Mind the Gap: Diagnosing Spatial Reasoning Failures in Multimodal Large Language Models

Ilias M. Stogiannidis[†] **Steven McDonagh**[†] **Sotirios A. Tsaftaris**[†]

[†]The University of Edinburgh, Edinburgh, UK

Abstract. Multimodal Large Language Models (MLLMs) perform well on many vision–language tasks, but this can hide fundamental spatial reasoning weaknesses. Existing benchmarks blur the line between static scene description and the dynamic mental simulation needed for rotation, folding, and perspective-taking. To target core spatial skills, we introduce a psychometrically grounded diagnostic suite of seven task splits adapted from cognitive assessments, spanning 2,855 controlled samples over synthetic and real images. Evaluating 32 state-of-the-art MLLMs reveals a sharp dissociation: models succeed on information extraction yet fall to near-chance on mental simulation. Longer reasoning chains reduce accuracy, models consistently confuse mirror reflections with rotations, and left–right perspective inversions remain unsolved across all 32 models. Crucially, scaling model parameters does not improve mental simulation: 3D mental rotation stays near chance for all architectures and sizes. Our analysis offers interpretable failure diagnostics and concrete directions for architectural and training improvements.

Correspondence: i.stogiannidis@ed.ac.uk

1 Introduction

Multimodal Large Language Models (MLLMs) have demonstrated strong performance across a broad range of vision–language tasks, from image captioning and visual question answering to complex multi-step reasoning (Liu et al., 2023; Dubey et al., 2024; Radford et al., 2021). Yet intelligence is multifaceted, and one dimension that remains notably underdeveloped is *spatial reasoning*—the cognitive ability to understand, visualise, and mentally manipulate the positions and orientations of objects in space (Zhang et al., 2025a; Shiri et al., 2024; Chen et al., 2024; Cheng et al., 2024). Spatial reasoning develops in humans between the ages of two and eleven (Hodgkiss et al., 2021), underpins the ability to navigate complex environments (Johnson, 1987; Newcombe and Huttenlocher, 2000), and is a prerequisite for embodied AI applications in robotics and autonomous navigation (Venkatesh et al., 2021).

Despite an expanding set of spatial reasoning benchmarks, current work offers little diagnostic clarity about *how* models fail. Foundational cognition tests find models performing near chance on simple tasks (Ramakrishnan et al., 2025; Jia et al., 2025), whilst psychometric studies uncover architectural limits on mental simulation (Xu et al., 2025; Li et al., 2025a; Zhang et al., 2025c). Similarly, 3D evaluations highlight egocentric–allocentric asymmetries (Ma et al., 2025; Li et al., 2025b), and mental simulation benchmarks show models reaching only 28% compared to humans' 77% (Huang et al., 2025; Jiang et al., 2025). Yet it remains unclear whether errors are consistent across architectures, whether extended reasoning helps or harms, and which concrete representational limitations underlie each distinct failure mode.

In this work, we present a diagnostic suite of seven psychometrically grounded task splits—mental rotation (easy and hard), paper folding, navigation, orientation, spatial relations, and perspective taking—each adapted from established cognitive assessments and covering controlled samples that span both synthetic and real-world images. We evaluate 32 state-of-the-art MLLMs across ten architectural families, from 500M to 241B parameters, including both open-source and closed-source models, and complement accuracy measurement with fine-grained error analyses—distractor selection rates, cross-model wrong-answer agreement, reasoning-length quantification, and controlled thinking-vs-instruct comparisons—that expose the specific failure modes underlying aggregate scores.

Our evaluation reveals a core gap in current capabilities. As shown in Figure 1, models perform well on tasks that extract information directly from images—Relations (μ = 70.0%) and Orientation (μ = 68.6%)—but accuracy collapses on tasks requiring mental simulation, with MRT Hard near chance (μ = 25.0%) across all 32 models. Navigation is a high-variance visual-grounding probe, with the best model at 95.8% and the worst at 18.2%. Longer chains of thought correlate with lower accuracy; models reliably confuse mirror reflections with rotations, and left–right perspective inversions are the most common error. Scaling does not fix these failures: a 7B-parameter model

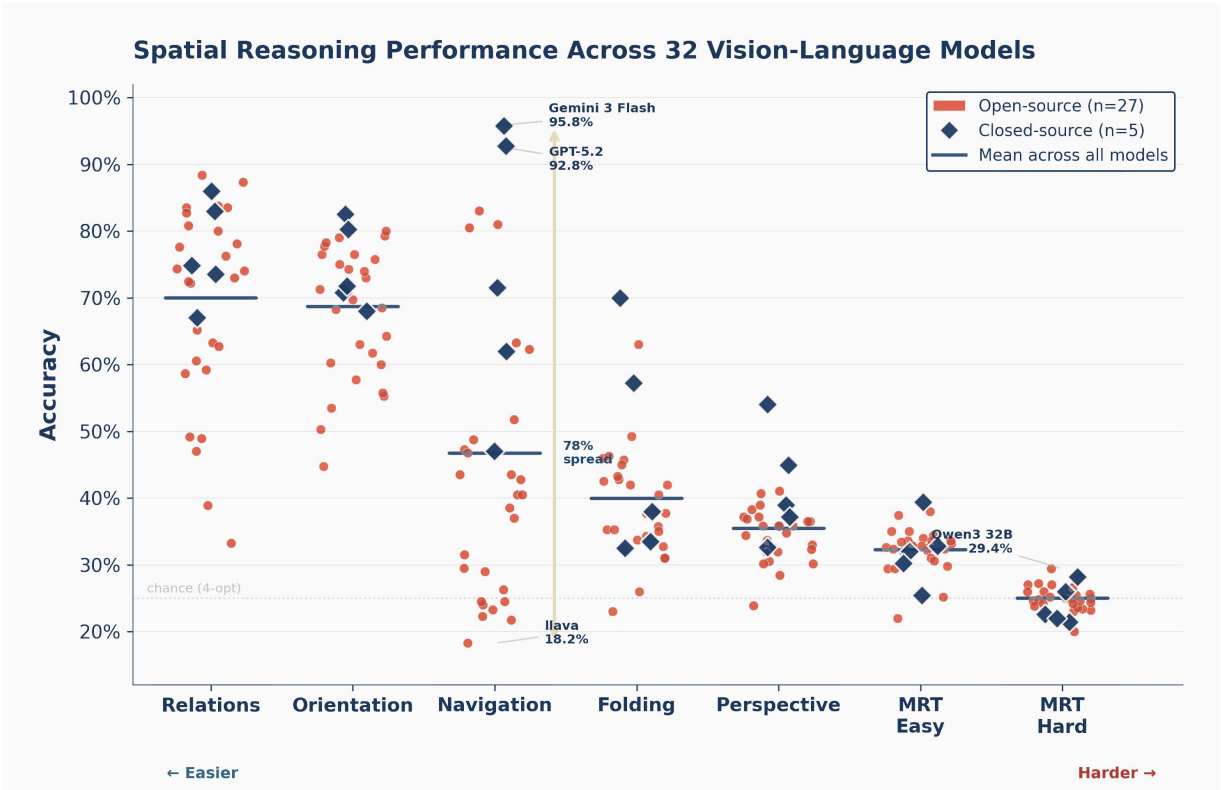


Figure 1 Performance of 32 MLLMs across seven spatial reasoning tasks. Each dot represents one model, coloured by source type (closed-source in coral, open-source in teal). Tasks are ordered by mean difficulty from left (easiest) to right (hardest). Navigation exhibits the widest spread (77.5pp), serving as a visual-grounding probe, while MRT Hard clusters all models near chance (~25%).

matches a 241B model on 3D mental rotation. Bridging this spatial reasoning gap will require qualitative changes to both architecture and training, including equivariant representations for 3D transformations and objectives that reward intermediate visual simulation rather than purely verbal reasoning.

2 Related Work

The spatial reasoning gap is well established: models reach only 25–60% accuracy on many benchmarks, while humans exceed 80%. However, most prior work is *descriptive*, cataloguing low accuracy, rather than *diagnostic*, explaining *how* models fail, whether architectures share failure modes, or whether extended reasoning helps or harms. Our study adds this diagnostic layer: we perform fine-grained error analyses, measure how reasoning length relates to correctness, and compare thinking-orientated vs. instruction-orientated models, evaluating 32 models across seven task splits derived from standard cognitive tests.

2.1 Foundational Spatial Cognition Assessments

Benchmarks show that spatial reasoning remains a core bottleneck for current MLLMs. Ramakrishnan et al. (2025) report that frontier models, despite strong maths and language skills, perform near chance on basic spatial tasks (navigation, distance estimation, mental rotation), arguing that embodied experience is needed for true spatial reasoning. Jia et al. (2025) similarly evaluate 36 MLLMs on psychology-grounded tasks: their best model scores 36 points below humans and only slightly above chance. Li et al. (2025a), using items from standard spatial aptitude tests and expert annotations to separate perceptual complexity from reasoning, test 14 MLLMs and 402 humans; human correctness tracks cognitive features, but MLLM correctness is almost random per instance. Xu et al. (2025) use a psychometric framework to isolate Gardner's five basic spatial abilities, finding that MLLMs mirror human hierarchies—strongest in 2D orientation, weakest in 3D mental rotation—but hit an architectural ceiling that blocks mental simulation. Zhang et al. (2025c) introduce SPHERE, a three-level benchmark from basic perception to complex reasoning, and show that basic spatial skills do not transfer: even with context, models rarely exceed 60% accuracy on reasoning tasks and exhibit strong viewpoint biases.

2.2 3D and Perspective-Based Reasoning

3DSRBench (Ma et al., 2025) is the first comprehensive 3D spatial reasoning benchmark, covering 12 question types on real and synthetic images. Using FlipEval (horizontal flips) and common vs. uncommon viewpoints, they report accuracy drops of up to 25 points on uncommon views. Orientation tasks are hardest, location tasks easier, indicating overfitting to training distributions rather than robust spatial reasoning. Wang et al. (2025d) extend this to 6D spatial reasoning with synthetic questions at five difficulty levels; their Relative Performance Dropping Rate shows models rely on 2D visibility instead of true 3D reasoning. Li et al. (2025b) study perspective taking and find a strong asymmetry: MLLMs handle egocentric tasks but nearly fail at allocentric ones, barely above chance. Their Multi-View Spatial Model, trained on multi-perspective data, improves performance by 46.24%, highlighting data gaps. Zhang et al. (2025d) analyse Frame of Reference ambiguity across languages and find MLLMs prefer specific conventions (English-like reflected transformations) yet stay inconsistent even when perspectives are explicit. Zhang et al. (2025b) introduce the first UAV spatial reasoning benchmark, showing absolute metric estimation is harder than relative positioning. Native 3D LLMs still do not outperform 2D MLLMs, suggesting explicit 3D input alone does not ensure reliable spatial understanding.

2.3 Mental Simulation and Transformation

Huang et al. (2025) introduce a taxonomy along Intrinsic/Extrinsic × Static/Dynamic axes with procedurally generated 3D transformation tasks. Across 28 models, average accuracy is 28.2% versus 76.8% for humans. Gemini-Flash performs better on dynamic than static tasks, suggesting heuristics rather than genuine spatial understanding. Errors mainly stem from reasoning, not perception, indicating a spatial-processing bottleneck rather than visual encoding. Wang et al. (2025a) programmatically create 1,180 items for four spatial sub-abilities (mental rotation, folding, penetration, animation), finding that ~60% of errors arise from perceptual and spatial transformation failures. Chain-of-Thought prompting increases the advantage of closed-source models while reducing open-source accuracy—the reverse of its usual effects in linguistic and mathematical reasoning. Li et al. (2025c) test ~4,000 visual simulation tasks, showing that models handle simple 2D transformations but fail on harder ones such as cube net folding and tangram assembly. Intermediate visual steps can *worsen* performance for models like Claude-3.5 and Gemini-2.0-Flash, suggesting that visual scaffolding disrupts their heuristic strategies. Using fill-in-the-blank evaluations, Jiang et al. (2025) probe full reasoning trajectories on spatial planning and 3D assembly. Even with detailed chains of thought, state-of-the-art models score 0% on standard assembly benchmarks, and accuracy falls from 70% to under 1% as blanks increase from one to four, revealing a sharp disconnect between verbose reasoning and true spatial competence.

2.4 Multi-View and Embodied Spatial Understanding

Yang et al. (2025) evaluate spatial reasoning on real-world indoor videos with configurational, measurement, and spatiotemporal tasks: humans reach 79% accuracy, while models plateau near 46%. Models detect objects but fail at spatial arrangement and geometry, indicating that spatial reasoning is harder than basic perception. They also show that models build local world representations but struggle to integrate them into global maps. Yin et al. (2025) study multi-view image collections and find standard MLLMs perform only slightly above chance on spatial consistency; providing ground-truth spatial maps can even hurt performance, whereas a map-then-reason framework—combining supervised fine-tuning and reinforcement learning—improves accuracy. Wang et al. (2025c) combine a bottom-up survey of 31 datasets with top-down cognitive theories to create 8,068 balanced spatial reasoning tasks, and find a strong correlation ($r = 0.902$) between spatial intelligence scores and robotic manipulation success, indicating that benchmark performance predicts embodied performance. Jung et al. (2024) analyze first-person orientation understanding and uncover strong orientation biases and labelling inconsistencies in current MLLMs; by emphasizing egocentric spatial supervision, their method improves orientation understanding without harming general performance, showing that targeted training can correct specific spatial reasoning weaknesses. Our work spans these levels—from scene-level perception through 3D mental transformation and perspective-taking—but shifts the focus from aggregate accuracy to diagnostic depth: granular error analysis, cross-model agreement on wrong answers, and the relationship between reasoning effort and correctness.

3 The Benchmark

Human spatial reasoning stems from cognitive abilities for navigating, manipulating, and understanding three-dimensional space (Hegarty, 2010; Darken et al., 1999; Wang and Spelke, 2002). Using tasks adapted from gold-standard psychometric assessments (Ekstrom and Harman, 1976; Peters et al., 1995; Shepard and Metzler, 1971), we systematically evaluate latent spatial intelligence factors across object-centric, egocentric, and allocentric

Table 1 Comparison against prior spatial-reasoning benchmarks. ✓ = addressed; – = not addressed. † Novel diagnostic introduced in our work.

Dimension	STARE	ViewSpatial	3DSR	MindCube	OmniSpatial	11-Plus	BSA	Sphere	Ours
Psychometrically grounded task design									
Cognitive-test adaptation	✓	✓	–	✓	–	✓	✓	–	✓
Perception. vs. simulation split †	–	–	–	–	–	–	–	–	✓
Evaluation breadth									
≥ 15 models	–	–	✓	✓	✓	✓	✓	✓	✓
≥ 5 arch. families	✓	✓	✓	✓	✓	–	✓	✓	✓
Cross-scale, sub-1B–200B+ †	–	–	–	–	–	–	–	–	✓
Fine-grained error analysis									
Distractor selection rates †	–	–	–	–	–	–	–	–	✓
Cross-model agreement †	–	–	–	–	–	–	–	–	✓
Reasoning-length quant.	–	–	✓	–	✓	✓	✓	–	✓
Thinking vs. Instruct †	–	–	–	–	–	–	–	–	✓

reference frames, including mental rotation, spatial visualization, perspective-taking, relational understanding, orientation, and egocentric navigation.

3.1 Mental Rotation and Visualisation

We evaluate mental rotation in MLLMs by adapting the classic Mental Rotation Test (MRT) (Cooper, 1975). The original MRT shows pairs of rotated 3D objects or letters and asks participants to distinguish identical shapes from mirror images (Shepard and Metzler, 1971), measuring accuracy and response time at 0°, 60°, 120°, and 180° (F Caissie et al., 2009). Our digital version preserves this protocol while tailoring it to MLLMs. We design five polycube shapes and generate images with a target shape in the top row and four candidates below: one is the target rotated by 0°, 60°, 90°, or 120°; the others are two mirrored versions at different rotations and one unrelated shape. To vary difficulty, *MRT-Hard* uses white shapes on blank backgrounds, while *MRT-Easy* uses coloured shapes on a 3D Cartesian grid and reduces choices to three by removing one mirrored candidate. Each subset has 500 test cases. We also assess more complex spatial reasoning with an adaptation of the Paper Folding Test (Ekstrom and Harman, 1976; McGee, 1979), a measure linked to success in spatially demanding domains (Carroll, 1993). Each item shows a square of paper folded once or twice (vertically, horizontally, or diagonally) and punched with one to three holes. The model must choose among three alternatives for the unfolded hole pattern, testing its ability to internalise sequential geometric operations and simulate their combined effects. This subset also has 400 test cases.

3.2 Spatial Relations

We assess scene understanding using a curated 370-question subset of the 2,000-question Spatial-Obj dataset (Shiri et al., 2024), a multiple-choice benchmark on spatial relationships in natural images. Built via a dual-stage annotation process, Spatial-Obj covers 36 relations, from basic positions (e.g. ‘ right of, ‘ above), to more complex interactions (e.g. ‘attached to’, ‘touch’, ‘overlapping’). The questions probe object localisation, orientation discrimination, and contextual spatial reasoning, providing a robust test of MLLMs’ grasp of relational spatial language in realistic scenes.

3.3 Orientation and Navigation

Next, we evaluate navigation and orientation, both crucial for real-world spatial intelligence. For navigation, we use the Maze-Nav component of SpatialEval (Wang et al., 2024), which tests models on paths through visual mazes of coloured blocks. Tasks include finding routes from start (S) to exit (E), counting direction changes, and describing spatial relations between key locations. We treat this split as a task-specific visual-grounding probe: it tests whether models condition route choices on the maze image, not whether they possess broad spatial-reasoning competence. For orientation, we used 400 binary questions from EgoOrientBench (Jung et al., 2024), which standardises spatial orientation evaluation in a camera-centric framework. An eight-class egocentric taxonomy (Left, Right, Front-Left, Back-Right, etc.) defines object orientations relative to the observer. This egocentric design improves the reliability of the evaluation and aligns with the need for MLLMs to work in real-world, user-centred settings such as robotics, where spatial understanding must be grounded in a human perspective. Each subset has 400 test cases.

Table 2 Per-task random-guessing baselines. We report every split against its own chance level because the answer spaces differ across tasks.

Split	Answer format	Random baseline
MRT Easy	3-way multiple choice	33.3%
MRT Hard	4-way multiple choice	25.0%
Folding	3-way multiple choice	33.3%
Navigation	4-way multiple choice	25.0%
Orientation	Binary yes/no	50.0%
Perspective	4-way multiple choice	25.0%
Relations	4-way multiple choice	25.0%

3.4 Perspective Taking

Finally, we evaluate the allocentric perspective-taking capabilities of MLLMs. Although humans generally find it relatively easy to mentally simulate visual experience from the point of view of another individual, this ability is substantially more challenging for MLLMs (Zhang et al., 2025c; Jia et al., 2025). For this subset, we select all questions from "What's UP" (Kamath et al., 2023) that explicitly require the adoption of another agent's visual or cognitive perspective, resulting in a total of 285 image–question pairs.

4 Experiments and Analysis

4.1 Setup

We evaluate a total of 32 MLLMs. Specifically, we consider the QwenVL3 series (Bai et al., 2025) with model sizes of 8 and 32 billion parameters, as well as mixture-of-experts (MoE) variants with 30B parameters (3B active per token) and 235B parameters (22B active per token). For each of these, we include both *instruct* and *thinking* versions. We further evaluate the InternVL3.5 series (Wang et al., 2025b) with 8, 14, and 38 billion parameters, together with MoE variants with 30B parameters (3B active per token) and 241B parameters (22B active per token). In addition, we include Llava 1.5 7B (Liu et al., 2024), LlavaNext 7B (Li et al., 2024), LLaVA-OneVision 7B , Idefics3 8B (Laurençon et al., 2024) and SmolVLM2 with 500M and 2.2B parameters (Marafioti et al., 2025). We also evaluate MiniCPM-V-4.5 9B (Yu et al., 2025), Llama 3.2-Vision (Meta AI, 2024) with 11B and 90B parameters, and Gemma 3 with 4B, 12B and 28B parameters. Finally, we include Llama 4 Scout and Llama 4 Maverick (Meta AI, 2025), as well as GLM-4.6V (Team et al., 2026). Among closed-source systems, we evaluate Gemini 3 Flash, GPT-5.2, Claude Sonnet-4.5, and Grok 4.1 Fast. We access proprietary models, Llama 4 Maverick and GLM-4.6V, via the OpenRouter API, whereas all other models are run on a compute cluster equipped with 8 NVIDIA H100 GPUs (full evaluation protocol in Appendix K).

4.2 The Perception–Simulation Dissociation

If models truly understand space, they should handle both describing a scene and transforming it. They do not. Table 3 shows that performance in *perceptual* tasks—Relations (μ = 70.0%) and Orientation (μ = 68.6%)—is moderate, but accuracy collapses in *simulation* tasks, falling to μ = 25.0% in MRT Hard, the chance baseline for a four-option task. The gap is not gradual; it is a cliff.

Navigation as Visual-Grounding Probe. Navigation splits models into two groups based on a single property: whether they ground route planning in the visual maze. Visually grounded models (e.g., Gemini-3-flash-preview, 95.8%) correctly trace the maze, while most lower-ranked models ignore the image and hallucinate generic routes. This yields a 77.5 pp spread and the strongest correlation with overall ranking (r = 0.91), making Navigation the most diagnostic *visual-grounding* split in the benchmark rather than standalone evidence of mental simulation.

Chance-Level Ceiling on Mental Rotation. MRT Hard tells the opposite story: no spread at all. All 32 models cluster between 20% and 29.4%, regardless of architecture, scale, or training data. This ceiling at chance level is the clearest evidence that 3D mental rotation is not a *difficulty* problem but a *capability* problem—one that indicates a failure of the current paradigm to support the necessary internal representations for spatial simulation, which scaling alone does not appear to rectify.

4.3 The Reasoning Length Paradox

If models lack spatial simulation capacity, it becomes pertinent to investigate whether symbolic reasoning can compensate. Empirical evidence suggests otherwise; deeper reasoning is often counterproductive. As Figure 2 shows, across all evaluation splits, incorrect responses use much longer chains of thought than correct ones: 849

Table 3 Comprehensive Benchmark Results for Spatial Reasoning in MLLMs (Sorted by Performance)

Model Name	Overall	Fold.	MRT-E	MRT-H	Nav.	Ori.	Pers.	Rel.
Closed Source / API Models								
Gemini-3 Flash Preview	**63.15%**	**70.00%**	**39.40%**	**28.20%**	**95.75%**	**82.50%**	**54.04%**	**85.95%**
GPT-5.2	51.63%	32.50%	25.40%	21.40%	92.75%	80.25%	38.95%	82.97%
Claude-Sonnet-4.5	51.52%	57.25%	32.00%	26.00%	71.50%	70.75%	37.19%	74.86%
Grok-4.1-fast	44.73%	38.00%	32.80%	22.00%	47.00%	71.75%	44.91%	67.03%
Qwen3-VL-235B (Think.)	56.50%	63.00%	33.00%	25.00%	83.00%	76.50%	37.19%	88.38%
Qwen3-VL-235B	55.80%	49.25%	38.00%	25.00%	81.00%	79.25%	41.05%	87.30%
Qwen3-VL-30B (Think.)	51.21%	46.25%	34.00%	27.00%	63.25%	75.75%	37.19%	83.78%
Qwen3-VL-30B	41.61%	23.00%	22.00%	20.00%	46.75%	76.50%	32.98%	80.81%
Qwen3-VL-32B (Think.)	48.72%	37.75%	32.40%	23.20%	62.25%	77.75%	33.68%	82.70%
Qwen3-VL-32B	53.91%	46.00%	32.80%	29.40%	80.50%	80.00%	32.63%	83.51%
Qwen3-VL-8B (Think.)	45.92%	33.75%	37.40%	23.40%	48.75%	74.25%	31.93%	78.11%
Qwen3-VL-8B	45.36%	37.75%	29.40%	26.60%	43.50%	78.25%	28.42%	80.00%
InternVL3.5-241B	49.35%	42.50%	35.00%	28.20%	47.25%	79.00%	38.25%	83.51%
InternVL3.5-38B	47.95%	45.75%	33.60%	25.60%	51.75%	73.00%	36.49%	77.57%
InternVL3.5-30B	44.83%	42.75%	32.20%	24.60%	40.50%	69.75%	35.79%	76.22%
InternVL3.5-14B	44.31%	42.00%	34.40%	24.80%	38.50%	71.25%	30.53%	74.32%
InternVL3.5-8B	43.89%	38.00%	32.60%	24.40%	43.50%	68.50%	32.98%	74.05%
GLM-4.6v	44.94%	33.50%	30.20%	22.60%	62.00%	68.00%	32.63%	73.51%
MiniCPM-V-4.5	43.78%	43.25%	32.80%	26.00%	37.00%	68.25%	32.28%	72.97%
Llama-3.2-90B Vision	43.71%	35.25%	33.40%	27.00%	31.50%	74.00%	40.70%	72.16%
Llama-3.2-11B Vision	38.14%	35.75%	35.00%	23.00%	24.50%	55.25%	36.84%	62.70%
Llama-4-Scout-109B-A17B	43.43%	45.00%	31.00%	23.80%	29.00%	75.00%	35.79%	72.43%
Llama-4-Maverick-402B-A17B	41.33%	42.00%	30.60%	26.00%	42.75%	57.75%	30.18%	65.14%
Gemma-3-27B	39.93%	40.50%	32.80%	27.20%	26.25%	60.25%	34.39%	63.24%
Gemma-3-12B	37.55%	35.25%	32.40%	24.20%	22.25%	60.00%	35.79%	58.65%
Gemma-3-4B	34.89%	26.00%	29.80%	26.00%	24.50%	55.75%	38.95%	48.92%
Idefics3-8B	38.49%	34.25%	33.20%	25.00%	24.00%	64.25%	34.74%	59.19%
LLaVA-OneVision-7B	38.39%	32.75%	29.40%	25.40%	29.50%	61.75%	35.79%	60.54%
LLaVA-1.5-7B	34.29%	35.00%	32.00%	25.20%	40.50%	44.75%	23.86%	38.92%
LLaVA-v1.6-Mistral-7B	32.15%	32.25%	25.20%	24.20%	18.25%	50.25%	30.18%	49.19%
SmolVLM2-2.2B	35.97%	31.00%	33.60%	23.60%	21.75%	63.00%	36.49%	47.03%
SmolVLM2-500M Video	32.78%	31.00%	33.60%	24.60%	23.25%	53.50%	31.93%	33.24%

versus 188 words on average, with the ratio approaching 3× on the Navigation and Folding splits. Qualitatively, when models are uncertain about a spatial task, they fail to converge on a coherent solution and instead *thrash*: cycling through incompatible hypotheses, repeatedly backtracking, hallucinating task-relevant details, and producing long post hoc rationalisations that culminate in an incorrect answer.

4.4 Does Extended Reasoning Training Improve Spatial Reasoning?

Four Qwen3-VL model families enable a direct controlled comparison between Thinking and Instruct variants on the same architecture family. Figure 3 visualises the per-split accuracy deltas; the results show that extended reasoning is not uniformly beneficial and interacts with scale, architecture, and post-training recipe.

Qwen3-VL-235B. Thinking (56.5%) and Instruct (55.8%) are essentially tied overall (+0.7 pp). Thinking excels on Folding (+13.8 pp) but Instruct wins on MRT Easy (−5.0 pp) and Perspective (−3.9 pp).

Qwen3-VL-30B. Thinking (51.2%) clearly outperforms Instruct (41.6%) by 9.6 pp overall, winning six of seven splits. The gains are especially pronounced on Folding (+23.3 pp) and Navigation (+16.5 pp)—the clearest case in our study where extended reasoning delivers a consistent benefit.

Qwen3-VL-32B. Instruct (53.9%) outperforms Thinking (48.7%) by 5.2 pp, winning four splits including Navigation (+18.3 pp) and MRT Hard (+6.2 pp). Extended reasoning *actively degrades* performance at this scale.

Qwen3-VL-8B. The two variants are essentially tied (45.9% vs. 45.4%), with each winning three of seven splits. The pattern is neither consistent nor monotonic with scale: gains appear at 30 B, vanish at 8 B and 235 B, and reverse at 32 B. We therefore avoid interpreting Thinking/Instruct differences as evidence for a sharp capacity threshold. The safer conclusion, reinforcing the reasoning-length findings above, is that textual Chain-of-Thought alone is

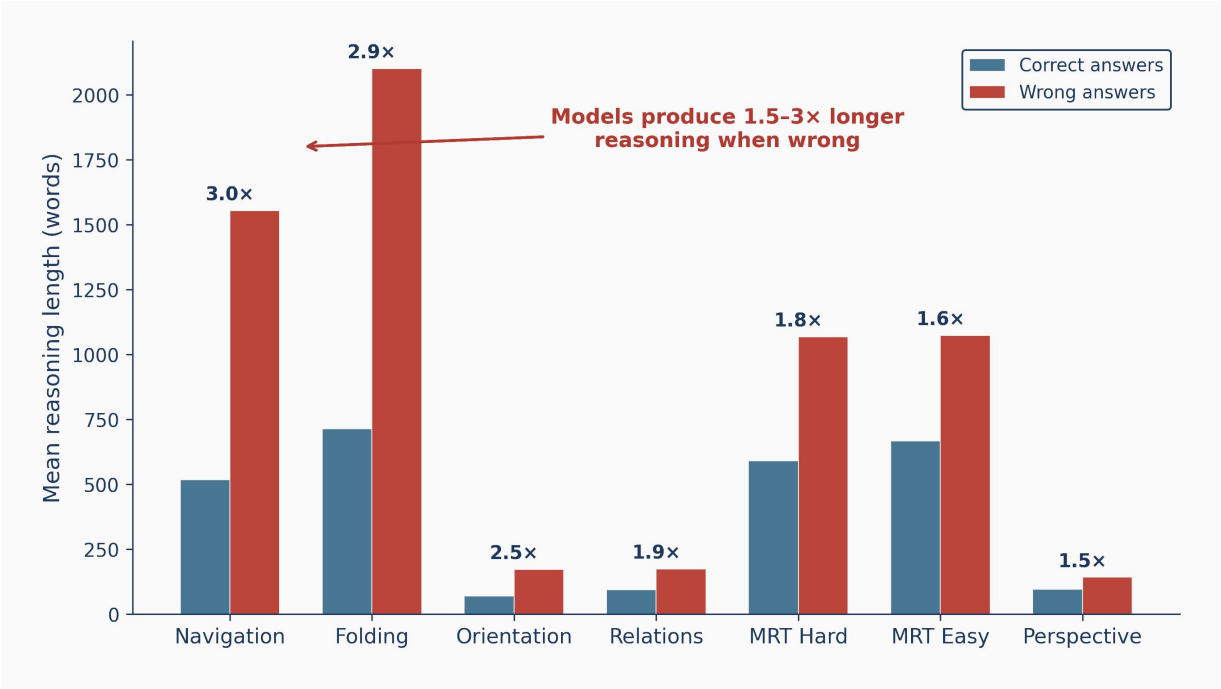


Figure 2 The reasoning length paradox. Across all seven splits, incorrect answers (coral) are accompanied by substantially longer chains of thought than correct answers (teal). The effect is most pronounced on Folding and Navigation, where incorrect reasoning is approximately 3× longer, indicating that models "thrash" through conflicting hypotheses when uncertain rather than reasoning toward the correct solution.

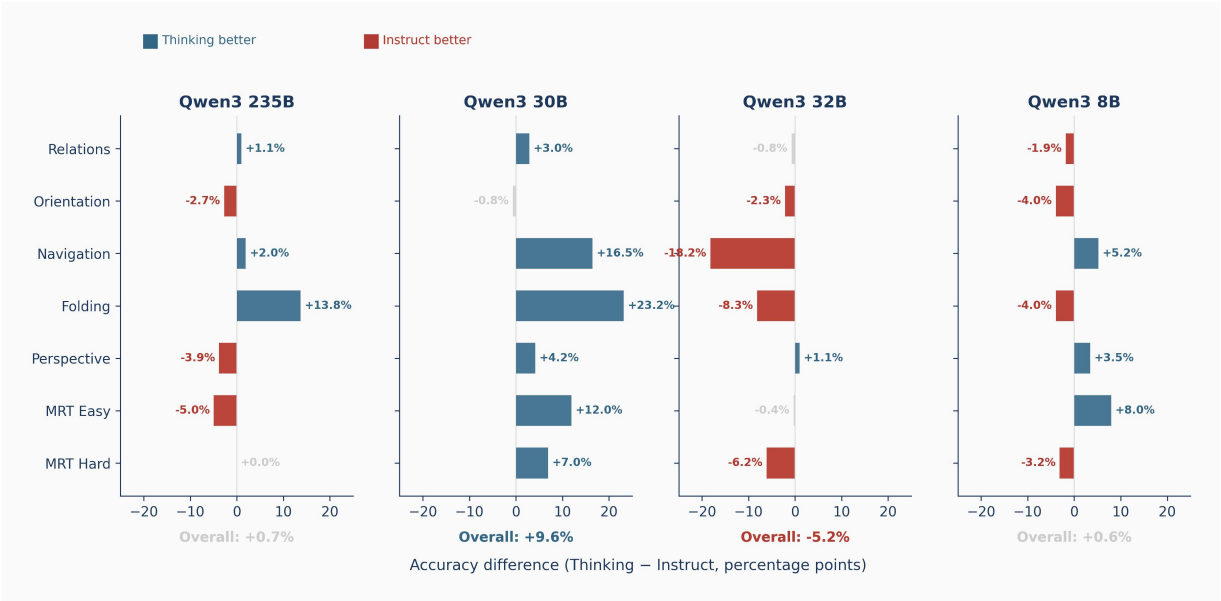


Figure 3 Per-split accuracy delta (Thinking − Instruct) for the Qwen3-VL family. Positive values favour the Thinking variant. The inconsistent pattern across scales confirms that extended chain-of-thought is not a reliable path to spatial competence.

not a reliable route to spatial competence: when the underlying representation cannot support mental simulation, additional reasoning tokens often accumulate errors rather than converge on the correct answer.

4.5 Granular Failure Analysis

The dissociation above shows that models fail in simulation; the error patterns show how. We conduct a fine-grained distractor and confusion analysis for all 32 models on the three hardest splits (Figure 4).

Chirality Confusion Dominates MRT Errors. Under a uniform distribution of errors, each incorrect option would

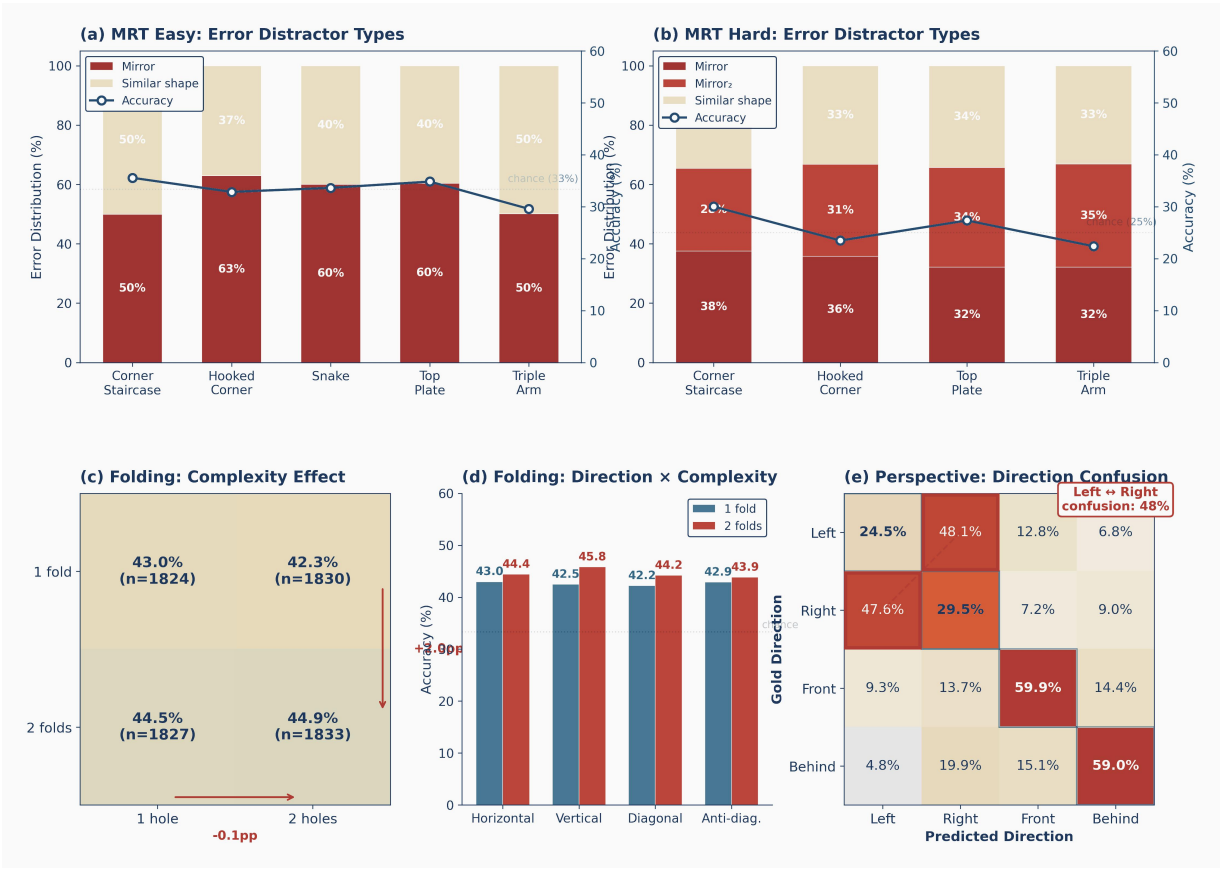


Figure 4 Granular error analysis for the three hardest task splits. (a–b) MRT distractor selection rates by polycube shape: mirror distractors (coral/pink) dominate over structurally dissimilar shapes (gold), indicating partial structural understanding but systematic chirality confusion. (c–d) Paper folding accuracy decomposed by fold count × hole count and by fold direction, revealing a relatively flat complexity landscape. (e) Perspective-taking confusion matrix: left–right confusions (highlighted) account for nearly half of all errors, while front–behind accuracy remains substantially higher.

be selected with equal frequency. Instead, mirror-reflected distractors capture 50–63% of errors on MRT Easy and 66–72% on MRT Hard. Models can reject shapes that are grossly different—their structural understanding is real—but they cannot distinguish an object from its reflection. The failure is specifically one of *chirality*: models perceive 3D structure, yet project it into a representation that is approximately symmetric under reflection.

Paper Folding Errors Are Invariant to Task Complexity. One might expect difficulty to increase with more folds or holes, but it does not. Accuracy stays within 42.3%–44.9% in all fold×holes (Figure 4c–d), and fold direction has little effect. This flat pattern suggests models do not track the sequential transformation but instead use a single shallow heuristic regardless of complexity.

Systematic Left–Right Inversion in Perspective-Taking. The perspective split produces the most striking pattern (Figure 4e). When the correct answer is "left," models respond "right" 48.1% of the time; when the correct answer is "right," they respond "left" 47.6% of the time. Lateral accuracy (24.5% and 29.5%) hovers at or below the chance level of four-options. In contrast, "Front and "behind," are identified at ~59%. The failure is diagnostic: models default to the camera-centric frame and cannot flip left and right when adopting another viewpoint. This single error mode accounts for nearly half of all perspective errors.

4.6 Cross-Model Error Agreement

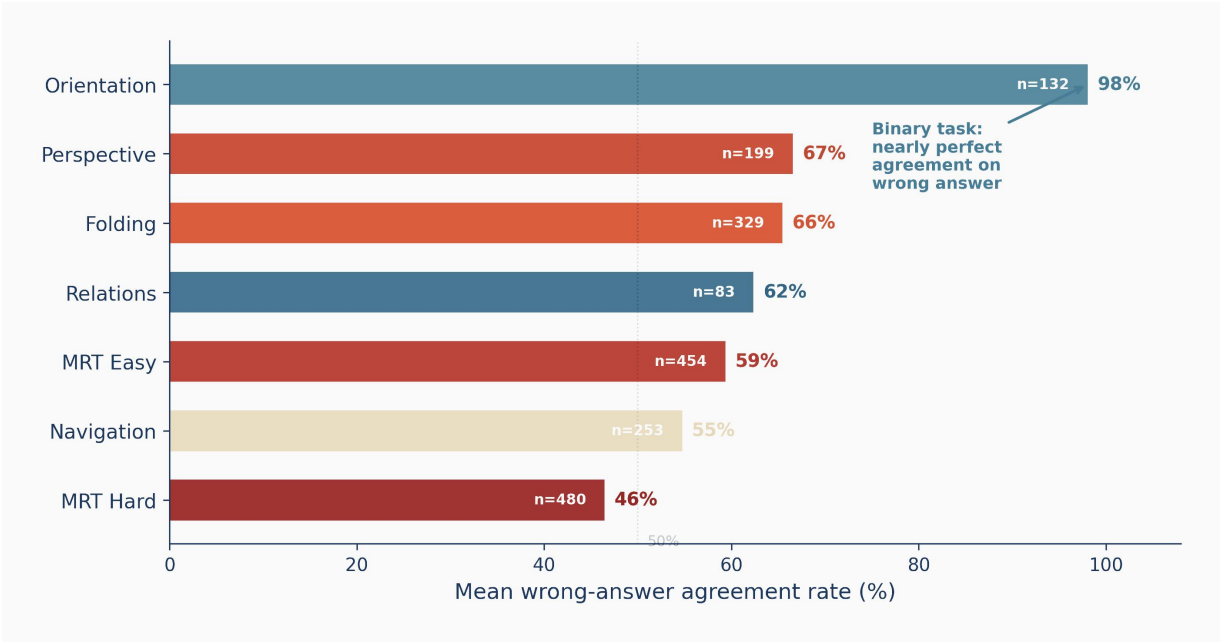
The preceding analysis shows *what* models confuse; we now ask whether those confusions are idiosyncratic or shared. Under a uniform error distribution, wrong-answer agreement would be expected at 33% for four-option tasks and 50% for three-option tasks by chance. Instead, observed rates substantially exceed these baselines (Table 4): Perspective reaches 66.6% against a 33% baseline and Folding 65.5% against a 50% baseline, indicating statistically consistent consensus on the same incorrect answer rather than independent random failure.

Answer Position Bias. For MRT Easy (three options: A, B, C), wrong answers skew toward option B, consistent

Table 4 Wrong-answer agreement rates across hard questions. "Mean agr." reports the average proportion of wrong-answering models that select the single most common incorrect option.

Split	N hard Qs	Mean agr.	>50% agr.	>70% agr.
Orientation	132	98.1%	132	132
Perspective	199	66.6%	169	112
Folding	329	65.5%	305	110
Relations	83	62.4%	63	34
MRT Easy	454	59.4%	317	87
Navigation	253	54.8%	129	38
MRT Hard	480	46.5%	126	10

Figure 5 Systematic wrong-answer agreement across splits. Higher agreement indicates that models converge on the same incorrect answer, revealing shared failure modes rather than random errors.



with a hedging bias toward middle options under uncertainty. For MRT Hard (four options: A–D), the bias is more distributed but option D is selected less frequently. However, per-question analysis indicates that errors are predominantly driven by genuine task difficulty rather than positional preference.

Wrong-Answer Agreement. Orientation shows near-perfect wrong-answer agreement (98.1%): when models err, nearly all select the same incorrect option, as expected given the binary nature of the task. Perspective and Folding also exhibit high systematic error rates (>65%), indicating that models consistently misinterpret the same spatial features rather than making random errors. MRT Hard has the lowest agreement (46.5%), as the four-option format and genuine confusion lead to a greater scattering of incorrect responses.

Error Taxonomy. Analysis across splits reveals seven recurring error types: *(1) Perspective transformation failure* (~36–50% of perspective errors): describing the scene from the camera view but labeling it as the observer view, ignoring the left–right mirror rule. *(2) 3D mental rotation breakdown* (~95% of MRT errors in weaker models): accurately describing structures but failing to track their appearance after rotation. *(3) Reference frame confusion* (~40% of relation errors): mixing camera-centric and object-centric frames within a single response. *(4) Visual grounding failure* (only in the weakest navigation models): producing plausible maze-navigation text that is unrelated to the image. *(5) Incomplete reasoning* (~10–15% across splits): truncating the reasoning chain or not reaching a conclusion. *(6) Structure misidentification* (~5–10%, mainly MRT): miscounting cubes or misidentifying connectivity, leading to incorrect answers. *(7) Ambiguity handling* (~15–20% of orientation errors): failing on genuinely ambiguous views where "front-left" vs. "left" is unclear.

These patterns show spatial failures are systematic, not random noise, and stem from how current architectures represent space.

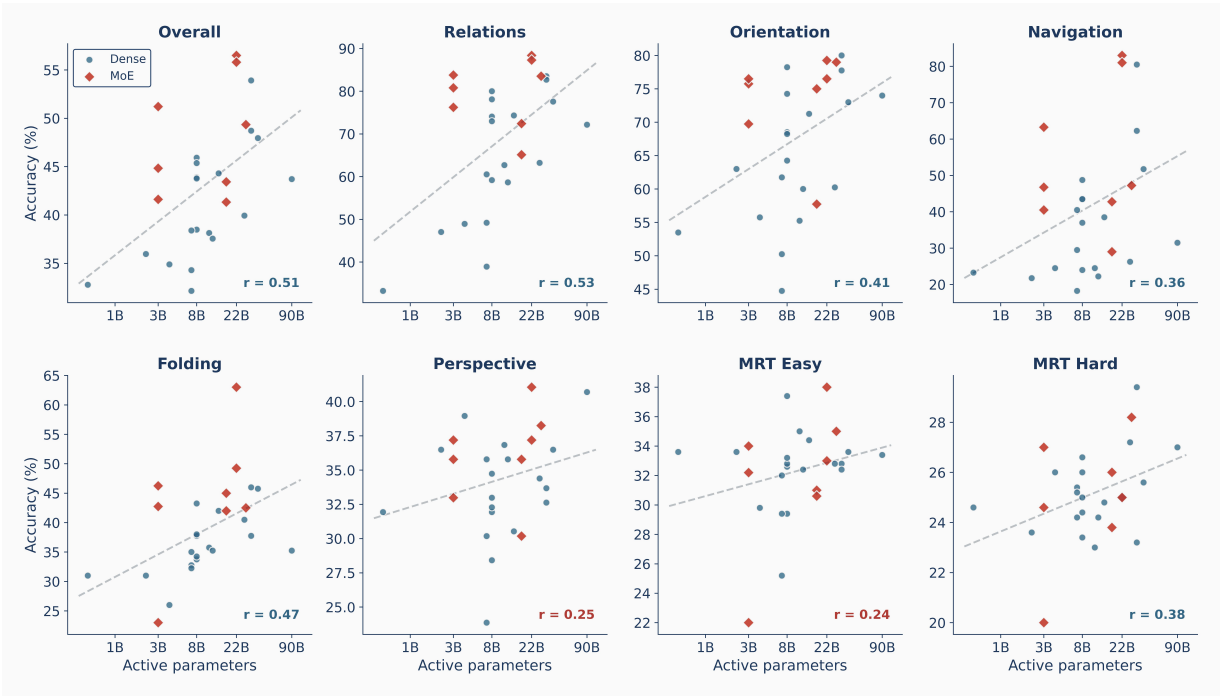


Figure 6 Scaling behaviour of 27 open-source models. Each panel plots per-split accuracy against log-scaled active parameters, distinguishing dense (circles) from MoE (triangles) architectures. Relations exhibits moderate scaling ($r = 0.53$), whereas MRT Hard is essentially flat ($r = 0.38$), confirming that 3D mental rotation does not improve with model size in the current regime.

4.7 Performance Does Not Scale with Model Size

Could the near-chance performance on mental simulation simply close with larger models? We relate performance to active parameter count across 27 open-source systems (Figure 6). For perceptual tasks, the answer is a qualified yes: Relations shows moderate scaling ($r = 0.53$), consistent with its reliance on object recognition. But for simulation tasks, the answer is no. MRT Hard is essentially flat ($r = 0.38$): a 7 B model and a 241 B model achieve comparable performance. Navigation sits between ($r = 0.36$), constrained less by model size than by the quality of the visual encoder.

What about proprietary data? Closed-source models outscore open-source ones overall (Mann–Whitney U, $p = 0.026$; $\mu = 51.2\%$ vs. 43.1%), but the advantage is almost entirely explained by Navigation ($\Delta = +32.1$ pp, $p = 0.008$). On mental rotation, there is no statistically significant difference. The implication is clear: the MRT wall is not a data problem or a compute problem. It is a *representation* problem – one that will require architectural interventions such as equivariant representations capable of modelling 3D transformations, rather than additional scale or proprietary training data.

5 Discussion

Our findings suggest several directions for closing the spatial reasoning gap.

Visual Reasoning over Textual Reasoning. Text-based Chain-of-Thought can hinder spatial tasks: models that reason longer perform worse, and extended thinking gives inconsistent gains across matched model variants. Providing intermediate visual information also brings inconsistent or negative gains (Li et al., 2025c), implying that the problem is not simply access to visual input but insufficient representational capacity to use it. A promising approach is to train models to generate and reason over intermediate visual representations—rotated views, annotated diagrams, or mental imagery sketches—that externalise rather than verbalise spatial transformations.

As a proof of concept, we additionally evaluated Chain-of-Visual-Thought (CoVT) (Qin et al., 2025), a model trained to generate intermediate visual signals during reasoning. CoVT shows a modest gain on MRT Hard relative to the 32-model pool (+3.0 pp; one-sided paired $p = 0.038$), suggesting that our diagnostic suite can identify a failure mode and motivate a targeted, measurable intervention. The effect is narrow: CoVT does not outperform the other models overall ($\Delta = -3.3$ pp; exact one-sided permutation test $p = 0.80$), and remaining per-task effects

are not significant ($p \geq 0.50$). We therefore treat this as preliminary evidence that visual intermediates may help mental rotation specifically, not as a general solution.

Chirality-Aware Training. The mirror-distractor bias on MRT (Figure 4a–b) indicates that visual encoders project 3D structure into representations that are approximately symmetric under reflection. Chirality-sensitive contrastive losses or architectural inductive biases that preserve handedness through the encoding pipeline may be needed to resolve this.

Reference Frame Grounding. The left–right confusion in perspective-taking (Figure 4e) and the affirmation bias in orientation share a common cause: models default to the camera-centric frame and cannot transform into an allocentric one. Embodied multi-viewpoint supervision or explicit reference-frame tokens at inference time could anchor spatial computations to the correct frame.

Targeted Curricula. Cross-split correlations (Appendix G) reveal three largely independent capability clusters: visual grounding (navigation, orientation, relations), object transformation (MRT, folding), and perspective-taking. Because gains in one cluster do not transfer to others, targeted spatial curricula with specialised data and objectives are preferable to monolithic training.

Beyond Scale. MRT Easy accuracy is essentially uncorrelated with model size ($r = 0.12$): a 241B-parameter model matches a 7B model, and closed-source systems offer no benefit. The bottleneck is architectural, not statistical: treating visual inputs as flattened, unordered token sequences lacks the inductive biases needed for coordinate-invariant reasoning. Current MLLMs excel at semantic recognition but lack equivariant 3D spatial representations. Further progress may require explicit geometric reasoning—equivariant networks, spatial transformers, or neuro-symbolic modules—to break the current performance ceiling.

6 Conclusion

Current MLLMs can describe static spatial relationships but cannot perform the dynamic mental simulations needed for rotation, folding, or perspective-taking. By separating *static spatial perception*—reading off what is visible—from *dynamic mental simulation*—mentally transforming what is seen—we reveal a dissociation that high aggregate scores hide. Across 32 models and 2,855 diagnostic items, the pattern is clear: strong scene description coexists with near-chance mental rotation, mirror confusions indicating absent chirality, left–right inversions revealing a fixed reference frame, and a paradox where more reasoning yields worse answers. Crucially, these failures do not improve with scale. The gap is not about model size or data, but about what representations can express. Advancing spatial intelligence requires architectures that move beyond description to maintain stable, consistent representations across egocentric, allocentric, and object-centric reference frames.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

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*, pages 14455–14465, 2024.

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. arXiv:2406.01584.

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. *The Open Psychology Journal, 2, Article 94-102.*, 2009. doi: https://psycnet.apa.org/doi/10.2174/1874350100902010094.

Mary Hegarty. Chapter 7 - components of spatial intelligence. In *The Psychology of Learning and Motivation*, volume 52 of *Psychology of Learning and Motivation*, pages 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.

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.

Xinmiao Huang, Qisong He, Zhenglin Huang, Boxuan Wang, Zhuoyun Li, Guangliang Cheng, Yi Dong, and Xiaowei Huang. Spatial-dise: A unified benchmark for evaluating spatial reasoning in vision-language models. *arXiv preprint arXiv:2510.13394*, 2025.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*, 2025.

Yulun Jiang, Yekun Chai, Maria Brbić, and Michael Moor. Marble: A hard benchmark for multimodal spatial reasoning and planning. *arXiv preprint arXiv:2506.22992*, 2025.

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.

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.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Qingtao Li, Zeyu Gao, Yan Xia, José Hernández-Orallo, Ivan Vulić, and Furu Wei. 11plus-bench: Demystifying multimodal llm spatial reasoning with cognitive-inspired analysis. *arXiv preprint arXiv:2508.20068*, 2025a.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models, 2025b. URL https://arxiv.org/abs/2505.21500.

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.

Linjie Li, Mahtab Bigverdi, Jiawei Gu, Zixian Ma, Yinuo Yang, Ziang Li, Yejin Choi, and Ranjay Krishna. Unfolding spatial cognition: Evaluating multimodal models on visual simulations, 2025c. URL https://arxiv.org/abs/2506.04633.

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)*, pages 26296–26306, June 2024.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6924–6934, 2025.

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.

Meta AI. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models — ai.meta.com. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/, 2024. [Accessed 07-02-2026].

Meta AI. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation — ai.meta.com. https://ai.meta.com/blog/llama-4-multimodal-intelligence/, 2025. [Accessed 07-02-2026].

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

M Peters, B Laeng, K Latham, M Jackson, R Zaiyouna, and C Richardson. A redrawn vandenberg and kuse mental rotations test: different versions and factors that affect performance. *Brain Cogn.*, 28(1):39–58, June 1995.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and XuDong Wang. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens, 2025. URL https://arxiv.org/abs/2511.19418.

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.

Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, and Vladlen Koltun. Does spatial cognition emerge in frontier models? In *The Thirteenth International Conference on Learning Representations*, 2025.

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.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Haochen Li, Jiale Zhu, Jiali Chen, Jiaxing Xu, Jiazheng Xu, Jing Chen, Jinghao Lin, Jinhao Chen, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Ruiliang Lyu, Shangqin Tu, Sheng Yang, Shengbiao Meng, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wei Jia, Wenkai Li, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyu Zhang, Xinyue Fan, Xuancheng Huang, Yadong Xue, Yanfeng Wang, Yanling Wang, Yanzi Wang, Yifan An, Yifan Du, Yiheng Huang, Yilin Niu, Yiming Shi, Yu Wang, Yuan Wang, Yuanchang Yue, Yuchen Li, Yusen Liu, Yutao Zhang, Yuting Wang, Yuxuan Zhang, Zhao Xue, Zhengxiao Du, Zhenyu Hou, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2026. URL https://arxiv.org/abs/2507.01006.

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.

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.

Siting Wang, Minnan Pei, Luoyang Sun, Cheng Deng, Kun Shao, Zheng Tian, Haifeng Zhang, and Jun Wang. Spatialviz-bench: An mllm benchmark for spatial visualization. *arXiv preprint arXiv:2507.07610*, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025b. URL https://arxiv.org/abs/2508.18265.

Wenqi Wang, Reuben Tan, Pengyue Zhu, Jianwei Yang, Zhengyuan Yang, Lijuan Wang, Andrey Kolobov, Jianfeng Gao, and Boqing Gong. Site: towards spatial intelligence thorough evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9058–9069, 2025c.

Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. Pulsecheck457: A diagnostic benchmark for comprehensive spatial reasoning of large multimodal models. *arXiv preprint arXiv:2502.08636*, 2025d.

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.

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

Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*, 2025.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe, 2025. URL https://arxiv.org/abs/2509.18154.

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*, September 2025a. doi: 10.48550/arXiv.2509.02359.

Weichen Zhang, Zile Zhou, Zhiheng Zheng, Chen Gao, Jinqiang Cui, Yong Li, Xinlei Chen, and Xiao-Ping Zhang. Open3dvqa: A benchmark for comprehensive spatial reasoning with multimodal large language model in open space. *arXiv preprint arXiv:2503.11094*, 2025b.

Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Junqi Zhao, Allison Koenecke, Boyang Li, and Lu Wang. SPHERE: Unveiling spatial blind spots in vision-language models through hierarchical evaluation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11591–11609, Vienna, Austria, July 2025c. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.568. URL https://aclanthology.org/2025.acl-long.568/.

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*, 2025d. URL https://openreview.net/forum?id=84pDoCD4lH.

Supplementary Material

The appendix is organised as follows. Section A documents task sources, modifications, and quality-control checks. Section B provides annotated visual examples for all seven task splits. Section C presents curated failure cases with model reasoning traces, illustrating the recurring error patterns identified in the main paper. Section D details per-split difficulty and discrimination power. Section E analyses per-question difficulty distributions. Section F examines reasoning behaviour—length, uncertainty language, and reasoning–answer alignment. Section G reports cross-split correlations and open- vs. closed-source comparisons. Section I provides granular scaling and error analysis. Section K documents the complete evaluation protocol for reproducibility.

A Task Sources, Modifications, and Quality Control

Table 5 summarises the provenance of each split, the modifications used to standardise it for MLLM evaluation, and the quality-control checks used before evaluation. The key design choice is that answer labels are rule-derived rather than manually inferred wherever possible: rotations, mirror distractors, fold operations, maze paths, and egocentric orientation labels are generated or verified from the construction rules. This reduces annotation ambiguity and ensures that model errors can be analysed against known spatial transformations.

Table 5 Task sources, modifications, and quality-control checks.

Split	Source	Modification	QC check
MRT Easy	Shepard–Metzler-style mental rotation	Five generated polycube families; 3 choices; coloured grid and reduced mirror distractors	Rule-derived rotations and mirror labels; visual inspection of generated renders
MRT Hard	Shepard–Metzler-style mental rotation	Same polycube families; 4 choices; white shapes on blank background	Rule-derived target/mirror/unrelated options; duplicate and invalid render checks
Folding	Paper Folding Test	Digital fold-and-punch items with one or two folds and one to three holes	Programmatic unfold simulation validates final hole pattern
Relations	Spatial-Obj subset	Curated 370 questions spanning 36 spatial relations	Source labels retained; malformed answer options filtered
Navigation	SpatialEval Maze-Nav	Maze route questions used as a visual-grounding probe	Path validity checked from start/exit coordinates and obstacle layout
Orientation	EgoOrientBench	400 binary egocentric orientation questions	Source taxonomy checked against eight camera-centric orientation classes
Perspective	What'sUp perspective subset	285 questions requiring another agent's viewpoint	Questions filtered for explicit allocentric perspective taking and unambiguous options

B Annotated Task Examples

C Illustrative Examples of Reasoning Failures

Each example below includes the task image, the model's reasoning, the correct answer, and a brief analysis. Failure cases use **coral** boxes; the contrasting **teal** box highlights a success case for comparison.

C.1 The Perspective Mirror-Inversion Problem

Our granular error analysis (Section J) reveals that models systematically confuse left and right when asked to adopt an alternative viewpoint. The model correctly identifies the spatial layout from the *viewer's* frame but fails to invert the lateral axis when projecting into the *subject's* frame.

(g) Perspective taking

Figure 7 Annotated examples for all seven task splits: (a) MRT Easy, (b) MRT Hard, (c) Paper Folding, (d) Spatial Relations, (e) Orientation, (f) Navigation, and (g) Perspective taking. The examples illustrate the visual input and answer-option structure for each split.

Example 1: Perspective Left/Right Reversal (GPT-5.2)

Task: Perspective **Pred:** B (right) **Gold:** D (left)

Q: Where is the boy from the man's perspective?

Model's Reasoning:

> *"The man is on the left side of the image, and the boy is walking beside him on the man's right-hand side (viewer's right)."*

Analysis: The model correctly perceives that the boy is on the *viewer's* right. However, the man walks **towards the camera**, so his left and right are mirrored relative to the viewer. The boy is on the man's **left**. This is the canonical perspective mirror-inversion error: the model maps viewer-frame coordinates directly onto the subject's frame without performing the requisite 180° reference-frame rotation.

Example 2: Object-Relation Perspective Transfer (GPT-5.2)

Task: Perspective **Pred:** B (right) **Gold:** C (left)

Q: Where is juice glass from woman's perspective?

Model's Reasoning:

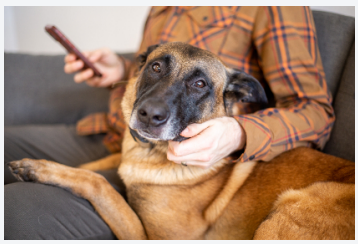
> *"The woman is positioned on the left side of the scene, leaning toward the table. The juice glass is placed to the right of the plate. From the woman's point of view as she faces the table, that puts the juice glass on her right side."*

Analysis: The model constructs a correct viewer-centric description—the glass *is* to the right of the plate from the camera's vantage. Yet since the woman faces *away* from the camera, the glass is on her **left**. The error is identical to Example 1: a failure to invert the lateral axis when transferring between reference frames.

C.2 Hand Laterality Confusion in Spatial Relations

A closely related failure mode emerges in the Relations split, where models must identify which hand a person uses to perform an action. This requires the same viewer-versus-subject frame resolution, but grounded in body-part laterality.

Example 3: Hand Laterality Inversion (GPT-5.2)



Task: Relations **Pred:** B (right) **Gold:** A (left)

Q: Which hand is the man using to pat the dog?

Model's Reasoning:

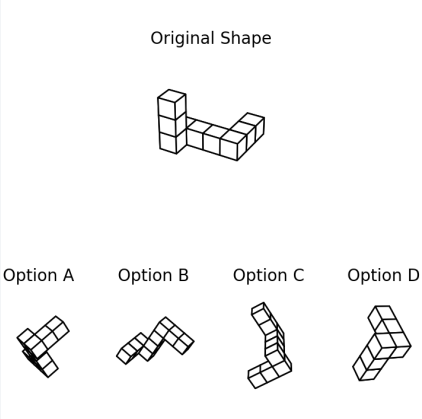
> *"The man is holding a phone with one hand and using the other hand to touch the dog's chin/neck. The hand patting the dog is the one not holding the phone, which is his right hand."*

Analysis: The exclusion logic is valid—the patting hand *is* the one not holding the phone—but the laterality label is wrong. Because the man faces the camera, his left hand appears on the *viewer's right*. The model conflates *image-space position* (right side of the frame) with *body-space identity* (right hand), reproducing the mirror-inversion error at the level of body parts.

C.3 Exhaustive Reasoning Without Spatial Comprehension

The most striking finding from our reasoning behaviour analysis (Section F) is that *longer reasoning does not guarantee correct answers*. The example below is the most extreme case in our dataset: nearly two thousand characters of meticulous structural analysis that still arrive at the wrong answer.

Example 4: Exhaustive Structural Analysis, Wrong Rotation (GPT-5.2)



Task: MRT Hard **Pred:** B **Gold:** C

Q: Which option is the original shape in a rotated orientation?
Reasoning length: 1,851 characters

Model's Reasoning (excerpt):

> *"1) Extract the structure of the original 3D polycube: a straight bar of 4 cubes along the top; from the second cube, a 1-cube drop, then one more cube extending outward. Total = 6 cubes.*
> *2) Key fingerprint features preserved under rotation: exactly one cube has 3 neighbours; the branching point is one in from an end. [. . . 1,200 more characters. . .]*
> *3) Option B preserves the same connectivity. Therefore, only Option B is a rotated version."*

Analysis: The model identifies rotation-invariant properties (cube count, connectivity fingerprint, branching topology) and checks each option against them—exactly the correct procedure. Yet the final answer is wrong. This reveals a fundamental **reasoning-perception decoupling**: the model articulates *what* should be invariant under rotation but cannot reliably *perceive* whether those invariants hold in a given 3D rendering. The reasoning is genuine spatial logic applied to faulty visual perception.

D Split Difficulty and Discrimination Power

This section provides the quantitative foundations for the difficulty hierarchy discussed in Section 4.2. Table 6 presents the complete difficulty statistics across all seven task splits. The rightmost column reports the Pearson correlation between each split and overall accuracy, indicating how well the split discriminates between strong and weak models.

Table 6 Split difficulty statistics across all 32 models. Range is reported in percentage points (pp). The correlation column (r) reports the Pearson correlation of each split's accuracy with overall accuracy.

Split	Mean Acc	Std Dev	Min	Max	Range	r (Overall)
Relations	70.0%	0.144	33.2%	88.4%	55.1pp	0.88
Orientation	68.6%	0.099	44.8%	82.5%	37.7pp	0.85
Navigation	46.7%	0.221	18.2%	95.8%	77.5pp	0.91
Folding	39.9%	0.099	23.0%	70.0%	47.0pp	—
Perspective	35.5%	0.053	23.9%	54.0%	30.2pp	—
MRT Easy	32.2%	0.035	22.0%	39.4%	17.4pp	0.39
MRT Hard	25.0%	0.020	20.0%	29.4%	9.4pp	0.29

Navigation ($r = 0.91$), Relations ($r = 0.88$), and Orientation ($r = 0.85$) are the strongest predictors of overall model ranking. MRT Easy ($r = 0.39$) and MRT Hard ($r = 0.29$) barely discriminate between models. The top-5 versus bottom-5 model gap further quantifies this: Navigation separates these groups by 60.9 percentage points, Relations by 42.2pp, but MRT Easy by only 2.9pp and MRT Hard by 1.1pp. Figure 8 and Figure 9 visualise these patterns across all 32 models.

E Per-Question Difficulty Analysis

Within each split, individual questions vary enormously in difficulty. Table 7 reports the distribution of per-question accuracy computed across all 32 models.

Relations and Orientation contain the most questions above 50% accuracy (287 and 268, respectively). MRT splits have almost none: 96% of MRT Easy and MRT Hard questions fall below 50% accuracy, meaning even the "easiest" MRT questions defeat more than half the models.

A total of 80 questions achieve 0% accuracy across all 32 models. Their distribution reveals which skills remain entirely beyond current model capabilities: Perspective accounts for 42 of these universally-hard questions (52.5%),

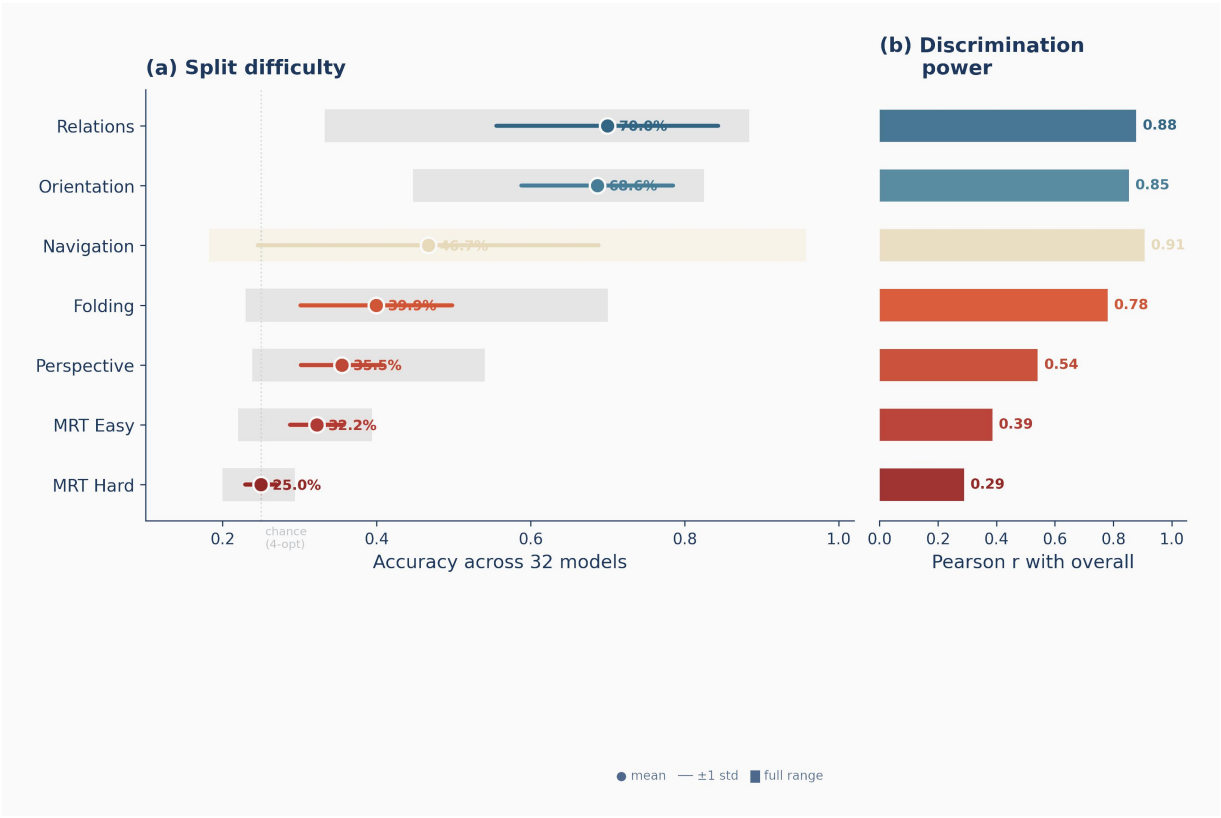


Figure 8 Split difficulty hierarchy. Each bar represents one model's accuracy on a given split, sorted by overall performance.

Table 7 Per-question difficulty distribution. Columns report the number of questions at which no model is correct (0%), fewer than 10% of models are correct (<10%), more than half of models are correct (>50%), and more than three-quarters are correct (>75%).

Split	N	Mean	Median	0%	<10%	>50%	>75%
Relations	370	0.664	0.742	1	9	287	170
Orientation	400	0.542	0.581	1	17	268	25
Navigation	400	0.457	0.419	0	5	147	43
Folding	400	0.376	0.355	0	5	71	0
Perspective	285	0.328	0.226	27	91	86	35
MRT Easy	500	0.281	0.258	0	64	46	0
MRT Hard	500	0.229	0.194	3	92	20	0

followed by MRT Easy (20), MRT Hard (11), Relations (3), Navigation (2), and Orientation (2). The universally-hard perspective questions are almost exclusively those requiring viewpoint transformation with left–right mirror inversion. At the opposite extreme, universally-easy questions (>90% accuracy) are dominated by relations questions about simple camera-relative positions and orientation questions about objects with clearly asymmetric front–back features.

F Reasoning Behaviour Analysis

F.1 Per-Model Reasoning Patterns

Models differ dramatically in verbosity. Qwen3-VL-8B-Thinking averages 4,411 words per response and achieves 45.9% accuracy. GPT-5.2 averages 500 words and achieves 51.6%. Gemini-3-flash-preview uses only 170 words on average yet leads the benchmark at 63.1%. At the small-model end, SmolVLM2-500M averages just 21 words, often producing bare single-letter answers with no articulated reasoning. These patterns indicate that verbosity and accuracy are not positively correlated. Within GPT-5.2, the wrong-to-correct reasoning length ratio reaches 3.84×, the highest among top-performing models, meaning it produces nearly four times more text when arriving at an

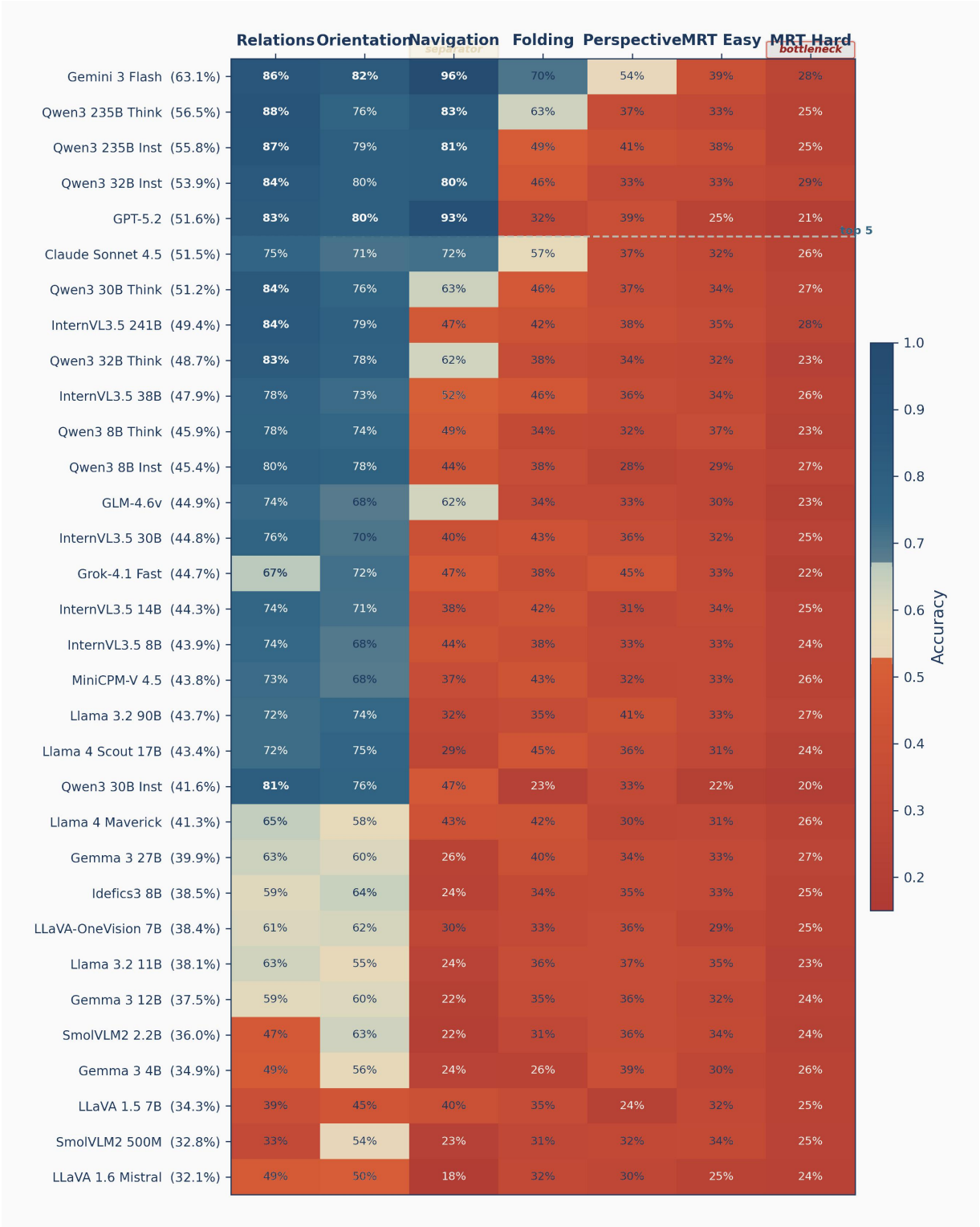


Figure 9 Model × split performance heatmap. Rows are ordered by overall accuracy (top = best). Columns are ordered by mean split difficulty (left = easiest).

incorrect answer.

Medium-difficulty questions (30–50% accuracy across models) elicit the longest average reasoning chains (1,004 words), while very-easy questions (>70% accuracy) elicit the shortest (108 words). Models reason most verbosely

in the "uncertain middle" where they attempt extended analysis but frequently fail.

F.2 Uncertainty and Confidence Language

Table 8 reports the frequency of uncertainty markers ("might", "perhaps", "not sure", "could be") and confidence markers ("clearly", "definitely", "certainly") stratified by split and correctness.

Table 8 Proportion of responses containing uncertainty or confidence markers, stratified by correctness.

Split	Unc. (Correct)	Unc. (Wrong)	Conf. (Correct)	Conf. (Wrong)
Folding	21.5%	30.5%	7.5%	9.6%
MRT Hard	42.6%	45.3%	14.2%	14.9%
Orientation	11.8%	16.7%	8.5%	8.5%
Relations	10.4%	9.2%	16.6%	7.3%

For most splits, wrong answers contain slightly more uncertainty language than correct ones, but the difference is small. On MRT, models express uncertainty nearly half the time regardless of outcome, reflecting genuine difficulty. On Relations, correct answers contain substantially more confidence language (16.6%) than wrong ones (7.3%), suggesting that when models are certain about a spatial relation, they tend to be correct.

F.3 The Reasoning–Answer Alignment Paradox

Qualitative inspection of model outputs reveals a systematic pattern in which models reason correctly at the feature level but reach incorrect spatial conclusions. In perspective questions, models correctly identify objects, people, and their positions in the image, then attempt the perspective transformation but fail to mirror left and right. The reasoning reads logically—"The man faces the camera. The boy is to the viewer's right. Since the man faces us, his right is our left. . . "—but the final answer either contradicts the chain of reasoning or applies the mirror rule inconsistently. In MRT questions, models accurately count cubes and describe connectivity patterns but lose track of how structures would appear after rotation. In orientation questions, models correctly identify visual features such as windshield shape and head position but sometimes map these to the wrong spatial direction.

G Cross-Split Correlations and Model Profiles

Figure 10 presents the pairwise Pearson correlation matrix between task splits computed across all 32 models. A clear block structure emerges: Navigation, Orientation, and Relations are highly correlated ($r = 0.71$–0.91), sharing a common requirement of basic visual grounding and spatial vocabulary. MRT Easy and MRT Hard are moderately correlated with each other ($r = 0.47$) but weakly correlated with other splits ($r = 0.10$–0.19), indicating that 3D mental rotation taps an essentially independent capability. Folding sits between these clusters, showing moderate correlations with both MRT ($r = 0.51$–0.58) and the visual grounding splits ($r = 0.47$–0.61). Perspective is weakly correlated with most splits ($r = 0.12$–0.46), suggesting it tests a relatively independent skill.

Every single model across all 32 tested exhibits the same worst split: MRT Hard. No model exists for which mental rotation is not the most difficult task, confirming that 3D mental rotation represents a fundamental architectural limitation rather than a model-specific weakness. Best-split profiles vary more: most models perform best on Relations, while several perform best on Orientation. Only Gemini-3-flash-preview and GPT-5.2 have Navigation as their best split, reflecting their uniquely strong visual grounding. The coefficient of variation (CV) across splits ranges from 0.23 (LLaVA-1.5-7B, most generalist) to 0.60 (Qwen3-VL-30B-A3B-Instruct, most specialist). Low-performing models tend to appear more "generalist" simply because they perform uniformly poorly, while high performers exhibit greater specialisation with strong scores on visual grounding splits but near-chance on MRT.

H Open-Source vs. Closed-Source Statistical Comparison

This section extends the aggregate comparison in Section 4.5 with a full per-split statistical breakdown. Table 9 summarises the five closed-source systems. Notably, the best open-source model (Qwen3-VL-235B-A22B-Thinking at 56.5%) outperforms three of the five closed-source systems.

Table 10 reports the per-split statistical comparison. The overall closed-source advantage ($\Delta = +8.1$pp; Mann–Whitney $U = 110$, $p = 0.026$; Cohen's $d = 1.18$) is almost entirely driven by Navigation ($\Delta = +32.1$pp, $p = 0.008$), the only split reaching high significance. Perspective also shows a significant gap ($p = 0.033$). MRT splits exhibit

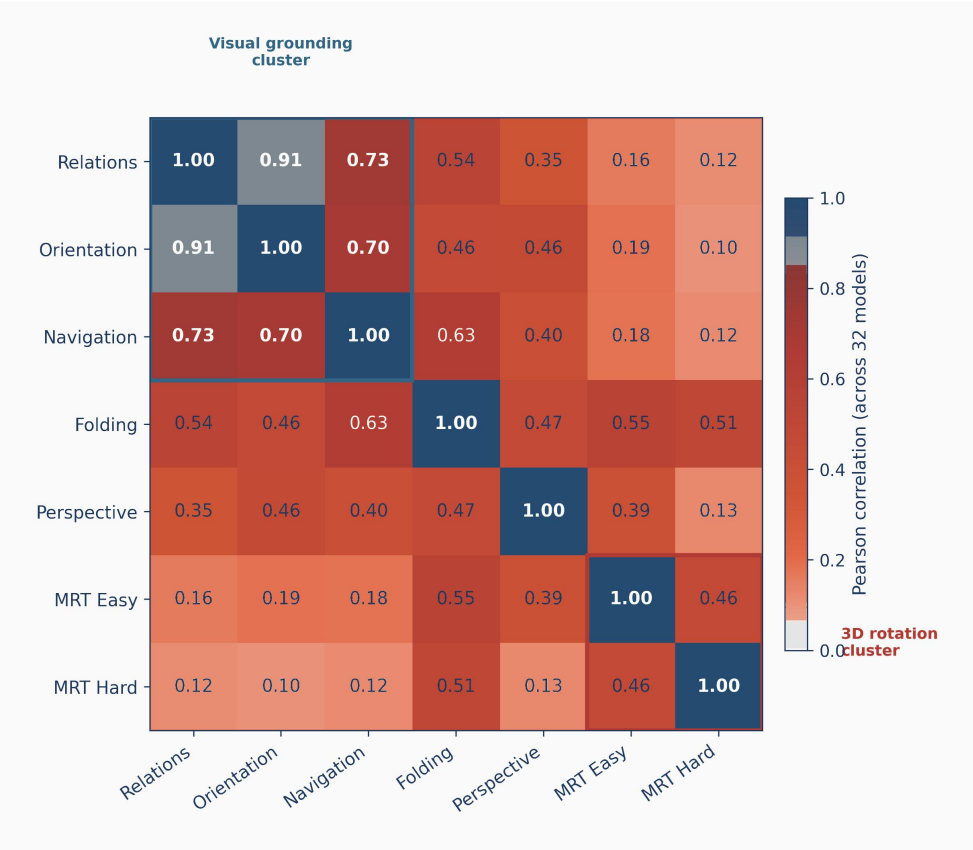


Figure 10 Cross-split Pearson correlation matrix. Warmer colours indicate stronger positive correlations.

Table 9 Closed-source model performance summary.

Model	Overall	Best Split	Worst Split
Gemini-3-flash-preview	63.1%	Nav (95.8%)	MRT-H (28.2%)
GPT-5.2	51.6%	Nav (92.8%)	MRT-H (21.4%)
Claude-Sonnet-4.5	51.5%	Rel (74.9%)	MRT-H (26.0%)
GLM-4.6v	44.9%	Rel (73.5%)	MRT-H (22.6%)
Grok-4.1-fast	44.7%	Ori (71.8%)	MRT-H (22.0%)

no closed-source advantage—in fact, a slight non-significant disadvantage—confirming that the MRT ceiling is architecture-independent and cannot be overcome by proprietary training data or inference pipelines.

Table 10 Per-split open-source vs. closed-source comparison. Significance: ** $p < 0.01$, * $p < 0.05$, ns = not significant.

Split	Open μ	Closed μ	Δ	p	Sig.
Navigation	41.7%	73.8%	+32.1pp	0.008	**
Perspective	34.3%	41.5%	+7.2pp	0.033	*
Relations	68.7%	76.9%	+8.2pp	0.299	ns
Orientation	67.5%	74.7%	+7.1pp	0.177	ns
Folding	38.8%	46.2%	+7.5pp	0.586	ns
MRT Easy	32.3%	32.0%	−0.3pp	0.533	ns
MRT Hard	25.1%	24.0%	−1.1pp	0.323	ns

Figure 11 provides a visual summary of the comparison across all splits.

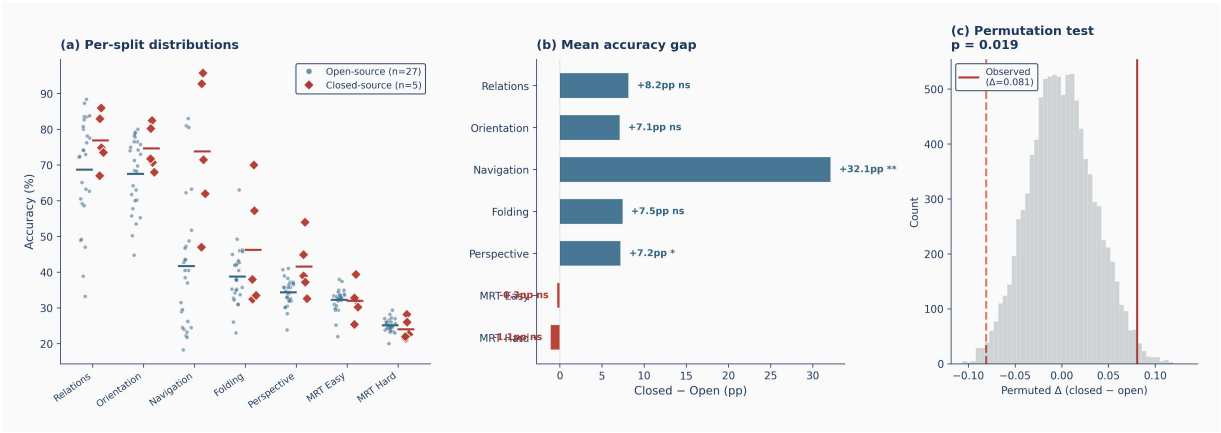


Figure 11 Open-source vs. closed-source performance comparison across splits.

I Detailed Scaling Analysis

Table 11 extends the scaling analysis presented in Section 4.5 by reporting Pearson correlations between model size and per-split accuracy for all 27 open-source models. The overall correlation with total parameters ($r = 0.68$) is stronger than with active parameters ($r = 0.51$), partly because MoE models—which have large total but smaller active parameter counts—tend to perform well.

Table 11 Pearson correlation between model size and per-split accuracy. Correlations are computed against $\log_{10}$ of total and active parameter counts for 27 open-source models.

Split	r (Total params)	r (Active params)
Relations	0.69	0.53
Folding	0.64	0.47
Navigation	0.57	0.36
Orientation	0.53	0.41
Perspective	0.35	0.25
MRT Hard	0.23	0.38
MRT Easy	0.12	0.24

Relations and Folding scale most strongly with model size, suggesting that increased capacity helps with spatial vocabulary and multi-step transformations. MRT Easy exhibits the weakest scaling relationship ($r = 0.12$ with total parameters), confirming that 3D mental rotation represents a capability that does not emerge with scale: a 241B-parameter model performs essentially the same as a 7B model on this split, making it the clearest example of a flat scaling curve in spatial reasoning.

J Granular Error Analysis by Spatial Dimension

This section complements the distractor analysis in Section 4.4 by examining which specific spatial dimensions are hardest for models within each task split. Figure 12 presents the full breakdown.

J.1 Perspective: Left–Right Confusion

Table 12 reports accuracy stratified by the gold-standard direction. Left–right discriminations are dramatically harder than front–back or vertical judgements: "left" accuracy is 16.4% and "right" is 19.3%, compared with 40.0–51.0% for front, behind, and above. The confusion matrix reveals a mirror-swap pattern: when the gold answer is "left," 64% of errors select "right," and conversely, when the gold answer is "right," 68% of errors select "left." Vertical relations (above, over) represent the easiest perspective judgements, likely because depth and height relations are preserved regardless of observer viewpoint.

J.2 Orientation: Cardinal vs. Diagonal Directions

Models correctly identify facing direction 47.3% of the time for "front" but only 33.0% for "front left." Cardinal directions (front: 47.3%, left: 43.4%, right: 41.3%, back: 40.4%) are consistently easier than diagonal directions

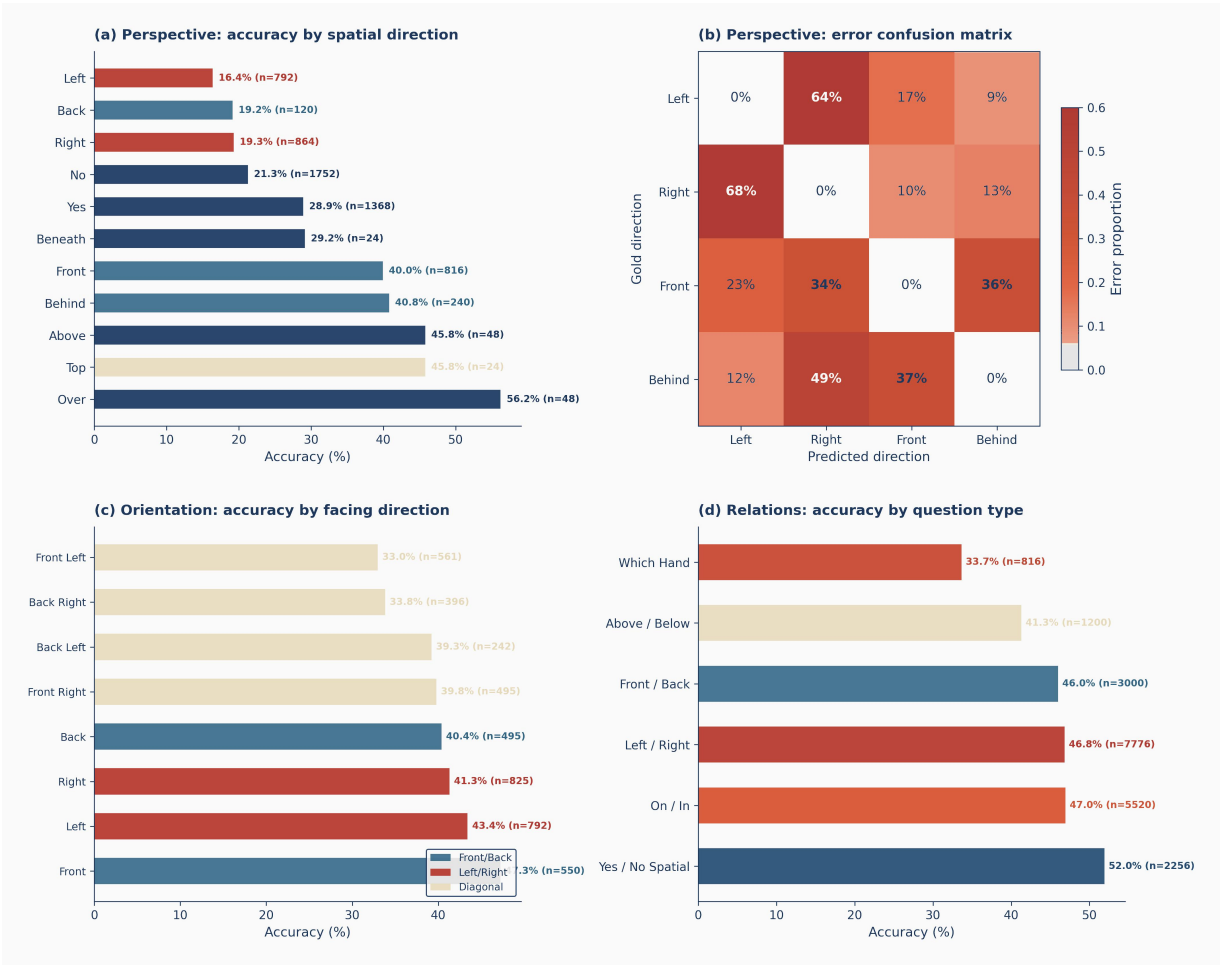


Figure 12 Granular error type analysis across three task splits, stratified by specific spatial dimension.

Table 12 Perspective accuracy by gold-standard direction.

Gold Direction	Accuracy	N
Over/Above	51.0%	96
Front	40.0%	816
Behind	40.8%	240
Right	19.3%	864
Left	16.4%	792

(front right: 39.8%, back left: 39.3%, back right: 33.8%, front left: 33.0%), suggesting that models can distinguish the four primary axes but struggle when an object faces between two axes.

A notable bias emerges: models are better at confirming true orientations (gold = yes: 44.9%) than rejecting false ones (gold = no: 35.1%), a 9.8pp gap indicating an affirmation bias in spatial judgements. Object type also affects performance: real animals with clear front–back asymmetry (e.g. tiger: 50.0%, rhinoceros: 48.5%, giraffe: 48.2%) are substantially easier to orient than toy vehicles with ambiguous features (e.g. toy boat: 28.5%, toy bus: 33.2%, toy motorcycle: 33.8%). This ~20pp gap reveals that models rely on learned associations between object categories and facing cues, and toy objects break these associations.

J.3 Relations: Hand Laterality

Table 13 reports accuracy by relation type. "Which hand" questions (e.g. "Which hand is the man using to touch the baby?") are the hardest relation type at 33.7%, nearly 20pp below yes/no spatial questions. Hand laterality requires both fine-grained visual grounding—identifying which hand is in use—and perspective-aware left–right assignment. Above/below relations (41.3%) are harder than left/right (46.8%), likely reflecting the difficulty of

judging vertical spatial relations in 2D images with ambiguous depth cues.

Table 13 Accuracy by relation type in the Spatial Relations split.

Relation Type	Accuracy	N
Yes/No spatial	52.0%	2,256
On/In	47.0%	5,520
Left/Right	46.8%	7,776
Front/Back	46.0%	3,000
Above/Below	41.3%	1,200
Which hand	33.7%	816

K Evaluation Protocol and Reproducibility

We document all design choices that affect measured accuracy, enabling exact replication and fair comparison with future work. The benchmark dataset is publicly available on Hugging Face as `stogian/srbenchv2`.

Prompting

System prompt. All models receive an identical system message:

> *"You are a spatial reasoning AI assistant specialized in analyzing, understanding, and solving problems involving spatial relationships, geometric transformations, and visual-spatial concepts."*

User prompt. The user turn contains the task image followed by the question text, drawn directly from the dataset's `question` field. Each question is self-contained: it states the task, presents the answer options (A–D), and asks the model to identify the correct one.

Chain-of-thought. All models (open-source and closed) receive the same CoT instruction appended to the user message:

> *"Let's think step by step. Provide a detailed reasoning process before giving the final answer within curly brackets, e.g. {A}."*

Thinking variants. For models with a native thinking mechanism (e.g. Qwen3-VL Thinking), no additional instruction is injected on top of CoT; the extended reasoning is produced by the model's own chain-of-thought tokens and the final answer is extracted from the terminal response text.

Image Handling

Images are passed as-is from the dataset's `image` field (PIL `RGB` objects). For open-source models, images are fed directly to the model's processor without any resizing, cropping, or tiling on our side; the processor applies the model's native patch encoding. For closed/API models, images are re-encoded as JPEG at quality 85 before base64 encoding to reduce payload size; no spatial information is lost at this quality level for our image types. Each question uses exactly one image; no multi-crop or multi-image variants are employed.

Decoding Parameters

Parameter	Open-source	Closed / API
Decoding	Greedy (`do_sample=False`)	Greedy (`temperature=0`)
Max new tokens	8,192	8,192
Samples per question	1	1
Random seed	123	123
Batch size	32	100 concurrent requests
Precision	`bfloat16`	API-managed

Greedy decoding is used throughout; no nucleus or top-k sampling is applied. For open-source models, all seeds (`random`, `numpy`, `torch`, `CUDA`, `PYTHONHASHSEED`, `CUBLAS_WORKSPACE_CONFIG`) are fixed to 123 to ensure fully deterministic outputs.

Output Parsing

The answer extractor applies a four-step priority cascade to map raw model output to a choice label {A, B, C, D}:

1. **Curly-bracket format** — matches `{A}` or `{**A**}` (the CoT target format).
2. **Mixed letter–word format** — matches patterns like `C. Left` and extracts the letter.
3. **Semantic prefix/suffix phrases** — e.g. "the answer is A", "I choose B", "C is correct".
4. **Direct bold/plain letter** — e.g. `**A**` or a bare standalone letter.

If none of the four steps yields a valid label, the response is marked as incorrect. We never re-prompt or retry; a single forward pass is all that is used per question.

Model Identifiers and Access Dates

Proprietary / API models (via OpenRouter at `openrouter.ai/api/v1`):

Model	OpenRouter / API identifier	Accessed
GPT-5.2	`openai/gpt-5.2`	Feb 2026
Gemini 3 Flash	`google/gemini-3-flash-preview`	Feb 2026
Claude Sonnet-4.5	`anthropic/claude-sonnet-4.5`	Feb 2026
Grok 4.1 Fast	`x-ai/grok-4.1-fast`	Feb 2026
Llama 4 Maverick	`meta-llama/llama-4-maverick-402b-17b-instruct-fp8`	Feb 2026
GLM-4.6V	`zai-org/GLM-4.6V-Flash`	Feb 2026

Open-source models (8× NVIDIA H100 80 GB SXM; `bfloat16`; Flash Attention 2 where supported; inference stack: `transformers` 4.51.x, `torch` 2.6.0+cu124, `flash-attn` 2.7.x):

Model (paper name)	HuggingFace identifier
Qwen3-VL-235B (Instruct / Think.)	`Qwen/Qwen3-VL-235B-A22B-Instruct`
Qwen3-VL-32B (Instruct / Think.)	`Qwen/Qwen3-VL-32B-Instruct`
Qwen3-VL-30B (Instruct / Think.)	`Qwen/Qwen3-VL-30B-A3B-Instruct`
Qwen3-VL-8B (Instruct / Think.)	`Qwen/Qwen3-VL-8B-Instruct`
InternVL3.5-241B	`OpenGVLab/InternVL3_5-241B-A28B-HF`
InternVL3.5-38B	`OpenGVLab/InternVL3_5-38B-HF`
InternVL3.5-30B	`OpenGVLab/InternVL3_5-30B-A3B-HF`
InternVL3.5-14B	`OpenGVLab/InternVL3_5-14B-HF`
InternVL3.5-8B	`OpenGVLab/InternVL3_5-8B-HF`
Llama-4-Scout	`meta-llama/Llama-4-Scout-109B-A17B-Instruct`
Llama-3.2-90B	`meta-llama/Llama-3.2-90B-Vision-Instruct`
Llama-3.2-11B	`meta-llama/Llama-3.2-11B-Vision-Instruct`
Gemma-3-27B	`google/gemma-3-27b-it`
Gemma-3-12B	`google/gemma-3-12b-it`
Gemma-3-4B	`google/gemma-3-4b-it`
MiniCPM-V-4.5	`openbmb/MiniCPM-V-4_5`
LLaVA-OneVision-7B	`llava-hf/llava-onevision-qwen2-7b-ov-hf`
LLaVA-v1.6-Mistral-7B	`llava-hf/llava-v1.6-mistral-7b-hf`
LLaVA-1.5-7B	`llava-hf/llava-1.5-7b-hf`
Idefics3-8B	`HuggingFaceM4/Idefics3-8B-Llama3`
SmolVLM2-2.2B	`HuggingFaceTB/SmolVLM2-2.2B-Instruct`
SmolVLM2-500M	`HuggingFaceTB/SmolVLM2-500M-Video-Instruct`

Thinking variants use the corresponding `-Thinking` suffixed checkpoint where available; all other settings are identical to the instruct variant.

L Use of Large Language Models

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