Mind the Gap: Benchmarking Spatial Reasoning in Vision-Language Models

Ilias Stogiannidis, Steven McDonagh, Sotirios A. Tsaftaris
The University of Edinburgh
Edinburgh, UK
{i.stogiannidis, s.mcdonagh, s.tsaftaris}@ed.ac.uk

Abstract

Vision-Language Models (VLMs) have recently emerged as powerful tools, excelling in tasks that integrate visual and textual comprehension, such as image captioning, visual question answering, and image-text retrieval. However, existing benchmarks for VLMs include spatial components, which often fail to isolate spatial reasoning from related tasks such as object detection or semantic comprehension. In this paper, we address these deficiencies with a multi-faceted approach towards understanding spatial reasoning. Informed by the diverse and multi-dimensional nature of human spatial reasoning abilities, we present a detailed analysis that first delineates the core elements of spatial reasoning: spatial relations, orientation and navigation, mental rotation, and spatial visualization, and then assesses the performance of these models in both synthetic and real-world images, bridging controlled and naturalistic contexts. We analyze 13 state-of-the-art Vision-Language Models, uncovering pivotal insights into their spatial reasoning performance. Our results reveal profound shortcomings in current VLMs, with average accuracy across the 13 models approximating random chance, highlighting spatial reasoning as a persistent obstacle. This work not only exposes the pressing need to advance spatial reasoning within VLMs but also establishes a solid platform for future exploration. Code available on GitHub[1] and dataset available on HuggingFace.[2]

1. Introduction

Vision-Language Models (VLMs) have recently emerged as powerful tools, excelling in tasks that integrate visual and textual comprehension such as image captioning, visual question answering, and image-text retrieval [18, 33]. Developed through extensive pretraining on diverse datasets, these models have demonstrated a remarkable ability to interpret complex interactions between visual information and language [47]. However, despite these advances, a critical capability remains largely unaddressed: spatial reasoning.

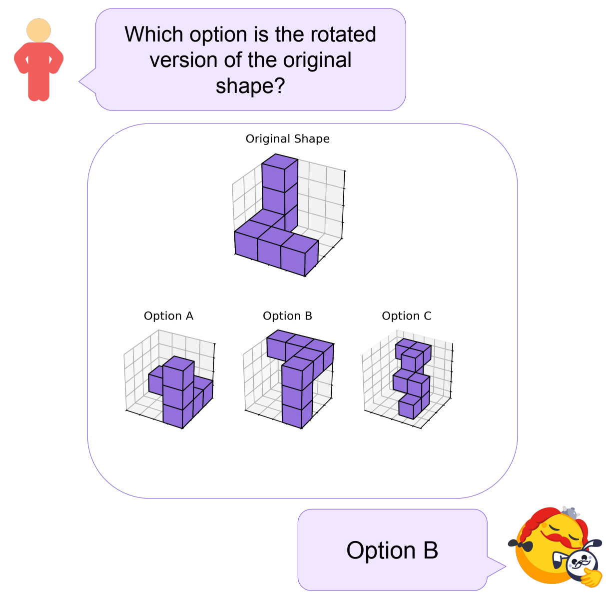

Figure 1. **Example of VLM Responses to Mental Rotation Tasks:** This example underscores a notable constraint of existing VLMs, which have difficulty in precisely understanding potential rotations of the objects shown, exposing a considerable deficiency in their spatial reasoning skills.

Spatial reasoning involves comprehending and analyzing the locations, orientations, and relations of objects within a scene—skills that are natural to humans but pose significant challenges for deep learning models [9, 11, 51]. By fostering strong spatial intelligence, deep learning models can improve their ability to make intricate decisions that necessitate an understanding of geometric transformations and spatial context. While current benchmarks for VLMs incorporate tasks related to spatial reasoning, they often prioritize elements such as object detection or semantic interpretation [35], leaving the core challenges of spatial cog-

[1] https://github.com/stogiannidis/srbench
[2] https : / / huggingface . co / datasets / stogian / srbench

nition underexplored. Additionally, many datasets focus exclusively on specific aspects of spatial reasoning, sometimes augmenting visual data with extra elements such as point clouds, scene graphs, or depth maps to deepen spatial context [9, 12]. The absence of a precise definition and structured evaluation approach for spatial reasoning in VLMs has hindered the creation of models capable of attaining human-level spatial reasoning.

Humans naturally excel at spatial reasoning, a cognitive faculty that manifests in development and underpins our ability to engage with and navigate intricate environments [26, 42]. This competence includes a variety of skills, from calculating distances and orienting things in three dimensions to combining various sensory inputs in a smooth manner. It has been established via rigorous empirical studies in cognitive science and neuroscience that our brain architecture is suited to process spatial relationships. These findings constantly confirm that spatial reasoning is deeply ingrained in our perceptual and motor systems and is not just essential for daily tasks [6, 24]. By attempting to understand and reproduce these natural human abilities, we can shed light on the shortcomings of current VLMs and drive the development of computer systems that exhibit closer to human-like spatial awareness.

Performant spatial reasoning is crucial beyond lab tests, enabling computational agents in fields such as robotics, autonomous navigation, and augmented reality to adapt to complex real-world environments [57]. It enhances VLMs for interpreting visual scenes anthropomorphically and drives advancements in technologies requiring precise spatial understanding. The present study aims to fill these gaps by a thorough and systematic analysis of spatial reasoning in the context of VLMs. We provide a thorough analysis that begins with the characterization of the basic components of spatial reasoning, namely spatial relations, orientation and navigation, mental rotation, and spatial visualization. After laying this conceptual foundation, we assess the models' performance on a wide range of visual stimuli, including both artificially generated images and images from naturalistic environments. By performing a discrete evaluation of each component, we gain a more nuanced understanding of the unique advantages and disadvantages that exist in current VLMs and help us understand their operational capabilities. In summary, our contributions are as follows:

- **Precise Definition of Spatial Reasoning and its components:** We formalize spatial reasoning in VLMs through key components: spatial relations (e.g., object positioning), orientation and navigation (e.g., directionality), mental rotation (e.g., invariant object recognition), and spatial visualization (e.g., transformation prediction).
- **Extensive Benchmark:** We present a novel spatial reasoning benchmark, combining programmatically generated, GenAI-synthesized, and real-world images to bridge controlled evaluation and real-world applicability.
- **Component-Specific Evaluation:** Our benchmark uniquely isolates spatial reasoning components, offering precise insights into VLMs' strengths and limitations.
- **Extensive Model Evaluation:** We assess 13 advanced VLMs, discovering that their typical spatial reasoning performance hovers around random chance, pointing out ongoing challenges.

2. Spatial Reasoning: Definition and Components

Spatial reasoning is a fundamental cognitive ability concerned with understanding and manipulating spatial relationships between objects and oneself within a space [32]. It encompasses the mental skills involved in visualizing, transforming, and reasoning about spatial information. This ability is not monolithic but rather comprises several distinct yet interrelated components [40]. In computer science, spatial reasoning is increasingly relevant to fields such as robotics, computer vision, and geographic information systems, where algorithms and systems must effectively process and interpret spatial data [52]. Understanding these components is crucial for designing effective computational models of spatial intelligence.

2.1. Mental Rotation

Mental rotation is a specific and well-studied aspect of spatial visualization, which refers to the ability to mentally rotate two-dimensional (2D) and three-dimensional (3D) objects [50]. It involves imagining an object as it would appear if rotated in space, and is often measured by tasks requiring individuals to determine if two objects are the same, but presented at different orientations [56]. Pellegrino et al. [45] showed that mental rotation is a distinct spatial ability, separate from other spatial skills such as spatial perception. Mental rotation tasks are widely used in cognitive psychology to study spatial processing and have implications for understanding sex-related differences in spatial abilities, as well as neural substrates of spatial cognition [63]. While closely related to spatial visualization, mental rotation is often considered a more constrained and specific type of spatial transformation, focusing primarily on rotational transformations.

2.2. Spatial Visualization

Spatial visualization goes beyond simple perception and involves the capacity to mentally manipulate and transform spatial information [38]. It is often defined as the ability to process complex spatial information and imagine spatial transformations, such as mentally rotating objects, folding shapes, or understanding movements through space [19]. Spatial visualization is considered a dynamic spatial skill,

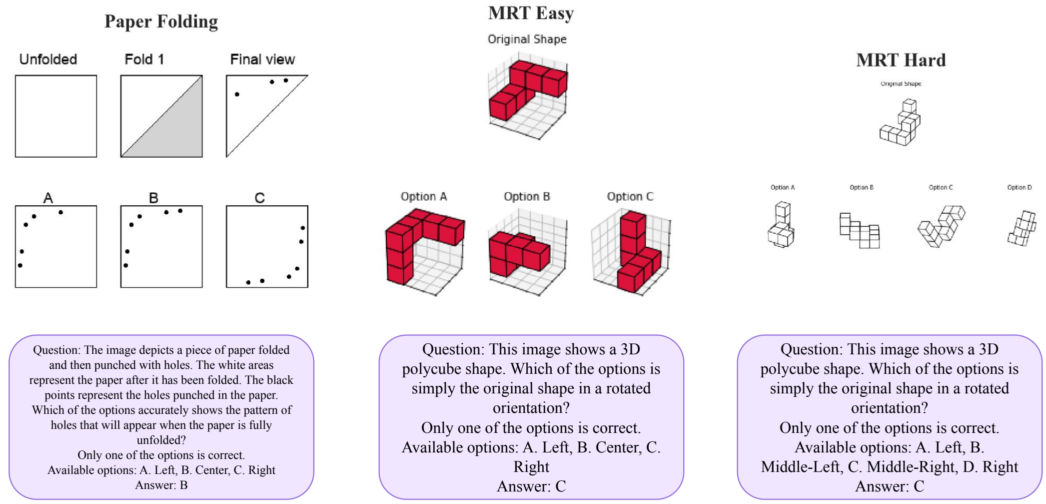

Figure 2. Images from our benchmark created algorithmically, drawing inspiration from cognitive tests. Left image depicts paper folding, the middle shows the easy MRT version, and the right displays the hard variant.

requiring the transformation of visual-spatial representations to derive new spatial configurations [36]. This component is particularly important in fields such as engineering and design, where professionals must mentally simulate the assembly or operation of complex systems. In computer graphics and virtual reality, spatial visualization skills are essential for creating immersive and interactive 3D environments, allowing users to navigate and manipulate virtual objects in a spatially meaningful way [5].

2.3. Spatial Orientation and Navigation

Spatial orientation is a complex cognitive function that enables us to navigate intricate surroundings by actively integrating various sensory inputs, such as visual, vestibular, proprioceptive, and auditory signals, to construct and continually refine internal spatial maps. Studies indicate that these maps can be divided into two main reference frames: *egocentric*, which records spatial data in relation to one's own body (e.g., "the book is on my left"), and *allocentric*, which organizes spatial data based on the relationships among landmarks in the environment, regardless of the observer's position [17, 60]. This duality facilitates flexible navigation; path integration allows individuals to compute their position by updating self-motion cues, and landmark-based encoding helps create a more stable, map-like allocentric space representation [28]. In robotics, these principles guide autonomous systems like Simultaneous Localization and Mapping (SLAM), where data from cameras, LiDAR, and inertial units are combined using probabilistic models for accurate real-time mapping and self-localization, even in GPS-denied or dynamic environments [54]. These interdisciplinary insights show that both egocentric and allocentric frameworks are crucial for human spatial cognition and adaptation, as well as for designing artificial systems interacting with complex environments.

2.4. Spatial Relations

Spatial relations encompass the ability to understand and reason about relationships between multiple objects within a space [43]. This involves processing different types of spatial relations, including topological relations (for example, adjacency, containment, connectivity), projective relations (for example, above / below, left / right, front / behind), and metric relations (e.g., distance, size, volume) [20]. Reasoning about spatial relationships is fundamental for tasks such as spatial planning, solving spatial puzzles, and understanding spatial analogies. In geographic information systems (GIS), the ability to computationally represent and reason with spatial relations is paramount for spatial queries, spatial analysis, and automated map interpretation [4]. The formalization of spatial relations in computational terms is an ongoing area of research in artificial intelligence.

Understanding these distinct yet interconnected components provides a more nuanced framework for both cognitive and computational investigations of spatial reasoning.

3. Data Creation

In contrast to previous benchmarks which test a few aspects of spatial reasoning, our benchmark draws inspiration from human cognition studies [17, 22, 60] and is designed to evaluate VLMs under a comprehensive set of the basic principles of human spatial reasoning.

3.1. Mental Rotation

To assess a VLM's capability to mentally rotate an object, we draw upon the mental rotation test [14], a test designed to measure this ability in humans. The Mental Rotation

Test (MRT) involves a participant comparing two 3D objects (or letters), often rotated along some axis, to determine if they are the same or mirror images [49]. Typically, pairs of images in the test are rotated by specific angles (e.g. 0°, 60°, 120°, or 180°). A fixed number of these pairs will feature identical images simply rotated, while others will be mirrored versions. Participants are evaluated by the researcher based on their speed and precision in differentiating between mirrored and non-mirrored pairs [7]. We proceed by creating similar images to those found in MRT [14], manually crafting five polycube shapes. Subsequently, we construct an image displaying the original shape in the top row, with four potential mirrored versions of that shape in the bottom row. Among these four options, one is the same shape rotated by either 0, 60, 90, or 120 degrees. The other three include two mirrored forms of the shape, each rotated, and one randomly selected, differently rotated object. The shapes are white and devoid of any background, forming what we term the *mrt-hard* subset. An example from *mrt-hard* is depicted in Figure 2 (right). Considering that these images might not offer sufficient visual cues to the model, we develop another test version with colored shapes and a 3D Cartesian grid as the background. Furthermore, we decrease the number of candidates to three by eliminating one mirrored version of the shape. This variant is referred to as the *mrt-easy* subset. We generate 200 images from both subsets. Figure 2 (middle) shows an example of *mrt-easy* images.

3.2. Spatial Visualization

The Paper Folding Test is a non-verbal reasoning assessment commonly used in psychometric evaluations. It assesses spatial visualization abilities, requiring individuals to mentally manipulate folded paper and identify the location of holes punched through it after unfolding [19]. This test is considered a measure of spatial orientation and visualization, distinct from spatial perception [38] . Performance on the Paper Folding Test is often correlated with success in fields requiring strong spatial reasoning skills, such as engineering, architecture, and design [8]. The format of the test typically involves a series of multiple-choice questions, each presenting a sequence of paper folds and hole punches, followed by several unfolded paper options, only one of which is correct. We replicate the paper-folding test by constructing an image where the top row displays a square representing a piece of paper being folded. To the right of this image, the folding process is visually depicted by redrawing the entire square and adding a line to indicate where the fold occurs, such as in the center, while shading the folded portion in light gray. Following this depiction, we puncture holes in the final folded view, marking them as black dots on the paper. Consequently, in the bottom row, three different unfolded paper images are shown, illustrating the unfolded paper with final hole positions. Among them, one is correct, another lacks a hole, and the last is accurately mirrored. We can produce an arbitrary number of folds and holes, but for simplicity, we restrict ourselves to experimenting with one or two folds and one to three holes. Moreover, we employ straightforward folding techniques, recursively folding the paper either vertically, horizontally, or diagonally at the center. Figure 2 (left) depicts an example of these images.

3.3. Spatial Relations

To assess VLMs' capability in recognizing spatial relationships, we employ a sample of the *Spatial-Obj* dataset [51]. This benchmark comprises 2,000 multiple-choice queries aimed at testing how effectively VLMs interpret spatial interactions between objects within images. Developed via a dual-phase annotation method, the dataset includes natural image-based inquiries concerning the spatial interplay of one or two objects. During the initial annotation phase, three annotators generated question-answer sets using templates, followed by a review and correction process by ten evaluators in the subsequent phase. Spatial-Obj encompasses 36 typical spatial relations (e.g., "right of", "left of", "attached to", "touching"). GPT-4o has been used to organize samples into visual categories such as object localization, orientation and direction, viewpoints, as well as positional/relational context, all presenting considerable challenges for VLMs.

3.4. Orientation and Navigation

Spatial orientation plays a crucial role in navigation-related activities. We assess this navigating ability using the *Maze-Nav* component of SpatialEval [59]. This dataset is crafted to test navigation capabilities within intricate environments resembling mazes, represented by colored blocks: green signifies the starting point (S), red marks the exit (E), black indicates walls that cannot be passed, white denotes walkable paths, and blue highlights the correct path from S to E. It is available in both textual (ASCII) and visual formats, with the goal being to navigate from S to E by following the blue path while using only up, down, left, and right directions. Tasks include answering queries such as counting the number of turns or identifying the spatial relationship between S and E, which are straightforward for humans but challenging for current VLMs. For evaluating orientation comprehension, we examine 400 question-and-answer pairs from *EgoOrientBench* [27]. The annotation method tackles inconsistencies in object orientation labeling within VLMs by creating a unified egocentric system. Utilizing current developments in embodied AI that emphasize user-centric viewpoints, the authors formulated an eight-class orientation taxonomy (Front, Back, Left, Right, Front-Left, Front-Right, Back-Left, and Back-Right) that

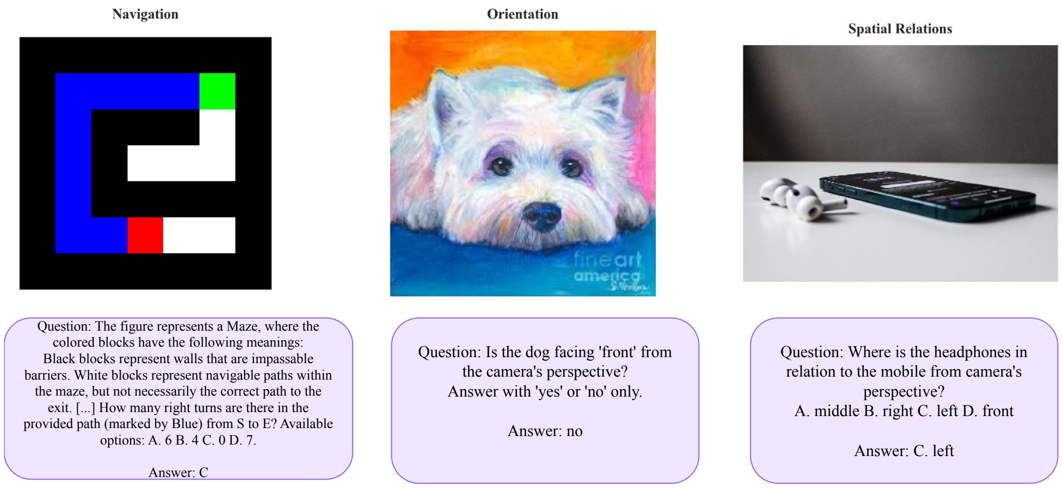

Figure 3. Examples sampled from other benchmarks are depicted as follows: Left is *Maze-Nav* [59], the center displays *EgoOrient-Bench* [27], and on the right is *Spatial-Obj* [51].

consistently positions objects with respect to the user (or camera) perspective. In addition, to facilitate easier evaluation, we sample only those questions whose answers are limited to "yes" or "no". This egocentric alignment not only addresses the problem of unclear orientation annotations that hinder VLMs' spatial understanding but also bolsters their use in real settings where user perspective is key. The approach is in line with the growing trend of egocentric datasets in AI, establishing a structured system that enhances orientation interpretation consistency in various applications.

4. Evaluation

In our experiments, we use PyTorch [44] and Hugging Face Transformers [62]. We evaluated the spatial intelligence of 13 prominent VLMs, which include both open-source and commercial variants. Among the commercial standards, we tested OpenAI's GPT-4o and o1 [2, 25]. The open-source models evaluated comprised Qwen2.5 3B and 7B [46], Llava 1.5 7B [34], LlavaNext 7B [31], InstructBlip 7B and 13B [15], Idefics 8B [30], SmolVLM 500M and 2B [37], Llama-3.2-Vision 11B [18], MiniCPM-V-2.6 8B [65], and InternVL2_5 8B and 26B [10]. All the model variants are instruction-tuned and greedy decoding [21] was equipped. To carry out these experiments, we use the Azure OpenAI API service for accessing OpenAI's models, and for the open-source models, we perform inference using 4$\times$ A100 40GB NVIDIA GPUs.

4.1. Benchmark Distribution

The benchmark dataset comprises a total of 1,800 image-question pairs. Of these, 22.2% (400 pairs) evaluate mental rotation ability, with equal distributions of 11.1% (200 pairs) from the classic Mental Rotation Test (MRT) and 11.1% (200 pairs) from the MRT easy set. For spatial visualization assessment, 11.1% (200 pairs) feature stimuli with either single or double folds and between one and three pierced holes. The remaining pairs are evenly distributed, with 22.2% (400 pairs) sampled from *Maze-nav* to evaluate navigation skills, 22.2% (400 pairs) from *Spatial-Obj* [51] to assess spatial relations comprehension. Finally, 22.2% (400 pairs) are equipped from *EgoOrientBench* [27] to evaluate orientation understanding.

4.2. Performance Evaluation

Overall Performance: Table 1 summarizes the performance of various models across multiple spatial reasoning tasks, including two variants of the Mental Rotation Test (MRT), Paper Folding, Spatial Relations, Navigation, and Orientation. The chance level for these tasks is 32.37%, which represents the expected accuracy if a model were to guess randomly, based on the number of choices available in each task. Notably, models such as InternVL2.5 (26B) and InternVL2.5 (8B) achieve the highest overall accuracy–48.83% and 47.72% respectively–surpassing this chance level by nearly 15%, while several models (e.g., InstructBLIP variants and MiniCPM-V 2.6) perform near or significantly below chance levels overall, suggesting that their broader spatial reasoning capabilities remain limited.

Spatial Relations: In the Spatial Relations task, models demonstrate strong performance, with InternVL2.5 (26B) reaching 70.75% and InternVL2.5 (8B) 64.50%, indicating that they effectively capture relational aspects within scenes.

Mental Rotation Tests (MRT): The MRT tasks reveal disparities. In the easier version MRT, MiniCPM-V-2.6 (8B) leads with 36.00%, yet its overall spatial reasoning

Model	Paper Folding ↑ (%)	MRT Easy ↑ (%)	MRT Hard ↑ (%)	Navigation ↑ (%)	Orientation ↑ (%)	Spatial Relations ↑ (%)	Overall ↑ (%)
Number of Instances	**200**	**200**	**200**	**400**	**400**	**400**	**1800**
SmolVLM (500M)	34.00	34.00	22.00	32.50	51.00	30.00	35.22
MiniCPM-V-2.6 (8B)	27.50	**36.00**	29.00	9.75	45.00	14.75	25.72
InternVL2.5 (8B)	41.50	30.00	24.50	33.25	69.00	64.50	47.72
Llava-1.5 (7B)	38.00	32.50	20.00	24.75	52.50	50.25	38.39
Idefics3 (8B)	42.00	31.00	**29.50**	**35.50**	63.00	58.50	46.28
InternVL2.5 (26B)	42.00	31.50	28.00	28.25	70.00	**70.75**	**48.83**
o1 (Undisclosed)	36.50	33.00	20.50	33.25	71.00	64.75	47.56
Qwen2.5VL (7B)	37.50	27.00	25.00	16.50	66.00	55.75	40.67
Qwen2.5VL (3B)	**46.00**	31.50	25.50	13.50	63.10	33.25	35.86
SmolVLM (3B)	35.50	32.00	29.00	29.75	51.75	38.50	37.39
Llama-3.2-Vision (11B)	26.00	29.50	24.00	26.25	48.00	48.75	36.17
InstructBLIP (7B)	36.50	32.00	21.00	10.50	51.00	17.50	27.50
InstructBLIP (13B)	38.50	35.60	23.40	12.50	53.00	20.50	29.94
GPT-4o (Undisclosed)	36.00	32.50	20.00	32.75	**72.50**	66.50	47.44
LlavaNext (7B)	27.00	32.50	26.50	27.00	52.00	52.50	38.78
Random	33.33	33.33	25.00	25.00	50.00	25.00	32.37

Table 1. **Accuracy of Models in Spatial Reasoning Tasks:** Presented are the accuracy percentages of different models across various spatial reasoning tasks, including Paper Folding, MRT Easy, MRT Hard, Navigation, Orientation, Relations, and Overall performance. The overall score is determined by taking a weighted average (based on the proportion of each subset) of these results, with the highest scores in each column emphasized in bold.

performance is low (25.72%). In the hard MRT variant, Idefics3 (8B) scores 29.50%, with most models clustering between 20.00% and 29.00%. This suggests that success in a simplified mental rotation task does not necessarily translate to robust spatial reasoning across tasks.

Paper Folding: The Paper Folding task yields modest scores (with Qwen2.5VL (3B) at 46% and both Idefics and InternVL2.5 (26B) around 42%), reinforcing the notion that many models struggle with visualizing object transformations.

Navigation and Orientation: Navigation scores, while variable, are measured against an expected chance level of 50% (because the tasks typically present two-choice questions). While certain models exceed this baseline, others—like Idefics3 (8B), achieving merely 35.50%—perform worse than random guessing. In the Orientation task, models such as GPT-4o and o1 achieve scores above 70%. This indicates that when visual cues like clear object outlines are present, some models can effectively discern orientation.

This raises a critical question: *Is the limitation in mental rotation due to the synthetic nature of the images lacking essential visual cues, or does it reflect a fundamental inability of VLMs to reason about spatial transformations?*

To explore this question, we examined a question-answer test centered around the use of generative AI (genAI)-produced images to evaluate mental rotation further. We developed a synthetic dataset employing generative models to examine mental rotation capabilities under controlled conditions (see Appendix 8). Initially, we crafted five meticulously designed image descriptions portraying objects in various orientations (e.g., facing right). These descriptions served as in-context learning prompts to Claude3.7 [1], accompanied by a list of objects, to stimulate the generation of 100 analogous image descriptions using the specified objects or their synonyms. These generated descriptions were then processed by two generative models: `Flux.1-dev` [29] and `Stable Diffusion 3.5 Large` [3], producing a total of 200 images. Following manual curation, we chose 80 question-answer pairs that necessitated inferring an object's orientation after a hypothetical rotation (refer to Figure 4 for illustrative examples).

As can be observed in Figure 5, performance on the synthetic mental rotation assessment varied considerably across models. Some models, particularly InternVL2.5 (26B) and InternVL2.5 (8B), consistently demonstrated higher overall accuracy compared to other models such as Llava-1.5 (7B) and SmolVLM (500M). Models like Llama-3-2 Vision (11B) and MiniCPM-V-2.6 (8B) exhibited relatively consistent performance between the GenAI-generated mental rotation task and the standard MRT. This suggests that performance differences are not solely due to the data type used. Choosing synthetic images similar to training data does not significantly improve performance, indicating a limitation in spatial reasoning. The enhancement observed in the GenAI-synthetic images may be attributed to their potential resemblance to the original training dataset of these models or reliance on other visual cues.

Summary: While tasks such as Spatial Relations and Orientation reveal certain strengths of current models, the struggles in Mental Rotation, Paper Folding, and Navigation tasks highlight significant gaps in spatial reasoning. The novel genAI mental rotation assessment provides additional insights into these limitations and offers a controlled environment to further probe the mechanisms behind spatial transformations in VLMs.

Question: If we turn the rabbit in the opposite direction, which way will it be facing?
Options:
A. Front
B. Right
C. Back
D. Left

Answer: B. Right

Figure 4. Example of a GenAI image shows a rabbit in a green meadow, challenging mental rotation skills. It asks VLMs to identify the rabbit's orientation after rotation, with "Right" as the correct choice.

5. Related work

5.1. Spatial Reasoning in Vision-Language Models

Recent progress in VLMs has significantly advanced multimodal comprehension, though explicit spatial reasoning still presents substantial challenges. Initially, VLMs were primarily tailored for overarching image understanding and captioning, frequently overlooking the intricate spatial relationships required for uses in robotics and augmented reality. As a countermeasure, various strategies have emerged, integrating spatial supervision into training datasets and model frameworks.

One area of research endeavors to create large-scale synthetic spatial reasoning datasets. For instance, [9] employs an automated 3D spatial VQA data production process to craft millions of region-conscious question–answer pairs derived from 2D imagery by building 3D scene graphs and incorporating metric depth estimation. This method enhances the training data with spatial annotations, thereby equipping the VLMs with improved spatial reasoning capabilities both qualitatively and quantitatively. In a similar vein, [12] expands on this concept by embedding region-level hints and relative depth data into the visual encoder. By integrating a depth-to-language connector and processing user-defined region proposals, the approach shows significantly improved outcomes on spatial reasoning benchmarks—even in intricate 3D environments. This design emphasizes the necessity of combining local region features with depth indicators to capture both broad contexts and precise spatial details.

Recent initiatives such as [53] focus on training VLMs in core 2D spatial tasks, enhancing skills such as direction interpretation, distance estimation, and localization, thus improving spatial reasoning. This approach suggests basic spatial skills lay the foundation for tackling complex challenges. Research into grounded and compositional strategies, such as multimodal spatial grounding, further improves alignment between visuals and language [48]. However, models still fall short of human-level reasoning, especially in dynamic environments, indicating a need for future exploration. Improving VLM spatial reasoning requires effective data curation and architecture. Despite progress through techniques like 3D annotations and depth features, achieving reliable human-level understanding in real-world applications needs further effort.

5.2. Spatial Reasoning in Humans

Spatial reasoning is a multifaceted cognitive ability that enables individuals to perceive, manipulate, and navigate space. Seminal work by Shepard and Metzler [49] introduced the mental rotation paradigm, laying the groundwork for subsequent studies that have refined our understanding of spatial cognition. Researchers such as Hegarty and Waller [23] and Newcombe and Huttenlocher [42] have differentiated between intrinsic skills (e.g., mental rotation and spatial visualization) and extrinsic skills (e.g., navigation and perspective-taking), establishing frameworks that underscore the link between early spatial abilities and later academic achievement in STEM domains [58].

Recent intervention studies demonstrate that targeted spatial training can enhance children's mathematical performance [13, 55]. In parallel, interdisciplinary research has applied computational and qualitative frameworks to model human spatial reasoning for applications in areas such as human–robot interaction and geographic information systems [39, 41]. These combined efforts affirm that spatial reasoning is not only a trainable and critical cognitive skill but also a pivotal foundation for solving real-world problems and advancing STEM education.

5.3. Spatial Reasoning Benchmarking

Benchmarking spatial reasoning capabilities is critical for evaluating the effectiveness of VLMs in real-world scenarios. Recent efforts have introduced dedicated benchmarks that focus on both qualitative and quantitative aspects of spatial understanding. For example, [12] not only improves model performance but also introduces a benchmark dataset comprising both qualitative and quantitative spatial reasoning tasks derived from indoor, outdoor, and simulated environments. This benchmark evaluates models on tasks such

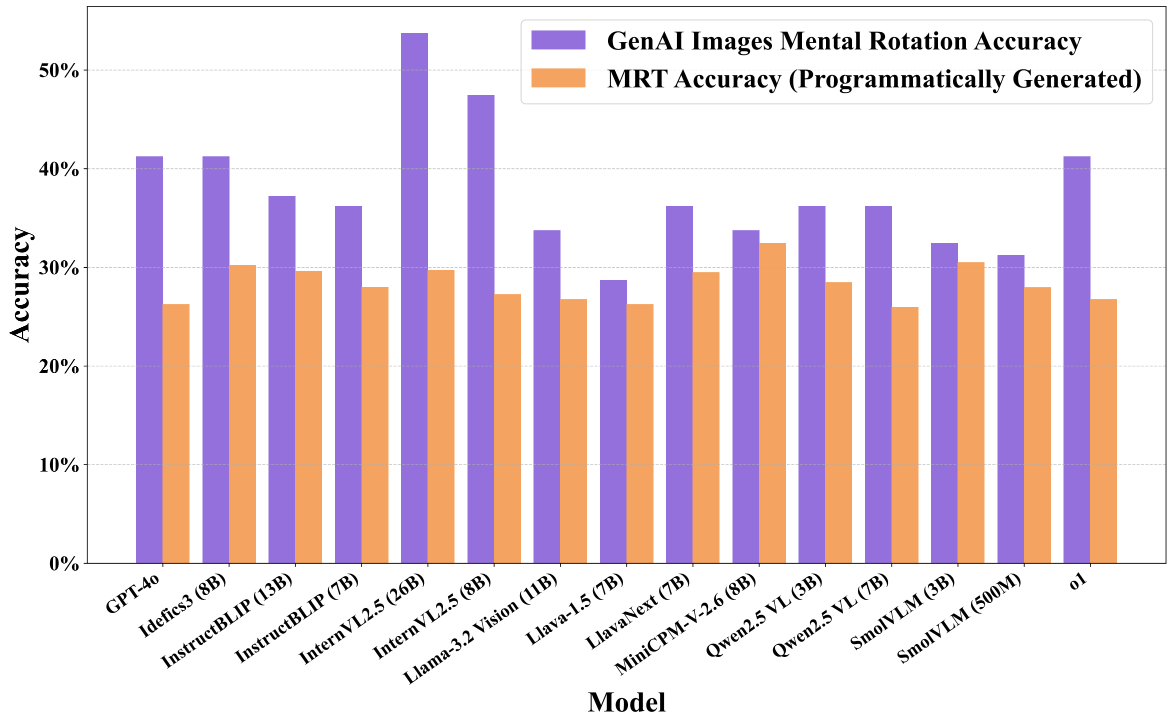

Figure 5. Comparison of Mental Rotation Images Accuracy (violet) and MRT Accuracy (teal) across various vision-language models. Models such as **InternVL2.5 (26B)** exhibit a significant gap between the two accuracy metrics, suggesting a stronger capability in interpreting mental rotation images compared to structured MRT tasks. In contrast, lower-performing models like **LLaVa-1.5 (7B)** show consistently weak performance across both categories. The overall trend indicates that while some models excel in image-based spatial reasoning, their understanding of abstract mental rotation tasks remains limited.

as determining relative positions (e.g., above, below, left, right) and measuring metric distances (e.g., direct, horizontal, vertical distances).

Other benchmarking approaches, such as those incorporated in [53], focus on isolating basic spatial capabilities (direction, distance, localization) and then composing these to solve more complex spatial problems. Meanwhile, grounded spatial reasoning evaluations in multi-modal settings assess a model's ability to align visual evidence with textual spatial descriptions [48]. These benchmarks not only serve to highlight the current limitations of VLMs but also provide clear metrics for tracking progress as new architectures and training strategies are developed.

Concurrent work [64] used human-applied psychometric tests to investigate spatial thinking in VLMs, with similar results. Their results demonstrate that VLMs underperform relative to humans on these tests, underscoring the need for further exploration of these models' spatial capabilities. Our approach diverges by incorporating real-world photographs alongside psychometric assessments. To ensure full control over the images and expand the test cases, we also introduce custom-developed psychometric assessments. Finally, we perform an ablation study on the image data to see if any important cues that are necessary for the model to perform these tasks are absent.

Collectively, these benchmarking efforts underscore the need for systematic evaluation of spatial reasoning. They provide a foundation for comparing diverse approaches and guiding future research toward achieving robust, human-level spatial understanding in VLMs.

6. Conclusion

This paper tackles the underexplored topic of spatial reasoning in VLMs by offering a clear definition, introducing a robust benchmark with both synthetic and real-world images, and evaluating 13 state-of-the-art VLMs. The key finding is stark: the majority of contemporary VLMs perform near random chance on spatial reasoning tasks on our benchmark; but some appear to perform better on tasks that use natural or generative images that may also contain other visual cues. Our work hence exposes a major limitation in VLMs' ability to achieve human-like visual understanding. This gap has significant implications, as spatial reasoning is vital for practical applications. Our work made a positive step in this direction; it didn't explore in depth which cues models use in natural images which can be seen as a limitation. Looking ahead, future research should continue to explore the interaction between spatial reasoning components and their respective roles in complex spatial tasks, as well as the reliance on other visual cues to reasoning, paving the way for more sophisticated and human-like spatial reasoning capabilities in artificial intelligence systems.

References

[1] The Claude 3 Model Family: Opus, Sonnet, Haiku. 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

[3] Stability AI. Introducing stable diffusion 3.5. https://stability.ai/news/introducing-stable-diffusion-3-5, 2024. Accessed: 2025-03-02. 6

[4] Brandon Bennett. Qualitative spatial reasoning: ontologies, granularity and containment. *International Journal of Human-Computer Studies*, 48(6):749–765, 1998. 3

[5] Doug A Bowman, Ernst Kruijff, Joseph J LaViola Jr, and Ivan Poupyrev. 3d user interfaces: Theory and practice. 2004. 3

[6] Neil Burgess. Spatial cognition and the brain. *Annals of the New York Academy of Sciences*, 1124(1):77–97, 2008. 2

[7] Andre F Caissie, Francois Vigneau, and Douglas A Bors. What does the mental rotation test measure? an analysis of item difficulty and item characteristics. *Open Psychol. J.*, 2 (1):94–102, 2009. 4

[8] John B Carroll. *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge university press, 1993. 4

[9] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 1, 2, 7

[10] 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. 5

[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, 2024. GSCC: 0000013 arXiv:2406.01584 Read_Status: Read Read_Status_Date: 2024-10-31T14:46:44.677Z. 1

[12] 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. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2025. 2, 7

[13] Yi-Ling Cheng and Kelly S Mix. Spatial training improves children's mathematics ability. *Journal of cognition and development*, 15(1):2–11, 2014. 7

[14] Lynn A Cooper. Mental rotation of random two-dimensional shapes. *Cogn. Psychol.*, 7(1):20–43, 1975. 3, 4

[15] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 5

[16] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023. 2

[17] Rudolph P Darken, Terry Allard, and Lisa B Achille. Spatial orientation and wayfinding in large-scale virtual spaces. *Presence: Teleoperators and Virtual Environments*, 8(6):3–6, 1999. 3

[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. 1, 5

[19] Ruth B Ekstrom and Harry Horace Harman. *Manual for kit of factor-referenced cognitive tests, 1976*. Educational testing service, 1976. 2, 4

[20] Christian Freksa, Alexander Klippel, Louisa Knuf, Bernhard Nebel, and Stefan Wölfl. Spatial relations among parts of objects: a survey. *KI-Künstliche Intelligenz*, 27:73–79, 2013. 3

[21] Ulrich Germann. Greedy decoding for statistical machine translation in almost linear time. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 72–79, 2003. 5

[22] Mary Hegarty. Chapter 7 - components of spatial intelligence. In *The Psychology of Learning and Motivation*, pages 265–297. Academic Press, 2010. 3

[23] Mary Hegarty and David Waller. A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, 32(2):175–191, 2004. 7

[24] Masud Husain and Parashkev Nachev. Space and the parietal cortex. *Trends in cognitive sciences*, 11(1):30–36, 2007. 2

[25] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. 5

[26] Mark Johnson. *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. University of Chicago Press, 1987. 2

[27] Ji Hyeok Jung, Eun Tae Kim, Seo Yeon Kim, Joo Ho Lee, Bumsoo Kim, and Buru Chang. Is' right'right? enhancing object orientation understanding in multimodal language models through egocentric instruction tuning. *arXiv preprint arXiv:2411.16761*, 2024. 4, 5

[28] Roberta L Klatzky. Spatial Cognition: An Interdisciplinary Approach to Representing and Processing Spatial Knowledge. 1998. 3

[29] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 6

[30] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions, 2024. 5

[31] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 5

[32] Marcia C Linn and Anne C Petersen. Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child development*, pages 1479–1498, 1985. 2

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1

[34] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, 2024. 5

[35] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024. 1

[36] Gordon D Logan. Spatial working memory, visual attention, and object identity in visual search. *Cognitive psychology*, 31(3):263–309, 1996. 3

[37] Andres Marafioti, Merve Noyan, Miquel Farré, Elie Bakouch, and Pedro Cuenca. Smolvlm - small yet mighty vision language model, 2024. 5

[38] Meredith G McGee. Human spatial abilities: psychometric studies and environmental, genetic, hormonal, and neurological influences. *Psychological bulletin*, 86(5):889, 1979. 2, 4

[39] Daniel R Montello. Scale and multiple psychologies of space. In *European conference on spatial information theory*, pages 312–321. Springer, 1993. 7

[40] Daniel R Montello. Cognitive components of environmental spatial cognition. *Handbook of spatial psychology*, pages 253–278, 2005. 2

[41] Reinhard Moratz and Thora Tenbrink. Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. *Spatial cognition and computation*, 6(1):63–107, 2006. 7

[42] Nora Newcombe and Janellen Huttenlocher. *Making space: The development of spatial representation and reasoning*. MIT Press, 2000. 2, 7

[43] Nora S Newcombe and Janellen Huttenlocher. Developing spatial competence. pages 55–84, 2000. 3

[44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703, 2019. 5, 2

[45] James W Pellegrino, Robert Kail, and Brian McCloskey. Cognitive correlates of spatial ability: Evidence for two distinct processes. *Intelligence*, 26(2):97–126, 1998. 2

[46] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. 5

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1

[48] Navid Rajabi and Jana Kosecka. Towards grounded visual spatial reasoning in multi-modal vision language models. In *ICLR 2024 Workshop on Data-centric Machine Learning Research (DMLR): Harnessing Momentum for Science*, 2024. 7, 8

[49] R N Shepard and J Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971. 4, 7

[50] Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971. 2

[51] Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Reza Haf, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21440–21455, Miami, Florida, USA, 2024. Association for Computational Linguistics. 1, 4, 5

[52] Oliviero Stock. Spatial and temporal reasoning. pages 925–972, 2008. 2

[53] Yihong Tang, Ao Qu, Zhaokai Wang, Dingyi Zhuang, Zhaofeng Wu, Wei Ma, Shenhao Wang, Yunhan Zheng, Zhan Zhao, and Jinhua Zhao. Sparkle: Mastering basic spatial capabilities in vision language models elicits generalization to composite spatial reasoning. *arXiv preprint arXiv:2410.16162*, 2024. 7, 8

[54] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic robotics*. MIT press, 2005. 3

[55] 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. 7

[56] Steven G Vandenberg and Allan R Kuse. Mental rotations of the alphabet letters. *Perceptual and motor skills*, 41(1): 259–264, 1975. 2

[57] Sagar Gubbi Venkatesh, Anirban Biswas, Raviteja Upadrashta, Vikram Srinivasan, Partha Talukdar, and Bharadwaj Amrutur. Spatial reasoning from natural language instructions for robot manipulation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11196–11202. IEEE, 2021. 2

[58] Jonathan Wai, David Lubinski, and Camilla P Benbow. Spatial ability for stem domains: aligning over 50 years of cu-

mulative psychological knowledge solidifies its importance. *Journal of educational Psychology*, 101(4):817, 2009. 7

[59] 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*, 37:75392–75421, 2025. 4, 5

[60] Ranxiao Wang and Elizabeth Spelke. Human spatial representation: insights from animals. *Trends Cogn. Sci.*, 6(9): 376, 2002. 3

[61] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2

[62] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771, 2019. 5, 2

[63] Meredith Wraga, Roger N Shepard, Patricia S Churchland, Souheil Inati, and Stephen M Kosslyn. Mental rotation of photographs of human bodies. *Attention, Perception, & Psychophysics*, 62:635–650, 2000. 2

[64] Wenrui Xu, Dalin Lyu, Weihang Wang, Jie Feng, Chen Gao, and Yong Li. Defining and evaluating visual language models' basic spatial abilities: A perspective from psychometrics. *arXiv preprint arXiv:2502.11859*, 2025. 8

[65] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 5

[66] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022. 2

Mind the Gap: Benchmarking Spatial Reasoning in Vision-Language Models

Supplementary Material

7. Data Creation

In our study, we developed Python scripts to generate image sets tailored for cognitive evaluations involving paper folding and mental rotation tests. The code and obtained images will be made publicly available upon acceptance.

For the folded paper images, we employed the Python Imaging Library (PIL) to simulate the visual process of **paper folding**. Initially, we draw a white paper measuring 100×100 pixels onto a 120×120 pixel canvas. Using predefined drawing routines and fold reflection rules, our script recursively applies vertical, horizontal, or diagonal fold lines, generating intermediate views of each fold stage through polygon clipping operations (Algorithm 1). Holes are randomly punched into the final folded polygon. The script then calculates their unfolded positions via reflections, effectively doubling layers with each fold. Finally, we produce candidate images, comprising the correctly unfolded paper and two distractors—these distractors either omit a hole, mirror positions, or slightly rotate them. All generated views and candidates are combined into a single composite image, clearly separating folding stages in the top row and randomized candidate options in the bottom row. The corresponding metadata for each image is recorded and stored in a JSONL file. To ensure clarity, our approach restricts folds to half of the paper at a time and avoids mixing diagonal folds with vertical or horizontal ones.

In parallel, we crafted a script for mental rotation tests (**MRT**) involving complex polycube shapes, utilizing matplotlib's 3D plotting capabilities (Algorithm 2). Our method begins by selecting polycube configurations ranging from basic shapes to intricate forms such as snake-like arrangements. For each configuration, the script generates a set of candidate images through transformations including rotations, mirrored reflections, or substitutions with visually similar shapes. Candidates consist of one correct rotation and two distractors—one involving mirrored rotations and the other substituting with a similar but incorrect shape. The final image arranges the original shape prominently at the top, with the candidates randomized below, deliberately concealing transformation details to assess mental rotation capabilities. Two variations are implemented: a challenging version mimicking traditional MRT setups in human cognition without visual aids and using rotations at angles of $60°$, $90°$, and $120°$ around all axes, and a simplified version with visual cues and rotations restricted to $-90°$, $90°$, or $180°$ along a single axis. Images and metadata are stored in JSONL files for structured evaluation.

Algorithm 1 Generate Folded Paper Test Image

1: **Initialize:** Define drawing routines and fold reflection rules.
2: **Recursive Folding:**
- Start with an initial paper polygon.
- For each fold, clip the polygon (using a fixed midpoint) and save the intermediate view.

3: Obtain the final folded polygon.
4: Generate random holes within the final polygon.
5: Compute unfolded hole positions via reflection (doubling layers per fold).
6: Create candidate images:
- **Correct:** Use unfolded holes.
- **Wrong:** Modify holes (e.g., remove one, mirror, or rotate).

7: Assemble a composite image with:
- Top row: Fold views.
- Bottom row: Candidate options (labeled A, B, C).

8: Save the composite image and record metadata.

Algorithm 2 Generate Mental Rotation Test Images

1: **Initialize:** Define `SHAPES` and `SIMILAR_MAPPING`.
2: **for** each image to generate **do**
3: Randomly select a shape and compute its vertices.
4: Generate three candidate transformations:
- **Correct:** Rotate the shape.
- **Wrong 1:** Mirror then rotate.
- **Wrong 2:** Use a visually similar shape then rotate.

5: Shuffle candidate order and record the correct option.
6: Plot the original shape and candidates in a grid.
7: Save the image and append metadata.
8: **end for**

8. GenAI Image Generation Process

To create the images for our ablation study, we manually handcrafted five exemplar image descriptions, each depicting an object positioned in a specific orientation within a photo-realistic setting (see Table 2). We then fed these examples into `Claude3.7`, a Large Language Model (LLM) by Anthropic, to generate additional descriptions featuring different objects. The model was prompted using the structured format shown in Table 3. This prompt is meticulously organized, beginning with clearly defined components such as the **Role** and **Objective**, which immediately establish

Examples
A photo-realistic image of a car, viewed from the front, with the wheels turned to the left, parked in a driveway, with a clear blue sky in the background.
A photo-realistic image of a mug, viewed from the side, with the handle on the right, filled with steaming coffee, on a wooden table, with a window in the background showing a sunny day.
A photo-realistic image of a laptop, viewed from the back, with the screen open.
A photo-realistic image of a bicycle, viewed from the side, with the front wheel turned to the right, parked on a cobblestone street, with a row of colorful houses in the background.
A photo-realistic image of a cat, viewed from the front, with the tail curled to the right, sitting on a windowsill, with a potted plant in the background.

Table 2. Photo-realistic Image Examples that were given to Claude 3.7 to generated similar ones.

Role: Text-Based Image Description Generation Assistant

Objective: To generate high-quality text image descriptions for generative models.

Input (Textual):

- A list of objects (provided as text by the user).
- Example image descriptions (provided as text by the user).

Output (Textual): New image descriptions (as text).

Task: Generate new text image descriptions that meet the following criteria:

1. **Style Mimicry:** Replicate the writing style, sentence structure, and vocabulary used in the example text descriptions.
2. **Object Novelty:** Feature the provided list of objects, ensuring they are different from objects explicitly mentioned in the example text descriptions.
3. **Setting Novelty:** Describe the objects in new and different settings or contexts compared to those presented in the example text descriptions.
4. **Logical Coherence & Realism:** Ensure all generated text descriptions are logically sound, realistic, and portray plausible scenarios in text. Avoid nonsensical or physically impossible descriptions in text.

Table 3. Image Description Generation Guidelines

the context and purpose of the task. It subsequently delineates the required **Input** and expected **Output** through bullet points, ensuring clarity and ease of understanding. The core of the prompt—the **Task**—is presented as a numbered list that outlines specific criteria, including style mimicry, object and setting novelty, and logical coherence and realism. This detailed segmentation not only clarifies the structure for the generated text but also systematically guides the generative process. Overall, the structured use of bold headings and lists enhances the prompt's effectiveness and user-friendliness by ensuring that all essential information is clearly highlighted and easily referenced. Alongside the five exemplars, we provide a compiled list of objects to give extra context and guidance to the model. The list is the following: `bicycle`, `motorcycle`, `scooter`, `car` `truck`, `bus`, `train`, `airplane`, `helicopter`, `boat`, `ship`, `dog`, `cat`, `bird`, `fish`, `rabbit`. Figure 6 depicts examples of the generated images from the diffusion models alongside the hand-curated questions and answers.

9. Implementation Details

For the evaluation of these 13 state-of-the-art (SOTA) models, we utilized PyTorch [44] (version 2.5.1) along with the HuggingFace[62] libraries: `transformers`, `diffusers`, and `accelerate`. Each model was loaded using `bfloat16` precision, and `Flash Attention 2` [16] was implemented wherever applicable to enhance inference speed. The experiments were executed on four A100 Nvidia GPUs, each with 40 GB of memory, using CUDA version 12.0. Models were used with a greedy decoding approach, providing their outputs immediately, and not utilizing any prompting methods such as Chain-of-Thought (CoT) [61] or Least-To-Most (LTM) [66] prompting. The results were matched using regular expressions,

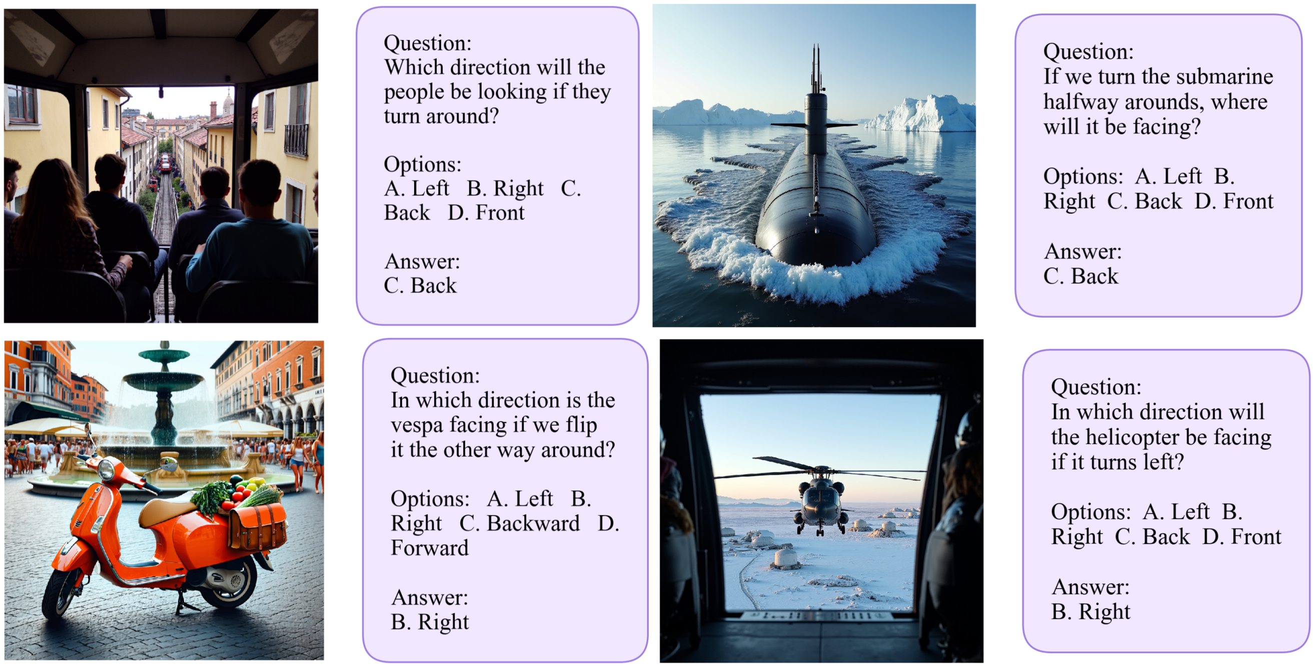

Figure 6. Example of GenAI-generated Images and Question-Answer pairs that were used for the mental rotation evaluation.

and responses that did not conform to the specified format were deemed incorrect.