MIND THE GAP: DIAGNOSING SPATIAL REASONING FAILURES IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) have captivated the research community by effectively merging visual and textual information, implying a holistic comprehension of the environment. These models find applications in tasks such as Image Captioning and Visual Question Answering, fostering the assumption that they perceive reality in a way similar to human cognition. However, this apparent understanding may be misleading. We argue that a critical component of comprehension—spatial reasoning—has been insufficiently addressed, as current benchmarks primarily test models' ability to identify object positions rather than evaluate genuine spatial logic. In this study, we aim to address this limitation. Drawing from the fundamental elements of human cognition, we developed a diagnostic framework designed to isolate the essential components of spatial reasoning: relational understanding, orientation, mental rotation, and visualization. We evaluated 17 state-of-the-art VLMs within both controlled synthetic settings and the complex variability of images captured in the real world. Results indicate a substantial gap in performance: the apparent competence of these models decreases significantly under spatial reasoning tasks that require any dynamic transformation and manipulation of spatial information. On average, their performance parallels random guessing, which highlights a major systematic weakness in spatial reasoning in current VLMs. In addition to providing evidence for this limitation, this study also provides the research community with a foundational framework for developing models that can accurately understand and reason about spatial properties in their environment.

1 INTRODUCTION

Vision-Language Models (VLMs) have demonstrated impressive proficiency across a broad spectrum of multimodal tasks, such as Image Captioning, Visual Question Answering, and text-image retrieval (Liu et al., 2023; Dubey et al., 2024; Radford et al., 2021). Leveraging extensive datasets in their pretraining phases, these models can effectively map the intricate interactions between visual and textual data. Nonetheless, one crucial facet of intelligence that remains notably deficient is **spatial reasoning**. This essential skill entails understanding object locations, orientations, and their interrelations within a scene—a capability that is instinctive to humans but poses a substantial challenge for modern deep learning models (Zhang et al., 2025b; Shiri et al., 2024; Chen et al., 2024a; Cheng et al., 2024).

Spatial reasoning is not a niche skill; it is fundamental to cognition. In humans, it develops between two and eleven years old Hodgkiss et al. (2021) and underpins our ability to navigate and interact with complex environments (Johnson, 1987; Newcombe & Huttenlocher, 2000). Cognitive science and neuroscience research confirms that spatial cognition is deeply engrained in our perceptual and motor systems (Burgess, 2008; Husain & Nachev, 2007). Bridging this gap in AI is crucial for moving beyond pattern recognition toward a more human-like understanding of the world. Moreover, performant spatial reasoning is a prerequisite for real-world applications such as robotics, autonomous navigation, and augmented reality, where agents must interact with and adapt to dynamic physical spaces (Venkatesh et al., 2021).

A key obstacle to progress has been the absence of a structured evaluation framework for spatial reasoning in VLMs. Current benchmarks often subsume spatial tasks under broader categories such

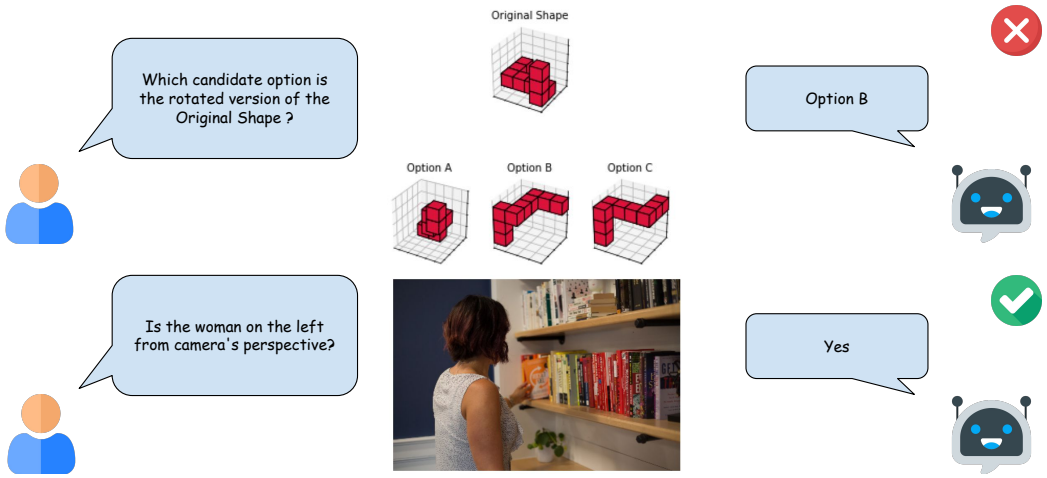

Figure 1: **Example of VLM responses to mental rotation and relation tasks.** Models frequently struggle to identify legitimate rotations, exposing deficiencies in tasks requiring dynamic transformations. However, they exhibit competence in discerning spatial relations present within the image.

as object detection or semantic interpretation, failing to isolate core cognitive challenges (Liu et al., 2024b). This lack of a targeted evaluation has hindered the development of models with robust spatial intelligence. To address this, we introduce a comprehensive benchmark grounded in foundational paradigms from cognitive psychology, designed to systematically evaluate the core facets of spatial reasoning (Bar-Hen-Schweiger & Henik, 2024). Our framework assesses performance across four distinct pillars: (1) mental rotation, (2) spatial visualization, (3) relational understanding, and (4) egocentric navigation.

Our evaluation of 18 current VLMs shows they handle static spatial reasoning fine, yet fail whenever the task requires simulating a transformation of the object to reach the answer. While leading models exhibit high accuracy on tasks involving spatial relations and orientation in static images, their performance collapses on tasks requiring mental manipulation (Fig. 1). Notably, on the mental rotation task, nearly all models—including state-of-the-art systems such as GPT-4o—perform at or below random chance. This points to a gap: models have trouble imagining or predicting spatial transformations.

In summary, our contributions are as follows:

- We introduce a new comprehensive benchmark for systematically evaluating dynamic spatial reasoning in VLMs, grounded in foundational paradigms from cognitive psychology.
- We evaluate 17 state-of-the-art models, revealing that their performance on many core spatial tasks is near random chance, highlighting a critical area for future research.
- We analyse dissociations between static scene understanding and dynamic reasoning, highlighting a fundamental weakness in current models.
- We find that scaling the model alone is insufficient to overcome the inability to perform simulative spatial reasoning (i.e., mentally transforming objects). Future progress will likely require new inductive biases that explicitly support simulation-style reasoning about object dynamics.

2 Constructing the Benchmark and Experimental Setup

Human spatial reasoning emerges from the intricate interplay of several cognitive abilities that allow us to navigate, manipulate, and understand our three-dimensional world (Hegarty, 2010; Darken et al., 1999; Wang & Spelke, 2002). Unlike previous benchmarks that evaluate isolated aspects of spatial cognition Ma et al. (2024); Kamath et al. (2023), our comprehensive evaluation framework systematically assesses Vision-Language Models across the fundamental interconnected pillars of

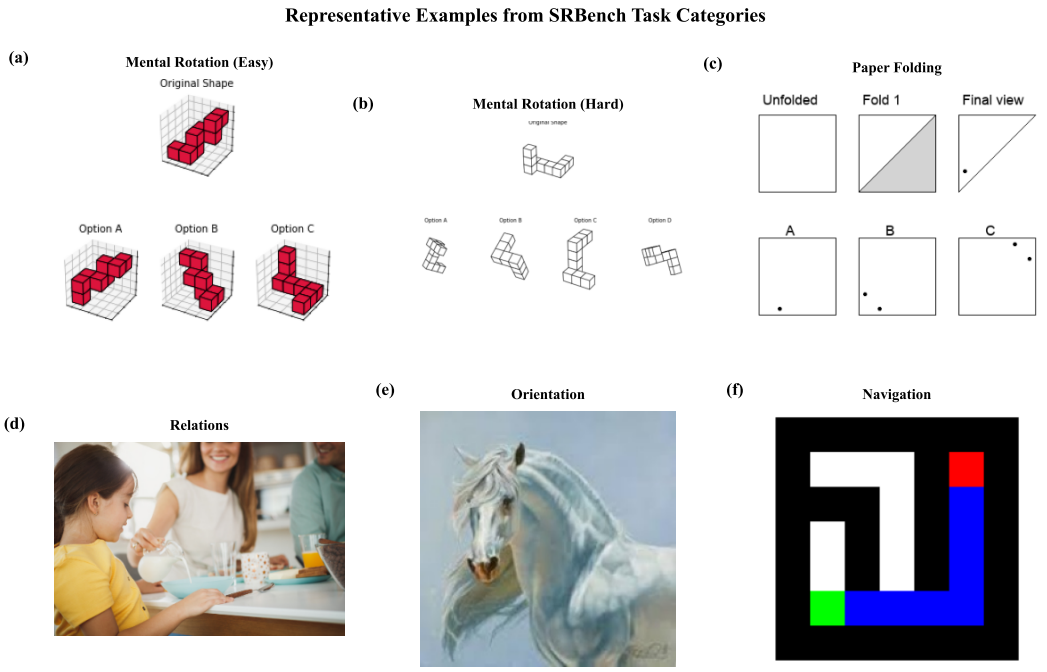

Figure 2: Representative examples from SRBench spatial reasoning tasks. (a-b) Mental rotation tasks with hard and easy difficulty levels. (c) Paper folding visualization task. (d) Spatial navigation with route planning. (e) Spatial orientation and perspective-taking. (f) Spatial relations between geometric elements. Each panel demonstrates the visual complexity and cognitive demands of the respective spatial reasoning category in the benchmark dataset.

human spatial reasoning: mental rotation, spatial visualization, relational understanding, and egocentric navigation.

2.1 Mental Rotation

We begin by evaluating a model's capability for mental rotation—the ability to mentally transform three-dimensional objects in space. Drawing from the seminal Mental Rotation Test (MRT) (Cooper, 1975), which has served as the gold standard for measuring this cognitive ability in humans for decades, we adapt this classical paradigm to modern VLMs.

The original MRT presents participants with pairs of 3D objects or letters, rotated along various axes, challenging them to distinguish between identical shapes and their mirror images (Shepard & Metzler, 1971). Human performance is typically assessed through both accuracy and response time at rotation angles of 0°, 60°, 120°, and 180° (F Caissie et al., 2009).

Our digital adaptation follows this established protocol while accommodating the unique characteristics of VLMs. We manually craft five distinct polycube shapes and construct test images featuring the target shape in the top row, accompanied by four candidate shapes below. Among these candidates, exactly one represents the original shape rotated by 0°, 60°, 90°, or 120°; the remaining three consist of two mirrored versions at different rotations and one randomly selected unrelated shape.

To systematically vary task difficulty, we develop two complementary variants. The *MRT-Hard* subset presents white shapes against blank backgrounds, offering minimal visual cues and posing a significant challenge to the model's internal spatial representations. Recognizing that this austere presentation might limit model performance, we create the *MRT-Easy* subset, which incorporates colored shapes positioned within a 3D Cartesian grid background and reduces the choice set to three candidates by removing one mirrored candidate. Each subset consists of 200 carefully designed test cases, as illustrated in Figure 2 (a and b).

2.2 Spatial Visualization

Beyond object rotation, spatial reasoning demands the ability to mentally simulate complex geometric transformations. We assess this through an adaptation of the Paper Folding Test (Ekstrom & Harman, 1976; McGee, 1979)—a psychometric instrument whose performance strongly correlates with success in spatially demanding fields such as engineering and architecture (Carroll, 1993).

Each instance presents a temporal sequence of transformations: a paper square undergoes one or two folds (vertical, horizontal, or diagonal), followed by punching one to three holes through the folded configuration. The model must then predict the resulting hole pattern when the paper is unfolded, selecting from three plausible alternatives. This task directly probes the model's capacity to internalize sequential geometric operations and mentally simulate their cumulative effects—a cornerstone of spatial visualization ability. The subset comprised 200 test cases, illustrated in Fig. 2 (c) as examples.

2.3 Spatial Relations

Understanding the relative positioning and interactions between objects forms the foundation of scene comprehension. We evaluate this critical capability using a curated sample from the SpatialObj dataset (Shiri et al., 2024), a rigorously constructed benchmark that contains 2,000 multiple choice queries regarding spatial relationships in natural images.

The authors generated this dataset employing an in-depth dual-stage annotation procedure, thoroughly encompassing 36 essential spatial relationships. These range from elementary positional notions such as 'right of' and 'above', to intricate geometric interactions including 'attached to,' 'touch', and 'overlapping'. The queries encompass diverse visual challenges including identification of precise object location, orientation discrimination, and contextual spatial reasoning, providing a robust assessment of how well VLMs comprehend relational spatial language in realistic visual scenarios. This subset contains 400 test cases, with examples shown in Fig. 2 (d).

2.4 Orientation and Navigation

Finally, we examine spatial reasoning within the critical domains of navigation and egocentric perspective taking, abilities essential for real-world spatial intelligence.

For navigation assessment, we employ the Maze-Nav component of SpatialEval (Wang et al., 2024), which challenges models to reason about paths through visual mazes represented by colored block configurations. Tasks include identifying routes from start (S) to exit (E) points, counting directional changes, and describing spatial relationships between key locations. While trivial for human spatial cognition, these challenges reveal significant limitations in current VLMs' navigational reasoning capabilities.

Complementing navigation assessment, we evaluate orientation understanding using 400 binary questions from EgoOrientBench (Jung et al., 2024). This benchmark addresses critical inconsistencies in spatial orientation evaluation by establishing a unified, camera-centric perspective framework. Through an eight-class egocentric taxonomy (Left, Right, Front-Left, Back-Right, etc.), it provides consistent object orientation definitions relative to the observer's viewpoint. This egocentric approach not only enhances evaluation reliability but also aligns with the increasing need for VLMs to operate effectively in user-centered, real-world applications, such as robotics, where spatial understanding must be grounded in human perspective. This subset contains 400 test cases, with examples shown in Fig. 2 (e and f).

2.5 Setup

Our experiments were conducted with PyTorch (Paszke et al., 2019) and Hugging Face Transformers (Wolf et al., 2020). We evaluated the spatial reasoning capabilities of 17 VLMs, which include open-source and commercial models. Specifically, from the commercial side, we included evaluations of OpenAI's GPT-4o and o1 (Achiam et al., 2023; Jaech et al., 2024). The open source model set consists of: QwenVL2.5 in sizes 3B, 7B, 32B, and 78B (Bai et al., 2025); Llava 1.5 7B (Liu et al., 2024a); LlavaNext 7B (Li et al., 2024); Idefics3 8B (Laurençon et al., 2024) and SmolVLM2

at 500M and 2.2B (Marafioti et al., 2025). Additionally, MiniCPM-V-2.6 8B (Yao et al., 2024); InternVL-3 models at 8B, 38B, and 78B (Chen et al., 2024b) and Gemma3 at 12B, and 28B. All models are instruction tuned and the experiments were conducted using greedy decoding (Germann, 2003) and Chain-of-Thought (Wei et al., 2022) prompting. For OpenAI's models, we used the Azure OpenAI API service, while for the open-source models, inference was performed using 2 $\times$ H200 140GB NVIDIA GPUs.

3 RESULTS

Our investigation into the spatial reasoning of contemporary VLMs reveals a compelling, two-part narrative. On one hand, models exhibit a promising, emergent ability to parse static visual scenes. On the other, this competence proves remarkably brittle, collapsing entirely when confronted with tasks that require dynamic mental manipulation. This core tension, explored below, points to a fundamental gap between superficial pattern recognition and robust spatial cognition.

Model	Paper Folding	MRT Easy	MRT Hard	Navigation	Orientation	Relations	Overall
Random	33.0	33.0	25.0	25.0	50.0	25.0	32.0
Idefics3 8B	35.0	28.0	22.5	30.5	64.0	59.8	43.35
InternVL-3 8B	27.0	33.0	28.5	18.3	69.5	66.3	43.64
InternVL-3 38B	42.5	**40.5**	**29.0**	43.0	**77.5**	73.5	55.00
InternV-L3 78B	43.5	34.5	23.0	**55.0**	74.2	**73.8**	**55.77**
MiniCPM-V 2.6	35.5	32.5	24.0	23.5	47.0	41.0	34.65
Qwen2.5-VL 3B	24.0	29.0	21.1	19.3	60.3	53.8	37.50
Qwen2.5-VL 7B	36.0	35.0	26.0	21.0	65.0	63.2	49.25
Qwen2.5-VL 32B	42.5	34.0	22.5	42.25	68.5	69.25	50.94
Qwen2.5-VL 72B	**45.0**	39.5	24.0	40.8	69.5	73.3	52.33
SmolVLM2 2.2B	35.0	31.9	11.0	17.8	65.2	43.4	36.38
SmolVLM2 500M	29.5	36.0	27.5	34.5	51.8	37.8	37.53
Gemma 3 12B	31.0	32.0	24.0	22.3	57.5	29.95	34.10
Gemma 3 27B	16.50	22.50	16.00	25.00	57.2	47.3	34.3
LLaVA-1.5 7B	36.0	35.5	25.5	36.0	52.8	31.4	37.1
LLaVA-NeXT 7B	25.0	34.5	27.0	20.3	53.1	48.3	36.29
o1 (Undisclosed)	36.0	33.0	20.5	33.3	71.0	64.8	47.05
GPT-4o (Undisclosed)	36.0	32.0	20.0	32.8	72.5	66.5	47.48

Table 1: Performance of models across various spatial reasoning tasks. The Random baseline is included for comparison. All scores are accuracy percentages. The best performance in each category is highlighted in **bold**.

3.1 THE FRAGILITY OF SPATIAL INTELLIGENCE: FROM STATIC COMPETENCE TO DYNAMIC COLLAPSE

At first glance, the models detailed in Table 3 demonstrate a solid grasp of basic spatial properties. On static tasks like **Orientation** and **Relations**, leading architectures such as InternVL-3 38B achieve high accuracy (77.5% and 73.5%, respectively), suggesting they can adeptly identify and relate objects in a fixed scene. This initial success, however, masks a profound underlying weakness.

This apparent competence is undermined when models must perform internal simulations of dynamic object transformations. On the **Mental Rotation Test (MRT Hard)**, a task requiring complex, multi-axis mental manipulation, performance plummets. The failure is not merely a gradual decline but a catastrophic collapse: most models do not outperform the random baseline. Most strikingly, even state-of-the-art models like GPT-4o score just 20%, performing significantly *worse* than random chance (25%). This deficit likely stems from factors such as biases in training data (lack of rotated/diagonal views, spurious correlations), pretraining focused on static descriptions over internal simulations, and limitations in architecture for encoding continuous 3D priors. This indicates that their success in static spatial tasks does not imply the capacity for simulating transformations; they have learnt to describe the world as it is, but cannot reliably reason about how it might change Newman et al. (2024); Li et al. (2025).

3.2 Scaling Laws and the Emergence of Articulated Reasoning

A central question is whether this cognitive deficit can be overcome by simply increasing model scale. Our findings affirm the powerful effect of scaling laws, which manifest not only in quantitative accuracy but also in the qualitative nature of the models' reasoning.

As we scale models from billions to tens of billions of parameters, a distinct shift in cognitive style emerges. Smaller models, like **InternVL3-8B**, tend to produce terse, direct answers, offering little insight into their decision-making process. Their larger counterparts, such as **InternVL-78B**, behave fundamentally differently. They engage in articulated, step-by-step reasoning, verbalizing their analysis of visual evidence and systematically evaluating options. This transition from opaque, "black-box" intuition to a more transparent, deliberative process suggests that scaling does not just improve accuracy—it unlocks more sophisticated and explicit reasoning pathways.

This qualitative evolution is mirrored by quantitative gains. Across the QwenVL2.5 and InternVL-3 families, models with tens of billions of parameters generally showcase much better performance compared to smaller ones (for example, **InternV-L3 78B** scores 55.77% versus 43.64% for the 8B variant). But scaling is not strictly monotonic: mid-size models sometimes beat larger ones (e.g., **InternVL-3 38B** outperforms 78B on the 'MRT hard' split), and we observe plateaus with little or no gain for some jumps, as depicted in Fig. 3. Given that model size typically covaries with various other elements, such as the training ensemble, objectives, data, and optimisation processes, it is not safe to assert that these effects arise solely due to the number of parameters. A plausible set of mechanisms that accompany scaling helps explain the qualitative shift. Larger parameter counts increase representational capacity, enabling models to internalize multi-step algorithms or templates for reasoning rather than relying on single-step heuristics. Larger models are also typically trained with more compute over longer runs on much bigger and more diverse corpora, raising the chance they encounter examples that demonstrate explicit, chain-of-thought–style analyses which they can imitate. These correlations necessitate controlled ablation studies to determine causality. Crucially, even the best and largest models still fail catastrophically on the hardest tasks, showing that scale alone does not resolve the underlying gaps in their reasoning toolkit.

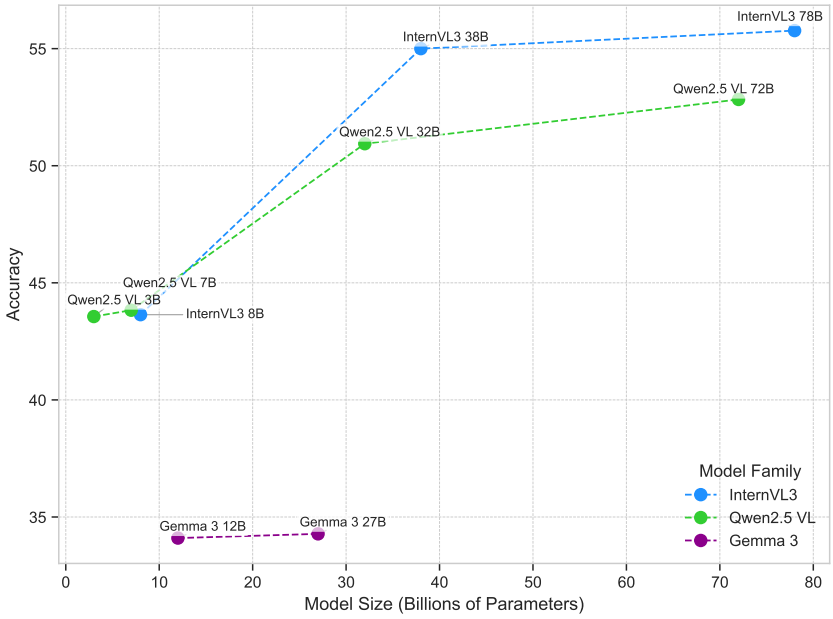

Figure 3: Model accuracy as a function of parameter count. While a positive trend exists within families, architectural differences create distinct performance tiers, highlighting that scale alone is not a panacea for complex reasoning failures

3.3 A Granular Dissection of Failure Modes

To understand the limits of scaling, we performed a granular analysis of common failure modes. This investigation revealed a consistent Achilles' heel across all models and scales: a fundamental difficulty in processing diagonal and rotational transformations, particularly evident in the 'MRT hard' and 'Paper folding' tasks.

3.3.1 Static Perception vs. Dynamic & Diagonal Reasoning

The most straightforward tasks reveal a foundational bias. In **Orientation** tasks, every model is more adept at identifying cardinal directions (e.g., "front") than diagonal directions (e.g., "front left"). This suggests an inbuilt preference for axis-aligned spatial judgments, a tendency that is highly likely attributable to biases in the training data.

This perceptual weakness extends to the **Relations** task, where reasoning accuracy declines. Through manual qualitative inspection of model responses, we observed that models reliably resolved queries involving static, unambiguous relationships (e.g., *"The bus is to the left of the building"*), yet their responses deteriorated noticeably for prompts involving agents performing actions (e.g., *"The man is **holding** the..."*). These observations suggest that while models can parse a static layout, they fail to build a robust model of interactions, a more complex and dynamic form of reasoning.

3.3.2 The Fragility of Abstract Spatial Transformation

Through qualitative analysis of model responses, we noticed that difficulties with dynamic operations become most apparent when models are required to mentally simulate transformations. In the **Paper Folding** subset, outputs were consistently more reliable for simple axis-aligned folds than for diagonal ones, and reasoning quickly broke down as additional folds were introduced. Responses often became inconsistent or contradictory when asked to track more than a couple of sequential folds, suggesting that the cognitive load of maintaining an object's state across transformations exceeds the models' effective reasoning capacity.

A similar pattern was visible in the more demanding **MRT** tasks. For example, the **InternVL-38B** model seemed most stable on objects of medium complexity, but its answers deteriorated as the number of polycubes increased.

Upon inspecting the model's responses, it was clear that for medium-complexity shapes, the model was able to correctly count the number of polycubes and their reasoning tended to be more stable. However, in more complex shapes, the model often failed to identify how many polycubes each shape was comprised of, leading to fabricated or inaccurate reasoning. This aligns with the impression that model strategies can handle limited complexity but degrade in a roughly predictable manner once that limit is crossed.

Collectively, these failures—from a bias against diagonals to an inability to track sequential rotations—highlight that the path to robust, human-like reasoning will require not just greater scale, but new architectural paradigms and new inductive biases designed to master the reasoning over the complexities of dynamic object transformations. This conclusion is supported by performance across task categories such as navigation and 'MRT easy'. Importantly, in the most difficult 'MRT hard' partition, models of the Qwen2.5-VL family exhibit sub-par performance, although the accuracy of the largest models shows a minor improvement. This suggests that while scaling generally improves analytical reasoning for moderately complex problems, it does not guarantee enhanced performance on tasks that may exceed the architectural limitations of the current models, where a more explicit reasoning process may not confer an advantage and could even be detrimental.

4 Related work

4.1 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 endeavours to develop comprehensive synthetic datasets for spatial reasoning on a large scale. For example, Chen et al. (2024a) utilizes an automated system to generate 3D spatial Visual Question Answering datasets, producing millions of comprehensive question-answer pairs from 2D images by constructing 3D scene graphs and applying metric depth estimation. This technique enriches the training datasets with spatial labels, thereby enhancing the spatial reasoning capability of Vision-Language Models (VLMs) both qualitatively and quantitatively. Similarly, Cheng et al. (2025) extends this approach by integrating region-level cues and relative depth information into the visual encoder. By incorporating a depth-to-language transformation module and accommodating user-specified region proposals, the method demonstrates substantial improvements in performance on spatial reasoning benchmarks—even in intricate 3D settings.

Recent studies delving into the underlying causes of these issues indicate that the challenge transcends merely insufficient datasets. According to Zhang et al. (2025a), expanding the amount of training data yields limited improvements, pointing to a fundamental structural bottleneck. Their research emphasizes that spatial comprehension is heavily dependent on the positional encodings within the visual encoder, while the language model plays a minor role in ultimate spatial assessments. This highlights a critical limitation in how visual data is formatted and integrated into the language module.

An alternative research thread delves into the internal workings of the model during the reasoning process. From the viewpoint of mechanistic interpretability, Chen et al. (2025) identified that errors in spatial tasks are highly associated with the misalignment of visual attention. Their findings indicate that although image tokens constitute most of the input, they receive relatively minimal attention (approximately 10%). Crucially, the model frequently neglects to concentrate on the pertinent objects or regions essential for producing a correct response. They propose a solution to be applied during inference, which adaptively sharpens or smoothens attention based on a model's confidence, thereby substantially enhancing performance without the need for retraining.

These findings are complemented by initiatives such as that of Tang et al. (2024), which 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 Rajabi & Kosecka (2024). However, models still fall short of human-level reasoning, especially in dynamic environments, indicating a need for future exploration. Improving VLM spatial reasoning requires not only effective data curation, but also critical architectural innovation. Despite progress through techniques such as 3D annotations and depth features, achieving reliable human-level understanding in real-world applications needs further effort.

4.2 Spatial Reasoning in Humans

Spatial reasoning is a multifaceted cognitive ability that enables individuals to perceive, manipulate, and navigate space. Seminal work by Shepard & Metzler (1971) introduced the mental rotation paradigm, laying the groundwork for subsequent studies that have refined our understanding of spatial cognition. Researchers such as Hegarty & Waller (2004) and Newcombe & Huttenlocher (2000) 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 (Wai et al., 2009).

More recent intervention studies demonstrate that targeted spatial training can enhance children's mathematical performance (Uttal et al., 2013; Cheng & Mix, 2014). 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 (Moratz & Tenbrink, 2006; Montello, 1993). 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.

4.3 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, Cheng et al. (2025) not only improve model performance but also introduce 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 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 Tang et al. (2024), 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 Rajabi & Kosecka (2024). Although these benchmarks are instrumental in highlighting the current limitations of VLMs and providing clear metrics for tracking progress, they fail to address the models' significant deficiencies in dynamic spatial reasoning.

A complementary line of work focusses on how VLMs handle the inherent ambiguity in spatial language, which arises from different Frames of Reference (FoR). Zhang et al. (2025c) introduced the COMFORT evaluation protocol to systematically assess how VLMs resolve these ambiguities. Using controlled 3D-rendered scenes, they test whether models can flexibly adopt different perspectives (e.g., the camera's, an observer's, or an object's intrinsic view). Their findings reveal that VLMs exhibit significant shortcomings: they show poor robustness, struggle to adopt alternative FoRs when prompted, and overwhelmingly default to English-centric conventions, even when tested in other languages. This approach highlights the need to evaluate not just geometric accuracy, but also the cognitive and cross-cultural dimensions of spatial reasoning.

Concurrent work Xu et al. (2025) 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 images alongside psychometric assessments, which provide richer coverage and precise control over reasoning tasks.

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.

5 Conclusion

This paper studies spatial reasoning in VLMs—the ability to infer, predict, and manipulate geometric relationships and transformations (rotation, translation, scaling, occlusion) from images—by providing a clear definition, a robust benchmark with synthetic and real-world images, and an evaluation of 17 state-of-the-art VLMs. We find a stark gap: while most VLMs handle tasks that infer information present in an image, their performance falls to near-random on tasks that require reasoning about transformations, revealing a major limitation with important practical consequences. Our work takes a step toward addressing this gap; future research should analyze which cues models use in natural images, introduce inductive biases that explicitly encode transformations, and design architectures or modules for object-centric representation and manipulation of transformations. Continued study of how spatial components interact and how other visual cues support reasoning will be crucial to achieving more human-like spatial reasoning in AI systems.

References

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.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*,

2025.

Moran Bar-Hen-Schweiger and Avishai Henik. Looking beyond seeing: Components of visual-spatial ability as an overarching process. *Acta Psychologica*, 251:104577, 2024. ISSN 0001-6918. doi: https://doi.org/10.1016/j.actpsy.2024.104577. URL https://www.sciencedirect.com/science/article/pii/S0001691824004554.

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

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

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*, pp. 14455–14465, 2024a.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. *arXiv preprint arXiv:2503.01773*, 2025.

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*, 2024b.

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, October 2024. URL http://arxiv.org/abs/2406.01584. GSCC: 0000013 arXiv:2406.01584 Read_Status: Read Read_Status_Date: 2024-10-31T14:46:44.677Z.

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.

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

Lynn A Cooper. Mental rotation of random two-dimensional shapes. *Cognitive psychology*, 7(1): 20–43, 1975.

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.

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.

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

André F Caissie, François Vigneau, and Douglas A Bors. What does the mental rotation test measure? an analysis of item difficulty and item characteristics. 2009.

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*, pp. 72–79, 2003.

Mary Hegarty. Chapter 7 - components of spatial intelligence. In *The Psychology of Learning and Motivation*, volume 52 of *Psychology of Learning and Motivation*, pp. 265–297. Academic Press, 2010. doi: https://doi.org/10.1016/S0079-7421(10)52007-3. URL https://www.sciencedirect.com/science/article/pii/S0079742110520073.

Mary Hegarty and David Waller. A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, 32(2):175–191, 2004. URL https://doi.org/10.1016/S0160-2896(03)00065-8.

Alex Hodgkiss, Katie A Gilligan-Lee, Michael SC Thomas, Andrew K Tolmie, and Emily K Farran. The developmental trajectories of spatial skills in middle childhood. *British Journal of Developmental Psychology*, 39(4):566–583, 2021.

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

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.

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

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.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? Investigating their struggle with spatial reasoning, October 2023. URL http://arxiv.org/abs/2310.19785. GSCC: 0000056 arXiv:2310.19785 Read_Status: Read Read_Status_Date: 2024-10-28T11:51:31.242Z.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions, 2024. URL https://arxiv.org/abs/2408.12637.

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.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. *arXiv preprint arXiv:2501.02189*, 2025.

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

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)*, pp. 26296–26306, June 2024a.

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?, 2024b. URL https://arxiv.org/abs/2307.06281.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. 3dsr-bench: A comprehensive 3d spatial reasoning benchmark. *arXiv preprint arXiv:2412.07825*, 2024.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Mark G McGee. Human spatial abilities: psychometric studies and environmental, genetic, hormonal, and neurological influences. *Psychological bulletin*, 86(5):889, 1979.

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

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.

Nora Newcombe and Janellen Huttenlocher. *Making space: The development of spatial representation and reasoning*. MIT press, 2000.

Kaleb Newman, Shijie Wang, Yuan Zang, David Heffren, and Chen Sun. Do pre-trained vision-language models encode object states? *arXiv preprint arXiv:2409.10488*, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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*, pp. 8748–8763. PmLR, 2021.

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. URL https://openreview.net/forum?id=gXprdL2EIW.

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

Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Gholamreza Haffari, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. *arXiv preprint arXiv:2411.06048*, 2024.

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.

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.

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)*, pp. 11196–11202. IEEE, 2021.

Jonathan Wai, David Lubinski, and Camilla P Benbow. Spatial ability for stem domains: aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of educational Psychology*, 101(4):817, 2009.

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, 2024.

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

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020. URL https://arxiv.org/abs/1910.03771.

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.

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.

Wanyue Zhang, Yibin Huang, Yangbin Xu, JingJing Huang, Helu Zhi, Shuo Ren, Wang Xu, and Jiajun Zhang. Why do mllms struggle with spatial understanding? a systematic analysis from data to architecture. *arXiv preprint arXiv:2509.02359*, 2025a.

Wanyue Zhang, Yibin Huang, Yangbin Xu, JingJing Huang, Helu Zhi, Shuo Ren, Wang Xu, and Jiajun Zhang. Why Do MLLMs Struggle with Spatial Understanding? A Systematic Analysis from Data to Architecture. (arXiv:2509.02359), September 2025b. doi: 10.48550/arXiv.2509.02359.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. In *The Thirteenth International Conference on Learning Representations*, 2025c. URL https://openreview.net/forum?id=84pDoCD4lH.

This appendix provides supplementary materials and detailed examples that support the main paper's findings. Section A offers a curated look at specific instances where leading models failed, revealing critical gaps in their spatial reasoning capabilities.

A Illustrative Examples of Reasoning Failures

These case studies are not mere errors; they are windows into the systematic shortcomings of current models, providing tangible evidence for the key phenomena discussed in our paper. Each example includes the model, the task, the question, the model's (often incorrect) reasoning, the correct answer, and a brief analysis of the failure.

A.1 The Perceptual-Abstractive Gap

This gap appears when a model can correctly identify individual objects and their properties but fails to synthesize them into a coherent, abstract understanding of the scene. It sees the trees but misses the forest.

A.1.1 Failure on Diagonal vs. Cardinal Orientations

Models show a distinct weakness in identifying diagonal orientations compared to simpler cardinal directions, pointing to a foundational bias for axis-aligned reasoning.

Example 1: Diagonal Orientation Failure

- **Model:** Gemma-3 27B-it
- **Task:** Orientation
- **Question:** Is the toy truck facing 'back left' from the camera's perspective? Answer with 'yes' or 'no' only.
- **Model's Prediction:** YES
- **Correct Answer:** NO

Model's Reasoning:

```
The camera is positioned to the side of the truck, and slightly
towards the front. The front of the truck (white cabin) is on
the left side of the image, and the back (green tank) is on
```

```
the right side of the image. Therefore, the truck is facing
towards the back and left from the camera's perspective.
```

Analysis: This is a classic failure of synthesis. The model correctly identifies the **perceptual ingredients** ('back' and 'left') but cannot combine them into the correct **abstract recipe** (the diagonal orientation 'back-left'). It treats the components as a simple checklist rather than an integrated spatial concept.

A.1.2 Failure on Dynamic Action-Based Relations

Models that succeed at identifying static relationships between objects often fail when the relationship is defined by a dynamic action performed by an agent.

Example 2: Action-Based Relation Failure

- **Model:** InternVL-3 8B
- **Task:** Relations
- **Question:** Is the person using the tool to hold the food on the left or the right side of the image?
- **Model's Prediction:** RIGHT
- **Correct Answer:** LEFT

Model's Reasoning:

```
The image shows a person holding a piece of food with a tool.
The food is located on the right side of the image.
```

Analysis: The model sees a static scene of objects—person, tool, food—but fails to comprehend the **action** ('using the tool to hold') that connects them. It correctly identifies the location of the food but completely misses the context of the question. It answers "where is the food?" instead of "where is the *act of holding* happening?", highlighting a critical gap in reasoning about dynamic events.

A.2 The Fragility of Abstract Spatial Transformation

This category of failure occurs when models are tasked with mentally simulating changes to an object's state, such as folding or rotating it. Their ability to perform these "mental gymnastics" is surprisingly brittle.

A.2.1 Exponential Failure in Sequential Transformation (Paper Folding)

The most dramatic failures occur when models must track an object's state through multiple sequential transformations. Performance collapses exponentially with each additional fold.

Example 3: 3-Fold Paper Folding Failure

- **Model:** InternVL-3 78B (Top Performer)
- **Task:** Paper Folding
- **Question:** A piece of paper is folded three times, and one hole is punched. Which option shows the unfolded paper?
- **Model's Prediction:** D
- **Correct Answer:** B

Model's Reasoning:

```
After the first fold, there are 2 layers. After the second fold,
there are 4 layers. After the third fold, there are 8 layers.
A hole is punched through all 8 layers. When unfolded, there
will be 8 holes. Option D shows 8 holes.
```

Analysis: This reasoning is a textbook case of the Perceptual-Abstractive Gap applied to transformation. The model demonstrates a shallow, **quantitative success** (correctly calculating $2^3 = 8$ holes) while utterly failing the core **qualitative, spatial task** of predicting the holes' symmetrical pattern. It found an answer with the right *number* of holes but in the completely wrong *configuration*, revealing that its 'reasoning' was a brittle shortcut, not a true mental simulation of unfolding the paper.

A.2.2 The Cognitive "Sweet Spot" in Mental Rotation

The top-performing model's accuracy on MRT tasks does not scale linearly with object complexity. It fails on overly simple shapes and overly complex ones, performing best on objects of medium complexity.

Example 4a: Failure on a Simple Shape

- **Model:** InternVL-3 78B
- **Task:** MRT Hard
- **Question:** (Image with a simple 5-cube shape)
- **Model's Prediction:** A
- **Correct Answer:** D

Model's Reasoning:

```
The original shape is composed of five red cubes. By rotating
the original shape 90 degrees to the right around the vertical
axis, it matches the configuration of Option A.
```

Analysis: The model's failure on a simple shape is revealing. It suggests that without sufficient complexity to anchor its analysis, the model resorts to **flimsy heuristics** that lead to error. It is akin to overthinking a simple problem and making a careless mistake because its more powerful, systematic reasoning pathways were not fully engaged.

Example 4b: Success on a Medium-Complexity Shape

- **Model:** InternVL-3 78B
- **Task:** MRT Hard
- **Question:** (Image with a more complex 7-cube shape)
- **Model's Prediction:** B
- **Correct Answer:** B

Model's Reasoning:

```
The original shape is an assembly of seven cubes. Let's trace
the key features. There is a central column of three cubes and
two arms protruding. By performing a 180-degree rotation around
the horizontal axis, the shape aligns perfectly with Option B.
```

Analysis: Here, with a moderately complex object, the model hits its stride. Its reasoning becomes more descriptive and structural ("central column," "two arms"), indicating a deeper level

of engagement. It successfully simulates the rotation because the problem is complex enough to demand **genuine spatial processing** but not so complex as to overwhelm it. This success, contrasted with the failure on the simpler shape, perfectly illustrates the non-linear "sweet spot" for its performance.

B USE OF LARGE LANGUAGE MODELS

In the preparation of this paper, Large Language Models (LLMs) were utilized to refine the text, improving clarity, grammatical precision, and stylistic flow without altering the substantive ideas or original authorship. This AI-assisted process enabled a more polished presentation of the research while maintaining academic integrity.