Understanding Depth and Height Perception in Large Visual-Language Models

Shehreen Azad[1*] Yash Jain [2] Rishit Garg[3] Yogesh S Rawat[1] Vibhav Vineet[2]

[1]Center for Research in Computer Vision, University of Central Florida;
[2]Microsoft Research; [3]Indian Institute of Technology, Kharagpur.

Project Page

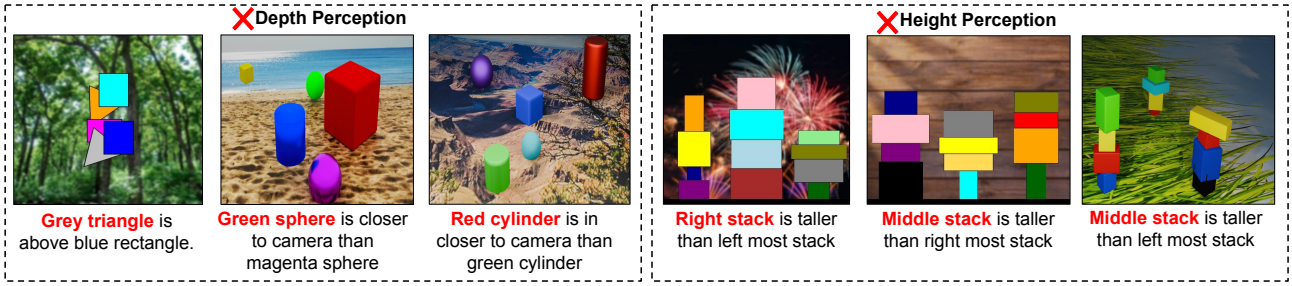

Figure 1. **Depth and height perception capability of existing VLM.** Here, we show failure cases of GPT-4V in understanding depth and height on GeoMeter, our proposed suite of benchmark datasets.

Abstract

Geometric understanding—including depth and height perception—is fundamental to intelligence and crucial for navigating our environment. Despite the impressive capabilities of large Vision Language Models (VLMs), it remains unclear how well they possess the geometric understanding required for practical applications in visual perception. In this work, we focus on evaluating the geometric understanding of these models, specifically targeting their ability to perceive the depth and height of objects in an image. To address this, we introduce GeoMeter, a suite of benchmark datasets—encompassing 2D and 3D scenarios—to rigorously evaluate these aspects. By benchmarking 18 state-of-the-art VLMs, we found that although they excel in perceiving basic geometric properties like shape and size, they consistently struggle with depth and height perception. Our analysis reveal that these challenges stem from shortcomings in their depth and height reasoning capabilities and inherent biases. This study aims to pave the way for developing VLMs with enhanced geometric understanding by emphasizing depth and height perception as critical components necessary for real-world applications.

1. Introduction

In recent years, the AI community has significantly focused on integrating visual and natural language inputs, notably in Visual Question Answering (VQA) systems. These systems analyze images and answer questions about them, showing substantial advancements in understanding basic visual concepts such as shape identification [19], object detection [40], and the spatial relationships [5, 17, 22] by using large Visual Language Models (VLMs). These models have excelled in processing complex text and visual inputs due to their strong visual understanding capability, leading to applications in image captioning, visual question answering, image text retrieval, and so on.

The ability to understand visual properties such as size, shape, depth, and height is fundamental to visual understanding, yet many existing Visual Question Answering (VQA) benchmarks [5, 11, 17, 22, 35] do not specifically focus on the depth and height perception capabilities of Vision Language Models (VLMs). Accurate perception of these dimensions is vital for practical applications like surveillance, navigation, and assistive technologies. The lack of accurate depth and height understanding in VLMs can lead to serious consequences, such as misjudging the proximity of objects, which could result in catastrophic outcomes in real-world scenarios.

Despite VLMs' abilities to recognize object shapes and sizes, their depth and height reasoning often relies on learned size/shape cues rather than actual spatial analysis, potentially influenced by biases from training data [16]. Alternatively, models might estimate the depth based on the

*Corresponding author. Shehreen.Azad@ucf.edu

apparent size of objects, without genuine inter-object reasoning. An example illustrated in Figure 1 shows how GPT-4V [27], one of the most popular closed-source VLMs, struggles with depth and height perception in images featuring basic 2D or 3D objects. The model is just randomly answering without correctly assessing the spatial relationship between the objects, relying on visual cues that conflict with their actual arrangement. Additional examples in Figure 1 further demonstrate GPT-4V's failures in perceiving depth and height. These limitations highlight the need to explore such shortcomings more thoroughly and develop targeted benchmarks and training strategies that can better equip VLMs to handle complex, real-world environments with accurate depth and height perception.

In this paper, we aim to evaluate the depth and height reasoning capabilities of Vision Language Models (VLMs) to identify their strengths and limitations in visual perception. While auxiliary sensors play a crucial role in depth estimation and other alternative methods of estimating depth and height may outperform visual language models (VLMs) in specific tasks, our research aims to assess the standalone capabilities of VLMs rather than suggesting their replacement. To achieve this, we design GeoMeter, a suite of synthetic benchmark datasets focusing on 2D and 3D scenarios, named GeoMeter-2D and GeoMeter-3D respectively. These probing datasets, feature basic shapes, such as rectangles, circles, cubes, and cylinders, and are crafted to test the visual reasoning capabilities of VLMs. The development of synthetic datasets is motivated by concerns about test-time data leakage, where large VLMs, trained on vast datasets, might encounter images during testing that they have already seen during training. We prioritize clean, programmatically generated data over mere size to ensure that the evaluation is not compromised by dataset familiarity. This controlled approach minimizes the risk of data leakage and enables a more focused and precise assessment of VLMs' understanding of depth and height, free from the confounding influence of real-world cues present in many publicly sourced datasets. To this end, our probing datasets consist of around 4k unique images and 11.2k image text pairs, designed to probe depth and height reasoning in VLMs.

We extensively analyze our proposed suite of benchmark datasets on 18 recent open-source and closed-source models for the VQA task. Our findings reveal several key insights: (1) While VLMs demonstrate basic geometric understanding, they struggle significantly with depth and height perception tasks. (2) Models generally show better depth perception than height, likely due to the more common and simpler depth cues like occlusion and perspective, which are prevalent in training datasets, making depth easier to process than the more complex cues required for accurate height estimation. (3) The lack of depth and height perception ability stems from the models' intrinsic visual reasoning abilities rather than the level of prompt detail. (4) Inherent biases are evident in models' responses when faced with advanced perception tasks.

Table 1. **Dataset statistics** of our proposed benchmark suites.

Dataset	Perception Task	Description	Images	Img-Text pairs
GeoMeter-2D	Depth	Determine which object is on the top.	1200	4800
	Height	Provide height ordering from shortest to tallest among given stacks	1200	
GeoMeter-3D	Depth	Determine which object is closer to the camera.	800	6400
	Height	Provide height ordering from shortest to tallest among given stacks	800	

Overall, our contributions can be summarized as follows:

- We examine the depth and height reasoning abilities of VLMs, identifying strengths, limitations, and areas for improvement in visual perception tasks.
- Our analysis spans 18 VLMs, both open-source and closed-source, revealing behavioral patterns and biases in depth and height perception.
- To support this evaluation, we introduce GeoMeter, comprising two datasets - GeoMeter-2D and GeoMeter-3D - that test VLMs on depth and height perception tasks.

2. Related Works

Visual Language Models (VLMs). The field of AI has undergone a significant transformation with the advent of vision language models (VLMs), which are trained on extensive multimodal datasets and are versatile across numerous applications [3, 10, 20, 21, 24, 28, 31, 33]. These models have shown remarkable performance in language and vision-related tasks, e.g. recognition, reasoning, etc. VLMs are models with a pre-trained LLM backbone and a vision encoder; which are aligned by using different methods. Recent closed-source VLMs such as GPT-4 [27], Gemini [13], Claude [2] showcase a strong potential for tasks that require understanding and processing information across different modalities. Additionally, various openly available VLMs such as LLaVA [24], LLaVA-NeXT [23], Bunny [14] etc. also have comparative performance with the closed-source models across different vision-language tasks. All of these VLMs are trained on massive amount of public and proprietary data, making them a strong performer of general reasoning.

Exploring Visual Reasoning Capability of VLMs. Previous works have extensively explored the spatial reasoning and object understanding capabilities of Vision Language Models (VLMs), probing their ability to grasp object-attribute relationships and spatial concepts like spatial reasoning through various benchmarks [11, 15, 17, 18, 22,

29, 30, 35, 36]. However, specific geometric properties such as depth and height perception have been largely under-explored. While there are benchmarks that assess geometric property understanding [6, 32, 38], they often rely on mathematical knowledge and do not directly probe these properties in the context of natural visual understanding. Moreover, many of the datasets used in these studies [11, 18, 22, 29, 35, 36] are curated from pre-existing datasets and/or the internet, which introduces the risk of data leakage during testing, making it difficult to assess VLMs' true capability for depth and height reasoning. Although synthetic datasets have been developed [17, 19]; they are not specifically tailored to tasks focusing on depth and height understanding, further limiting their effectiveness in thoroughly evaluating these advanced visual concepts. Our proposed benchmark suite addresses this gap by offering image-text pairs that target depth and height perception, without relying on mathematical reasoning, providing a more focused assessment of VLMs in this area.

3. Benchmark

Our proposed suite of benchmark datasets consist of GeoMeter-2D, and GeoMeter-3D datasets that are designed to test model performance on depth and height perception tasks, utilizing unique identifiers as diverse query attributes for question generation. Table 1, and Figure 2 respectively show the dataset statistics and sample images of our proposed datasets. More samples from each dataset is given in the appendix. In the following sections, we describe the detailed data generation process for our benchmark.

3.1. Datasets

The generation process is divided into Image generation (Section 3.1.1) and Question generation (Section 3.1.2).

3.1.1. Image Generation

Our proposed synthetic datasets are divided into two categories - *Depth* and *Height*, with each image containing a real-world scene background to add realism while maintaining controlled, programmatically generated content. We generate images in two variety of scene density - 3 shapes and 5 shapes, with each shape having one unique identifier which is used as query attribute to refer to that certain object while probing the VLMs' depth and height perception.

GeoMeter-2D dataset includes 2400 *images and* 4800 *unique questions*, and is designed to test depth and height perception through basic 2D shapes. The *Depth* category features overlapping 3 or 5 geometric shapes, like rectangles, triangles, and circles, positioned to create depth illusions. Ground truth for these images is stored in a scene graph that annotates each object's shape, size, color, and spatial positioning, including depth ordering through directed edges connecting overlapping objects. Each object is assigned a unique identifier based on color and numeric labels. For the *Height* category, we generated scenes featuring sequentially labeled 3 or 5 towers, each consisting of four stacked rectangles. Each tower was created by randomizing the height and width of the individual rectangles to add variability to the scene. The bottom-most rectangle in some images is placed on a black strip representing an elevated platform, making the tower effectively shorter by one rectangle in actual height but visually elevated. These images are categorized into two subgroups: *w/ step* for towers placed on a platform and *w/o step* for towers directly placed on the ground. This setup allows VLMs to be rigorously tested on height comparison tasks, requiring them to correctly interpret both the visual cues of the towers' absolute and relative heights, and the additional complexity introduced by the raised platforms. Each object in the scene is uniquely identified by its color and label, and the scene graph provides the ground truth, detailing the size, position, and elevation of each tower.

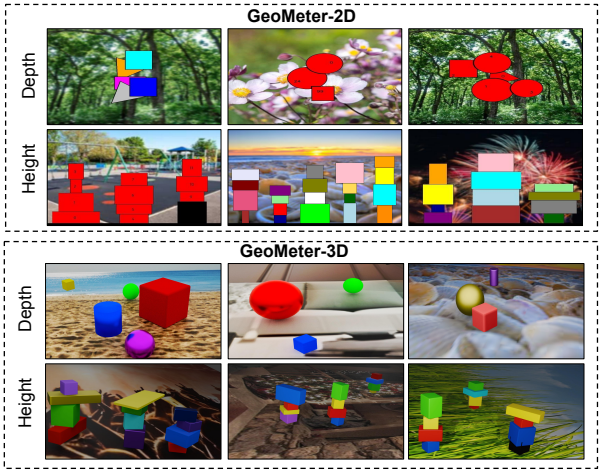

Figure 2. **Samples from the proposed suite of benchmark datasets.** Here each samples are shown with random query attributes- color and numeric label for GeoMeter-2D; and color and material for GeoMeter-3D dataset.

GeoMeter-3D dataset consists of 1600 *images and* 6400 *unique questions*, created based on the existing CLEVR dataset [17]. Scenes are generated using Blender [9], with random jittering of light and camera positions to ensure variety. Objects in these scenes are annotated using a scene graph, which records each object's shape, size, color, material (shiny "metal" or matte "rubber"), and position on the ground plane. The *Depth* category includes randomly placed 3 or 5 cubes, spheres, and cylinders with distinct colors and materials as unique identifiers. These shapes are colored from a palette of eight colors and two materials, with increased horizontal and vertical margins than original CLEVR images between objects to reduce ambiguous spatial relationships. The scene graph captures all

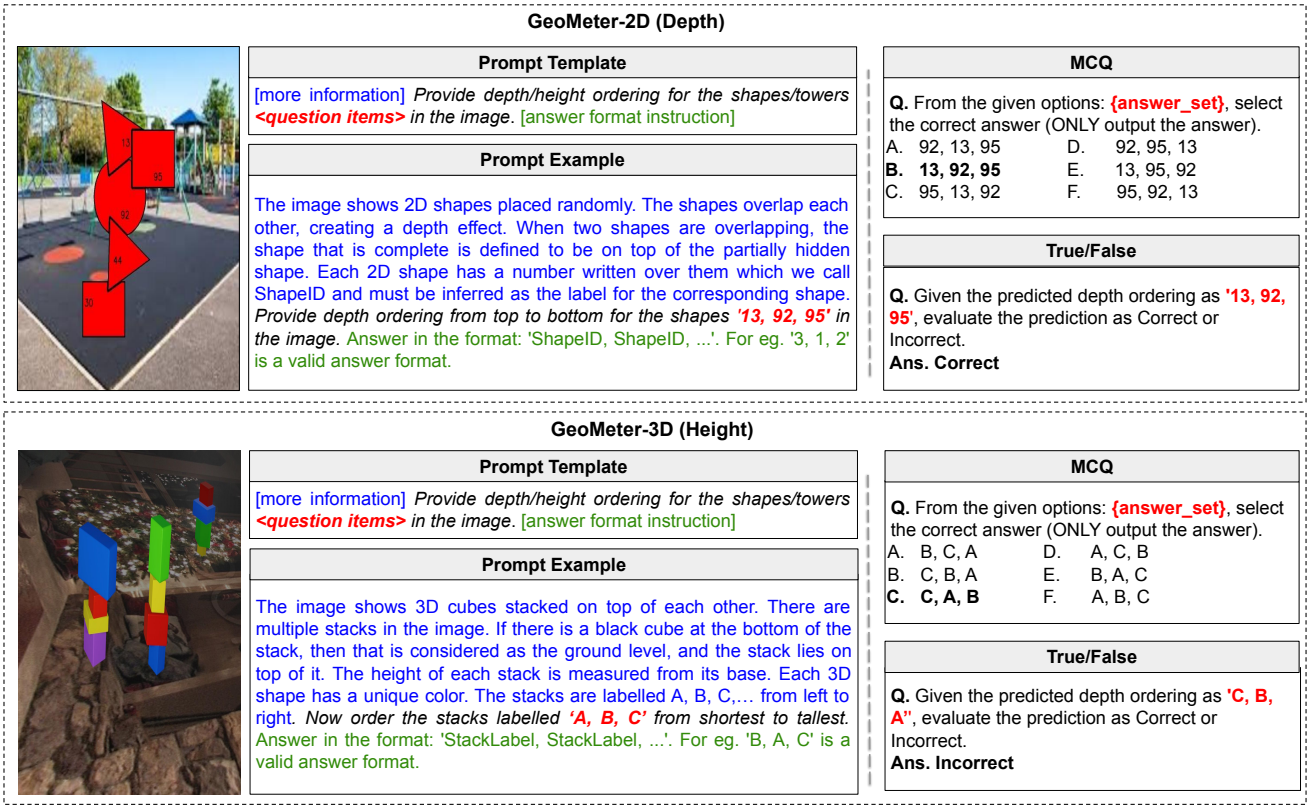

Figure 3. **Sample image-text pair from the datasets.** Here, prompt template shows the basic template for each image-text pair in our datasets, where the prompt example is the actual prompt for the image. The prompt example is appended with either MCQ or True/False type question.

ground-truth information required to evaluate depth perception tasks, such as object distances and relative positions. For the *Height* category, same as the GeoMeter-2D dataset's height category setup, we created scenes with 3 or 5 towers, each consisting of four cubes stacked on top of each other. We created a base tower mesh and randomized each cube's size, color, and material (either shiny "metal" or matte "rubber") for every image. Same as GeoMeter-2D, in some scenes, the bottom-most cube is black and matte, representing an elevated platform. The ground truth for these images is represented in the scene graph, detailing the exact size, position, and elevation of each tower.

3.1.2. Question Generation

The method used for generating questions is consistent across all our proposed datasets. Each question is a *Description prompt* appended with an *Answer format instruction*. The description prompt contains some general information about the scene providing semantic cues to the given image; followed by the actual question and answer format instruction. For example, *"[more information] Provide depth/height ordering for the shapes <question items> in the image. [more information]"* is a descriptive prompt. This is followed by *"From the given options: <answer set>, select the correct answer [more information]."* which is an answer format instruction.

The *question items* is a list containing *<query attribute>* appended by *<shape>*. Here *<query attributes>* is one of the unique identifiers of the dataset. For example in the question item *"green metal cube"*, *"green metal"* is the *<query attribute>* and *<cube>* is the shape. The *answer set* contains all possible valid values (*<query attribute>* + *<shape>*) to that given prompt. To generate both the question items and answer set, we read through the scene graph and run depth-first search on it to generate valid unambiguous values of object-pair relationship. For each image, there are two types of questions - MCQ and True/False. Some example prompts along with their corresponding image is shown in Figure 3.

4. Experimental Setup

4.1. Vision Language Models

We perform our benchmark evaluation on 18 state-of-the-art visual-language models. All of our chosen VLMs are trained on very large (public and/or proprietary) datasets. The selected VLMs can be categorized into 14 open-source and 4 closed-sourced models.

Table 2. **Performance comparison of the studied models on proposed datasets.** The reported results are averaged across depth and height category, query attributes and scene density with top scores in bold. Average denotes average performance of both datasets. Here, T/F denotes True/False type questions.

	Model	GeoMeter-2D		GeoMeter-3D		Average	
		MCQ	T/F	MCQ	T/F	MCQ	T/F
Open	LLaVA 1.5 7B [23]	**28.8**	50.5	28.0	49.8	28.4	50.2
	LLaVA 1.5 13B [23]	17.8	52.5	29.0	51.3	23.4	51.9
	LLaVA 1.6 Mistral 7B [25]	22.1	52.2	26.7	48.7	24.4	50.5
	LLaVA 1.6 Vicuna 7B [25]	17.1	51.7	28.6	50.0	22.9	50.9
	LLaVA 1.6 Vicuna 13B [25]	28.2	**54.2**	32.5	52.7	30.4	**53.5**
	Bunny-v1.0-3B [14]	24.1	50.1	17.1	37.1	20.6	43.6
	Bunny-v1.0-4B [14]	24.2	52.6	19.9	39.3	22.1	46.0
	Bunny-v1.1-4B [14]	26.6	52.3	26.9	44.4	26.8	48.4
	Bunny-Llama-3-8B-V [14]	27.9	50.2	26.9	43.2	27.4	46.7
	Fuyu-8B [4]	8.6	53.0	19.4	43.2	14.0	48.1
	InstructBLIP-Flan-T5-XL [10]	10.8	47.4	37.5	52.1	24.2	49.8
	InstructBLIP-Vicuna-7B [10]	28.3	49.0	38.1	53.8	**33.2**	51.4
	LLaMA-Adapter-v2-Multimodal [12]	22.9	48.8	32.7	52.4	27.8	50.6
	MiniGPT-4 [39]	25.0	50.4	**39.4**	**56.3**	32.2	53.4
Closed	GPT-4V [1]	25.5	54.0	35.2	50.5	30.4	52.3
	GPT-4o [1]	**30.8**	**56.7**	**38.5**	**52.4**	**34.7**	**54.6**
	Claude 3 Opus [2]	29.0	51.9	36.2	49.9	32.6	50.9
	Gemini 1.5 Pro [13]	28.8	54.5	36.5	51.0	32.7	52.8
	Human evaluators	**91.0**	**99.0**	**90.5**	**97.0**	**90.8**	**98.0**

Open-source models. We evaluate a range of open-source multimodal models: *LLaVA* [23, 24] and *LLaVA-NeXT* [25], that combine CLIP [28] as a visual encoder with Vicuna [7] language decoder; *Fuyu-8B* [4], an efficient model that integrates image patches directly into the transformer without an image encoder; *Bunny* [14], a versatile model family offering various combinations of vision encoders and LLM backbones; *InstructBLIP* [10] builds on BLIP-2 [20] for visual instruction tuning; *LLaMA-Adapter* [12], a parameter-efficient visual instruction model; and *MiniGPT-4* [39], which aligns BLIP-2 with Vicuna via a projection layer.

Closed-source models. We evaluated several closed-source multimodal models: *GPT-4* [27], which includes GPT-4V and GPT-4o variants, both pre-trained on large datasets and fine-tuned with RLHF [8]; *Claude 3 Opus* [2], demonstrating competitive performance; and *Gemini 1.5 Pro* [13] noted for surpassing GPT-4V across multiple benchmarks.

4.2. Human Evaluators

We conducted a preliminary human evaluation on our benchmark datasets, involving three evaluators who assessed 100 uniformly sampled data points from each subcategory. Evaluators were shown one image and prompt at a time, with tasks including multiple-choice and true/false questions focused on depth and height ordering (example illustrated in Figure 3). Their responses were compared to ground truth to calculate accuracy scores, establishing a baseline for human performance.

4.3. Evaluation Metrics

We evaluate our benchmark on the task of visual question answering (VQA), with accuracy being the performance metric on MCQ and True/False type questions. Evaluation is done across query attributes and scene density for probing the VLMs' depth and height perception.

4.4. Implementation Details

All models are used in accordance to the provided evaluation code and model weights. The closed-source models were accessed through APIs provided through a paywall by the corresponding developing teams. For MCQ, the order of the given options are randomly generated, and ground truth is always randomly placed in one of those options. We have implemented already established practices [26, 34] for creating options in multiple choice questions, randomizing both the position and the quantity of these options (up to 120 choices), and ensuring variability in the correct answer's location. For the True/False questions, the ground truth is randomly selected between True and False.

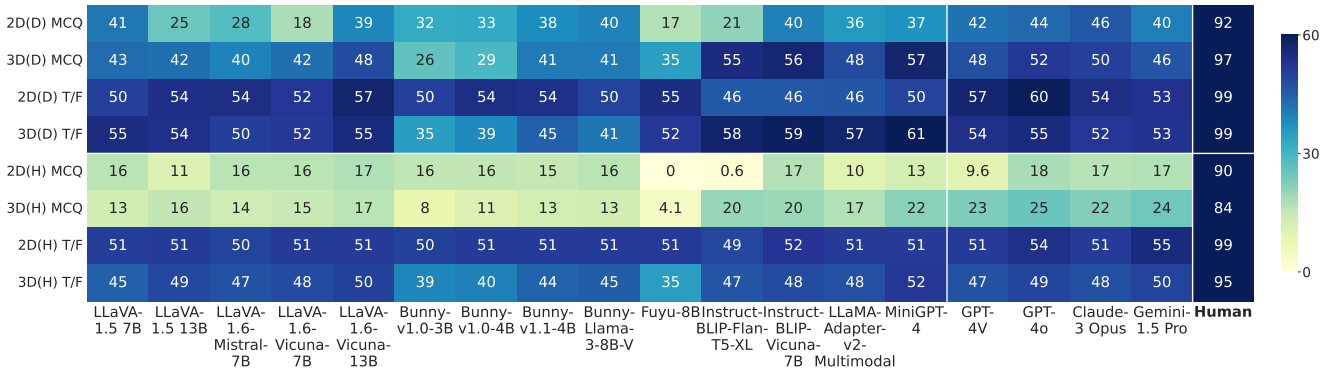

Figure 4. **Depth and height perception performance** on the proposed GeoMeter-2D and GeoMeter-3D dataset on MCQ and True/False (T/F) questions. D and H respectively denote depth, height performance. For example, 2D(D) MCQ and 2D(H) MCQ corresponds to respectively GeoMeter-2D depth and height performance on MCQ questions. Y-axis denotes the average performance across shape and query attributes and X-axis denotes all the evaluated models. Darker color denotes better performance.

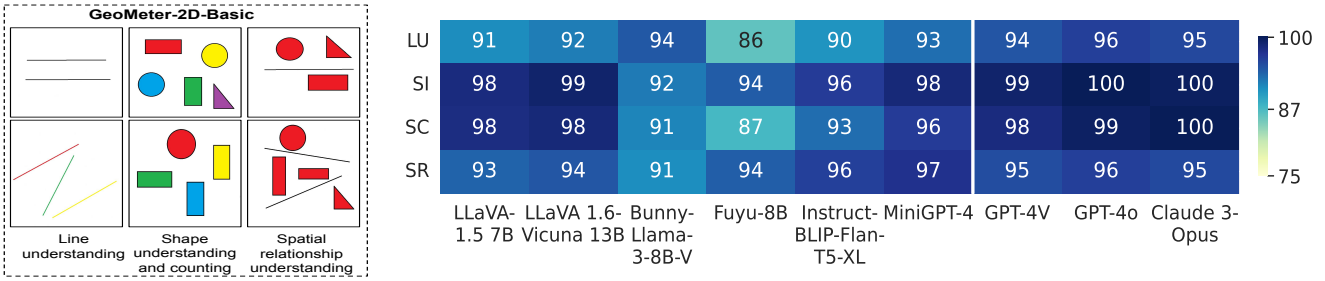

Figure 5. **Model behavior on basic understanding of shapes and size** on our created GeoMeter-2D-Basic dataset (samples on the *left*). Performance of selected models on this dataset is shown in *right*. Here, LU, SI, SC and SR respectively denote line understanding, shape identification, shape counting and spatial reasoning. Y-axis denotes performance accuracy of different categories and X-axis denotes evaluated models. Darker color denotes better performance.

4.5. Results

The performance of the selected models and human evaluators on the VQA task for MCQ and True/False type questions on the proposed benchmark datasets are shown in Table 2, where each model's performance represents the average accuracy of depth and height perception across all different query attributes and scene density. Depth and height category wise results are presented in Figure 4. Additional results across all query attributes and scene density are reported in the supplementary.

5. Analysis and Discussion

5.1. Model Behavior Analysis

Human evaluations confirm tasks are straightforward. Despite the seemingly straightforward nature of depth and height perception tasks for humans, current Vision Language Models (VLMs) struggle to achieve comparable performance. Our initial human evaluations on our datasets show consistently high accuracy in both depth and height perception tasks (Table 2, Figure 4), demonstrating that humans can effortlessly solve these tasks. In contrast, VLMs exhibit significant limitations. This performance discrepancy highlights that while these tasks may appear trivial from a human perspective, they pose substantial challenges for foundation models. Moreover, the human evaluation serves as a baseline, indicating that these tasks should be within the capability of an advanced AI system. This clear gap in model performance underscores critical limitations in VLMs' visual reasoning, revealing that the models are not yet equipped to handle even elementary geometric understanding without additional sensory input.

Models show basic visual reasoning capability but struggles in advance perception tasks. We developed a specialized dataset called *GeoMeter-2D-Basic* containing 30 image-text pairs (some samples shown in Figure 5 *left*) to evaluate the fundamental visual reasoning capabilities of Vision Language Models (VLMs). This dataset focuses on basic geometric tasks like line understanding, shape recognition, shape counting, and assessing spatial relationships between shapes. The initial assessments using MCQs demonstrate high performance by models on these basic tasks, as detailed in Figure 5 *right*. Despite this proficiency in simple visual properties, results from Figure 4 highlight

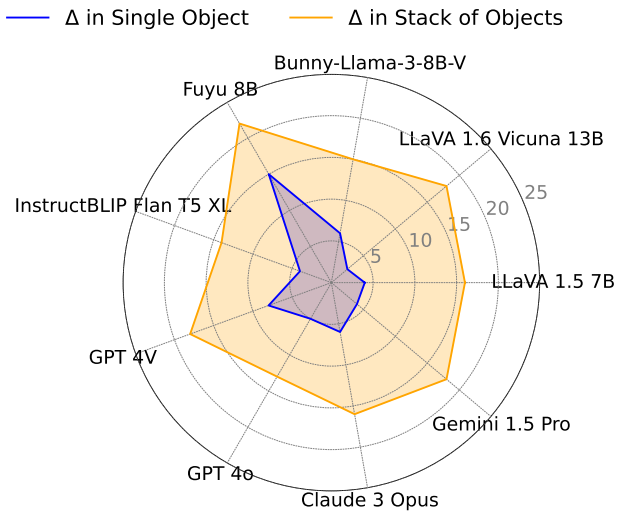

Figure 6. **Height perception is more challenging in stacked object arrangements than depth.** Here, Δ denotes performance gap between depth and height perception, which grows even larger with stacked arrangement of objects, as opposed to single objects. This suggests that while models struggle with height perception in general, stacked objects further degrade their performance.

that these same models struggle significantly with depth and height perception tasks involving similar shapes. This discrepancy underscores the benchmark's value in identifying gaps in VLMs' capabilities to handle more complex spatial reasoning, beyond mere shape recognition.

Height perception poses greater challenges than depth perception, especially in stacked object arrangements. The superior performance of models in depth perception tasks, as compared to height perception (Figure 4 row 1,2 vs row 5,6) is likely due to the availability of simpler depth cues—such as occlusion and perspective—in training datasets, which are relatively easy for VLMs to interpret. In contrast, we hypothesize height perception is more complex, requiring analysis of vertical object placement and relationships between object sizes in stacked arrangements. To further support our hypothesis, our analysis of single and stacked objects from the GeoMeter-3D dataset reveals (Figure 6) that while the performance gap between depth and height tasks is minor for single objects, there is a substantial decline in performance for height tasks with stacked objects. This pattern suggests that height perception, especially with multiple objects stacked vertically, poses a greater challenge for VLMs than depth perception.

Models' limitation is due to inherent reasoning capability and not insufficient prompt detail. To provide models with additional contextual information regarding visual cues with the help of intermediate reasoning, we implemented chain-of-thought prompting following [37]. Chain of thought prompting enhances problem-solving by guiding models through logical reasoning steps, similar to human

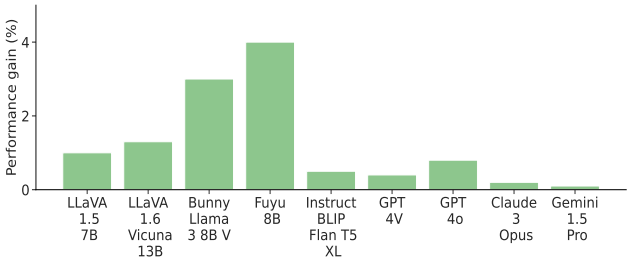

Figure 7. **Performance gain with chain of thought prompting over standard prompting** on subset of GeoMeter-3D dataset.

cognitive processes.

To evaluate its effectiveness, we selected a subset (100 image-text pairs) from the GeoMeter-3D dataset's depth category and created chain-of-thought prompts with intermediate reasoning steps. Testing top-performing models with these prompts showed only slight performance improvements (Figure 7), despite the highly detailed nature of these prompts (Figure 8). This marginal improvement suggesting that the models may already perform some internal reasoning with standard prompts. The findings indicate that limited depth and height perception performance is due to inherent model limitations in spatial understanding, underscoring the need for architectural advancements rather than solely relying on prompt engineering.

5.2. Model Bias Analysis

We conducted further analysis on the type of prompts to study any inherent biases in the models could be influencing their performance on MCQ and True/False type questions on a smaller subset (1600 image-text pairs uniformly selected from the depth and height categories) of the GeoMeter-3D dataset.

Some open-source models are more biased towards picking True over False than others. The performance of some open-source models on True/False questions tends to hover around 50% (Table 2), suggesting they might not be effectively distinguishing between true and false statements, potentially defaulting to random guesses. This is highlighted by experiments showing similar outcomes (Figure 9 *left*) when ground truth is random versus always set to "True," and a significant performance decline when it is always "False," indicating a bias towards predicting "True." This bias toward "True" may arise from imbalances in training data, where models are overexposed to affirmative statements or lack sufficient counterexamples of false statements. As a result, rather than demonstrating genuine understanding, these models often rely on heuristic patterns or shortcuts. Furthermore, this behavior highlights a deeper issue: the models' inability to engage in more nuanced decision-making or reasoning under uncertainty. True/False questions, though simple in format, test models' grasp of

Original Prompt

The image shows 3D shapes placed randomly. From the camera viewpoint distance some shapes are in front and some are at the back, creating a depth effect. Each 3D shape has a color and an associated material which we call Color and Material and must be inferred as the label for the corresponding shape. *Provide depth ordering from front to back for the shapes **'red cube, purple cylinder'** in the image.* Answer in the format: 'ShapeColor shape, ShapeColor shape, ...'. For eg. 'red cube, blue sphere, green cylinder is a valid answer format.

Chain of Thought Prompt

The image shows 3D shapes placed randomly. From the camera viewpoint distance some shapes are in front and some are at the back, creating a depth effect. Each 3D shape has a color and an associated material which we call Color and Material and must be inferred as the label for the corresponding shape.

Let's think step by step.
Step 1: Identify the Shapes and Their Colors: Observe the image carefully and list the 3D shapes along with their colors. For example: "I see a red cube, a purple cylinder, and a yellow sphere."
Step 2: Determine the Depth Ordering of the Shapes: Focus on the relative positions of the shapes from the camera viewpoint. Look for visual cues such as overlapping shapes, size differences due to perspective, and shadows. For example: "The red cube is in front of the yellow sphere. The purple cylinder is behind the yellow sphere."
Step 3: Provide the Depth Ordering from Front to Back: Based on the observations from Step 2, arrange the shapes in order from the closest to the furthest from the camera. For example: "The depth ordering from front to back is: red cube, yellow sphere, purple cylinder."
Final Answer: Format the final answer as specified in the prompt. For example: "red cube, purple cylinder"

*Provide depth ordering from front to back for the shapes **'red cube, purple cylinder'** in the image.* Answer in the format: 'ShapeColor shape, ShapeColor shape, ...'. For eg. 'red cube, blue sphere, green cylinder is a valid answer format.

Figure 8. **Prompt engineering using chain of thought prompting.** Here the intermediate reasoning steps introduced in the engineered prompts of the GeoMeter-3D dataset is denoted by a dashed box.

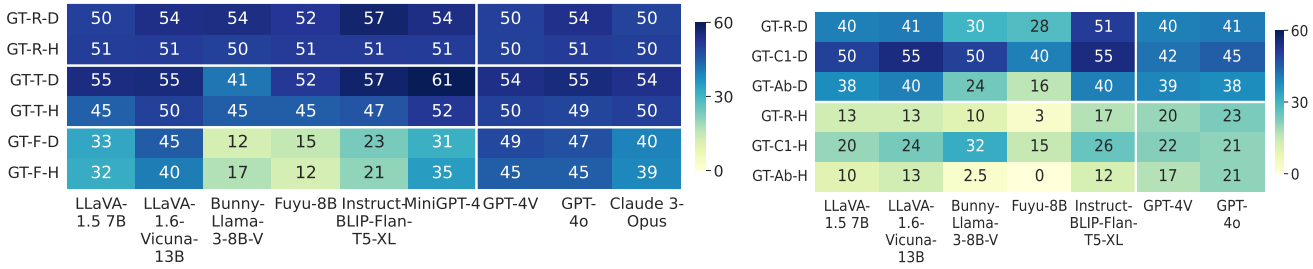

Figure 9. **Model bias analysis.** *Left:* Effect of ground truth value in True/False questions. GT-R denotes randomly set ground truth between true and false; whereas GT-T/F denotes ground truth always true or always false. *Right:* Effect of ground truth ordering in choices of MCQs. GT-C1 and GT-Ab denotes ground truth being choice 1 and not present respectively. The Y-axis denotes the average performance and X-axis denotes all the evaluated models. Darker colors denote better performance.

logical consistency and factual correctness - an area where many open-source models falter. By exposing such tendencies, this evaluation method provides valuable insight into where these models need refinement, particularly in developing the capacity for more context-driven predictions.

Some open source models are more biased towards picking the first choice in case of MCQ. Experiments reveal that while closed-source models show consistent performance across various MCQ ground truth placements, open-source models exhibit a significant bias toward selecting the first option, particularly when the ground truth is positioned as the first choice (Figure 9 *right*). This bias could stem from the way training data is structured, where the first choice is frequently correct or if the models encounter more examples with answers listed early in the sequence, leading models to develop a preference for selecting it. Their performance drops notably when the correct answer is absent, suggesting these models struggle with identifying "None of the above" options and may rely on heuristics rather than actual reasoning, leading to random selections. This reflects a limitation in their reasoning abilities, as they likely rely on pattern recognition rather than genuine understanding of the question and its context, which suggests that open-source models may lack sophisticated decision-making processes, opting for shortcuts when faced with challenging questions.

6. Conclusion

Our study highlights significant challenges in the depth and height reasoning capabilities of current Vision Language Models (VLMs). While these models demonstrate basic geometric understanding and spatial reasoning, they consistently struggle with more complex visual tasks, particularly depth and height perception, which remains underdeveloped. These shortcomings are not resolved by improved prompting alone, indicating an intrinsic limitation in the models' visual reasoning abilities. While our benchmark provides significant insights into the limitations of VLMs' perception abilities, it remains limited in scope. Expanding this benchmark to incorporate broader geometric reasoning, temporal dynamics, and higher-order reasoning tasks could offer a more comprehensive evaluation. Future work should focus on developing more targeted training strategies and benchmarks that address these perceptual weaknesses.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5

[2] Anthropic. Introducing claude, 2023. Accessed on 28 May 2024. 2, 5

[3] Shehreen Azad, Vibhav Vineet, and Yogesh Singh Rawat. Hierarq: Task-aware hierarchical q-former for enhanced video understanding. *arXiv preprint arXiv:2503.08585*, 2025. 2

[4] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. 5

[5] 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. *arXiv preprint arXiv:2401.12168*, 2024. 1

[6] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *arXiv preprint arXiv:2105.14517*, 2021. 3

[7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. 5

[8] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017. 5

[9] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 3

[10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 5

[11] Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. Why is winoground hard? investigating failures in visuolinguistic compositionality. *arXiv preprint arXiv:2211.00768*, 2022. 1, 2, 3

[12] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 5

[13] Google. Gemini: A family of highly capable multimodal models, 2024. 2, 5

[14] Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*, 2024. 2, 5

[15] Siyuan Huang, Haonan Chang, Yuhan Liu, Yimeng Zhu, Hao Dong, Peng Gao, Abdeslam Boularias, and Hongsheng Li. A3vlm: Actionable articulation-aware vision language model. *arXiv preprint arXiv:2406.07549*, 2024. 2

[16] Bargav Jayaraman, Chuan Guo, and Kamalika Chaudhuri. D\'ej\a vu memorization in vision-language models. *arXiv preprint arXiv:2402.02103*, 2024. 1

[17] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 1, 2, 3

[18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 2, 3

[19] Alexander Kuhnle and Ann Copestake. Shapeworld-a new test methodology for multimodal language understanding. *arXiv preprint arXiv:1704.04517*, 2017. 1, 3

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2, 5

[21] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, pages 237–255. Springer, 2025. 2

[22] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 1, 2, 3

[23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 2, 5

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2, 5

[25] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 5

[26] Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. *arXiv preprint arXiv:2405.12209*, 2024. 5

[27] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024. 2, 5

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 5

[29] Madeline Schiappa, Raiyaan Abdullah, Shehreen Azad, Jared Claypoole, Michael Cogswell, Ajay Divakaran, and Yogesh Rawat. Probing conceptual understanding of large visual-language models, 2024. 3

[30] Madeline Chantry Schiappa, Shehreen Azad, Sachidanand Vs, Yunhao Ge, Ondrej Miksik, Yogesh S Rawat, and Vibhav Vineet. Robustness analysis on foundational segmentation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1786–1796, 2024. 3

[31] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019. 2

[32] Kai Sun, Yushi Bai, and Nianyi Lin. Advancing geometric problem solving: A comprehensive benchmark for multimodal model evaluation. *arXiv preprint arXiv:2404.05091*, 2024. 3

[33] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 2

[34] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022. 5

[35] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 1, 3

[36] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 3

[37] 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. 7

[38] Jiaxin Zhang, Zhongzhi Li, Mingliang Zhang, Fei Yin, Chenglin Liu, and Yashar Moshfeghi. Geoeval: Benchmark for evaluating llms and multi-modal models on geometry problem-solving. *arXiv preprint arXiv:2402.10104*, 2024. 3

[39] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 5

[40] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023. 1