3DAXISPROMPT: PROMOTING THE 3D GROUNDING AND REASONING IN GPT-4O

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) exhibit impressive capabilities across a variety of tasks, especially when equipped with carefully designed visual prompts. However, existing studies primarily focus on logical reasoning and visual understanding, while the capability of MLLMs to operate effectively in 3D vision remains an ongoing area of exploration. In this paper, we introduce a novel visual prompting method, called 3DAxisPrompt, to elicit 3D understanding capabilities of MLLMs in real-world scenes. More specifically, our method leverages the 3D coordinate axis and masks generated from the Segment Anything Model (SAM) to provide explicit geometric priors to MLLMs and then extend their impressive 2D grounding/reasoning ability to real-world 3D scenarios. Besides, we also provide a thorough investigation of the potential visual prompting formats and conclude our findings to reveal the potential and limits of 3D understanding capabilities in GPT-4o. Finally, we build evaluation environments with four datasets, *i.e.* ShapeNet, ScanNet, FMB, and nuScene datasets, covering various 3D tasks. Based on this, we conduct extensive quantitative and qualitative experiments, which demonstrate the effectiveness of the proposed method. Overall, our study reveals that GPT-4o, with the help of 3DAxisPrompt, can effectively perceive an object's 3D position in real-world scenarios. Nevertheless, a single prompt engineering approach does not consistently achieve the best outcomes for all 3D tasks. This study highlights the feasibility of leveraging MLLMs for 3D vision grounding/reasoning with prompt engineering techniques.

1 INTRODUCTION

In recent years, significant advancements and breakthroughs have been made in large language models (LLMs) (Brown et al., 2020; Chowdhery et al., 2022; Touvron et al., 2023; OpenAI et al., 2024). By aligning the representations with visual (and other) encoders, LLMs have been extended to multimodal large language models (MLLMs[1]) (GeminiTeam, 2024; OpenAI, 2024), which are capable of handling richer visual modalities. These studies have attracted significant interest from researchers, with numerous works continuously being proposed to enhance the reasoning capabilities of MLLMs in various aspects. For example, Yang et al. (2023a) leverage the SoM prompting to enable the visual grounding of GPT-4v, and Wu et al. (2024b) achieve accurate object detection with MLLMs in a Chain-of-Thought (Wei et al., 2023) manner. By leveraging the advanced reasoning capabilities of the language model component, MLLMs have been explored for perception and interaction with a variety of applications.

However, existing MLLMs are mainly pretrained with 1D data (*e.g.* texts) and 2D data (*e.g.* images), while the real-world challenges are inherently spatial and require spatial grounding in the context of 3D scenes. In this context, a critical question emerges:

Do vision-language-based MLLMs possess the capability for 3D grounding and reasoning?

Although lots of studies have explored the application of MLLMs in 3D scenarios, these works have not directly leveraged the 3D grounding and reasoning capabilities of MLLMs. For example, some work (Wen et al., 2024; Cui et al., 2024) apply MLLMs in the field of autonomous driving, however, they primarily leverage MLLMs for decision-making rather than 3D scene understanding.

[1] also known as large multimodal models (LMMs).

Besides, PointLLM (Xu et al., 2023) empowers MLLMs to understand 3D point clouds with additional point-text instruction training. Although this solution aligns high-level representations of points and texts, it only supports specific comprehension tasks (*e.g.* classification and captioning for single 3D objects) and does not truly activate the fine-grained 3D perception capabilities of MLLMs. Overall, existing studies do not fully answer the question we raised, and further in-depth exploration is required.

In this paper, we aim to investigate how to extend the exceptional 1D/2D grounding and reasoning capabilities of MLLMs into the 3D world space without further fine-tuning. Based on this requirement and the inspiration of visual prompts, *e.g.* (Yang et al., 2023a; Wu et al., 2024b), we propose a new prompting mechanism, called 3DAxisPrompt. Specifically, given the point cloud of a real scene, we first embed the 3D coordinate axis and meshes in this scene in an automatic manner to provide the 3D geometric priors. Then calibrated scene will be rendered into observation images from different angles. Furthermore, to introduce object-level semantic cues, we overlay the masks generated by the Segment Anything Model (SAM) (Kirillov et al., 2023) with numerical or alphabetic marks, similar to SoM (Yang et al., 2023a). In this way, we can extend MLLMs impressive 2D grounding/reasoning capabilities to real-world 3D scenarios.

Besides, for the first time, we present a comprehensive exploration of potential visual prompt formats, such as coordinate axis, masks, bounding boxes, marks, color highlights, *etc.*, in MLLMs for 3D understanding. Based on our investigation, we also conclude some findings to reveal the potential and limits of 3D understanding capabilities in MLLMs. For example, multi-view visual prompting cannot directly activate the 3D reasoning capabilities of MLLMs, but tri-view prompting can. Finally, we construct evaluation environments using four datasets—ShapeNet, ScanNet, FMB, and nuScene—covering a range of 3D tasks. We then conduct extensive quantitative and qualitative experiments, demonstrating the effectiveness of the proposed approach.

Overall, our objective is not to achieve perfect zero-shot performance with GPT-4o, but to explore its limitations and potential in zero-shot inference for 3D grounding/reasoning. We expect that future improvements to the MLLMs will lead to further quantitative gains on the actual tasks. To summarize, our main contributions are:

- We propose a visual prompt scheme called 3DAxisPrompt. By inserting the 3D coordinate axis in a real scene, the proposed 3DAxisPrompt can elicit the 3D grounding and reasoning capabilities in GPT-4o, such as 3D localization and planning.
- We provide the first comprehensive investigation of the potential visual prompt formats of MLLMs for 3D understanding. Besides, we conclude our findings to reveal the potential and limits of 3D understanding capabilities in GPT-4o.
- We conduct extensive experiments on a wide range of tasks, including indoor and outdoor 3D localization, route planning, and robot action prediction. These results demonstrate the proposed 3DAxisPrompt can effectively enhance 3D understanding capabilities in GPT-4o.

2 Related Work

LLMs and MLLMs. Significant progress has been witnessed in LLMs (Chowdhery et al., 2022; Touvron et al., 2023; Zhang et al., 2022; OpenAI et al., 2024). Trained on internet-scale data, LLMs are effective commonsense reasoners (Zhao et al., 2023). MLLMs (Liu et al., 2023a; Lu et al., 2024; Bai et al., 2023) integrate vision encoders (Radford et al., 2021) into LLMs, allowing them to reason over visual input directly. State-of-the-art MLMMs like GPT-4V, Gemini (GeminiTeam, 2024), Claude (The), and GPT-4o (OpenAI, 2024) have excelled in general vision-language tasks (Wu et al., 2023; Yang et al., 2023c; Fu et al., 2023). Leveraging the advanced vision-language reasoning ability, the exploration has been made of MLMMs in perception and interacting with the physical world (Lu et al., 2024), including autonomous driving (Wen et al., 2024; Cui et al., 2024), anomaly detection (Cao et al., 2023), robotic control and learning (Collaboration et al., 2024; Brohan et al., 2023), which requires fine-grained 3D spatial grounding that remains to be explored (Chen et al., 2024). To promote the connection of MLMMs to the real physical world (Chen et al., 2024), we aim to find a strategic prompting method to elicit and promote the 3D grounding and reasoning in MLMMs regarding a real 3D world, such as to reason about the 3D location of an object.

Visual prompting. Prompt engineering has emerged as a promising approach to improve MLLMs across multiple domains, such as in-context learning (Brown et al., 2020; Dong et al., 2024), Chain-of-Thought and Tree-of-Thought (Wei et al., 2023; Yao et al., 2023). Consequently, numerous prompting methods have been developed to improve visual grounding in MLLMs. Colorful prompting tuning (CPT) (Yao et al., 2022) overlays color-based co-referential markers in both images and text and enables strong few-shot and even zero-shot visual grounding capabilities. RedCircle (Shtedritski et al., 2023) guides the vision model to an enclosed region by adding a red circle. Blur Reverse Mask (Yang et al., 2023b) blurs the area outside the target mask to leverage the precise mask annotations to reduce focus on weakly related regions while retaining spatial coherence. These two methods promote fine-grained visual grounding. Furthermore, Lei et al. (2024) enhances the vision-language coordination by SCAFFOLD prompting that scaffolds coordinate on images. Set-of-Mark (Yang et al., 2023a) add a set of visual marks on top of image regions. Both these two methods indicate the emergent 2D spatial grounding (Mitra et al., 2024; Islam et al., 2023) in MLLMs, including 2D position and relation inference. To enhance the 3D spatial grounding, Nasiriany et al. (2024) propose an iterative prompting method (PIVOT) to infer the robot action considering spatial relation. COARSE CORRESPONDENCES (Liu et al., 2024) prompts the MLLMs to elicit the 3D spacetime understanding. These two methods concentrate on 3D spatial relation instead of 3D spatial position, showing limited performance in instance-level tasks that demand precise 3D localization and recognition. Our study strives to extend the 2D spatial grounding (Lei et al., 2024; Yang et al., 2023a) to 3D grounding by formulating a visual prompting method, promoting spatial position inference in MLLMs.

GPTs and grounding. Generative Pretrained Transformers (GPTs) (Brown et al., 2020; OpenAI et al., 2024) have led to a breakthrough in the realm of natural language processing. As a leading LMM, GPT-4V has significantly expanded the boundaries of MLLMs capabilities and shown abilities to understand visual annotations (Yang et al., 2023c) and solve visual reasoning tasks, such as web navigation (Yan et al., 2023a; Zheng et al., 2024), autonomous driving (Wen et al., 2024; Cui et al., 2024), and medicine diagnostics (Yan et al., 2023b; Liu et al., 2023b). Furthermore, GPT-4o (OpenAI, 2024) is the latest development in a string of innovations to MLLMs, which has shown significant performance in multiple tasks (Joe et al., 2024; Wu et al., 2024a; Shahriar et al., 2024; Hu et al., 2024). Our study is to explore a prompting method to promote the 3D spatial grounding in MLLMs. Since GPT-4V is proven to outperform the other models in visual grounding when equipped with visual prompts (Yang et al., 2023a) and GPT-4o shows significant improvement in 3D spacetime understanding (Liu et al., 2024), we believe the GPT-4o can present representative 3D spatial grounding abilities in MLLMs and conduct our experiments and analysis using the GPT-4o.

3 3DAXISPROMPT

Unlike previous 2D visual prompts that primarily focus on planar object relationships, we aim to introduce a 3D prompts method to enable effective 3D spatial reasoning and grounding in GPT-4o for real 3D environments. In this section, We revisit approaches for incorporating 3D information into visual prompts and propose an effective method 3DAxisPrompt, to enhance 3D spatial information through visual prompts.

3.1 PROBLEM FORMULATION

The goal of 2D visual prompts is to enhance the MLLMs's understanding of visual information by adding auxiliary information to the original images. This can be expressed by the following equation:

$$T^o = \mathcal{F}(T^i, VP(I)), \tag{1}$$

where $T^o = \left[t_1^o, \ldots, t_{l_o}^o\right]$ represents textual output with a length of l_o from a foundation multimodal language model $\mathcal{F}$. This output is generated given a task textual description T^i and a visual prompt $VP(I)$ derived from an observation image I.

However, directly annotating and representing real 3D scenes is a more demanding task compared to 2D prompts, as it requires consideration of spatial depth, occlusions, and intricate object relationships (Liu et al., 2024). A common approach is to utilize multiview images instead of original 3D representations while adding corresponding annotations to the 2D images. Since GPT-4V has been shown to significantly outperform other MLLMs in grounding ability when visual prompts are added (Yang et al., 2023a), we employ GPT-4o as $\mathcal{F}$ in this work.

Meanwhile, unlike previous visual prompt approaches that solely add 2D spatial information, we discovered that when GPT-4o is challenged with both the point cloud p^i provided in text format and visual prompts, it can recognize the text file as the point cloud p^i and reason about spatial positions based on the input sequence T^i, the observation image I, and the point cloud p^i.

We experimented with various visual prompt formats to determine the optimal way to transform an input image I into a marked image I^m with 3D cues. After evaluating different 3D cue representations through spatial reasoning tasks, we propose the 3DAxisPrompt framework, as illustrated in Figure 1.

Given a point cloud as input, 3DAxisPrompt adds the 3D axis to the point cloud and renders observation images from multiple views of the point data. For each view, SAM (Kirillov et al., 2023) is used to highlight the boundary of the region of interest and overlay the mark. Consequently, the observation I becomes an image sequence $I^m_j = [I^m_1, \ldots, I^m_j]$. Formally, Equation 3.1 becomes:

$$T^o = \mathcal{F}(T^i, p^i, \underbrace{3DAxis(I)}_{I^m_j}). \tag{2}$$

By incorporating the 3D axis and overlaying marks and contours onto the rendered observation image of a point cloud, the 3DAxisPrompt enables GPT-4o to perform 3D spatial grounding tasks such as localization, route planning, and robot action prediction.

In the following sections, we will delve into exploring the impact of adding various 3D visual cues on GPT-4o's spatial grounding and reasoning capabilities. Subsequently, we will conduct quantitative experiments to assess the performance of the proposed 3DAxisPrompt framework.

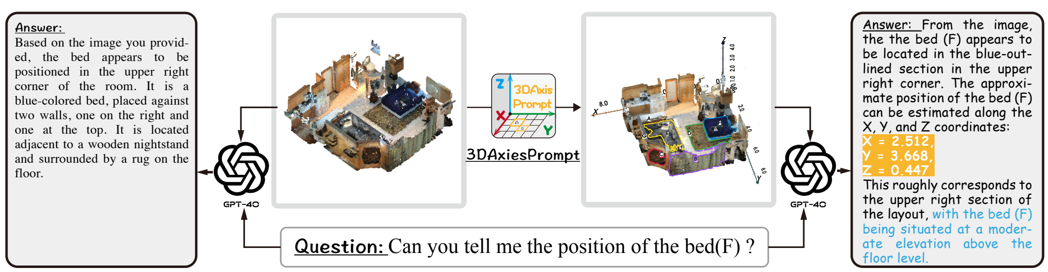


Figure 1: Comparing standard GPT-4o and its combination with 3DAxisPrompt. It shows that the proposed 3DAxiesPrompt helps GPT-4o to reason about the 3D spatial position. We highlight the differences between our method and the standard one.

3.2 INVESTIGATION ON ENCODING 3D CUES

Visual prompts, such as marks (Yang et al., 2023a; Liu et al., 2024), masks (Yang et al., 2023b), colors (Yao et al., 2022), scaffolding points (Lei et al., 2024), arrows (Nasiriany et al., 2024), and red circles (Shtedritski et al., 2023), have been shown to provoke 2D spatial grounding in GPT-4o. These visual prompts can be seen as integrating spatial information into images for grounding in image-text pairs (Brown et al., 2020; Li et al., 2022), leading to 2D spatial grounding. To extend 2D spatial grounding to 3D space, we propose encoding additional 3D cues into observation images to trigger 3D perception in GPT-4o. Based on this, we explore effective methods for representing these 3D cues.

3D axis integration in scenes. We found that adding a 3D axis to the point cloud of a 3D instance and rendering observation images with the x, y, and z axes as visual prompts enables GPT-4o to reason about 3D positions, as shown in Figure 2. This approach allows GPT-4o to associate semantics with spatial locations defined by the 3D axis, thereby facilitating 3D spatial grounding.

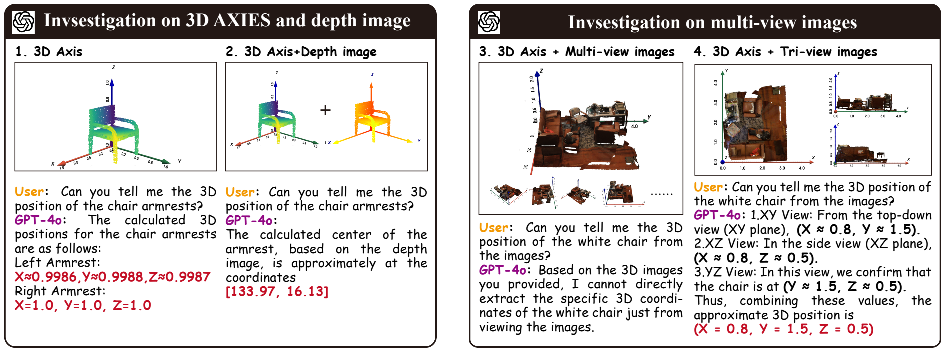


Figure 2: Investigation on encoding 3D cues in visual prompts. We present some examples of the investigations on the 3D Axis, depth image, multi-view images, and tri-view images. Depth image and multi-view images fail to provoke the 3D spatial position inference.

Depth compensation. Although 3D Axis prompts enable basic spatial grounding, the spatial positions inferred by GPT-4o lack accuracy, especially along the depth direction. We further explore potential solutions to compensate for the missing dimensions, including leveraging RGB-D images as visual input, as shown in Figure 2. More results are presented in Appendix A1. In conclusion, none of these depth compensation methods yielded satisfactory results. While GPT-4o can recognize depth images and surface color as depth or distance information, the depth and 2D positions are predicted separately, indicating a lack of interaction between them.

3D coordinates information. Based on our findings during depth compensation, we believe that encoding all 3D information solely within visual prompts is overly challenging (Liu et al., 2024). Additional 3D cues are necessary beyond just visual prompts. Furthermore, we discovered that GPT-4o can recognize point clouds formatted as coordinates in the input text, as demonstrated in Appendix A. However, when these points are combined with a 3D Axis visual prompt, GPT-4o effectively incorporates them for reasoning about 3D spatial positions. Consequently, we consider the point cloud in text format to be an essential input for the model.

Multiview and tri-view images. Inspired by Structure from Motion (SFM) (Schönberger & Frahm, 2016), which can reconstruct 3D structures from a series of 2D images, and tri-plane methods (Shue et al., 2023), which decompose a 3D scene into three distinct 2D projections, we further investigate the multiview and tri-view images of an actual scene. As shown in Figure 2, we render the images with the 3D axis of the actual scene from different angles. Additional results are provided in Appendix A1. Our findings indicate that the multi-view image sequence can only trigger 3D spatial grounding in GPT-4o when combined with text-formatted point clouds. In contrast, the tri-view images successfully provoke 3D spatial grounding in GPT-4o even without the text-formatted point cloud input. However, when reasoning about complex scenes, tri-view encounters significant occlusion issues, leading to considerable inaccuracies.

Based on these aforementioned findings, we incorporate the 3D Axis into the 3D scene and render observation images from various angles as the visual prompts.

3.3 Investigation on mark formats

We explored two methods for overlaying marks on visual prompts. The first method involves adding 2D marks directly onto the observation image, while the second method inserts 3D marks into the 3D space and then renders the observation image with these marks.

2D marks. The 2D marks are obtained using SAM to segment the objects of interest in the observation image. We consider two types of 2D marks: those on top-view images and those on perspective images, as illustrated in Figure 3. We also evaluate four main 2D mark formats—point, polygon, mask, and bounding box (see Appendix A2). Our empirical study indicates that all mark formats, when combined with the 3D Axis, successfully elicit 2D spatial grounding in GPT-4o.

3D marks. For 3D marks, we investigate the use of 3D bounding boxes and 3D edge points, as shown in Figure 3. We evaluate four types of 3D markers: marks, Axis-Aligned Bounding Boxes

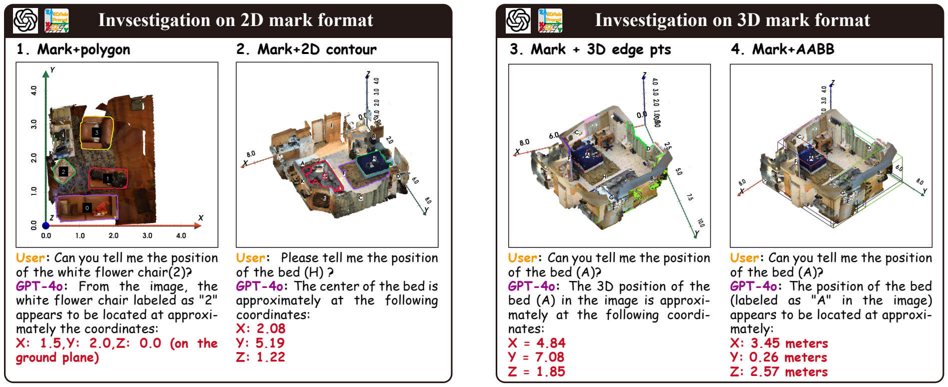


Figure 3: Some examples of the investigation on 2D and 3D mark formats. All the mark formats successfully provoke the 3D spatial position reasoning.

(AABB), Oriented Bounding Boxes (OBB), and 3D edge points. The 3D edge points are filtered from the input point cloud based on their normals. The visual results demonstrate that all the 3D marks successfully elicit 3D spatial grounding in GPT-4o. Additional results are provided in Appendix A3.

In conclusion, using both 2D and 3D marks in visual prompts can effectively elicit 3D spatial position reasoning in GPT-4o. To determine the optimal mark format, we evaluate all the mark formats in the following section. The quantitative results indicate that both the combination of (mark + 3D edge points) and (mark + 2D contour) perform better than the others, with the 2D contour outperforming the 3D edge points. This underscores the importance of object contours in visual prompts for 3D spatial position reasoning. Additionally, we employ multiview images instead of tri-views to mitigate the occlusion problem.

4 Experiments

4.1 Experimental Setup

Implementation. Our method does not require model training. However, due to the limited and costly GPT-4o API quota, we must exhaustively send 3DAxisPrompt-augmented images to the ChatGPT interface. To efficiently manage experiments and evaluations, we employ a divide-and-conquer strategy, opening a new chat window for each scene to prevent context leakage. All reported results are obtained in a zero-shot manner.

Benchmarks. Given the limited GPT-4o quota, we could not fully evaluate the validation set for each task. Instead, we randomly selected 20 scenes from each test dataset as validation data. We aimed to cover as many diverse scenes as possible across all datasets to preserve their original diversity. For each instance, we applied the 3DAxisPrompt to the observation images of the point cloud using our custom toolbox.

4.2 Quantitative Results

Indoor localization. On the indoor localization task (shown in Figure 1), we evaluate the localization errors of the 3DAxisPrompt on the subset of the Scannet (Dai et al., 2017) to fully analyze mark formats shown in Figure 3. Also, we integrate the Chain-of-Thought (CoT) (Mitra et al., 2024) with the proposed 3DAxiesPrompt and provide the additional coordinate of a nearby object to append *let's think step by step*. No previous work has presented localization errors related to 3D spatial grounding. We use the Normalized Root Mean Squared Errors (NRMSE) to quantify the spatial localization errors, as defined in Equation 4.2:

$$\text{NRMSE} = (\sum_{j=1}^{N} \frac{\sum_{i=1}^{n_j} \mathcal{D}(\hat{x_i}, x_i)}{n_j \cdot max(x_i)})/N \tag{3}$$

where $\hat{x_i}$ is the predicted position of the object i in scene j while x_i is the ground-truth position. n_j is the total number of the objects in scene j, and N is the total number of scenes selected for evaluation. $\mathcal{D}$ is the function to measure the distance between the predicted position $\hat{x_i}$ and the ground-truth position x_i. Two types of distance measurement function $\mathcal{D}$ are selected, including the distance to the object center (To center) and the distance to the bounding box (AABB) (To bbx). We use the Euclidean distance to measure the to-center distance. As for the to-bbx distance, we calculate the minimum distance from the predicted position to the AABB of the object.

Table 1: Main quantitative results of indoor localization on ScanNet dataset.

Mark Type	Prompt Elements	ScanNet	
		To center	To bbx
3D Mark	Mark	0.333	0.216
	Mark+OBB	0.350	0.231
	Mark+AABB (red)	0.376	0.219
	Mark+AABB (colors)	0.311	0.207
	Mark+3D edge points	0.305	0.205
2D Mark	2D contour (colors)	0.320	0.175
	Mark+2D contour (colors)	**0.271**	**0.138**
	Mark+2D contour (colors) + CoT	**0.219**	**0.115**

We present the quantitative results of the indoor localization in Table 1. It can be seen that besides the CoT, the combination of the mark and 2D contour achieves the best performance with a 7% decline in to-bbx distance errors compared to the Mark visual prompts. When combined with the CoT, the 3DAxiesPrompt achieves a 19% improvement on to-center distance. In the 3D mark, the (mark + 3D edge points) outperforms the others, and the performance of the OBB, AABB, 3D edge points, and 2D contour gradually improves. The bounding box, 3D edge point, and contour are the same in some ways because they all intend to depict the boundary of each instance region in a scene. This rend shows the importance of highlighting the instance boundary in visual prompts. Also, compared to a single color (red), highlighting each object boundary using different colors sees a 7% decline in to-center distance errors.

Table 2: Quantitative results of route planning, outdoor localization, and robot action prediction.

Task	Specification	ScanNet	nuScenes		FMB
		Success rate	To center	To bbx	Success rate
Route Planning	From door to chair	80%	n/a	n/a	n/a
	From door to bed	100%	n/a	n/a	n/a
	From door to desk	70%	n/a	n/a	n/a
	From couch to bed	90%	n/a	n/a	n/a
	From door to chair	60%	n/a	n/a	n/a
	Average	79%	n/a	n/a	n/a
Outdoor Localization	Vehicle	n/a	0.306	0.165	n/a
	Vegetation	n/a	0.283	0.143	n/a
Robot Action Prediction	Grasp	n/a	n/a	n/a	72.5%
	Release	n/a	n/a	n/a	62.5%

Route planning. We evaluate the indoor route planning abilities on the subset of Scannet (Dai et al., 2017), as shown in Figure 4. Following PIVOT (Nasiriany et al., 2024), we evaluate the performance via whether the navigation successfully reaches the destination. We select some common tasks that happen frequently in real life.

The quantitative results of the route planning task are demonstrated in Table 2. The 3DAxisPrompt achieves an average success rate of 79%, proving that the indoor route planning ability in GPT-4o is equipped with 3DAxisPrompt. However, when encountered with objects densely located together, such as chairs, the 3DAxisPrompt is more likely to fail (70% in from door to desk when needing to detour past many chairs).

Outdoor localization. We evaluate the outdoor localization of the 3DAxisPrompt on the subset of the nuScenes (Caesar et al., 2020) dataset. The point cloud is very sparse in nuScenes, so we choose the two types of obstacles frequently encountered in autonomous driving, including vehicle and vegetation, as shown in Figure 5. We use the same merits defined in Equation 4.2 to quantify the performance.

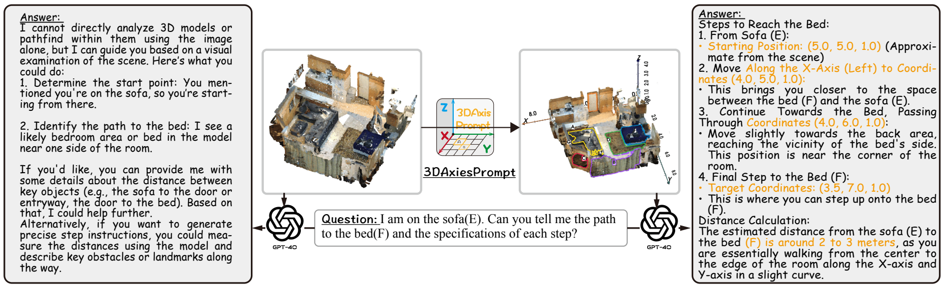


Figure 4: Experiments on route planning. It shows that our method 3DAxiesPrompt helps GPT-4o to plan the route based on spatial localization. We highlight the differences between our method and the standard one.

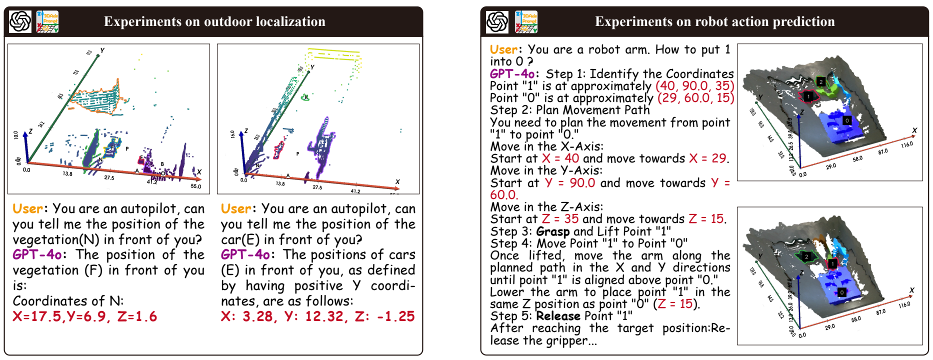


Figure 5: Some examples of the experiments on outdoor localization and robot action prediction.

The quantitative results of the outdoor localization are shown in Table 2. The localization performance of the vehicle is better than that of the vegetation.

Robot action prediction. In addition to localization and navigation tasks, we also examine the 3DAxisPrompt for robot action prediction on the subset of the robot control dataset FMB (Luo et al., 2024). There is no point data in the FMB, so we transform the RGBD images to point clouds according to the camera intrinsic as the evaluation data, as shown in Figure 5. The task is to predict the action to place the object onto the target destination, assuming the GPT-4o is a robot arm. Two types of actions are evaluated separately, namely grasp and release, because these two actions are the central part of the robot's grasping task. We evaluate the performance by determining whether the orders can complete the mission.

Table 2 presents the quantitative results. Equipped with the 3DAxisPrompt, GPT-4o can complete simple robot action prediction tasks.

Coarse object generation. We also evaluate the 3DAxiesPrompt for coarse object generation task on Shapenet (Chang et al., 2015) dataset, as shown in Figure 6. Some keypoints of an object are marked and predicted using the 3DAxiesPrompt. Then, a coarse object skeleton is constructed based on the answers.

4.3 Ablation Study

We conduct an ablation study on elements that may affect the GPT-4o to 'read' the coordinates from the 3D Axis, including the number of muti-view images and the axis elements.

The number of images. We conduct the ablation study on the number of observation images through the indoor localization tasks on the subset of the Scannet (Dai et al., 2017) dataset. The experimental results are shown in the line graph of Figure 7. A trend can be noticed that by increasing the number of scene views, the localization errors gradually decrease. The eight observation images outperform the others and achieve a 41% improvement compared to a single image.

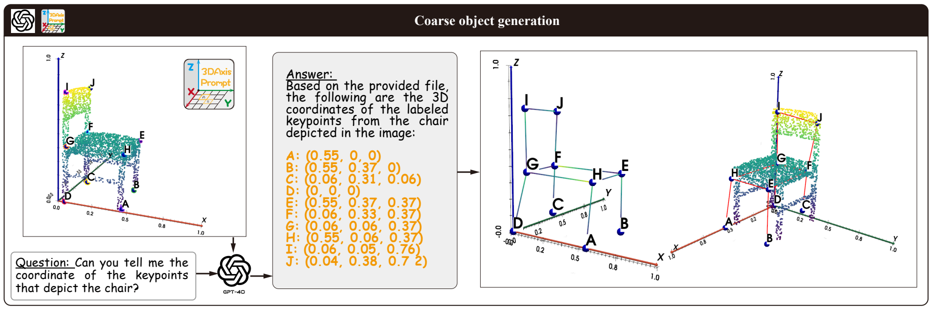


Figure 6: Coarse object generation on Shapenet dataset. It shows that based on our method, GPT-4o can reason about the keypoints that can represent the skeleton of an object.

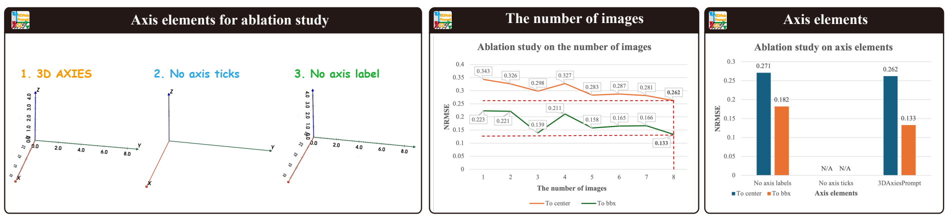


Figure 7: The axis elements considered for ablation study and the results of the number of images and the axis elements.

Axis elements. The elements of the axis, including the axis ticks and labels, are studied as shown in Figure 7. From the quantitative results shown in the histogram of Figure 7, we can see that the 3DAxisPrompt fails to provoke the spatial position reasoning without the axis ticks. Also, the axis label is essential, without which the errors of the to-bbx distance will increase by 37%.

5 Discussion and Conclusion

Where dose the 3D spatial grounding comes from? Our understanding is derived from experimental observations. We hypothesize that the 3D Axis offers essential scale information and spatial cues that serve as a foundation for localization. Interestingly, even without the 3DAxisPrompt, GPT-4o can make rough estimates of distances between objects when provided with an observation image of a real scene. However, by incorporating the 3D Axis, these estimates become more precise, as the axis ticks unify the units of measurement, allowing for a more accurate perception of distance. Additionally, the axis origin and direction act as reference points, supporting the localization process. In this way, the 3DAxisPrompt reinforces 3D spatial grounding by offering crucial 3D cues.

The essential factors in 3DAxisPrompt. The axis ticks and the highlighted contour of an object in the observation images are essential in 3DAxisPrompt. More specifically, the axis ticks provide an essential ruler to measure the world, while the contours marked in the observation images can significantly improve the localization performance. Also, we find that the localization performance can be further enhanced if given the precise coordinates of the objects (reference points) around the queried one. We think this is the same as human perception; the additional reference point makes the coordinate easier to read.

Conclusion. In this paper, we propose a visual prompt scheme called 3DAxisPrompt for MLLMs, particularly GPT-4o, aimed at enhancing 3D spatial grounding. By overlaying visible 3D axis, markers, and region edges on observation images from different angles, 3DAxisPrompt enables tasks like localization and spatial reasoning. Our study shows how various 3D visual prompts help GPT-4o interpret 3D space, with qualitative results indicating fine-grained perception and reasoning in real-world scenarios. We hope this work inspires future research on applying MLLMs to real-world interactions and advancing AI in everyday life.

REFERENCES

The claude 3 model family: Opus, sonnet, haiku. URL https://api.semanticscholar.org/CorpusID:268232499.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023. URL https://arxiv.org/abs/2307.15818.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

Yunkang Cao, Xiaohao Xu, Chen Sun, Xiaonan Huang, and Weiming Shen. Towards generic anomaly detection and understanding: Large-scale visual-linguistic model (gpt-4v) takes the lead, 2023. URL https://arxiv.org/abs/2311.02782.

Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository, 2015. URL https://arxiv.org/abs/1512.03012.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities, 2024. URL https://arxiv.org/abs/2401.12168.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. URL https://arxiv.org/abs/2204.02311.

Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Mart'in-Mart'in, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open x-embodiment: Robotic learning datasets and rt-x models, 2024. URL https://arxiv.org/abs/2310.08864.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 958–979, 2024.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes, 2017. URL https://arxiv.org/abs/1702.04405.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning, 2024. URL https://arxiv.org/abs/2301.00234.

Chaoyou Fu, Renrui Zhang, Zihan Wang, Yubo Huang, Zhengye Zhang, Longtian Qiu, Gaoxiang Ye, Yunhang Shen, Mengdan Zhang, Peixian Chen, Sirui Zhao, Shaohui Lin, Deqiang Jiang, Di Yin, Peng Gao, Ke Li, Hongsheng Li, and Xing Sun. A challenger to gpt-4v? early explorations of gemini in visual expertise, 2023. URL https://arxiv.org/abs/2312.12436.

GeminiTeam. Gemini: A family of highly capable multimodal models, 2024. URL https://arxiv.org/abs/2312.11805.

Danqing Hu, Bing Liu, Xiaofeng Zhu, and Nan Wu. The power of combining data and knowledge: Gpt-4o is an effective interpreter of machine learning models in predicting lymph node metastasis of lung cancer, 2024. URL https://arxiv.org/abs/2407.17900.

Ashhadul Islam, Md. Rafiul Biswas, Wajdi Zaghouani, Samir Brahim Belhaouari, and Zubair Shah. Pushing boundaries: Exploring zero shot object classification with large multimodal models, 2023. URL https://arxiv.org/abs/2401.00127.

Elphin Tom Joe, Sai Dileep Koneru, and Christine J Kirchhoff. Assessing the effectiveness of gpt-4o in climate change evidence synthesis and systematic assessments: Preliminary insights, 2024. URL https://arxiv.org/abs/2407.12826.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. Scaffolding coordinates to promote vision-language coordination in large multi-modal models, 2024. URL https://arxiv.org/abs/2402.12058.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training, 2022. URL https://arxiv.org/abs/2112.03857.

Benlin Liu, Yuhao Dong, Yiqin Wang, Yongming Rao, Yansong Tang, Wei-Chiu Ma, and Ranjay Krishna. Coarse correspondence elicit 3d spacetime understanding in multimodal language model, 2024. URL https://arxiv.org/abs/2408.00754.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023a. URL https://arxiv.org/abs/2304.08485.

Zhengliang Liu, Hanqi Jiang, Tianyang Zhong, Zihao Wu, Chong Ma, Yiwei Li, Xiaowei Yu, Yutong Zhang, Yi Pan, Peng Shu, Yanjun Lyu, Lu Zhang, Junjie Yao, Peixin Dong, Chao Cao, Zhenxiang Xiao, Jiaqi Wang, Huan Zhao, Shaochen Xu, Yaonai Wei, Jingyuan Chen, Haixing Dai, Peilong Wang, Hao He, Zewei Wang, Xinyu Wang, Xu Zhang, Lin Zhao, Yiheng Liu, Kai Zhang, Liheng Yan, Lichao Sun, Jun Liu, Ning Qiang, Bao Ge, Xiaoyan Cai, Shijie Zhao, Xintao Hu, Yixuan Yuan, Gang Li, Shu Zhang, Xin Zhang, Xi Jiang, Tuo Zhang, Dinggang Shen, Quanzheng Li, Wei Liu, Xiang Li, Dajiang Zhu, and Tianming Liu. Holistic evaluation of gpt-4v for biomedical imaging, 2023b. URL https://arxiv.org/abs/2312.05256.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024. URL https://arxiv.org/abs/2403.05525.

Jianlan Luo, Charles Xu, Fangchen Liu, Liam Tan, Zipeng Lin, Jeffrey Wu, Pieter Abbeel, and Sergey Levine. Fmb: a functional manipulation benchmark for generalizable robotic learning. *arXiv preprint arXiv:2401.08553*, 2024.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models, 2024. URL https://arxiv.org/abs/2311.17076.

Soroush Nasiriany, Fei Xia, Wenhao Yu, Ted Xiao, Jacky Liang, Ishita Dasgupta, Annie Xie, Danny Driess, Ayzaan Wahid, Zhuo Xu, Quan Vuong, Tingnan Zhang, Tsang-Wei Edward Lee, Kuang-Huei Lee, Peng Xu, Sean Kirmani, Yuke Zhu, Andy Zeng, Karol Hausman, Nicolas Heess, Chelsea Finn, Sergey Levine, and Brian Ichter. Pivot: Iterative visual prompting elicits actionable knowledge for vlms, 2024. URL https://arxiv.org/abs/2402.07872.

OpenAI. Hello gpt-4o, 2024. URL https://openai.com/index/hello-gpt-4o/.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4104–4113, 2016. doi: 10.1109/CVPR.2016.445.

Sakib Shahriar, Brady Lund, Nishith Reddy Mannuru, Muhammad Arbab Arshad, Kadhim Hayawi, Ravi Varma Kumar Bevara, Aashrith Mannuru, and Laiba Batool. Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency, 2024. URL https://arxiv.org/abs/2407.09519.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms, 2023. URL https://arxiv.org/abs/2304.06712.

J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20875–20886, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Licheng Wen, Daocheng Fu, Xin Li, Xinyu Cai, MA Tao, Pinlong Cai, Min Dou, Botian Shi, Liang He, and Yu Qiao. Dilu: A knowledge-driven approach to autonomous driving with large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Yang Wu, Shilong Wang, Hao Yang, Tian Zheng, Hongbo Zhang, Yanyan Zhao, and Bing Qin. An early evaluation of gpt-4v(ision), 2023. URL https://arxiv.org/abs/2310.16534.

Yiqi Wu, Xiaodan Hu, Ziming Fu, Siling Zhou, and Jiangong Li. Gpt-4o: Visual perception performance of multimodal large language models in piglet activity understanding, 2024a. URL https://arxiv.org/abs/2406.09781.

Yixuan Wu, Yizhou Wang, Shixiang Tang, Wenhao Wu, Tong He, Wanli Ouyang, Jian Wu, and Philip Torr. Dettoolchain: A new prompting paradigm to unleash detection ability of mllm. *arXiv preprint arXiv:2403.12488*, 2024b.

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

An Yan, Zhengyuan Yang, Wanrong Zhu, Kevin Lin, Linjie Li, Jianfeng Wang, Jianwei Yang, Yiwu Zhong, Julian McAuley, Jianfeng Gao, Zicheng Liu, and Lijuan Wang. Gpt-4v in wonderland: Large multimodal models for zero-shot smartphone gui navigation, 2023a. URL https://arxiv.org/abs/2311.07562.

Zhiling Yan, Kai Zhang, Rong Zhou, Lifang He, Xiang Li, and Lichao Sun. Multimodal chatgpt for medical applications: an experimental study of gpt-4v, 2023b. URL https://arxiv.org/abs/2310.19061.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v, 2023a. URL https://arxiv.org/abs/2310.11441.

Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting, 2023b. URL https://arxiv.org/abs/2306.04356.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision), 2023c. URL https://arxiv.org/abs/2309.17421.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023. URL https://arxiv.org/abs/2305.10601.

Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models, 2022. URL https://arxiv.org/abs/2109.11797.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022. URL https://arxiv.org/abs/2205.01068.

Zirui Zhao, Wee Sun Lee, and David Hsu. Large language models as commonsense knowledge for large-scale task planning, 2023. URL https://arxiv.org/abs/2305.14078.

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. Gpt-4v(ision) is a generalist web agent, if grounded, 2024. URL https://arxiv.org/abs/2401.01614.

A Appendix

A1 Investigation on encoding 3D cues

We present more investigation of encoding 3D cues in the visual prompts, as shown in Figure A1. It can be noted that GPT-4o can identify the distance information encoded on the object surface, but this extra information is not used to reason about spatial location. Also, the points.txt can be understood only if with the 3D Axis visual prompts. Both the RGB-D and depth information are not integrated into spatial reasoning.

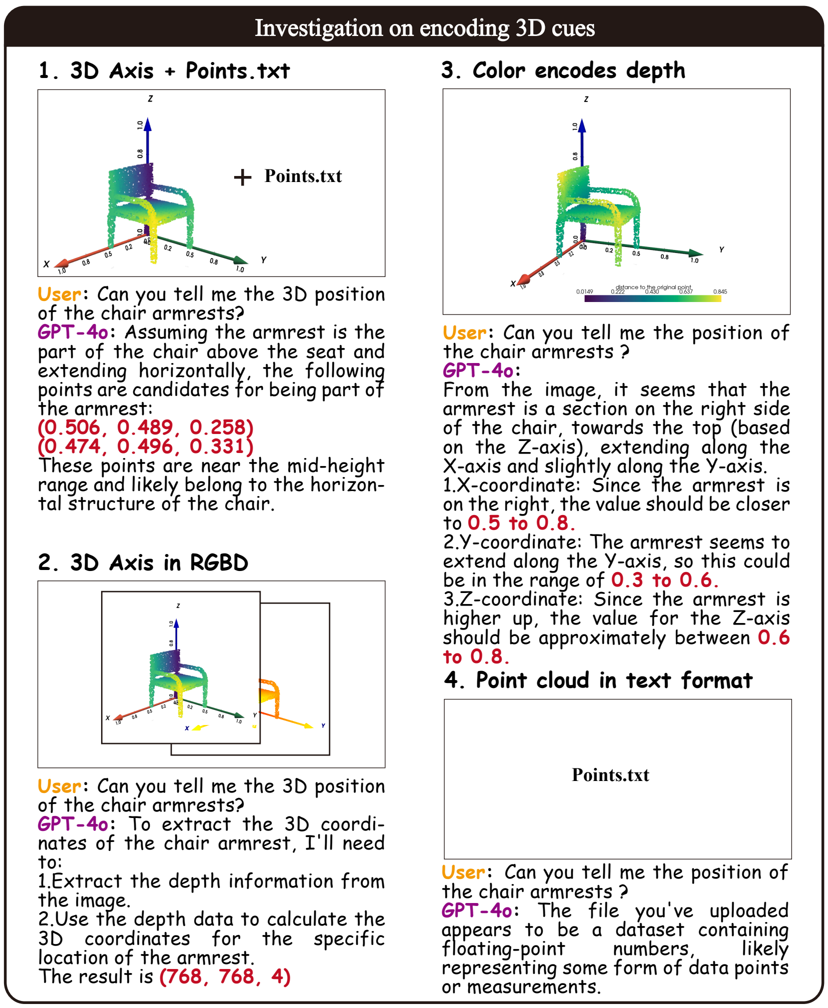


Figure A1: Investigation on encoding 3D clues in visual prompts.

We present more investigation of multi-view images, as shown in Figure A2. The multi-view images can only promote the sptial localization with the points.txt file. The tri-view images can promote spatial localization, but the objects are easy to block, as shown in Figure A2. Voxel can also represent a 3D scene, but it defects the spatial localization.

A2 Investigation on 2D mark format

We present more investigation of 2D mark formats, as shown in Figure A3 and A4. All the 3D mark formats can promote the spatial localization.

A3 INVESTIGATION ON 3D MARK FORMAT

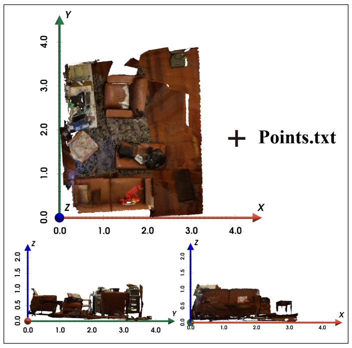
We present more investigation of 3D mark formats, as shown in Figure A5. All the 3D mark formats can promote the 3D spatial localization.

A4 LIMITATIONS

Even though the evaluation proves that the 3DAxisPrompt can promote the 3D spatial grounding in GPT-4o on some tasks, we have to admit that the performance is not perfect. When the objects are too small to be identified, or the boundaries are not clear enough, the performance will significantly drop. Moreover, we find that the GPT-4o still struggles to read the information encoded in the 3D Axis when the objects are far away from the 3D Axis.

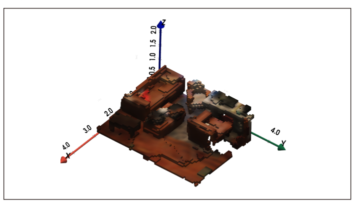
Investigation on multi-view images

1. 3D Axis+Tri-view+Points.txt



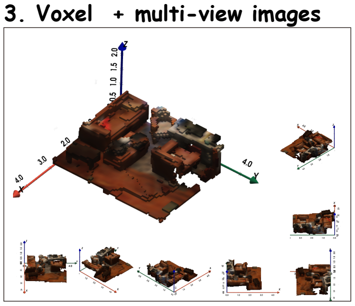
User: Can you tell me the 3D position of the white chair from the images?
GPT-4o: Final Estimated 3D Position of the White Chair:
X ≈ 0.5 meters
Y ≈ 3 meters
Z ≈ 0.8 to 1 meter

2. Voxel



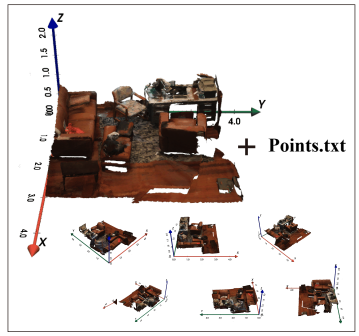
User: Can you tell me the 3D position of the white chair from the images ?
GPT-4o: The image you uploaded appears to represent a 3D voxel map or a block-style rendering. You asked for the position of a white chair in the 3D space shown, but without additional details on how this 3D structure is defined or segmented, it's hard to extract precise coordinates directly from the image alone.

3. Voxel + multi-view images



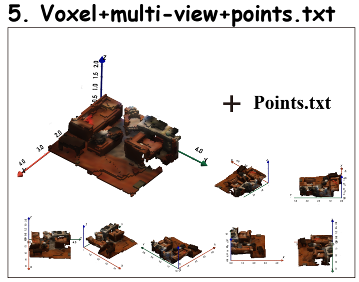
User: Can you tell me the 3D position of the white chair from the images ?
GPT-4o: To determine the 3D position of the white chair from these images, I would need a method to compute it based on the image data, such as image processing techniques for identifying the chair in each image and using the position from different angles to triangulate the 3D position.

4. 3D Axis+Multi-view+Points.txt



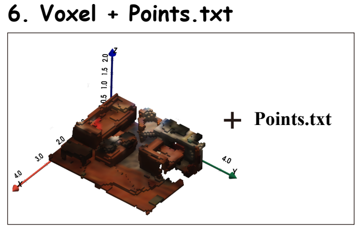
User: Can you tell me the 3D position of the white chair from the images?
GPT-4o: Based on the data from the text file, one of the likely 3D positions for the white chair could be found from the following coordinates:
X: 1.884, Y: 3.538, Z: 1.803

5. Voxel+multi-view+points.txt



User: Can you tell me the 3D position of the white chair from the images ?
GPT-4o: In the 3D views, we can use the X, Y, and Z axes to describe the chair's position. It appears to be in the lower left region in most of the views. From the images:
In the top-down view (image 2), the white chair is roughly around the coordinates
X: 0.5 - 1.0, Y: 2.0 - 2.5.The Z-axis height seems to be around 0.5 in several views (images 1 and 6).

6. Voxel + Points.txt

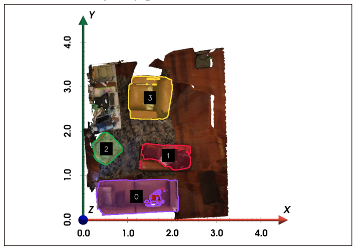


User: Can you tell me the 3D position of the white chair from the images ?
GPT-4o: Based on the data in the file, the white chair seems to be positioned around the following 3D coordinates:
x=0.471
y=0.491
z=0.311

Figure A2: Investigation on multi-view images.

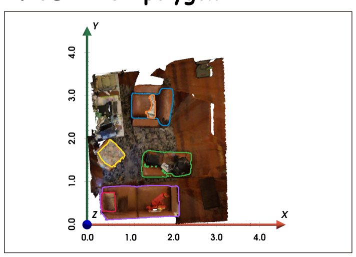
Investigation on 2D mark (I)

1. Mark+polygon+mask



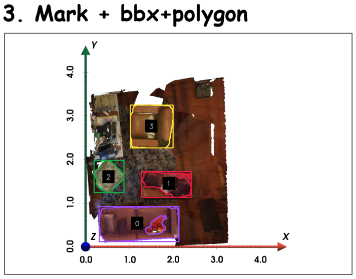
User: Can you tell me the position of the white flower chair (2) from the image?
GPT-4o: From the image, the white flower chair marked with "2" is located at approximately the following position on the X, Y, Z axes:
X-axis: Around 1.0
Y-axis: Around 2.0
Z-axis: 0 (since it's resting on the ground plane)

2. 3D Axis +polygon



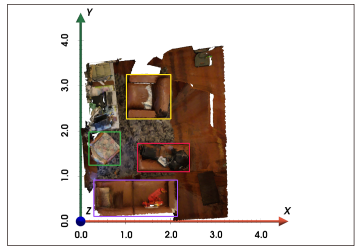
User: Can you tell me the position of the white flower chair (2) from the image?
GPT-4o: The position of the white flower chair (labeled as 2) is located near the lower left side of the image. The chair seems to be within the range of:
X-coordinate: around 1.0
Y-coordinate: around 1.5
Z-coordinate: 0.0 (ground level)

3. Mark + bbx+polygon



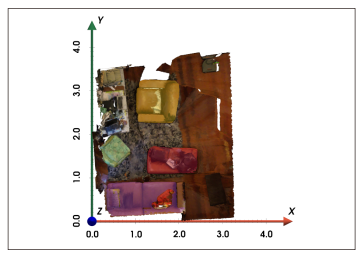
User: Can you tell me the position of the white flower chair (2) from the image?
GPT-4o: From the image, the white flower chair (labeled with "2" in a green box) appears to be positioned around the following coordinates:
X-coordinate: Between 1.0 and 2.0
Y-coordinate: Between 1.0 and 2.0
Z-coordinate: Close to 0.0, as the Z-axis seems to represent the ground plane.
Thus, the approximate position of the white flower chair is at
(X: 1.5, Y: 1.5, Z: 0.0).

4. 3D Axis+bbx



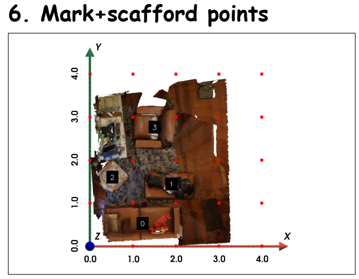
User: Can you tell me the position of the white flower chair (2) from the image?
GPT-4o: Based on the image you provided, the white flower chair is marked within a green bounding box. The approximate position of the white flower chair:
X-axis: Between 0.8 to 1.4 units.
Y-axis: Between 1.8 to 2.3 units.
Z-axis: Relatively close to the ground, at about 0 units.

5. 3D Axis +mask



User: Can you tell me the position of the white flower chair (2) from the image?
GPT-4o: From the provided image, the white flower chair (2) appears to be located near the bottom left corner of the room. The approximate position of the white flower chair to be around the following coordinates:
X ≈ 0.7 units
Y ≈ 0.9 units
Z ≈ 0.0 units

6. Mark+scafford points



User: Can you tell me the position of the white flower chair (2) from the image?
GPT-4o: Based on the image you provided, the white flower chair seems to be located near the bottom left corner of the room. Estimating its position from the coordinates grid on the floor, **it appears to be at approximately (X, Y) coordinates of around (1.0, 1.0) to (1.5, 1.5)**

Figure A3: Investigation on 2D marker format (I).

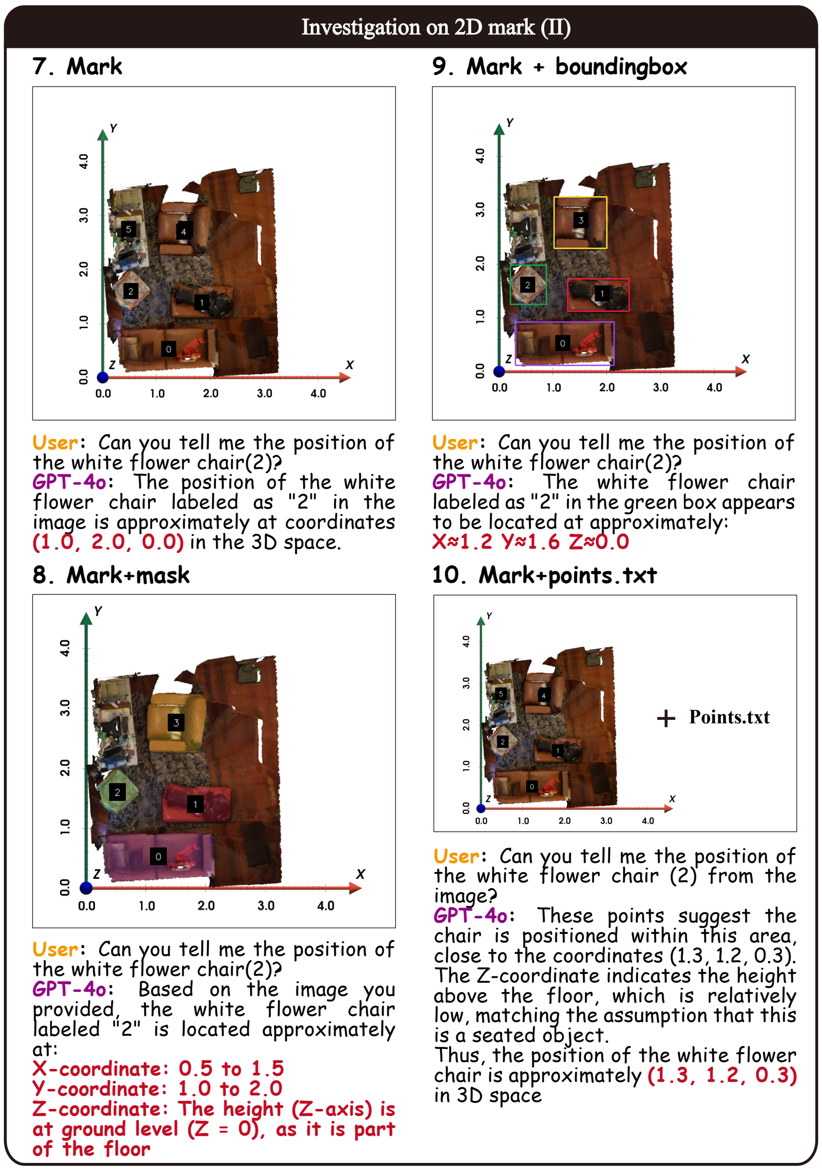
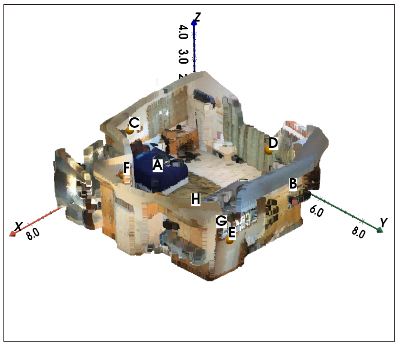


Figure A4: Investigation on 2D marker format (II).

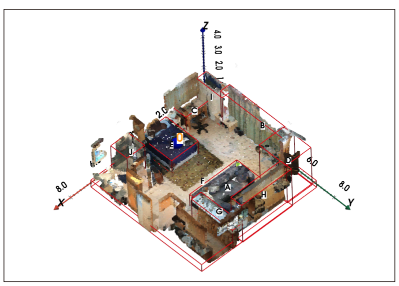
Investigation on 3D mark

1. Mark



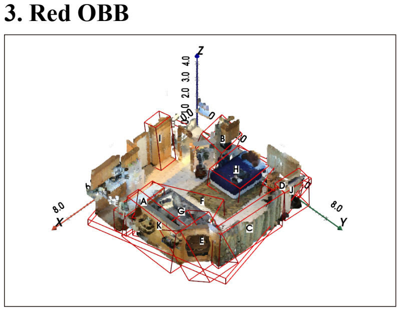
User: Can you tell me the position of the bed (A)?
GPT-4o: The 3D position of the bed (marked as A) is approximately located at coordinates
X = 0.0, Y = 2.3, Z = 1.7,
based on the data from the provided 3D scene file

2. Red AABB

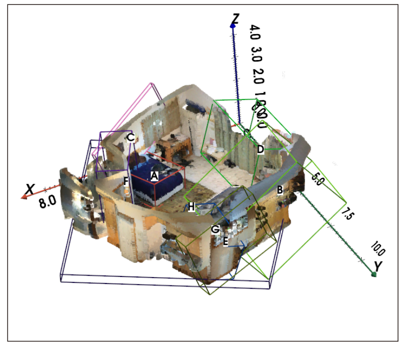


User: Can you tell me the coordinate of the bed (E) ?
GPT-4o: The coordinates for the bed, labeled as "E" in the image, are approximately:
(3.08, 1.49, 0.76).
These coordinates are extracted from the dataset based on object "E". It seems to refer to a point near the middle of the bed, with the dimensions indicating a placement within the room.

3. Red OBB

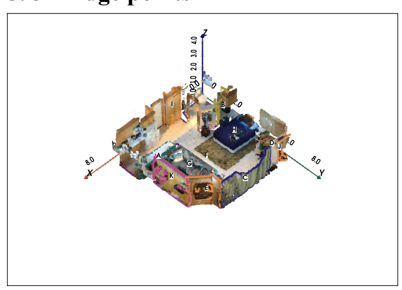


4. Mark+OBB



User: Can you tell me the position of the bed (A)?
GPT-4o: The 3D position of the bed (A) in the scene is approximately located at coordinates
X = 3.7, Y = 6.67, and Z = 1.22

5. 3D Edge points



User: Can you tell me the position of the bed (H) ?
GPT-4o:
In the image, point H is visually located on the bed, in the upper right corner near the coordinate axes. By cross-referencing the scene data file, it seems the relevant line for point H could correspond to:
X = 5.96
Y = 2.35
Z = 0.17

User: Please tell me the position of H?
GPT-4o: The three-dimensional coordinates of object H are approximately:
x=5.299
y=0.634
z=1.448
These coordinates indicate the position of object H within the scene.

Figure A5: Investigation on 3D marks.