recognizing essential objects like the printer. The VLM thus demonstrates a profound ability to explore and interact with an open-set world, leveraging solely image-based observations to understand and act within complex 3D scenes, as shown in Figure 4.

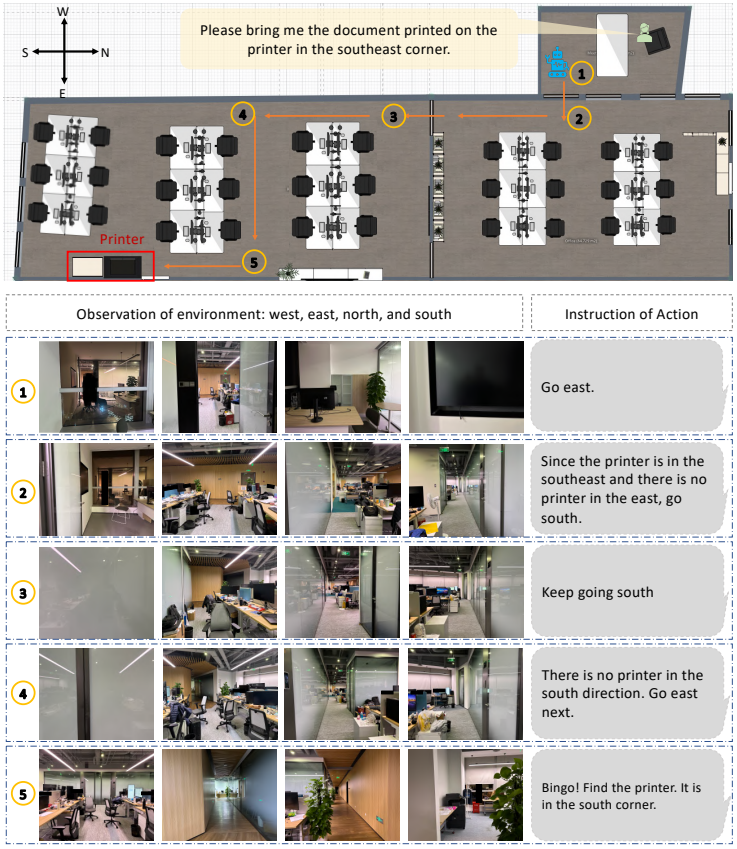


Fig. 4: Visualization of navigation in real world. The top section introduces the navigation task and provides an overview of an office setting. Subsequent rows feature observations of the environment, with GPT-4v-generated instructions on the right. The visualization concludes with the VLM successfully locating the printer, thereby accomplishing the task in an unfamiliar environment.

4.5 Ablation Study

In this section, we present an ablation study designed to evaluate the influence of varying grid line prompts on the selection of camera poses, subsequently affecting the VLMs' capacity for 3D scene understanding. Specifically, we randomly select a subset comprising 20% of the scenes from the ScanQA validation dataset as the basis for our investigation, focusing on the 3D question answering task.

Effect of viewpoints number. Our investigation into the influence of observation views on the performance of our system, herein referred to as Agent3D-Zero, is meticulously documented in Table 4. This analysis prompts the VLM to interpret 3D scenes using selections of 6, 12, and 24 images, with the choice in image count being constrained by the computational capabilities of the VLM.

Table 4: Effect of the number of viewpoints on 3D QA. This experiment is evaluated on the validation set of ScanQA dataset.

# Viewpoints	BLEU-1	METEOR	ROUHE-L	CIDEr	EM
6	17.1	12.8	28.4	50.8	13.1
12	23.3	15.0	35.3	67.9	16.9
24	**34.1**	**16.5**	**40.3**	**82.0**	**21.1**

Table 5: Effect of the density of lines in Set-of-Line prompt on 3D QA. Here, "-" indicates the GPT-4V fails to output the results.

Line density	BLEU-1	METEOR	ROUGE-L	CIDEr	EM
0x0	-	-	-	-	-
4x4	23.2	14.2	34.0	66.0	16.9
8x8	**34.1**	**16.5**	**40.3**	**82.0**	**21.1**
16x16	-	-	-	-	-

Our findings indicate a direct correlation between the number of images and the performance outcome, where an increase in image count leads to notably enhanced results. This trend underscores the potential of augmenting input imagery to significantly improve the efficacy of our proposed approach.

Effect of line density. We further investigate the role of line density within the Set-of-Line prompts utilized by Agent3D-Zero, with results detailed in Table 5. The application of line density is subject to the recognition abilities of the VLMs, imposing a practical limit to the density achievable (16x16). Despite this constraint, our results demonstrate that a higher density of auxiliary lines correlates with more precise 3D scene comprehension. The absence of dense visual prompts (0x0) significantly hampers the VLMs' ability to accurately determine camera poses, highlighting a notable limitation in the current capabilities of Agent3D-Zero for precise and mathematical pose estimation.

5 Conclusion

In this work, we introduce Agent3D-Zero, a pioneering framework that utilizes Vision-Language Models for zero-shot understanding and interaction within 3D environments. Through the strategic selection of diverse observational viewpoints and the incorporation of custom-designed visual prompts, Agent3D-Zero facilitates a nuanced and integrated perception of 3D scenes. Our experiments underscore the transformative potential of VLMs in redefining 3D scene analysis, emphasizing the effectiveness of multi-viewpoint synthesis and visual prompts in augmenting model capabilities. This advancement propels us towards the realization of intelligent systems proficient in comprehending and navigating the real world akin to human interaction.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62332016).

References

1. 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 **35**, 23716–23736 (2022)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
3. 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. pp. 19129–19139 (2022)
4. Banerjee, S., Lavie, A.: 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. pp. 65–72 (2005)
5. Cadena, C., Dick, A.R., Reid, I.D.: Multi-modal auto-encoders as joint estimators for robotics scene understanding. In: Robotics: Science and systems. vol. 5 (2016)
6. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
7. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Driess, D., Florence, P., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. arXiv preprint arXiv:2401.12168 (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. pp. 202–221. Springer (2020)
9. Chen, F.L., Zhang, D.Z., Han, M.L., Chen, X.Y., Shi, J., Xu, S., Xu, B.: Vlp: A survey on vision-language pre-training. Machine Intelligence Research **20**(1), 38–56 (2023)
10. Chen, G., Zheng, Y.D., Wang, J., Xu, J., Huang, Y., Pan, J., Wang, Y., Wang, Y., Qiao, Y., Lu, T., et al.: Videollm: Modeling video sequence with large language models. arXiv preprint arXiv:2305.13292 (2023)
11. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478 (2023)
12. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. Journal of Machine Learning Research **24**(240), 1–113 (2023)
13. 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. arXiv preprint arXiv:2210.11416 (2022)

14. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. pp. 303–312 (1996)
15. 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 conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
16. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13142–13153 (2023)
17. Deng, J., Zhang, S., Dayoub, F., Ouyang, W., Zhang, Y., Reid, I.: Poifusion: Multi-modal 3d object detection via fusion at points of interest. arXiv preprint arXiv:2403.09212 (2024)
18. Ding, Z., Han, X., Niethammer, M.: Votenet: A deep learning label fusion method for multi-atlas segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22. pp. 202–210. Springer (2019)
19. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: The kitti vision benchmark suite. URL http://www. cvlibs. net/datasets/kitti **2**(5) (2015)
20. Gong, T., Lyu, C., Zhang, S., Wang, Y., Zheng, M., Zhao, Q., Liu, K., Zhang, W., Luo, P., Chen, K.: Multimodal-gpt: A vision and language model for dialogue with humans. arXiv preprint arXiv:2305.04790 (2023)
21. Guo, Z., Huang, Y., Hu, X., Wei, H., Zhao, B.: A survey on deep learning based approaches for scene understanding in autonomous driving. Electronics **10**(4), 471 (2021)
22. 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 **36** (2024)
23. Horaud, R., Hansard, M., Evangelidis, G., Ménier, C.: An overview of depth cameras and range scanners based on time-of-flight technologies. Machine vision and applications **27**(7), 1005–1020 (2016)
24. Hua, B.S., Truong, Q.T., Tran, M.K., Pham, Q.H., Kanezaki, A., Lee, T., Chiang, H., Hsu, W., Li, B., Lu, Y., et al.: Shrec'17: Rgb-d to cad retrieval with objectnn dataset. In: Proc. Eurograph. Workshop 3D Object Retrieval. pp. 25–32 (2017)
25. Jatavallabhula, K.M., Kuwajerwala, A., Gu, Q., Omama, M., Chen, T., Maalouf, A., Li, S., Iyer, G., Saryazdi, S., Keetha, N., et al.: Conceptfusion: Open-set multimodal 3d mapping. arXiv preprint arXiv:2302.07241 (2023)
26. Kalkofen, D., Mendez, E., Schmalstieg, D.: Comprehensible visualization for augmented reality. IEEE transactions on visualization and computer graphics **15**(2), 193–204 (2008)
27. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision (2021)
28. Lemmens, M.: Airborne lidar sensors. GIM international **21**(2), 24–27 (2007)
29. Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., Ye, J., Chen, H., Xu, G., Cao, Z., et al.: mplug: Effective and efficient vision-language learning by cross-modal skip-connections. arXiv preprint arXiv:2205.12005 (2022)
30. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)

31. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
32. Li, Z., Zhang, C., Wang, X., Ren, R., Xu, Y., Ma, R., Liu, X.: 3dmit: 3d multi-modal instruction tuning for scene understanding. arXiv preprint arXiv:2401.03201 (2024)
33. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
34. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
35. Naeem, M.F., Khan, M.G.Z.A., Xian, Y., Afzal, M.Z., Stricker, D., Van Gool, L., Tombari, F.: I2mvformer: Large language model generated multi-view document supervision for zero-shot image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15169–15179 (2023)
36. Naseer, M., Khan, S., Porikli, F.: Indoor scene understanding in 2.5/3d for autonomous agents: A survey. IEEE access **7**, 1859–1887 (2018)
37. OpenAI: Gpt-4 technical report (2023)
38. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022)
39. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
41. 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)
42. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multimodal model. arXiv preprint arXiv:2311.03356 (2023)
43. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. In: European Conference on Computer Vision. pp. 146–162. Springer (2022)
44. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
45. Wang, Z., Huang, H., Zhao, Y., Zhang, Z., Zhao, Z.: Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. arXiv preprint arXiv:2308.08769 (2023)
46. Wu, W., Yao, H., Zhang, M., Song, Y., Ouyang, W., Wang, J.: Gpt4vis: What can gpt-4 do for zero-shot visual recognition? arXiv preprint arXiv:2311.15732 (2023)
47. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1912–1920 (2015)
48. Xu, X., Qiu, J., Wang, X., Wang, Z.: Relationship spatialization for depth estimation. In: European Conference on Computer Vision. pp. 615–637. Springer (2022)

49. Yang, J., Zhang, H., Li, F., Zou, X., Li, C., Gao, J.: Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. arXiv preprint arXiv:2310.11441 (2023)
50. Yang, J., Zhang, H., Li, F., Zou, X., Li, C., Gao, J.: Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. arXiv preprint arXiv:2310.11441 (2023)
51. Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: Deep modular co-attention networks for visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6281–6290 (2019)
52. Zhang, R., Han, J., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Gao, P., Qiao, Y.: Llama-adapter: Efficient fine-tuning of language models with zero-init attention. arXiv preprint arXiv:2303.16199 (2023)
53. Zhang, S., Deng, J., Bai, L., Li, H., Ouyang, W., Zhang, Y.: Hvdistill: Transferring knowledge from images to point clouds via unsupervised hybrid-view distillation. International Journal of Computer Vision pp. 1–15 (2024)