3DSRBench: A Comprehensive 3D Spatial Reasoning Benchmark

Wufei Ma Haoyu Chen[†] Guofeng Zhang Yu-Cheng Chou
Jieneng Chen Celso de Melo[°] Alan Yuille
Johns Hopkins University [†]Carnegie Mellon University [°]DEVCOM Army Research Laboratory

Abstract

3D spatial reasoning is the ability to analyze and interpret the positions, orientations, and spatial relationships of objects within the 3D space. This allows models to develop a comprehensive understanding of the 3D scene, enabling their applicability to a broader range of areas, such as autonomous navigation, robotics, and AR/VR. While large multi-modal models (LMMs) have achieved remarkable progress in a wide range of image and video understanding tasks, their capabilities to perform 3D spatial reasoning on diverse natural images are less studied. In this work we present the first comprehensive 3D spatial reasoning benchmark, 3DSRBench, with 2,772 manually annotated visual question-answer pairs across 12 question types. We conduct robust and thorough evaluation of 3D spatial reasoning abilities by balancing data distribution and adopting a novel FlipEval strategy. To further study the robustness of 3D spatial reasoning w.r.t. camera 3D viewpoints, our 3DSRBench includes two subsets with 3D spatial reasoning questions on paired images with common and uncommon viewpoints. We benchmark a wide range of open-sourced and proprietary LMMs, uncovering their limitations in various aspects of 3D awareness, such as height, orientation, location, and multi-object reasoning, as well as their degraded performance on images from uncommon 6D viewpoints. Our 3DSRBench provide valuable findings and insights about future development of LMMs with strong spatial reasoning abilities. Our project page is available here.

1. Introduction

Recent large multi-modal models (LMMs) [1, 4, 50] have achieved significant improvements in a wide range of image and video understanding tasks, such as image captioning [2, 34], visual question answering [23, 27, 38, 54], visual grounding [60], decision making [10, 32, 41], and action recognition [42, 59]. Notably, the spatial reasoning ability [16, 27, 29, 56], *i.e.*, parsing 2D and 3D spatial relationships between objects, serves as a crucial foundation for various high-level reasoning and interaction in downstream tasks. Studying the spatial reasoning ability of current LMMs will help us identify specific types of factual errors, uncover their fundamental limitations, and inform targeted improvements to further advance current LMMs.

Prior datasets [27, 29, 30, 35] studying spatial relationships often focused on relationships w.r.t. the viewer, *e.g.*, object A is to the left of object B from the viewer's perspective. We regard these as 2D spatial relationships as they can be captured merely from 2D bounding boxes of the objects (see Fig. 2b). They neglect 3D spatial relationships in the 3D world space or those from an object's perspective. Capturing 3D spatial relationships between objects in the images would help LMMs understand and predict the interactions between objects, and enable a broader range of applications in 3D, *e.g.*, robotics and embodied AI.

To study how LMMs can capture 3D spatial relationships, previous works often exploited synthetic environments and generated images with 3D ground-truths [57, 58]. Visual question-answer pairs were automatically synthesized by applying pre-defined rules to the known 3D scene graphs and object attributes. The synthetic images exhibit a significant domain gap with natural images and lacked the diversity and richness in real-world. More recent works [16] explored real datasets with 3D annotations, *e.g.*, Omni3D [9]. However, images in these datasets are limited to specific domains, such as indoor rooms and self-driving scenes. In general, visual question-answer pairs generated with rule-based methods from 3D annotations (i) limit the scope of theirs datasets to a small set of rigid object categories, and (ii) cannot enable a fine-grained and robust evaluation of 3D spatial relationships that can only be achieved with human annotated datasets (see Sec. 3.1).

In this work we present the first comprehensive 3D spatial reasoning benchmark, *3DSRBench*, that features 2,772 3D spatial reasoning questions from 12 question types on diverse and open-vocabulary entities, including rigid objects, humans, animals, and implicit concepts, such as logo on a car or arrow on a billboard. We manually annotate 2,100 visual question-answer pairs on natural images from the MS-COCO dataset [36], covering 12 subtypes of ques-

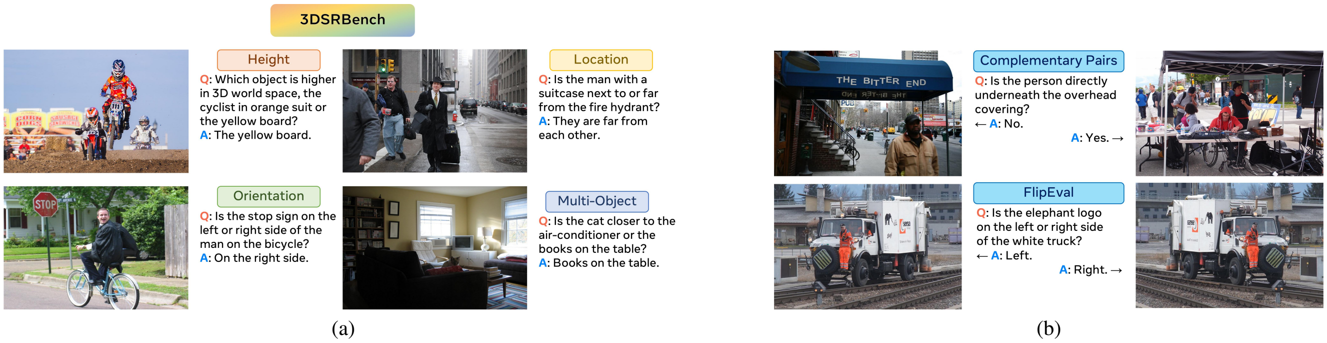

3DSRBench
Height
Q: Which object is higher in 3D world space, the cyclist in orange suit or the yellow board?
A: The yellow board.
Location
Q: Is the man with a suitcase next to or far from the fire hydrant?
A: They are far from each other.
Orientation
Q: Is the stop sign on the left or right side of the man on the bicycle?
A: On the right side.
Multi-Object
Q: Is the cat closer to the air-conditioner or the books on the table?
A: Books on the table.
(a)
Complementary Pairs
Q: Is the person directly underneath the overhead covering?
← A: No.
A: Yes. →
FlipEval
Q: Is the elephant logo on the left or right side of the white truck?
← A: Left.
A: Right. →
(b)

Figure 1. **Overview of our 3DSRBench. (a)** Example questions from the four main types of 3D spatial reasoning questions, *i.e*., height, location, orientation, and multi-object reasoning. **(b)** To enable a robust evaluation of the 3D spatial reasoning capabilities, we collect complementary images that lead to opposite answers given the same question and adopt a novel FlipEval strategy to remove left/right biases in 3D with paired VQAs (see Sec. 3.4).

tions from 4 main categories, *i.e*., height, location, orientation, and multi-object reasoning. Each category of questions focus on different combinations of 3D properties, such as object 3D location, 3D ground plane, camera extrinsic calibration, and/or object 3D poses. Examples from each question category are presented in Figure 1a.

Another challenge of 3D spatial reasoning arises from the 6D viewpoint of the camera, *i.e*., the 3D location and 3D orientation from which we are viewing the 3D scene. As shown in Fig. 3, 3D spatial reasoning questions can be easier for common 6D viewpoints, *e.g*., ones positioned at the eye level with natural viewing angles, while being more challenging for other uncommon viewpoints. Although uncommon viewpoints are less populated in most image datasets, cameras in embodied AI and robotics are often positioned in these uncommon viewpoints. Hence it is of crucial importance for LMMs to retain good 3D spatial reasoning performance for both common and uncommon viewpoints. To fairly compare the 3D spatial reasoning capabilities of LMMs w.r.t. different camera viewpoints, we annotate another 672 visual question-answer pairs on multi-view synthetic images rendered from the HSSD dataset [31].

Besides benchmarking a wide variety of open-sourced and proprietary LMMs, our 3DSRBench serves as an important diagnosis benchmark for developing 3D spatially intelligent LMMs. Inspired by previous studies on 3D awareness of visual foundation models [19, 43], our 3DSRBench takes one step further and evaluates LMMs on fundamental 3D spatial reasoning questions, which provide valuable insights regarding the 3D awareness of visual encoders [13, 24, 33, 47, 48] and the 3D reasoning abilities of language models [17, 18, 55, 61]. Such results would shed light on downstream tasks that build on 3D spatial reasoning, such as automatic navigation and robotic manipulation.

To enable a comprehensive and robust evaluation of 3D spatial reasoning abilities, 3DSRBench adopts several key designs: (1) balanced data distributions in multiple aspects, such as balanced answer distribution and complementary images pairs that lead to opposite answers given the same question (see Fig. 1b); (2) avoiding questions with shortcuts or trivial answers; and (3) a novel FlipEval strategy for robust evaluation of 3D spatial reasoning abilities.

Our 3DSRBench significantly advances the evaluation of 3D spatial reasoning abilities and provide valuable findings and insights about the future development of LMMs. We benchmark a wide variety of open-sourced and proprietary LMMs on 3DSRBench and study their 3D spatial reasoning abilities w.r.t. different types of 3D awareness. We further investigate how various visual encoder designs and scaling of language model sizes can benefit 3D spatial reasoning abilities. Moreover, with the paired images in 3DSRBench-`synthetic`, we analyze the robustness of 3D spatial reasoning abilities w.r.t. uncommon camera 6D viewpoints. Lastly, by analyzing failure modes of state-of-the-art LMMs, we highlight limitations of current LMMs and discuss possible future improvements. Experimental results on different splits of our 3DSRBench provide valuable findings and insights that will benefit future research on 3D spatially intelligent LMMs.

2. Related Works

Spatial reasoning. Early works [27, 29, 30, 35] studying spatial reasoning focused on spatial relationships w.r.t. the viewer, *e.g*. left/right relationships from the viewer's perspective. We regard these as 2D spatial relationships as they can be derived merely from 2D bounding boxes of the objects. To study how LMMs can perceive and understand 3D spatial relationships, previous datasets often adopted synthetic environments, *e.g*., Blender, with controllable simulation and 3D groundtruths for automatic question-answer generation [14, 57, 58]. However, synthetic images in these datasets exhibit a large domain gap with natural images and it remains unclear if insights and findings from these datasets would generalize to the real image domain. More recent works, such as SpatialRGPT [16] and Cambrian-1 [54], built on existing datasets with 3D annotations [8, 9, 11, 21, 51, 52] and generated visual question-answer pairs with pre-defined rules. Despite the improved image quality, they are essentially limited to a small number of rigid object categories in Omni3D [9] and the automatically generated VQAs are subject to shortcuts and biases. To enable a comprehensive and robust evaluation of the 3D spatial reasoning capabilities, we manually annotate visual question-answer pairs on diverse and open-vocabulary entities, such as logos on a car or arrows on the billboard, enforcing balanced data distributions in multiple aspects and avoiding questions with shortcuts or trivial answers.

3D awareness of visual foundation models. With the recent advancements in large multi-modal models [37–39], there has been a rising interest in applying these LMMs to a broader range of tasks, such as chatting about human poses [20], embodied question answering [46], and robotic manipulation [25, 26]. Notably, these tasks involve reasoning and interacting with the 3D scenes, which largely builds on the 3D awareness of vision encoders. Previous works studied the 3D awareness of visual foundation models by adopting proxy tasks, such as part correspondence [19] and pose estimation [43], and quantitatively evaluating the 3D awareness with linear probing. Our work can be considered as one step further — studying the 3D recognition and reasoning capabilities of LMMs by benchmarking their performance on fundamental 3D spatial relationship questions. Future research on downstream tasks, such as automatic navigation and robotic manipulation, could refer to the findings in our 3DSRBench and adopt LMMs with better 3D spatial reasoning capabilities.

3. 3DSRBench

In this section we introduce 3DSRBench for comprehensively analyzing the 3D spatial reasoning capabilities of LMMs. We start by presenting the design considerations in Sec. 3.1, *i.e.*, how these design choices lead to a robust and valuable evaluation of 3D spatial reasoning capabilities. Then we show the four main question types in Sec. 3.2, as well as the challenges in each type of questions. Next we introduce the three splits of 3DSRBench and their scopes in Sec. 3.3. In Sec. 3.4 we present our evaluation strategies, including CircularEval and FlipEval. Please refer to Sec. A in supplementary materials where we provide details of our data collection and summary statistics of 3DSRBench.

3.1. Design of 3DSRBench

When developing 3DSRBench, we incorporate the following four key designs to enable a robust and valuable evaluation of 3D spatial reasoning capabilities. **First, our 3D spatial reasoning questions are based on open-vocabulary entities.** Previous spatial reasoning benchmarks [12, 54] largely relied on existing datasets with 3D annotations [9], which limited their scope to a small number of rigid object categories. In our 3DSRBench, we annotate 3D spatial reasoning questions across a broad range of open-vocabulary entities (see Fig. 1), enabling a thorough analysis of the 3D awareness and 3D reasoning capabilities of LMMs over diverse, commonly encountered real-world objects. **Next, we avoid questions with shortcuts or trivial answers.** For instance, objects higher in 3D space are usually higher in 2D space. We collect diverse VQAs and avoid those with clear shortcuts (see Fig. 2a). Also, when comparing which of the two objects has a smaller 3D distance to a third anchor object, we avoid the cases when there is a significant gap between the two distances, which lead to trivial answers. **Moreover, we implement a balanced data distribution in various aspects,** such as a roughly same number of yes/no answers and complementary image pairs [23] that lead to opposite answers given the same 3D spatial reasoning question (see Fig. 1b). This effectively removes priors in the answer distribution, *e.g.*, pedestrians are often located lower than street lights, or the fact that objects higher in 3D space are also higher in 2D image plane. This design ensures that models cannot exploit biases or shortcuts for a higher benchmark performance. **Lastly, we adopt special evaluation strategies for robust evaluation,** including previous CircularEval [40] and our novel FlipEval (see Sec. 3.4).

3.2. Question Types

We present the 4 types of 3D spatial reasoning questions in our 3DSRBench. We discuss why they are challenging for LMMs and what kinds of 3D awareness and 3D spatial reasoning are needed to succeed in each type of questions. We present an overview of the 4 question types in Tab. 1.

Height questions. For height-related question, we study if models can determine which of the two given objects is positioned higher in the 3D world space. To correctly answer the questions, a model must (i) calibrate camera extrinsics, such as roll and pitch rotations, and then (ii) detect 3D locations of the objects in the 3D world space. This task poses a significant challenge for large multi-modal models as these fine-grained 3D knowledge are hard to derive from the weak language supervision in standard multi-modal pretraining. In Figure 2a we illustrate two examples of height questions. Notice how different pitch rotations of the camera, *i.e.*, viewing from above in the left figure and viewing upward in the right figure, play a crucial role to determine the final answer. In both examples, relying solely on the 2D locations within the image plane or the 3D locations in the camera coordinate system would lead to incorrect answers.

Type	# Subtypes	Camera	Loc.	Orient.	Reasoning
Height	1	✓	✓		+
Location	3		✓		+
Orientation	3		✓	✓	+
Multi-Object	5		✓	✓	++

Table 1. **Overview of the 4 main types of 3D spatial reasoning questions** and what kinds of 3D awareness and spatial reasoning are needed to answer each types of questions.

Location questions. There are three subtypes of location-related questions, *i.e*., determining (i) if two objects are next to or far from each other, (ii) which of the two objects is closer to the camera, and (iii) if an object is directly above or underneath another object. Models must not only ground the 2D locations of the objects, but also understand the depth of field presented in the image. Consider the location question in Fig. 1a. Although the 2D locations of the man and the hydrant are close, they are in fact far away from each other in the 3D space. Humans can determine the answer by estimating a rough depths of the two objects, or from other visual cues, such as how the pedestrian walk leads towards the vanishing point. Other examples include the top two questions in Fig. 1b, which also require an understanding of the depth field.

Orientation questions. Orientation-related questions study the 3D spatial reasoning that involves estimating the 3D orientation of an object. These questions are divided into three subtypes: determining which "side" of an object faces the camera, whether an object is in front of or behind another, and if an object is positioned on the left or right side of another. Unlike previous 2D spatial reasoning questions [12] that focus on spatial relationships w.r.t. the viewer's perspective, our orientation-related questions emphasize spatial relationships from the object's perspective. As demonstrated in Fig. 2b, 2D spatial reasoning questions can be addressed by analyzing objects' 2D locations and depths. Meanwhile, our orientation questions require estimating objects' 3D orientation and perform 3D spatial reasoning across various dimensions of 3D information.

Multi-object reasoning questions. Multi-object reasoning questions consider the 3D spatial relationships between multiple objects, such as asking which side of an object is facing another object, or with three objects, asking which of the given objects is facing towards or closer to the third object. These questions require more advanced 3D awareness than simpler 3D concepts such as "closer" (to the camera) or "higher", and require more complex 3D spatial reasoning, such as comparing distances between multiple objects from multi-step 3D computation.

3.3. Benchmark Splits

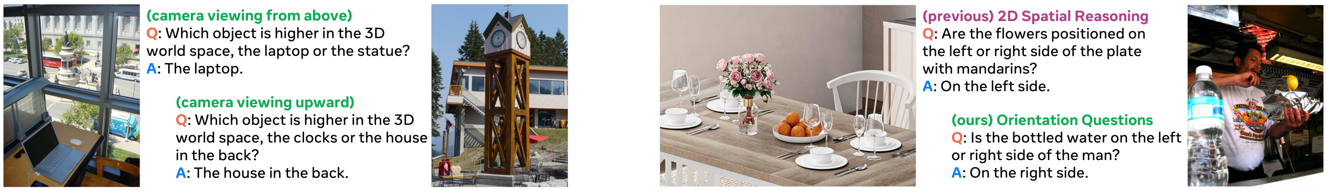
Our 3DSRBench is composed of three splits, a `real` split with 2,100 3D spatial reasoning questions on MS-COCO images [36] and two `synthetic` splits with 672 questions on synthetic images rendered with 3D scenes in HSSD [31]. We evaluate the standard 3D spatial reasoning capabilities of LMMs on visual question-pairs from the `real` split, and with the `synthetic` split, we study the robustness of 3D spatial capabilities w.r.t. common and uncommon camera 6D viewpoints by analyzing the gap between the `synthetic-common` and `synthetic-uncommon` splits.

With the HSSD 3D scenes and controllable photorealistic rendering, we obtain multi-view images of the same 3D scene, each rendered with a common and an uncommon viewpoint. We ask the same 3D spatial reasoning question regarding the two images and study if models can obtain the correct answers on common and uncommon camera 6D viewpoints. We define "common" viewpoints as 6D camera poses with zero roll rotation, small pitch rotation, and taken from the height of a human, simulating the typical perspective when people take pictures. Conversely, "uncommon" viewpoints include 6D poses with noticeable roll rotation, large pitch rotation, or perspectives taken close to the ground or from a high location. The two synthetic splits are denoted by `synthetic-common` and `synthetic-uncommon` and examples from the two splits are demonstrated in Fig. 3. Notice how the answers by GPT-4o are correct when shown the image from a common camera 6D viewpoint and wrong when prompted from an uncommon viewpoint, despite both images present a clear view of the 3D scene and humans can derive the correct answers without any difficulty.

3.4. Evaluation

Since all 3D spatial reasoning questions in 3DSRBench have two or four answer choices, we formulate these questions as multiple choice questions with two or four options. To accommodate the free-form answers predicted by pre-trained LMMs, we follow [40] and adopt LLM-involved choice extraction to obtain the predicted label. To enable a robust evaluation of various 3D spatial reasoning capabilities, we adopt the following two designs during testing:

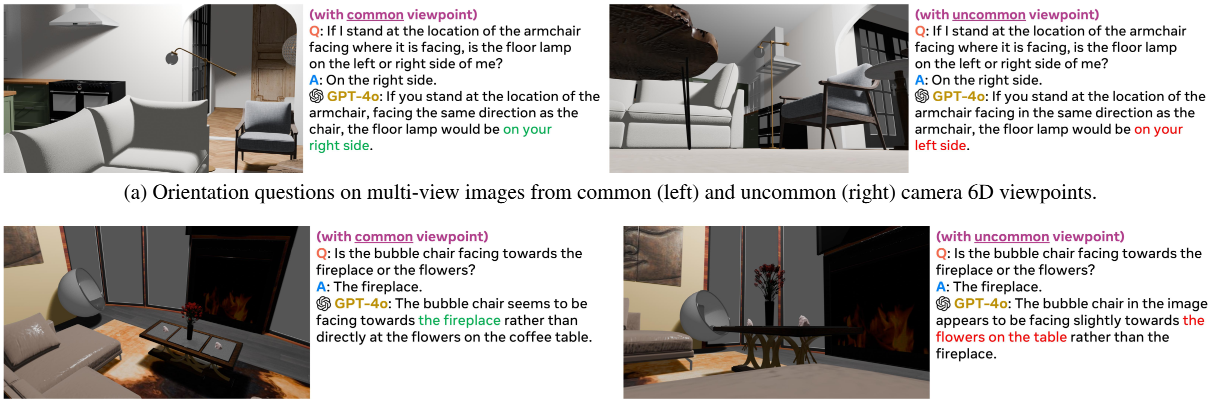
CircularEval [40]. To avoid the bias of choice ordering and the influence of random guessing for multiple choice questions, we adopt CircularEval [40] for more robust benchmark performance. Specifically we feed each question into the LMM two or four times, each with a different ordering of the answer choices. The LMM is considered successful in answering this question only if the predicted answer is correct for all passes.


(camera viewing from above)
Q: Which object is higher in the 3D world space, the laptop or the statue?
A: The laptop.
(camera viewing upward)
Q: Which object is higher in the 3D world space, the clocks or the house in the back?
A: The house in the back.
(previous) 2D Spatial Reasoning
Q: Are the flowers positioned on the left or right side of the plate with mandarins?
A: On the left side.
(ours) Orientation Questions
Q: Is the bottled water on the left or right side of the man?
A: On the right side.

(a) Height questions with different camera pitch rotations.

(b) Comparison between 2D and 3D spatial reasoning questions.

Figure 2. **Challenges of 3D spatial reasoning questions in our 3DSRBench. See Sec. 3.2. (a)** Height questions requires 3D spatial reasoning over a combination of camera extrinsics and object 3D locations. Notice how different camera pitch rotations play a crucial role to determine the final answer. **(b)** Previous 2D spatial reasoning questions can be addressed by analyzing objects' 2D locations and depths, while our orientation questions require complex 3D spatial reasoning on objects' 3D orientations and 3D locations.


(with common viewpoint)
Q: If I stand at the location of the armchair facing where it is facing, is the floor lamp on the left or right side of me?
A: On the right side.
GPT-4o: If you stand at the location of the armchair, facing the same direction as the chair, the floor lamp would be on your right side.
(with uncommon viewpoint)
Q: If I stand at the location of the armchair facing where it is facing, is the floor lamp on the left or right side of me?
A: On the right side.
GPT-4o: If you stand at the location of the armchair facing in the same direction as the armchair, the floor lamp would be on your left side.
(with common viewpoint)
Q: Is the bubble chair facing towards the fireplace or the flowers?
A: The fireplace.
GPT-4o: The bubble chair seems to be facing towards the fireplace rather than directly at the flowers on the coffee table.
(with uncommon viewpoint)
Q: Is the bubble chair facing towards the fireplace or the flowers?
A: The fireplace.
GPT-4o: The bubble chair in the image appears to be facing slightly towards the flowers on the table rather than the fireplace.

(a) Orientation questions on multi-view images from common (left) and uncommon (right) camera 6D viewpoints.

(b) Multi-object reasoning questions on multi-view images from common (left) and uncommon (right) camera 6D viewpoints.

Figure 3. **Examples of the paired visual question-answer data in our 3DSRBench-`synthetic`. (a)** Example questions from the four main types of 3D spatial reasoning questions. **(b)** To enable a robust evaluation of the 3D spatial reasoning capabilities, we collect complementary images that lead to opposite answers given the same question and adopt a novel FlipEval strategy (see Sec. 3.4).

FlipEval. Following the left-right biases discussed in [49], we further propose a novel *FlipEval* to remove left/right biases in 3D with paired visual question-answer pairs. By applying horizontal flip to an image, we obtain a new visual question. The answer would generally remain the same, such as for location and height questions, but when it involves 3D spatial relationships such as "left" and "right", the answer would change. We illustrate this idea in Fig. 1b, where the elephant logo is the on the left of the truck but changes to the right after image flipping. FlipEval effectively removes left/right biases in 3D spatial relationship, such as driver often sitting on the left side of the car or most people holding tools in their right hands. Lastly FlipEval also avoids the influence of random guessing and enriches the image distribution in our 3DSRBench.

4. Experiments

We first introduce our experimental settings in Sec. 4.1. Next in Sec. 4.2 we benchmark various LMMs on 3DSRBench. We further study how various model designs, *i.e.*, choice of visual encoders and scaling of language models, attribute to the 3D spatial reasoning abilities. Then we evaluate various LMMs on our 3DSRBench-`synthetic` and analyze the robustness of LMMs w.r.t. uncommon camera viewpoints in Sec. 4.3. Lastly we present some failure cases of GPT-4o and Gemini 2.0 in Sec. 4.4, highlighting limitations of current state-of-the-art LMMs and discussing possible future improvements.

4.1. Experimental Settings

With our 3DSRBench, we study: (1) standard 3D spatial reasoning abilities, by benchmarking various LMMs on 3DSRBench-`real` with VQAs on real images from MS-COCO [36], and (2) robustness of 3D spatial reasoning abilities w.r.t. uncommon camera viewpoints, by analyzing the performance gap between the two 3DSRBench-`synthetic` splits with common and uncommon viewpoints.

Testing data augmentation. We develop rule-based methods to augment the annotated visual question-answer pairs and obtain a larger number of testing data with a balanced and rich set of 3D spatial relationships. For instance, given a question asking which object is higher in the 3D world space, we generate a new question asking which

Model		3DSRBench-`real`				
		Overall	Height	Loc.	Orient.	Multi.
Baselines						
Random		20.9	25.0	25.0	16.8	20.1
Random++		45.8	50.0	50.0	41.7	45.0
Human		95.7	92.9	96.4	97.7	94.9
Open-sourced						
LLaVA-v1.5-7B [37]	13	38.1	39.1	46.9	28.7	34.7
Cambrian-1-8B [54]	11	42.2	23.2	53.9	35.9	41.9
LLaVA-NeXT-8B [39]	6	48.4	50.6	59.9	36.1	43.4
InternVL2.5-8B [15]	4	50.9	45.9	68.1	38.7	43.3
QWen2.5-VL-7B [7]	6	48.4	44.1	62.7	40.6	40.5
Specialist						
SpatialLLM [45]	9	44.8	45.8	61.6	30.0	36.7
SpatialRGPT [16]	14	32.7	55.9	39.0	27.8	20.0
SpatialRGPT *w/* depth [16]	6	48.4	55.9	60.0	34.2	42.3
SpatialReasoner [44]	1	60.3	52.5	75.2	55.2	51.8
Proprietary						
Claude-3.5V-Sonnet [4]	7	48.2	53.5	63.1	31.4	41.3
Gemini-2.0-Flash [22]	5	49.8	49.7	68.9	32.2	41.5
Gemini-2.0-Flash-bbox [22]	8	47.5	45.2	66.5	27.7	41.4
Gemini-2.0-Flash-think [22]	3	51.1	53.0	67.1	35.8	43.6
GPT-4o-mini [28]	12	39.7	44.3	52.4	21.0	36.5
GPT-4o [28]	10	44.2	53.2	59.6	21.6	39.0
QWenVLMax [7]	2	52.0	45.1	70.7	37.7	44.8

Table 2. **Experimental comparison of state-of-the-art large multi-modal models on our 3DSRBench.** Results show that state-of-the-art LMMs exhibit limited 3D spatial reasoning capabilities. Please refer to Sec. 4.2 for detailed analyses.

object is lower in the 3D world space. We further adopt FlipEval that augments the question set by horizontally flipping the images. This leads to a total of 5,250 questions on MS-COCO images, *i.e.*, 3DSRBench-`real`, and 1,692 questions on synthetic images, *i.e.*, 3DSRBench-`synthetic`.

Evaluation. To evaluate the correctness of free-form answers, we follow MMBench [40] and use exact matching to parse choice labels, or LLM-assisted evaluation, *e.g.*, with `gpt-4`, when matching fails. We further adopt CircularEval [40] that repeats a question N times, each with a different ordering of the choices. N is the number of choices.

4.2. Results on 3D Spatial Reasoning Abilities

We benchmark a wide range of open-sourced and proprietary LMMs on our 3DSRBench-`real` and analyze 3D spatial reasoning abilities on different types of questions. We consider three baseline results: (i) **random**: a simple baseline that predicts random answers for all visual questions. (ii) **random++**: a stronger random baseline that predicts consistent answers given different choice orders of a same visual question in CircularEval. (iii) **human**: a human-level performance established by human evaluators that did not participate in the data annotation process. We report the full results in Tab. 2.

We make the following observations: (i) **State-of-the-art LMMs have limited 3D spatial reasoning capabilities**, as found by low performance achieved by state-of-the-art open-sourced and proprietary LMMs, falling far behind human-level performance. (ii) **Scaling laws for LMMs are not effective for 3D spatial reasoning.** Results show that despite significant more training data and computation spent on the proprietary LMMs, they demonstrate limited advantages over open-sourced counterparts, featuring high-quality data with efficient training setups. Standard scaling laws demonstrate diminishing returns for 3D spatial reasoning abilities and we believe more effective approaches, *e.g.*, 3D-aware data, architecture, and training, would be necessary to significantly advance 3D spatial reasoning.

Design choices of visual encoder. We study how design choices of visual encoders can benefit 3D spatial reasoning abilities. Built on `LLaVA-v1.5-7B` [38], we experiment on a range of models with different choices of visual foundation models, *i.e.*, CLIP [48], MAE [24], DINOv2 [47], SAM [33], or model designs, *i.e.*, mixed encoders and visual projectors. Results in Tab. 3 show that with mixed encoders, DINOv2 can improve the overall 3D spatial reasoning abilities of LMMs, specifically for orientation and multi-object reasoning questions that build heavily on object 3D orientations. We also notice significant improvements for height questions when adopting MAE and SAM as vision encoder, suggesting that having richer visual features could help localize objects better. With spatial vision aggregator (SVA) [54], we can further improve the LMM with mixed encoder from 37.2% to 37.8%, demonstrating that fusing the semantic features with 3D-aware features from DINOv2 would benefit subsequent reasoning.

Scaling of language model size. We study how the scaling of language model, *i.e.*, in terms of the number of parameters, helps improve the 3D spatial reasoning abilities of LMMs. We consider two series of open-sourced LMMs, QWen2.5 [7] and InternVL2.5 [15], with a range of language model sizes from 0.5B to 72B. From the results in Fig. 4, we see that the scaling of language model sizes effectively improves the 3D spatial reasoning abilities of LMMs. Larger language models with more parameters exhibit enhanced reasoning abilities. They better capture 3D-aware information from the visual features and perform more complicated 3D spatial reasoning. However, given the importance of 3D spatial reasoning in a broad range of applications, scaling up language model size is highly inefficient — LMMs with over 70B parameters exceed the computation capacity of common robotics or embodied AI systems and significantly limit the model throughput.

4.3. Robustness to Uncommon Camera Viewpoints

We study the robustness of 3D spatial reasoning abilities w.r.t. common and uncommon viewpoints. We eval-

LLM	Vision Encoder	Connector	3DSRBench				
			Mean	Height	Loc.	Orient.	Multi.
Baseline							
Vicuna-v1.5-7B [37, 61]	CLIP-L14-336 [48]	2xMLP	36.8	38.5	**46.4**	27.7	31.8
Mixed Encoders							
Vicuna-v1.5-7B [37, 61]	CLIP-L14-336 [48] + DINOv2-L14-224 [47]	2xMLP	37.2	45.9	42.2	**28.7**	33.6
Vicuna-v1.5-7B [37, 61]	CLIP-L14-336 [48] + MAE-H14 [24]	2xMLP	33.1	42.7	39.2	26.1	27.5
Vicuna-v1.5-7B [37, 61]	CLIP-L14-336 [48] + SAM-L [33]	2xMLP	27.9	44.6	34.4	16.5	21.5
Connectors							
Vicuna-v1.5-7B [37, 61]	CLIP-L14-336 [48] + DINOv2-L14-224 [47]	SVA [54]	**37.8**	**46.0**	43.1	26.5	**35.9**
Vicuna-v1.5-7B [37, 61]	CLIP-L14-336 [48] + MAE-H14 [24]	SVA [54]	34.1	45.3	38.6	25.3	30.2

Table 3. **Experimental results on LMMs with various vision encoder setups.** We use `LLaVA-v1.5-7B` as the baseline model and studies how vision encoders with different features contribute to the final 3D spatial reasoning abilities of LMMs.

Model	3DSRBench-`synthetic-common`					3DSRBench-`synthetic-uncommon`					Rel. Drop
	Overall	Height	Loc.	Orient.	Multi.	Overall	Height	Loc.	Orient.	Multi.	δ
Open-sourced											
LLaVA-v1.5-7B [38]	42.0	40.0	50.6	20.8	47.6	38.0	41.0	43.6	17.9	45.2	-9.5%
Cambrian-1-8B [54]	48.1	37.5	56.1	39.6	47.6	39.9	35.0	45.7	29.2	41.9	-17.0%
LLaVA-NeXT-8B [39]	45.5	65.0	57.9	10.4	50.0	36.8	47.5	44.5	7.3	46.0	-19.1%
Proprietary											
Qwen-VL-Plus [6]	30.7	35.0	37.8	30.2	20.2	21.0	15.0	25.0	22.9	16.1	-31.6%
Qwen-VL-Max [6]	55.2	62.5	69.5	31.2	52.4	48.6	52.5	**59.8**	24.0	51.6	-12.0%
Claude-Sonnet [5]	47.4	47.5	58.5	26.0	49.2	39.4	**60.0**	48.2	16.7	38.7	-16.9%
Gemini-1.5-Flash [53]	44.6	57.5	59.8	13.5	44.4	37.7	42.5	45.7	11.5	46.0	-15.6%
Gemini-1.5-Pro [53]	**59.9**	65.0	69.5	**50.0**	**53.2**	**49.5**	42.5	52.4	**40.6**	**54.8**	-32.2%
GPT-4o-mini [28]	46.5	47.5	53.7	36.5	44.4	40.3	42.5	43.9	33.3	40.3	-13.3%
GPT-4o [28]	51.2	**70.0**	**70.1**	17.7	46.0	44.3	**60.0**	58.5	15.6	42.7	-13.5%

Table 4. **Experimental results on our 3DSRBench-`synthetic-common` and 3DSRBench-`synthetic-uncommon`.** We study the robustness of 3D spatial reasoning capabilities of LMMs by analyzing the performance gap between the two splits with images from the same 3D scene but from "common" and "uncommon" viewpoints. We find that LMMs does not generalize well to images with 6D camera viewpoints less represented in their training set. See Sec. 4.3 for detailed discussions.

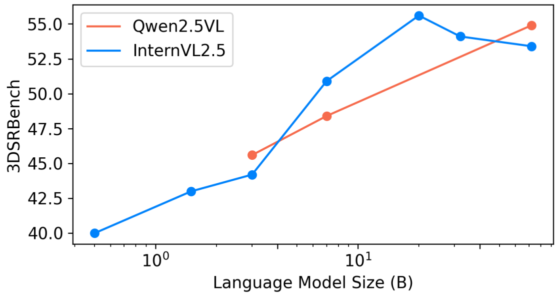


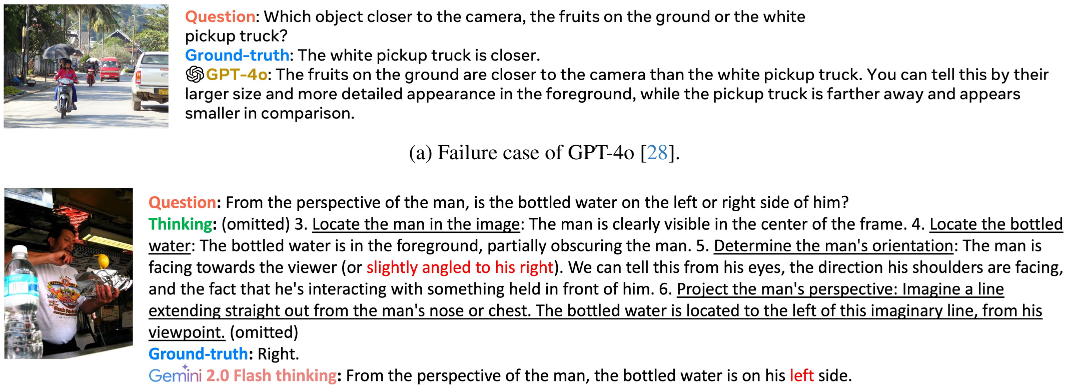
Figure 4. **Scaling of language model sizes.** Results show that scaling language model sizes can effectively improve the 3D spatial reasoning abilities of LMMs. However, with a 72B language model and a 6B vision encoder, InternVL2.5 still falls far behind human-level performance by more than 40%.

uate a variety of open-sourced and proprietary LMMs on our 3DSRBench-`synthetic-common` and 3DSRBench-`synthetic-uncommon` splits and analyze the relative performance drop, given by

$$\delta = \frac{\text{Acc}_{\text{uncommon}} - \text{Acc}_{\text{common}}}{\text{Acc}_{\text{common}}}$$

As demonstrated by the results in Tab. 4, all LMMs exhibit significantly degraded performance when generalizing from common to uncommon viewpoints, *e.g.*, a 13.5% drop in accuracy for GPT-4o [28], a 32.2% drop for Gemini-1.5-Pro [53], and a 19.1% drop for LLaVA-NeXT-8B [39]. We visualize two failure examples of GPT-4o in Fig. 3, showing how it is capable of predicting the correct answer when prompting with an image from a common viewpoint but fails when asked the same question with an image rendered from an uncommon viewpoint of the exact same scene.

We attribute such degraded performance in uncommon viewpoints to two factors: (i) *image domain gap due to different camera viewpoint distributions* between LMM training data and our 3DSRBench-`synthetic-uncommon`, and



Question: Which object closer to the camera, the fruits on the ground or the white pickup truck?
Ground-truth: The white pickup truck is closer.
GPT-4o: The fruits on the ground are closer to the camera than the white pickup truck. You can tell this by their larger size and more detailed appearance in the foreground, while the pickup truck is farther away and appears smaller in comparison.

(a) Failure case of GPT-4o [28].

Question: From the perspective of the man, is the bottled water on the left or right side of him?
Thinking: (omitted) 3. Locate the man in the image: The man is clearly visible in the center of the frame. 4. Locate the bottled water: The bottled water is in the foreground, partially obscuring the man. 5. Determine the man's orientation: The man is facing towards the viewer (or slightly angled to his right). We can tell this from his eyes, the direction his shoulders are facing, and the fact that he's interacting with something held in front of him. 6. Project the man's perspective: Imagine a line extending straight out from the man's nose or chest. The bottled water is located to the left of this imaginary line, from his viewpoint. (omitted)
Ground-truth: Right.
Gemini 2.0 Flash thinking: From the perspective of the man, the bottled water is on his left side.

(b) Failure case of Gemini 2.0 Flash thinking [22].

Figure 5. **Failures cases of GPT-4o [28] (top) and Gemini 2.0 Flash thinking [22] (bottom) on our 3DSRBench. (a) GPT-4o:** GPT-4o does not have an explicit 3D representation, *e.g.*, metric depth, and resort to visual cues to compare the distance, which leads to a wrong answer. **(b) Gemini 2.0 Flash thinking:** In this example Gemini 2.0 Flash thinking successfully breaks down the 3D spatial reasoning question into small and tractable steps. However, without explicit 3D representations, the model cannot perform reliable 3D spatial reasoning and predicts a wrong answer. See Sec. E in supplementary materials for more failure cases of the two models.

(ii) *state-of-the-art LMMs adopt an implicit representation of 3D scenes*. They are heavily built on the scaling law of data-driven approaches and lack explicit 3D representations that enable reliable 3D spatial reasoning. Despite the success of data-driven methods on a range of academic and empirical benchmarks, they face severe challenges generalizing to less represented data, which in our case, are images from uncommon camera 6D viewpoints.

These findings show that 3D spatial reasoning abilities of state-of-the-art LMMs are not robust to uncommon camera viewpoints. **This largely limits their applicability to various downstream applications in robotics and embodied AI.** Cameras mounted on robot arms or embodied AI systems are often positioned in uncommon locations and orientations as used in our study (see Fig. 3). On the one hand impressive advancements achieved by state-of-the-art LMMs in standard spatial reasoning benchmarks [27, 29, 35] may not generalize to downstream tasks; on the other hand, significantly degraded performance in uncommon viewpoints raises serious concerns about AI safety [3].

4.4. Failure Cases

We present two failure cases of GPT-4o [28] and Gemini 2.0 Flash thinking [22] in Fig. 5. In Fig. 5a we see that GPT-4o cannot perform rigorous 3D spatial reasoning and resort to various visual cues for reasoning. This is because GPT-4o lacks explicit 3D representations, *e.g.*, metric depth, that limits its ability to perform complex 3D spatial reasoning. In Fig. 5b, Gemini 2.0 Flash thinking successfully breaks down the 3D reasoning question into small and tractable steps. However, without explicit 3D representations, the model cannot perform reliable 3D spatial reasoning step-by-step. Despite the good thinking, the model fails to follow the planning and predicts a wrong answer.

We argue that for 3D spatial reasoning problems, models must not only have strong visual encoders to parse 3D-aware features, but also build a powerful reasoning model on various 3D information. Although scaling language model size leads to stronger reasoning abilities (see Fig. 4), a lack of explicit 3D representations would fundamentally limit models' abilities to solve complex 3D spatial reasoning questions that require multi-step 3D computations.

5. Conclusions

In this work we study the 3D spatial reasoning capabilities of LMMs. We introduce a new benchmark, 3DSRBench, by manually annotating 2,100 visual question-answer pairs on natural images from MS-COCO, featuring diverse and open-vocabulary entities and a balanced data distribution for robust evaluation. To study the robustness of 3D spatial reasoning capabilities w.r.t. camera 6D viewpoints, we further annotate 672 visual question-answer pairs on synthetic multi-view images, each with a common and an uncommon camera viewpoint. We benchmark a wide variety of open-sourced and proprietary LMMs on our 3DSR-Bench, studying various 3D spatial reasoning capabilities, *e.g.*, height, location, orientation, and multi-object reasoning, as well as the robustness of these LMMs to uncommon camera viewpoints. We also study how various designs of visual encoders and scaling of language models benefit 3D spatial reasoning. Experimental results on 3DSRBench provide valuable findings and insights to develop LMMs with strong 3D spatial reasoning abilities, as well as selecting LMMs for downstream applications that require robust 3D spatial reasoning.

Acknowledgements

We would like to thank Yiyan Li, Lizhi Ma, and the anonymous reviewers for their helpful comments and suggestions. Wufei Ma and Alan Yuille acknowledges support from ONR with N00014-23-1-2641 and ARL award with W911NF2320008.

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. 1

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022. 1

[3] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016. 8

[4] Anthropic. Claude 3.5 Sonnet. `https://www.anthropic.com/news/claude-3-5-sonnet`, 2024. 1, 6

[5] Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. Accessed: Dec 2024. 7

[6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 7

[7] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6, 1

[8] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 3

[9] Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3D: A large benchmark and model for 3D object detection in the wild. In *CVPR*, 2023. 1, 3

[10] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *CoRL*, 2023. 1

[11] 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. 3

[12] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024. 3, 4

[13] Jieneng Chen, Qihang Yu, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Vitamin: Designing scalable vision models in the vision-language era. In *CVPR*, 2024. 2

[14] Zhenfang Chen, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B Tenenbaum, and Chuang Gan. Comphy: Compositional physical reasoning of objects and events from videos. *arXiv preprint arXiv:2205.01089*, 2022. 2

[15] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 6

[16] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. In *NeurIPS*, 2024. 1, 3, 6

[17] 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)*, 2023. 2

[18] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2

[19] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *CVPR*, 2024. 2, 3

[20] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J. Black. Chatpose: Chatting about 3d human pose. In *CVPR*, 2024. 3

[21] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 3

[22] Google. Gemini, 2024. Accessed: Dec 2024. 6, 8, 1

[23] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 1, 3

[24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 2, 6, 7, 1

[25] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023. 3

[26] Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv preprint arXiv:2409.01652*, 2024. 3

[27] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 1, 2, 8

[28] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6, 7, 8, 1, 5

[29] 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, 8

[30] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's" up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*, 2023. 1, 2

[31] Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X Chang, and Manolis Savva. Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. In *CVPR*, 2024. 2, 4, 1

[32] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. In *CoRL*, 2024. 1

[33] 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 *ICCV*, 2023. 2, 6, 7, 1

[34] 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. 1

[35] Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14963–14973, 2023. 1, 2, 8

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 4, 5

[37] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 3, 6, 7, 1

[38] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 6, 7

[39] 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. 3, 6, 7

[40] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2025. 3, 4, 6, 1

[41] Taiming Lu, Tianmin Shu, Alan Yuille, Daniel Khashabi, and Jieneng Chen. Generative world explorer. *arXiv preprint arXiv:2411.11844*, 2024. 1

[42] Wufei Ma, Kai Li, Zhongshi Jiang, Moustafa Meshry, Qihao Liu, Huiyu Wang, Christian Häne, and Alan Yuille. Rethinking video-text understanding: Retrieval from counterfactually augmented data. In *ECCV*, 2024. 1

[43] Wufei Ma, Guanning Zeng, Guofeng Zhang, Qihao Liu, Letian Zhang, Adam Kortylewski, Yaoyao Liu, and Alan Yuille. Imagenet3d: Towards general-purpose object-level 3d understanding. *arXiv preprint arXiv:2406.09613*, 2024. 2, 3

[44] Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan Yuille. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. *arXiv preprint arXiv:2504.20024*, 2025. 6

[45] Wufei Ma, Luoxin Ye, Celso de Melo, Alan L Yuille, and Jieneng Chen. Spatialllm: A compound 3d-informed design towards spatially-intelligent large multimodal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2025. 6

[46] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Sasha Sax, and Aravind Rajeswaran. Openeqa: Embodied question answering in the era of foundation models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[47] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 6, 7, 1

[48] 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 *ICML*, 2021. 2, 6, 7

[49] Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S. Ryoo, and Tsung-Yu Lin. Learning to localize objects improves spatial reasoning in visual-llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12977–12987, 2024. 5

[50] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1

[51] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb,

and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*, 2021. 3

[52] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015. 3

[53] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 7

[54] 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. In *NeurIPS*, 2024. 1, 3, 6, 7

[55] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2

[56] Xingrui Wang, Wufei Ma, Zhuowan Li, Adam Kortylewski, and Alan L Yuille. 3d-aware visual question answering about parts, poses and occlusions. *Advances in Neural Information Processing Systems*, 36:58717–58735, 2023. 1

[57] Xingrui Wang, Wufei Ma, Zhuowan Li, Adam Kortylewski, and Alan L Yuille. 3d-aware visual question answering about parts, poses and occlusions. *NeurIPS*, 2024. 1, 2

[58] Xingrui Wang, Wufei Ma, Angtian Wang, Shuo Chen, Adam Kortylewski, and Alan Yuille. Compositional 4d dynamic scenes understanding with physics priors for video question answering. *arXiv preprint arXiv:2406.00622*, 2024. 1, 2

[59] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 1

[60] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023. 1

[61] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS*, 2023. 2, 7