SITE: towards Spatial Intelligence Thorough Evaluation

Wenqi Wang[1] Reuben Tan[2] Pengyue Zhu[1] Jianwei Yang[2]
Zhengyuan Yang[2] Lijuan Wang[2] Andrey Kolobov[2] Jianfeng Gao[2*] Boqing Gong[1*]
[1]Boston University, [2] Microsoft Research, Redmond
SITE-project-page

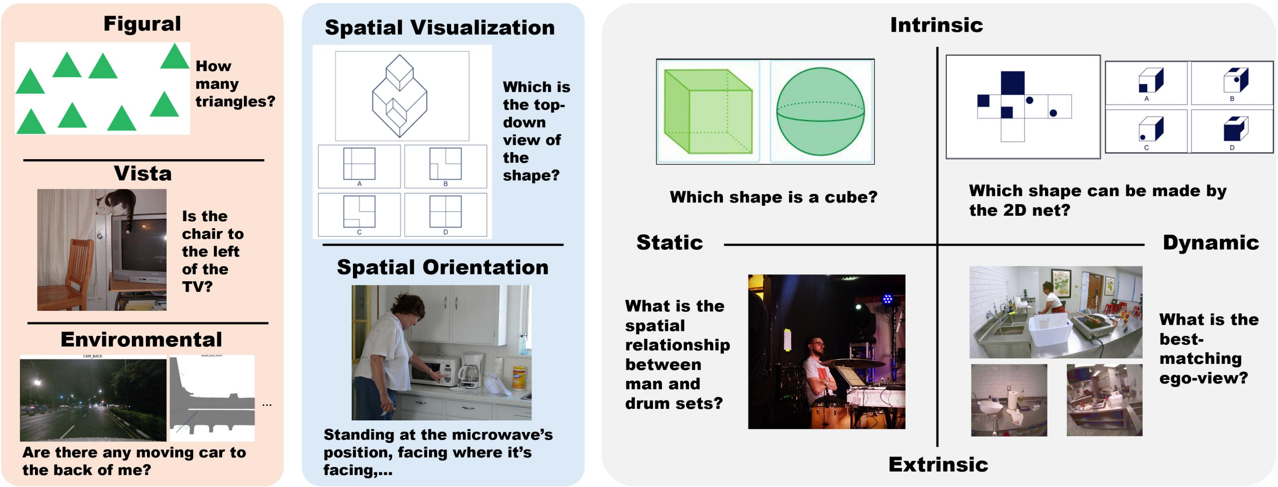
Figural
How many triangles?
Vista
Is the chair to the left of the TV?
Environmental
Are there any moving car to the back of me?
Spatial Visualization
Which is the top-down view of the shape?
A
B
C
D
Spatial Orientation
Standing at the microwave's position, facing where it's facing,...
Intrinsic
Which shape is a cube?
Which shape can be made by the 2D net?
Static
Dynamic
What is the spatial relationship between man and drum sets?
What is the best-matching ego-view?
Extrinsic

Figure 1. We introduce SITE, a comprehensive benchmark for evaluating large vision-language models' spatial intelligence (SI). Three SI classification systems drawn from cognitive science, corresponding to the three panels, drive the design of SITE.

Abstract

*Spatial intelligence (SI) represents a cognitive ability encompassing the visualization, manipulation, and reasoning about spatial relationships, underpinning disciplines from neuroscience to robotics. We introduce **SITE**, a benchmark dataset towards SI Thorough Evaluation in a standardized format of multi-choice visual question-answering, designed to assess large vision-language models' spatial intelligence across diverse visual modalities (single-image, multi-image, and video) and SI factors (figural to environmental scales, spatial visualization and orientation, intrinsic and extrinsic, static and dynamic). Our approach to curating the benchmark combines a bottom-up survey of existing datasets and a top-down strategy drawing upon three classification systems in cognitive science, which prompt us to design two novel types of tasks about view-taking and dynamic scenes. Extensive experiments reveal that leading models fall behind human experts, especially in spatial orientation, a fundamental SI factor. Moreover, we demonstrate a positive correlation between a model's spatial reasoning proficiency and its performance on an embodied AI task.*

1. Introduction

This work introduces **SITE**, a novel benchmark dataset towards Spatial Intelligence Thorough Evaluation, to assess the visuospatial ability of large vision-language models (VLMs). We achieve this by borrowing three classification systems about spatial intelligence (SI) from the cognitive science literature [8, 34, 47, 54, 56] to analyze vision-language tasks derived from 30 computer vision datasets. This process highlights a gap in existing benchmarks, leading us to design new tasks focusing on spatial orientation (view-taking) in static and dynamic contexts. We standardize all tasks to ease evaluation using a multiple-choice visual question answering (VQA) format.

SI represents a cognitive capacity encompassing the visualization, manipulation, and reasoning about spatial re-

[*] Equal advising

lationships [21]. Figure 1 shows some SI tests at different scales (left panel), that require spatial visualization independent of one's viewpoint and spatial orientation due to changes of viewpoints (middle), extrinsic, and dynamic (right). SI is essential for many professions, including architecture, engineering, and the arts, and SI is visual — some works in cognitive science use SI and "visuospatial abilities" interchangeably [62]. While the computer vision community has addressed some components of SI through tasks such as object detection [16, 22, 43, 73, 78], object referring [33, 51, 76, 79], 2.5D visual relationships [63], and counting/localization in VQA [4, 29, 52], progress has occurred mainly within individual domains or datasets. However, the advent of general-purpose large VLMs necessitates a unified testbed incorporating these diverse tasks and some uncaptured aspects of SI. This work presents such a benchmark and a systematic approach to building it.

The role of SI in AI models is highly tied to its role in human perception and reasoning. Kant argued that "space is the form of outer intuition", asserting that our perception of the physical world is structured by spatial relations between objects [32]. This concept is equally crucial for AI models, which have to develop SI to navigate and interact effectively in complex environments, especially with regard to tasks including but not limited to object manipulation [7, 13, 35, 65] and navigation [3, 24, 28, 61]. In this work, we specifically evaluate large VLMs. These models have demonstrated impressive capabilities in visual reasoning and question answering, thus positioning them as key components for perception and reasoning in embodied AI agents and robotics [35, 84]. However, their ability to perform fine-grained spatial reasoning remains relatively limited and only partially assessed. We expect this work will facilitate a comprehensive and holistic evaluation of VLMs across as broad a spectrum of SI as possible.

The proposed SITE benchmark poses a stark contrast to several similar efforts about SI. Table 1 summarizes the key differences. The visual part in SITE combines natural images, synthetic images, multiple views, and videos, while the existing ones contain only natural images [49, 66] or videos [75]. More importantly, these benchmarks measure limited aspects of spatial intelligence. For instance, CVBench [66] lacks viewpoint transformation. 3DSR-Bench [49] is limited to single-image questions, overlooking spatial reasoning in dynamic contexts. While VSI-Bench [75] uses videos to evaluate VLMs' capability to perform spatial reasoning across time, it is constrained to only indoor scenes. Finally, while counting and localization frequently appear in VQA benchmarks [4, 29, 46], tasks that require reasoning across multiple viewpoints remain largely unaddressed. These gaps manifest significant challenges in comprehensively evaluating VLMs' spatial intelligence.

To address the gaps, we approach the curation of SITE from two complementary paths: *Bottom-Up* and *Top-Down*. In the *bottom-up* path, we survey 30 representative datasets and systematically extract vision-language tasks after careful filtering. The filtering comprises two phases. We first prompt large language models (LLMs) using the language part of the tasks to reduce costs, and we then filter the surviving tasks by jointly screening their vision-language modalities. Finally, we identify six core categories from the tasks. This bottom-up approach gives rise to 6,943 tasks, including 3,135 image-based QA pairs and 3,808 video-based QA pairs. The *top-down* approach draws upon three classification systems of SI from the cognitive science literature, capturing SI's primary factors from different perspectives: scales (figural, vista, and environmental), view-taking (spatial visualization and orientation), intrinsic vs. extrinsic structures, and static vs. dynamic scenes. Investigation shows that the tasks resulting from the bottom-up path underrepresent the view-taking and dynamic factors, so we design two novel types of tasks using the Ego-Exo4D dataset [23], which is rich in camera views of dynamic events. SITE unions the bottom-up and top-down tasks and standardizes them for the ease of evaluations, with a total of 8,068 tasks, covering 30 existing benchmark datasets and 1 newly annotated dataset.

Through this systematic approach, we provide a compact and comprehensive benchmark to analyze VLM's spatial intelligence. We make the following key contributions.

1) Comprehensive Spatial Intelligence Benchmark: We systematically analyze existing datasets and benchmarks, extracting and pooling relevant tasks towards a comprehensive SI evaluation dataset that covers a broad range of visuospatial reasoning tasks.
2) Cognitive Science Inspired Taxonomy and New Tasks: We refer to not one but three SI classification systems grounded in cognitive science when building our dataset. The ample references reveal a need for view-taking and dynamic tasks, following which we design two novel types of tasks to close the gap.
3) Evaluation of Leading VLMs. We use SITE, the resulting benchmark dataset, to extensively assess state-of-the-art VLMs. Results show that existing VLMs especially struggle with spatial orientation tasks, falling significantly behind human performance.
4) Evident Correlation Between SI and Embodied AI. Finally, we empirically demonstrate that VLMs with high visuospatial ability also perform well on robot manipulation, with a correlation coefficient of 0.902.

2. Related Work

SI in cognitive science. SI has historically been studied as the interaction of multiple sensory modalities, including vision, touch, and hearing [14]. Out of these modalities, vision has been recognized as the dominant modality

	SITE (ours)	VSI-Bench [75]	3DSRBench [49]	CVBench [66]	SpatialEval [71]
Input	natural/synthetic image, video	video	image	image	image
Scale	figural, vista, environmental	environmental	vista	vista	figural, vista
Spatial Visualization	✓	✗	✗	✗	✗
Spatial Orientation	✓	✓	✓	✗	✓
Dynamic	✓	✓	✗	✗	✗
Intrinsic	✓	✗	✗	✗	✓

Table 1. SITE *vs.* similar efforts on benchmarking spatial intelligence. Besides being more diverse and comprehensive than existing datasets, SITE also introduces a structured classification system to better analyze spatial reasoning capabilities.

facilitating sensory integration [62], making visual-spatial ability the primary focus of mainstream spatial intelligence assessments. Psychologists and cognitive scientists have attempted to define and decompose spatial ability through factor-analytic studies [8, 25, 47, 54, 56, 68] based on numerous paper-and-pencil tests [15]. Several core factors have consistently emerged in these studies, including Spatial Visualization, Spatial Relations, and Spatial Orientation [47, 54, 56]. Later, Carroll et al. [8] introduced an additional factor: Visual Memory. Uttal et al. [68] and Hegarty and Waller [25] proposed new classification models for spatial intelligence. Due to variations in experimental paradigms and analytical methodologies, the definition of spatial ability remains highly inconsistent—a widely acknowledged consensus in cognitive science and psychology [69]. In Section 3, we will further elaborate on these concepts.

SI in computer vision. Spatial visual reasoning has long been an active research topic in computer vision. Initial efforts have mostly focused on constructing large-scale image-based datasets [4, 29, 31, 45], that include spatial visual reasoning tasks, to evaluateVQA approaches. Notable examples such as VQA [4], GQA [29], VSR [45], and CLEVR [31] incorporate spatial reasoning tasks, often evaluating a model's ability to reason about spatial relationships between objects within an image [4]. However, these datasets generally focus on relatively simple and straightforward tasks, such as verifying the correctness of spatial relationships between objects in an image [4]. This limitation negatively affects their ability to evaluate more complex aspects of spatial intelligence. Meanwhile, advancements in 3D vision and autonomous driving have significantly enriched spatial visual task datasets. Pioneering datasets such as ScanNet [12] and NuScenes [59] provide high-quality 3D annotations, thus facilitating the construction of spatial reasoning benchmarks [5, 9, 53, 59], which leverage multi-image and multi-view inputs to increase spatial task complexity. However, these benchmarks are still largely about static scenes. In contrast, our work on SITE aims to mitigate this limitation by providing a systematic evaluation of spatial intelligence at multiple scales and of both dynamic and static scenes.

SI for benchmarking VLM models. As mentioned earlier, many recent VLM benchmarks [18, 19, 42, 46, 80] have acknowledged the importance of spatial intelligence and included relevant evaluation questions. However, these benchmarks are often limited since spatial reasoning is generally treated as one among many tasks, scattered across other evaluation tasks of comprehension, reasoning, and perception, such as OCR and Math Reasoning. Existing image-based VLM benchmarks including MME [18], MMBench [46], SpatialEval [71] and CVBench [66] comprise tasks including but not limited to object counting, localization, and question answering about spatial relationships. The Blink [20] dataset is similar in nature to our SITE where it introduces evaluation tasks that involve spatial reasoning from multiple viewpoints and perspectives. Given the inherent difficulty in equipping VLMs with the capability to perform effective spatial reasoning, Cheng *et al.* [11] propose a new data curation pipeline that leverages 3D scene annotations as well as a module for integrating depth information into VLMs. Similarly, several video understanding benchmarks for VLMs also include tasks related to spatial reasoning. Notable examples of such datasets include MLVU [83], MVBench [42], and VideoMME [19]. However, similar to the image counterparts, these video benchmarks do not systematically isolate spatial intelligence as a core focus of their evaluations. Additionally, recent works such as 3DSRBench [49] and SpatialEval [71], heavily emphasize their evaluations on different aspects of spatial reasoning but are primarily limited to single-image evaluations. Similarly, VSI-Bench [75] incorporates video-based spatial tasks but remains constrained to indoor environments. Despite the contributions of these works, the introduced benchmarks generally do not explicitly evaluate VLMs' spatial reasoning abilities in a structured and comprehensive manner while our SITE aims to bridge this gap by evaluating spatial intelligence across multiple aspects and diverse visual context.

3. What Makes Spatial Intelligence (SI)?

We aim to cover various factors of SI comprehensively, and yet the first challenge we encounter is the lack of a consensus on the categorization of SI [69]. To the best of our

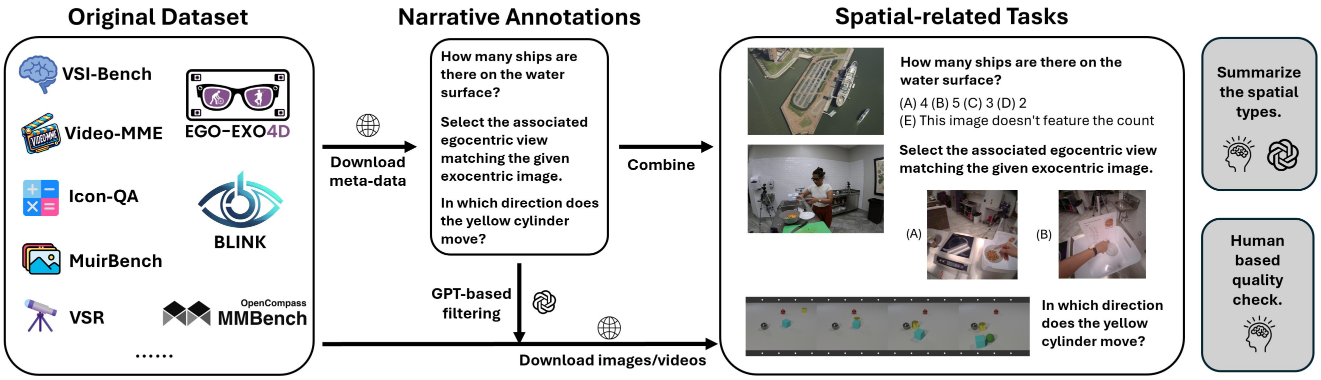
Original Dataset
VSI-Bench
Video-MME
Icon-QA
MuirBench
VSR
EGO-EXO4D
BLINK
OpenCompass MMBench
......
Download meta-data
Narrative Annotations
How many ships are there on the water surface?
Select the associated egocentric view matching the given exocentric image.
In which direction does the yellow cylinder move?
GPT-based filtering
Download images/videos
Combine
Spatial-related Tasks
How many ships are there on the water surface?
(A) 4 (B) 5 (C) 3 (D) 2
(E) This image doesn't feature the count
Select the associated egocentric view matching the given exocentric image.
(A)
(B)
In which direction does the yellow cylinder move?
Summarize the spatial types.
Human based quality check.

Figure 2. **Data collection pipeline for the *bottom-up* part of our benchmark.** We conduct a large-scale effort to select image and video-based benchmarks that may contain SI tasks before using the GPT-4o model to filter out irrelevant evaluation samples. Finally, we generate 6 coarse categories and perform stratified sampling to obtain an even distribution over all SI categories.

understanding, at least three major classification systems about SI exist in cognitive science, and we refer to them all when building our benchmark.

Figural, vista, and environmental SI. Hegarty *et al*. study SI at figural, vista, and environmental scales [27, 57] and note that "processing spatial information at different scales of space involves different brain structures and mechanisms," inspiring us to include visual inputs at all these scales. Figural space is small relative to the human body and can be apprehended from a single viewpoint. The following cognitive tests are in this space: mentally folding paper, transforming shapes, and solving mazes. Vista space, such as single rooms and town squares, is larger than the body and remains apprehensible from a single viewpoint. Environmental space contains an individual, referring to cities, neighborhoods, *etc*., which one must navigate to apprehend. Figure 1(left) presents examples at the three scales.

Spatial visualization & orientation. Hegarty and Waller compiled a summary of SI factors identified from primary cognitive studies [26], including spatial visualization (VZ), spatial orientation (SO), kinesthetic imagery [56], relational reasoning [47], visual memory [8], and others. Of these, VZ and SO are considered fundamental (see Figure 1(middle) for examples). Spatial visualization refers to the ability to mentally manipulate, rotate, or invert objects independently of one's own perspective, while spatial orientation involves imagining object appearances from differing observer viewpoints. It is important to note that these factors are derived from cognitive tests conducted prior to 2000, which primarily assess SI at figural and vista scales. It is also worth noting the extensive suite of cognitive tests about SI designed by Eliot and Smith [15].

2x2 classification. Uttal *et al*. categorize SI using a 2x2 classification system that relies on two distinctions, intrinsic vs. extrinsic and static vs. dynamic. Figure 1(right) demonstrates this system. Intrinsic information refers to an object's defining features, parts, and the relationships among the parts. Extrinsic information, conversely, concerns the spatial relationships between objects within a group or their relation to a broader framework. Static tasks are about fixed spatial information (*e.g*., counting chairs in a dining room), while dynamic tasks involve movement and transformation (*e.g*., mental rotation of a 3D shape).

In what follows, we employ all three categorization systems to drive our overarching design complementarily; meanwhile, we rely on them to balance and examine the coverage of the resulting benchmark from different aspects. Utilizing three rather than one categorization marks a significant difference between our work and most existing benchmarking efforts [18, 40, 42, 46, 49, 59, 75].

4. Data

We adopt a two-stage approach to constructing SITE. To begin, in Section 4.1, we compile a list of evaluation tasks from a suite of benchmarks that are focused on evaluating existing AI models' SI. Through a series of analysis and filtering steps, we construct a unified multiple-choice QA benchmark in a bid to standardize evaluations of SI in VLMs. Next, using the empirical analysis performed earlier as well as the classification framework introduced in Section 3, we identify two underrepresented aspects of SI that are largely unaddressed by existing spatial reasoning benchmarks. To address these gaps, we propose new evaluation tasks to enable a more comprehensive and holistic evaluation of VLMs' spatial reasoning capabilities in Section 4.2. Finally, we describe the curation of our final SITE dataset in Section 4.3.

4.1. Data Collection

Given the large number of isolated and fragmented evaluation benchmarks that focus on different aspects of SI, collecting a large-scale and comprehensive evaluation benchmark presents a significant challenge. With this challenge in mind, we propose a systematic data collection pipeline designed to filter and sample tasks relevant to SI in a structured manner, as Figure 2.

Data collection and filtering. To begin, we manually select a set of representative VQA datasets that potentially cover some aspects of SI (Figure 2). The initial collection of datasets comprises 22 image datasets and 8 video datasets, where examples of the former include VSR [45] as well as CV-Bench [66], and the latter include VSI-Bench [75] and VideoMME [19], respectively. We provide a detailed description of our selected benchmarks in the supplemental material. We note that we focus exclusively on the validation and test splits for each dataset. While the aforementioned datasets are generally large, they often only contain a subset of evaluation samples that are related to SI. Thus, it is necessary for us to filter out evaluation samples that are not relevant to the goal of evaluating spatial reasoning in VLMs. To filter out irrelevant evaluation samples, we adopt different strategies based on the existing annotations in each dataset. For datasets that contain category labels for the evaluation samples, we only retain evaluation samples that fall under categories relevant to spatial reasoning while discarding the rest. For datasets without predefined labels, we use a pretrained Large Language Model (LLM) to perform filtering of their constituent evaluation samples as described below.

LLM filtering. To further refine the quality and relevancy of the collected evaluation samples, we employ a filtering process by leveraging GPT-4o [2], a powerful LLM, to classify the evaluation samples (Figure 2). A major issue we faced in curating a comprehensive spatial reasoning benchmark is that different datasets use different labels for their evaluation samples, even though they might be evaluating similar aspects of SI. We tackle this issue by prompting the LLM to generate 6 coarse categories of tasks pertaining to SI. Specifically, we amass a set of original question labels from existing datasets and create a prompt for the LLM along with one to two sample datapoints as context for generating the task categories. The six coarse-level spatial intelligence categories are as follows: Counting and Existence (**Count.**), Spatial Relationship Reasoning (**Rel.**), Object Localization and Positioning (**Loc.**), 3D Information Understanding (**3D Inf.**), Multi-View Reasoning (**MultiV.**), and Movement Prediction and Navigation (**Mov.**). We proceed by classifying the valid spatial intelligence evaluation samples across all datasets into the abovementioned six categories.

Despite this categorization, many evaluation samples lack task-type labels or have noisy annotations. To refine the classification, we conduct a filtering stage using GPT-4o. We design a prompt template by incorporating key textual dataset information from each evaluation sample (e.g., questions, answers, options, and descriptions) along with carefully selected example data. The LLM is then queried to determine: (1) Whether the task pertains to spatial intelligence; (2) If so, which coarse category it falls under. We provide an example of the prompt template in the supplemental material.

Statistic	Number
Total questions	8,068
- 4-choice questions	5,019 (62.2%)
- 2-choice questions	1,573 (19.5%)
- 3/5/6-choice questions	1,476 (18.3%)
- Questions with annotations	6,943 (86.1%)
- Questions newly annotated	1,125 (13.9%)
Source datasets	31
- Existing image datasets	22
- Existing video datasets	8
- Our newly annotated datasets	1
Number of images	13,172
Number of videos	3,808

Table 2. **SITE benchmark statistics.**

Statistics and Reform QA types. After undergoing LLM-based filtering, the resulting dataset is reduced to 223,083 task examples, comprising 206,887 image-based and 16,196 video-based QA pairs relevant to spatial intelligence. We conduct a statistical analysis of the data across the coarse-level spatial categories and observe a significant data imbalance. Specifically, Relationship Reasoning (**Rel.**) and Counting and Existence (**Count.**) problems dominate the evaluation. In contrast, tasks such as Multi-View Reasoning (**MultiV.**) are underrepresented to a large degree. Additionally, due to the diverse data sources, the QA formats can vary considerably across different datasets. To ensure consistency and ease of evaluation, we standardize all tasks to a multiple-choice QA format, where we reformulate open-ended QA tasks accordingly.

4.2. Novel Proposed Tasks

To apply the cognitive science-based classification framework introduced in Section 3, we also conduct a fine-grained manual annotation of the filtered SI tasks. Following the classification workflow illustrated in Figure 2, we obtained the distribution of the collected dataset under different classification systems, as provided in the appendix. Our analysis reveals that there is a significant lack of tasks involving perspective transformations, which indicates that most existing SI evaluation tasks are constrained to reasoning from a fixed camera viewpoint. However, perspective transformation is especially essential for spatial reasoning in real-world scenarios such as navigation and route planning, where humans are naturally able to interpret spatial relationships from multiple viewpoints. To bridge this gap, we introduce two novel tasks specifically designed to evaluate **extrinsic-static** and **extrinsic-dynamic** spatial reason-

	View Association		Frames Reordering	
Model	ego2exo	exo2ego	ego2exo	exo2ego
Baseline & Upperbound on small subuset				
Random	0.0	0.0	0.0	0.0
Human performance	100	100	98	96
InternVL-2.5-8B	11.11	-5.56	-6.67	-6.67
GPT-4o	28.89	37.77	20.00	11.11
Open-source				
LLAVA-OneVision-0.5B	0.46	3.38	3.61	3.85
LLAVA-OneVision-7B	21.80	10.10	2.01	-4.41
Phi-3.5-Vision-4B	5.09	2.11	4.42	-0.28
InternVL-2.5-4B	1.85	-2.11	8.43	-4.79
InternVL-2.5-8B	-5.56	5.91	5.22	-0.66
QWen2.5-VL-3B	-0.93	0.84	3.61	-2.54
QWen2.5-VL-7B	5.09	-3.80	7.63	4.23
Proprietary				
Gemini-1.5-pro	15.30	-1.27	6.83	-4.04
GPT-4o	35.70	20.70	-2.01	-5.16

Table 3. **Evaluation on our proposed View Association and Frames Reordering tasks.** Dark red: Best among all models, light red: Best among open-source models.

ing under perspective-transformed conditions.

Data Preparation. To assess models' perspective transformation abilities in real-world scenarios, we leverage videos and annotations from the Ego-Exo4D [23] dataset. Ego-Exo4D provides 5,035 takes across eight real-world scenarios, where each take includes at least one egocentric view and multiple exocentric views captured synchronously. This multi-view setting provides a strong foundation for our task design, enabling a robust evaluation of spatial intelligence across diverse viewpoints.

Ego-exo View Association. As Figure 3 shows, this extrinsic-static task requires VLMs to associate egocentric and exocentric views of the same visual scene. Given an image with an exocentric viewpoint, a model must select the best-matching egocentric image from a set of candidates. Conversely, if an egocentric image is given, the model has to select the best-matching exocentric image. To construct this task, we utilize fine-grained key step annotations provided in the original dataset and sample multiple frames around each key step. Then, we rely on qualified human annotators to select challenging distractor frames.

Shuffled Frames Reordering. In this extrinsic-dynamic task, a model has to infer the correct temporal order from multiple viewpoints. Specifically, given a video clip, we extract the start and end frames from the egocentric video as reference points. Within this segment, we also sample 4 frames that capture key motion events from the exocentric views before shuffling them randomly. Finally, the model must predict the correct sequence by reasoning about motion dynamics across space and time. This task is also evaluated in reverse, by requiring models to predict the correct temporal order of egocentric frames based on the provided exocentric views. To ensure interpretability and feasibility of our proposed frame-reordering task, we also incorporate another round of human annotations to remove ambiguous and extremely difficult cases.

4.3. SITE

To ensure a balanced representation of various SI factors, we perform stratified sampling on the collected data from Section 4.1 to achieve an even distribution across different spatial categories while maintaining diversity. Including the two newly proposed tasks, our final benchmark consists of 8,068 QA pairs, including 4,260 image-based QA pairs and 3,808 video-based QA pairs, covering 30 existing benchmark datasets and the Ego-Exo4D dataset. Table 2 shows some statistics.

5. Experiments

5.1. Benchmark Evaluation

Evaluation Models. We evaluate 9 state-of-the-art VLMs that accept both image and video inputs, covering a diverse range of model architectures and parameter scales. From open-sourced models, we select LLAVA-OneVision [41], InternVL-2.5 [10], Qwen2.5-VL [6], and Phi-3.5V [1]. For proprietary models, we evaluate GPT-4o [2] and Gemini 1.5 [64]. To ensure standardized and reproducible evaluation, we utilize the lmms-eval [40] framework for benchmarking all models.

Evaluation Metrics. To ensure consistent and reliable evaluation of VLM-generated responses, we employ an LLM as part of an automated evaluation pipeline. In this study, we use GPT-4o for assessing model outputs. Since our benchmark follows a multiple-choice QA format with options of different lengths, the chance probability of guessing the correct options varies across questions. To mitigate this bias, we adopt a ***Chance-Adjusted Accuracy*** ($\mathcal{CAA}$) metric, which adjusts accuracy scores by accounting for the probability of random guesses, providing a more accurate measure of the model's true reasoning ability beyond chance:

$$\mathcal{CAA} = \left(\sum_{i=1}^{N} X_i - \sum_{i=1}^{N} \frac{1}{n_i} \right) / \left(N - \sum_{i=1}^{N} \frac{1}{n_i} \right) \quad (1)$$

where N, n_i, and X_i are the total number of questions, number of answer choices for the i-th multiple-choice question, and an indicator variable, respectively. If the model correctly answers the i-th question, we set $X_i = 1$. Otherwise, we set $X_i = 0$. $\mathcal{CAA} = 1$ when all predictions are correct ($X_i = 1$ for all i), indicating perfect performance; $\mathcal{CAA} = 0$ when the preditions perform no better than random guessing ($\sum X_i = \sum_{i=1}^{N} \frac{1}{n_i}$); $\mathcal{CAA} < 0$ reflects performace worse than random guess. This adjustment ensures

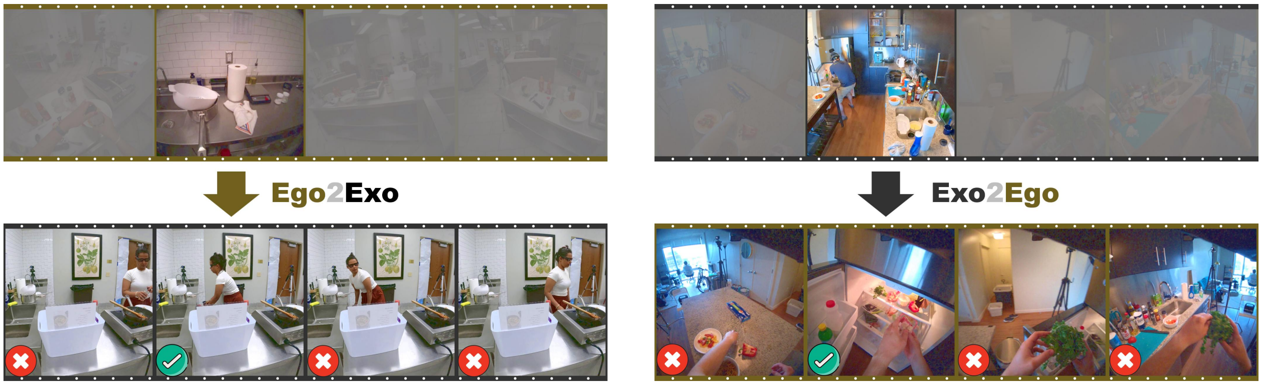

Figure 3. **Ego-Exo view association tasks.** The goal of this task is to pick the correct exocentric view given the egocentric view of a visual scene or vice versa.

Model	Overall	Count	Loc	3D Inf	MultiV	Rel	Mov
Random	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Tiny Subset							
Human	67.5	66.0	83.3	54.7	87.5	73.0	52.5
InternVL-2.5-8B	34.3	48.5	46.8	9.32	8.51	45.6	23.7
GPT-4o	35.6	42.4	51.2	11.0	17.8	42.7	19.5
Open-source							
LLAVA-OV-0.5B	18.4	28.0	32.3	5.67	3.77	30.6	4.70
LLAVA-OV-7B	30.2	51.8	38.5	22.4	9.40	55.3	9.18
Phi-3.5-Vision	21.8	33.2	34.0	11.7	3.33	32.8	11.7
InternVL-2.5-4B	29.4	47.9	32.9	11.4	3.94	47.2	22.9
InternVL-2.5-8B	32.8	47.1	37.0	23.2	9.05	47.6	28.7
Qwen2.5-VL-3B	29.5	45.6	37.5	13.2	7.14	45.6	18.8
Qwen2.5-VL-7B	31.4	52.6	44.1	9.42	1.08	51.5	18.9
Proprietary							
Gemini-1.5-Pro	32.5	48.0	45.8	25.3	5.33	48.8	18.4
GPT-4o	37.8	44.6	56.0	26.9	22.0	54.6	18.4

Table 4. **Performance comparison on the full SITE benchmark.** Dark red: Best, light red: Best among open-source models.

that model performance is evaluated as the improvement beyond random chance, providing a fairer comparison across questions with different numbers of answer options.

Baseline and Upper Bound. We use random chance as the baseline performance, where the expected score is **0**, following the $\mathcal{CAA}$ metric in equation 1. For the upper bound, we evaluate on six task categories, randomly sampling 50 QA pairs per task (see the **Tiny Subset** group in Table 4). We then conduct human performance evaluation by selecting 7 human participants to complete the benchmark. $\mathcal{CAA}$ scores from all participants are computed and averaged to derive the final upper bound performance.

5.2. Results on SITE

View Association. We begin by reporting the results of our evaluation on our proposed tasks of view association and frames reordering in Table 3. To benchmark the difficulty of these two tasks, we conduct a human study and the results demonstrate that humans are actually very adept at understanding and reasoning about 3D visual scenes from both egocentric and exocentric viewpoints. We randomly sampled 30 questions from each of the two views and found that human participants achieved perfect accuracy (100%). However, even state-of-the-art proprietary and open-sourced models such as GPT-4o and InternVL-2.5-8B achieve very low performances compared to the human level. This trend is also corroborated by the accuracy obtained by the abovementioned models on the full split of the view association task where GPT-4o and Gemini-1.5-pro achieve an average of 28.20% and 7.03%, respectively. Interestingly, while VLMs such as LLaVA-OneVision and Qwen2.5-VL have been demonstrated to perform well on conventional image and video question answering benchmarks like MME and VideoMME, their perception capabilities do not translate well to reasoning about spatial relationships and information beyond just perception. For instance, we observe that highly capable VLMs such as LLAVA-OneVision-7B and QWen2.5-VL-7B achieve very low average $\mathcal{CAA}$ of 15.95% and 0.65% on this view association task. We hypothesize that the lack of training data involving viewpoint transformations in large vision-language datasets—where most reasoning tasks rely directly on the camera's perspective—is the primary cause. The severe data imbalance observed during our dataset collection further supports this hypothesis.

Frames Reordering. Furthermore, we also observe a similar performance trend on our proposed task of temporal frames reordering (Table 3 right). Once again, the results show that humans are able to understand the temporal occurrence of events from different viewpoints, where they achieve close to perfect accuracy of 98% and 96% in the ego2exo and exo2ego directions, respectively. Interestingly, the GPT-4o model experiences a sharp drop in performance on this task, as compared to the task of view association. This might suggest that the GPT-4o model is not able to understand the mapping between different viewpoints of temporal events. Additionally, it is notable that the large-scale and proprietary models are severely underperforming open-sourced VLMs such as InternVL-2.5-8B and QWen2.5-VL-7B. In fact, the Qwen2.5-VL-7B model achieves the best average performance of 5.93%. One possible reason underlying this result is that the Qwen2.5-VL model is trained on video grounding data, which helps the model to learn a more effective understanding of time and consequently, the temporal order of events in videos.

Evaluation on the full SITE. We report our evaluation of state-of-the-art open-sourced and proprietary models in Table 4. For more detailed analysis, we break down the results of the evaluation across the six coarse categories of spatial intelligence tasks, as discussed in Section 3. To begin, the performance achieved by various open-sourced and proprietary models is consistent with our observations in Table 3. There is a large performance gap between the accuracy obtained by humans and state-of-the-art VLMs, which suggests that simply scaling up the amount of supervised fine-tuning (SFT) and instruction following multimodal data for pretraining may be inadequate in helping these VLMs to acquire effective spatial intelligence. It is also notable that humans only achieve an overall $\mathcal{CAA}$ score of 67.5%, which hints at the difficulty of our proposed SITE. Interestingly, humans perform significantly worse on counting(e.g., counting in a long video), 3D understanding(e.g., inferring the camera's transformation matrix), and movement prediction(e.g., navigating in a long video) as compared to the other three categories. This result might be due to humans' attention bottlenecks in tracking multiple objects [17], explaining why counting moving objects and tracking spatial transformations across time is very challenging.

However, the best-performing VLM GPT-4o still underperforms the human performance by $\sim$32%. The Qwen2.5-VL-7B and InternVL-2.5-8B models lead in overall accuracy with 31.4% and 32.8%, respectively. Despite containing much fewer model parameters, these open-sourced models perform competitively with GPT-4o and Gemini-1.5-Pro. Interestingly, we observe that Qwen2.5-VL-7B performs the best among all open-sourced VLMs on localization. This might be due to the pretraining recipe of Qwen2.5-VL-7B which also includes image and video grounding tasks [6]. We also see that multi-view reasoning is especially challenging for VLMs in general, where the performance obtained by GPT-4o is lower than that of human performance by over 70% on the tiny subset. On the full split, all of the state-of-the-art VLMs obtain $\mathcal{CAA}$ of less than 10%. One possible reason for the low performance is that these VLMs are generally not trained with different viewpoints for the same image or video.

From these empirical results, we also observe the benefits of using larger models. As evidenced by the consistent performance gains obtained by LLaVA-OneVision-7B and Qwen2.5-VL-7B over their smaller counterparts, VLMs with a higher number of parameters generally are able to perform spatial reasoning of visual scenes and environments much more effectively. However, it is notable that 3D understanding scores are consistently low across all VLMs, with most models scoring below 15%, indicating a persistent challenge in understanding depth(e.g., reasoning the depth relationships between objects) and three-dimensional spatial transformations(e.g., folding a 2D grid into a cube).

5.3. Spatial Intelligence on Downstream Tasks

Model	L2 Dist ↓	SR (%) ↑	CAA ↑
LLaVA-OneVision-0.5B	0.268 ± 0.241	0.0	18.4
LLaVA-OneVision-7B	0.142 ± 0.172	0.0	30.2
Qwen2.5-VL-3B	0.139 ± 0.153	0.0	29.5
Qwen2.5-VL-7B	**0.030** ± **0.040**	**38.0**	**31.4**

Table 5. **Correlation between SI and robotics manipulation on Libero Spatial.** The Pearson correlation coefficient between the negated mean L2 distance and CAA score is 0.902.

To understand why spatial intelligence is important, we conduct a toy experiment where we evaluate multiple VLMs of varying sizes on other real-world embodied tasks using the LIBERO-Spatial [44] dataset. Our goal is to analyze the relationship between performance on spatial intelligence benchmarks and a model's capability to perform well in real-world tasks. Thus, we fine-tune and evaluate both variants of the LLaVA-OneVision and Qwen2.5-VL model variants under the few-shot setting. Specifically, we only use 40-160 trajectories from each task in the spatial suite to train each model, but we do not notice much difference. We use a constant learning rate of 2e-5 and fine-tune each model for 30 epochs. In Table 5, we report the mean L2 distance between the final positions of the target object and robot arm effector across all episodes as well as the overall success rate. The negated mean L2 distance and CAA scores on our SITE benchmark across all VLMs have a positive Pearson Correlation Coefficient of 0.902. Notably, we also observe that the Qwen2.5-VL-7B model achieves a success rate of 38% while the others fail completely. These results suggest that pretraining data recipes and scale are both important for improving spatial intelligence in VLMs. We provide a more detailed correlation analysis in the appendix, comparing the impact of different benchmarks on embodied tasks. Importantly, these results indicate that AI agents have to possess a high degree of spatial intelligence to reason and interact effectively in the physical world.

6. Conclusion

In this work, we introduce SITE, a comprehensive benchmark focused on evaluating VLMs' ability to perform visuospatial reasoning. We pull tasks from 30 existing datasets and then design two novel types of tasks for view-taking and dynamic scenarios. Evaluation on SITE demonstrates a huge gap between humans and state-of-the-art VLMs. Moreover, we empirically demonstrate the positive correlation between the performance of VLMs on SITE and a robot manipulation task.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 6

[2] 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, 6

[3] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015. 2, 3

[5] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[6] 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, 8

[7] Homanga Bharadhwaj, Jay Vakil, Mohit Sharma, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4788–4795. IEEE, 2024. 2

[8] John B. Carroll. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge University Press, 1993. 1, 3, 4

[9] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 202–221. Springer, 2020. 3

[10] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198, 2024. 6

[11] 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. 3

[12] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 3

[13] Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396*, 2021. 2

[14] John Eliot. About spatial intelligence: I. *Perceptual and Motor Skills*, 94(2):479–486, 2002. 2

[15] John Eliot and Ian Macfarlane Smith. *An International Directory of Spatial Tests*. NFER-Nelson; Distributed in the USA by Humanities Press, Windsor, Berkshire; Atlantic Highlands, N.J., 1983. 3, 4

[16] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010. 2

[17] Steven L Franconeri, George A Alvarez, and Patrick Cavanagh. Flexible cognitive resources: competitive content maps for attention and memory. *Trends in cognitive sciences*, 17(3):134–141, 2013. 8

[18] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 3, 4, 1

[19] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 24108–24118, 2025. 3, 5, 1

[20] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision (ECCV)*, pages 148–166. Springer, 2024. 3, 1

[21] Howard E Gardner. *Frames of mind: The theory of multiple intelligences*. Basic books, 2011. 2

[22] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 2

[23] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19383–19400, 2024. 2, 6, 1

[24] Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Wang. Vision-and-language navigation: A survey of tasks, methods, and future directions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022. 2

[25] M Hegarty. A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, page 175–191, 2004. 3

[26] Mary Hegarty and D Waller. Individual differences in spatial abilities. *The Cambridge handbook of visuospatial thinking*, pages 121–169, 2005. 4

[27] Mary Hegarty, Daniel R. Montello, Anthony E. Richardson, Toru Ishikawa, and Kristin Lovelace. Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. *Intelligence*, page 151–176, 2006. 4

[28] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10608–10615. IEEE, 2023. 2

[29] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3, 1

[30] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. *arXiv preprint arXiv:1704.04497*, 2017. 1

[31] 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 (CVPR)*, pages 2901–2910, 2017. 3, 1

[32] Immanuel Kant. *Critique of Pure Reason*. Penguin Classics, 2003. 2

[33] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 2

[34] Hyunjae Kim, Yookyung Koh, Jinheon Baek, and Jaewoo Kang. Exploring the spatial reasoning ability of neural models in human iq tests. *Neural Networks*, 140:27–38, 2021. 1

[35] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. In *Conference on Robot Learning (CoRL)*, pages 2679–2713. PMLR, 2025. 2

[36] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018. 1

[37] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*, 2023. 1

[38] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[39] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024. 1

[40] Bo* Li, Peiyuan* Zhang, Kaichen* Zhang, Fanyi* Pu, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Accelerating the development of large multimoal models. *arXiv preprint arXiv:2407.12772*, 2024. 4, 6

[41] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6

[42] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206, 2024. 3, 4, 1

[43] 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 *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 2

[44] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:44776–44791, 2023. 8

[45] Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023. 3, 5, 1

[46] 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 *European Conference on Computer Vision (ECCV)*, pages 216–233. Springer, 2024. 2, 3, 4, 1

[47] David F. Lohman. Spatial abilities as traits, processes, and knowledge. In *Advances in the Psychology of Human Intelligence*, pages 181–248. Lawrence Erlbaum Associates, Inc., 1988. 1, 3, 4

[48] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*, 2021. 1

[49] Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Celso M de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. *arXiv preprint arXiv:2412.07825*, 2024. 2, 3, 4, 1

[50] 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. 1

[51] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer vision and Pattern Recognition (CVPR)*, pages 11–20, 2016. 2

[52] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3195–3204, 2019. 2

[53] Damiano Marsili, Rohun Agrawal, Yisong Yue, and Georgia Gkioxari. Visual agentic ai for spatial reasoning with a dynamic api. *arXiv preprint arXiv:2502.06787*, 2025. 3

[54] Mary G. McGee. Human spatial abilities: Psychometric studies and environmental, genetic, hormonal, and neurological influences. *Psychological Bulletin*, 86(5):889–918, 1979. 1, 3

[55] Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, et al. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *arXiv preprint arXiv:2408.02718*, 2024. 1

[56] William B. Michael, J. P. Guilford, Benjamin Fruchter, and Wayne S. Zimmerman. The description of spatial-visualization abilities. *Educational and Psychological Measurement*, 17(2):185–199, 1957. 1, 3, 4

[57] Daniel R Montello. Scale and multiple psychologies of space. In *European conference on spatial information theory*, pages 312–321. Springer, 1993. 4

[58] Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. Synthesize diagnose and optimize: Towards fine-grained vision-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13279–13288, 2024. 1

[59] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 4542–4550, 2024. 3, 4

[60] Arijit Ray, Jiafei Duan, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. Sat: Spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024. 1

[61] Dhruv Shah, Błażej Osiński, Sergey Levine, et al. Lmnav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on robot learning (CoRL)*, pages 492–504. PMLR, 2023. 2

[62] Priti Shah and Akira Miyake. *The Cambridge handbook of visuospatial thinking*. Cambridge University Press, 2005. 2, 3

[63] Yu-Chuan Su, Soravit Changpinyo, Xiangning Chen, Sathish Thoppay, Cho-Jui Hsieh, Lior Shapira, Radu Soricut, Hartwig Adam, Matthew Brown, Ming-Hsuan Yang, et al. 2.5 d visual relationship detection. *Computer Vision and Image Understanding*, 224:103557, 2022. 2

[64] 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. 6

[65] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024. 2

[66] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:87310–87356, 2024. 2, 3, 5, 1

[67] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*, 2024. 1

[68] David H. Uttal, Nathaniel G. Meadow, Elizabeth Tipton, Linda L. Hand, Alison R. Alden, Christopher Warren, and Nora S. Newcombe. The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin*, page 352–402, 2012. 3

[69] David H Uttal, Nathaniel G Meadow, Elizabeth Tipton, Linda L Hand, Alison R Alden, Christopher Warren, and Nora S Newcombe. The malleability of spatial skills: a meta-analysis of training studies. *Psychological bulletin*, 139(2): 352, 2013. 3

[70] Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024. 1

[71] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:75392–75421, 2025. 3, 1

[72] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*, 2023. 1

[73] Yu Xiang, Wonhui Kim, Wei Chen, Jingwei Ji, Christopher Choy, Hao Su, Roozbeh Mottaghi, Leonidas Guibas, and Silvio Savarese. Objectnet3d: A large scale database for 3d

object recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 160–176. Springer, 2016. 2

[74] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024. 1

[75] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 10632–10643, 2025. 2, 3, 4, 5, 1

[76] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4683–4693, 2019. 2

[77] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024. 1

[78] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2636–2645, 2020. 2

[79] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision (ECCV)*, pages 69–85. Springer, 2016. 2

[80] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *International Conference on M achine Learning (ICML)*. PMLR, 2024. 3, 1

[81] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134, 2019. 1

[82] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024. 1

[83] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 3, 1

[84] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning (CoRL)*, pages 2165–2183. PMLR, 2023. 2