Defining and Evaluating Visual Language Models' Basic Spatial Abilities: A Perspective from Psychometrics

Wenrui Xu[1†], Dalin Lyu[1†], Weihang Wang[1†], Jie Feng[2], Chen Gao[3*], Yong Li[2,3*]
[1]School of Architecture, Tsinghua University
[2]Department of Electronic Engineering, Tsinghua University
[3]BNRist, Tsinghua University
{xwr23,lvdl24,wang-wh22}@mails.tsinghua.edu.cn
{fengjie,chgao96,liyong07}@tsinghua.edu.cn

Abstract

The Theory of Multiple Intelligences underscores the hierarchical nature of cognitive capabilities. To advance Spatial Artificial Intelligence, we pioneer a psychometric framework defining five Basic Spatial Abilities (BSAs) in Visual Language Models (VLMs): *Spatial Perception*, *Spatial Relation*, *Spatial Orientation*, *Mental Rotation*, and *Spatial Visualization*. Benchmarking 13 mainstream VLMs through nine validated psychometric experiments reveals significant gaps versus humans, with three key findings: 1) VLMs mirror human hierarchies (strongest in 2D orientation, weakest in 3D rotation) with independent BSAs; 2) Many smaller models surpass larger counterparts, with Qwen leading and InternVL2 lagging; 3) Interventions like CoT and few-shot training show limits from architectural constraints, while ToT demonstrates the most effective enhancement. Identified barriers include weak geometry encoding and missing dynamic simulation. By linking Psychometrics to VLMs, we provide a comprehensive BSA evaluation benchmark, a methodological perspective for embodied AI development, and a cognitive science-informed roadmap for achieving human-like spatial intelligence.

1 Introduction

Visual Language Models (VLMs) excel in a wide range of specific tasks (Hong et al., 2023). However, achieving human-like spatial intelligence for embodied AI applications such as visual navigation and embodied Q&A remains a challenge (Durante et al., 2024; Duan et al., 2022). Recent studies reveal that even advanced models like GPT-4o fail basic 2D spatial reasoning tasks that humans solve effortlessly (Tang et al., 2024).

Theory of Multiple Intelligences (Bornstein and Gardner, 1986), which is widely accepted across disciplines, posits that human intelligence is hierarchical, with general intelligence (g) supported by subordinate intelligences including spatial intelligence. Spatial intelligence is also structured hierarchically with several basic spatial abilities (BSAs), which are crucial for advanced intelligence and provide a comprehensive framework for evaluation.

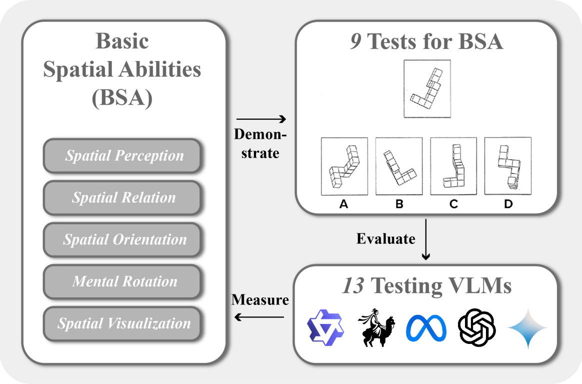


Figure 1: BSA measuring framework for VLMs[1].

However, existing studies assessing the spatial abilities of VLMs lack a solid theoretical foundation and focus on isolated abilities without a comprehensive framework, making it challenging to compare results across different studies or uncover potential interconnections between abilities. Moreover, most research lacks human baselines, leaving the gap between VLMs and human unexplored.

To address these gaps, we propose a psychometric framework using standardized human experiments to systematically evaluate VLMs' BSAs, thereby establishing benchmarks and pathways for enhancing spatial intelligence in AI systems.

2 Related works

2.1 Psychometric Studies on Human's BSAs

Human spatial intelligence, defined as the ability to mentally model and manipulate spatial environ-

† These authors contributed equally to this work.
* Corresponding authors.

[1] Image source: Vandenberg and Kuse (1978).

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11571–11590
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

ments (Bornstein and Gardner, 1986), has been extensively studied since Spearman (1904) first proposed it. Research then diverges into two streams (Porat and Ceobanu, 2023):

Psychometric Classification. The first stream establishes hierarchical intelligence models, from general intelligence to domain-specific intelligence like spatial intelligence (Geary, 2022; Spearman, 1904). These frameworks enable systematic measurement of spatial abilities through standardized tests, continually refining ability subtypes and assessment methods.

Interdisciplinary Mechanisms. The second stream integrates evolutionary psychology, developmental studies, and cognitive neuroscience to explore the origins and mechanisms of spatial abilities, complementing psychometric taxonomies.

Psychometric consensus defines spatial intelligence through three hierarchical levels (Caemmerer et al., 2020; McGrew, 2009; Johnson and Bouchardjr, 2005; Linn and Petersen, 1985; Maccoby and Jacklin, 1974; Michael et al., 1957; Thurstone, 1950): (1) **General Intelligence** (g): cross-domain cognitive processes, such as attentional control (Kane and Engle, 2002), neural integration (Jung and Haier, 2007), and cellular processes (Geary et al., 2021), which impact cross-domain learning and performance. (2) **Domain-Specific Intelligence**: Abilities in particular domains that share common features, with spatial intelligence representing a distinct set of abilities (Vernon, 1965), as formalized in models such as the CHC theory (Carroll, 1993; Horn, 1968; Cattell, 1963). (3) **Basic Spatial Abilities** (BSAs): Decomposing spatial intelligence into measurable subskills: Spatial Perception (SP), Spatial Relation (SR), Spatial Orientation (SO), Mental Rotation (MR), and Spatial Visualization (SV).(Johnson and Bouchard, 2007; Hegarty et al., 2006; Maier, 1996; Voyer et al., 1995; Halpern, 1992; Pellegrino et al., 1984).

We focus on Level 3, adopting validated human experiments to evaluate VLMs' BSAs.

2.2 Evaluation of VLMs' BSAs

Recent advances in VLMs have spurred evaluations of their spatial abilities, yet existing studies remain fragmented (Table 2). Most prior work focuses on text-based LLMs, assessing abstract spatial relations through verbal descriptions (Yamada et al., 2024), inherently neglecting visual-spatial processing. Emerging VLM evaluations primarily target specialized 3D tasks (e.g., robotic trajectory labeling (Sharma, 2023), indoor scene captioning (Fu et al., 2024)), which partially engage Spatial Perception, Spatial Relation, and Spatial Orientation.

Table 1: Decomposed BSA and the corresponding tests.

Type	Definition	Tests
Spatial Perception	The ability to perceive horizontal and vertical orientations without interference from miscellaneous information.	SVT
Spatial Relation	The ability of recognizing relationships between parts of an entity.	NCIT DAT:SR R-Cube-SR
Spatial Orientation	The ability to navigate or enter a given spatial state.	MRMT
Mental Rotation	The ability to mentally rotate 3D objects.	MRT PSVT:R
Spatial Visualization	The ability to mentally manipulate and transform 2D and 3D objects.	SBST R-Cube-Vis

Table 2: Existing studies testing LLMs and VLMs' spatial abilities.

Related Work	VLM	Tested Spatial Abilities				
		SP	SR	SO	MR	SV
Fu et al. (2024)	Yes	✓	✓	✓		
Tang et al. (2024)	Yes	✓	✓	✓		
Sharma (2023)	Yes		✓	✓		
Hong et al. (2023)	Yes	✓	✓	✓		
Bang et al. (2023)	Yes		✓	✓		
Yamada et al. (2024)	No		✓			
Momennejad et al. (2023)	No		✓	✓		
Cohn et al. (2023)	No		✓		✓	
This study	Yes	✓	✓	✓	✓	✓

However, critical gaps persist: (1) **Theoretical Disconnect**: Tasks lack theoretical framework grounding, preventing direct comparison with human cognition. For instance, robotic trajectory tests (Sharma, 2023) conflate spatial reasoning with action planning. (2) **Limited Scope**: Most studies omit Mental Rotation and Spatial Visualization (Table 2, MR/SV columns). (3) **Benchmark Absence**: No study systematically maps VLM performance to hierarchical BSAs or provides human baselines.

This study addresses these limitations by adopting a comprehensive psychometric evaluation framework. In contrast to previous work that assessed only subsets of BSAs (e.g., Fu et al. (2024): SP+SR+SO), we evaluate all five BSAs using standardized psychometric tests, which are widely validated and recognized, enabling comparisons across different models and between human and AI.

3 Methodology

3.1 Definition and Composition of Basic Spatial Abilities

Based on Psychometric theories (Maier, 1996) and existing VLM studies, this study identified five key dimensions of Basic Spatial Abilities (Pawlak-Jakubowska and Terczyńska, 2023), decomposing the concept as a comprehensive whole, as shown in Table 1. To carry out a complete spatial ability evaluation of VLMs, the study selected nine specific classic psychometric tests, as shown in Figure 2, designed to cover all five aspects of BSAs, enhancing the persuasiveness of the test. Human experiment results from existing studies were included as benchmarks for comparison.

3.2 Tests for Basic Spatial Abilities

The goal of the employed tests, on the one hand, is to evaluate the spatial abilities of different VLMs and to compare them with those of humans. On the other hand, by breaking down the basic abilities of spatial intelligence, we aim to identify the deficiencies in current VLMs and provide a necessary foundation for future research aimed at enhancing these abilities. To achieve this goal, the nine selected test question sets were carefully screened, ensuring that each is currently available, clearly assesses a specific BSA, contains complete questions and answers, and provides widely recognized and reproducible human performance results. These questions were not created by the researchers but are grounded in solid psychometric theoretical foundations. The study posits that using as many classic tests as possible can more authentically and accurately reflect the BSAs of VLMs. Therefore, arbitrarily deleting tests to balance the weights of the five core spatial abilities was deemed inappropriate.

Tests adopted are presented in the format of multiple-choice or true/false questions. In each specific test, after briefly explaining the test's content and the ways of answering, we provide the questions in the form of images.

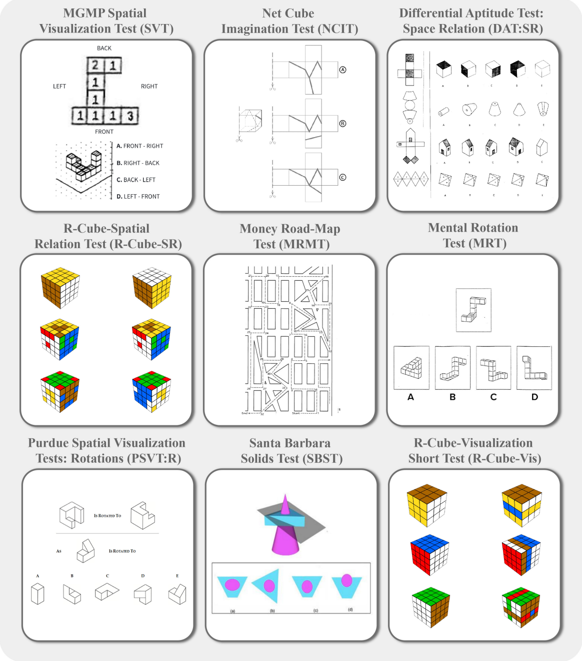

MGMP Spatial Visualization Test (SVT)
BACK
LEFT
RIGHT
FRONT
A. FRONT - RIGHT
B. RIGHT - BACK
C. BACK - LEFT
D. LEFT - FRONT
Net Cube Imagination Test (NCIT)
Differential Aptitude Test: Space Relation (DAT:SR)
R-Cube-Spatial Relation Test (R-Cube-SR)
Money Road-Map Test (MRMT)
Mental Rotation Test (MRT)
A
B
C
D
Purdue Spatial Visualization Tests: Rotations (PSVT:R)
Is Rotated To
As
Is Rotated To
A
B
C
D
E
Santa Barbara Solids Test (SBST)
(a)
(b)
(c)
(d)
R-Cube-Visualization Short Test (R-Cube-Vis)

Figure 2: Examples of classic BSA Tests[2].

3.2.1 Spatial Perception Tests

MGMP Spatial Visualization Test (SVT). SVT, originally developed for the Middle Grades Mathematics Project (MGMP), is also known as the "Lappan Test" and is composed of 32 multiple choice items, each with five options (Erkek et al., 2011; Ben-Chaim et al., 1986). The test utilizes the combination and transformation of square-cube buildings. The subject is expected to imagine the 2D flat view, the 3D corner view, and the "map plan", which is a numeric cube description of the base of the building. Questions include imagining the conversion between 2D and 3D views, the final appearance when some cubes are altered, and calculating the number of cubes used in a building.

3.2.2 Spatial Relation Tests

Net Cube Imagination Test (NCIT). NCIT is based on the conversion between 2D and 3D cubes with lines drawn on inner or outer faces (Pawlak-Jakubowska and Terczyńska, 2023). Each of the 16 tasks has three options. The first eight items require expanding the cube into a flat shape, while the rest involve the reverse development of the cube.

Differential Aptitude Test: Space Relation (DAT:SR). DAT:SR is part of the Differential Aptitude Test, measuring the ability to relate two and three-dimensional worlds (Katsioloudis et al., 2014; Bennett et al., 2012). The test consists of 40 items and focuses on folding and unfolding 3D

[2] Image source: Erkek et al. (2011); Ben-Chaim et al. (1986); Pawlak-Jakubowska and Terczyńska (2023); Katsioloudis et al. (2014); Bennett et al. (2012); Fehringer (2023); Friedman et al. (2020); Vingerhoets et al. (1996); Vandenberg and Kuse (1978); Peters et al. (1995); Bodner and Guay (1997); Maeda et al. (2013); Hegarty et al. (2009); Fehringer (2021).

geometric shapes. When provided 2D flat unfolded figures, subjects are required to choose the 3D restored form from four options and vice versa.

R-Cube-Spatial Relation Test (R-Cube-SR). R-Cube-SR uses Rubik's cubes with six different colors on each side as rotated visual materials (Fehringer, 2023). For each of the 48 items, two cubes are shown in the corner view, the right of which may be the possible rotated result of the left cube. The subjects are required to give a true/false answer with a limited view of the colored sides. The test was created in plain and pattern versions.

3.2.3 Spatial Orientation Tests

Money Road-Map Test (MRMT). The Standardized Road-Map Test of Direction Sense (Friedman et al., 2020; Vingerhoets et al., 1996) commonly known as MRMT, requires allocentric to egocentric right/left discrimination. Subjects follow a dashed path with 32 right/left turns on a map of an abstract city and are required to indicate the direction taken at each turn according to the facing direction. The turn types include the ones that require no rotation, a standard rotation of approximately 90°, and an irregular rotation between 90°-180°.

3.2.4 Mental Rotation Tests

Mental Rotation Test (MRT). A classic test developed by Vandenberg and Kuse (1978) takes the direct way of rotating 3D geometric figures made by cubes. Subjects are required to choose from the four options, two correct rotated reproductions of the target figure. The revised version of 24 items (Peters et al., 1995) is utilized.

Purdue Spatial Visualization Tests: Visualization of Rotations (PSVT:R). Developed by Bodner and Guay (1997), PSVT:R is an independent extended version of the Purdue Spatial Visualization Test, which requires matching between the target and the rotated objects. For each item, a geometric shape is rotated first to show the target rotation pattern. Subjects are required to choose from the five rotated views the correct rotated result of another shape under the same pattern. Maeda et al. (2013) further advanced the test to 30 test items, 17 of which involve asymmetrical shapes.

3.2.5 Spatial Visualization Tests

Santa Barbara Solids Test (SBST). SBST involves geometric solids intersected by a cutting plane, and the task is to imagine the 2D cross section of the solids and choose from the four options, the two correct answers (Hegarty et al., 2009). Two varying parameters result in the different difficulty of the 30 items: geometric complexity and cutting plane orientation. Simple solids, joined solids, and embedded solids take up one third of the items respectively, while orthogonal and oblique cutting planes divide the items at the same time.

R-Cube-Visualization Short Test (R-Cube-Vis). Similar to the R-Cube-SR test, R-Cube-Vis items consist of two juxtaposed Rubik's cubes (Fehringer, 2021). However, instead of rotating as a whole, the composing cubes can be rotated as well. Subjects are required to decide the possibility of the left cube rotated into the right one in the 60 items. Based on the size of the cube, the number of the rotated elements, and the ways they are rotated, the items are divided into six difficulty levels.

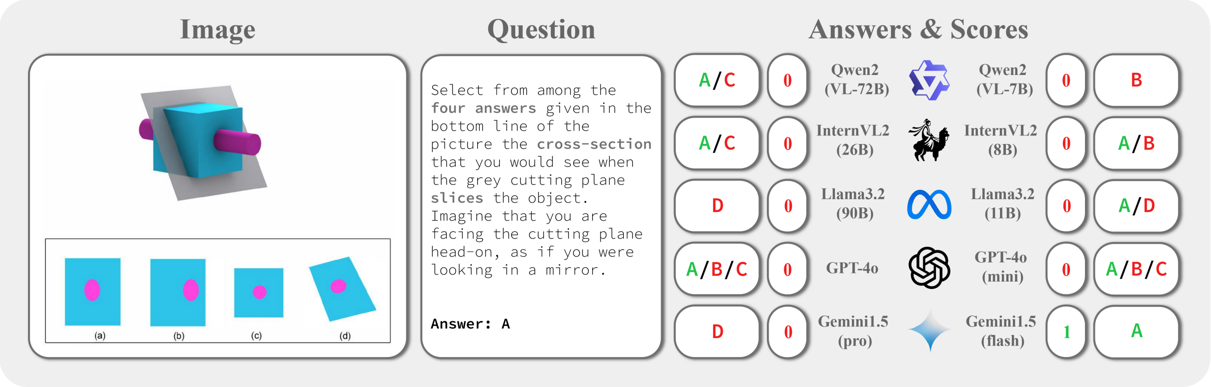
4 Experiments

4.1 Settings

Models and APIs. For BSA evaluation, we tested 13 mainstream open-source and commercial models, showcasing the full spectrum of the current major models' BSAs. For commercial VLMs, we used GPT-4o, GPT-4o mini, and GPT-4 Turbo from OpenAI (Yang et al., 2023) and Gemini-1.5-pro, Gemini-1.5-flash, and Gemini-1.5-flash-8b from Google (Gemini Team, 2024). For open-source models, we tested Qwen2-VL-72B, Qwen2-VL-7B (Bai et al., 2023), InternVL2-Llama3, InternVL2-26B, InternVL2-8B (Chen et al., 2024b), Llama-3.2-11B, Llama-3.2-90B (Llama Team, AI @ Meta, 2024) (the Qwen2 and Llama models are instruct variant). To carry out large-scale automatic tests, we used APIs from four platforms, including SiliconFlow, DeepInfra, OpenAI, and Google.

The temperature parameter are set uniformly to 0 for all models to ensure deterministic outputs and comparability across the tested models in the main experiments. Supplementary experiments are condcuted to investigate the impact of varying temperature on model performance. Generally, most models showed a decline in performance as the temperature increased, while Qwen2-VL-7B-Instruct and GPT-4o mini demonstrated a slight improvement at moderate temperature levels (Table 8).

Data Processing and Evaluation Metrics. A total of 312 questions are fed to each model. Among the questions, SP accounts for 10.26% (32 questions); SR accounts for 33.33% (104 questions); SO accounted for 10.26% (32 questions); MR ac-



Select from among the four answers given in the bottom line of the picture the cross-section that you would see when the grey cutting plane slices the object. Imagine that you are facing the cutting plane head-on, as if you were looking in a mirror.

Answer: A

A/C	0	Qwen2 (VL-72B)	Qwen2 (VL-7B)	0	B
A/C	0	InternVL2 (26B)	InternVL2 (8B)	0	A/B
D	0	Llama3.2 (90B)	Llama3.2 (11B)	0	A/D
A/B/C	0	GPT-4o	GPT-4o (mini)	0	A/B/C
D	0	Gemini1.5 (pro)	Gemini1.5 (flash)	1	A

Figure 3: An example of different VLMs' answers and scores to a particular Santa Barbara Solids Test question[3].

counts for 17.31% (54 questions); and SV accounts for 28.84% (90 questions). For the evaluation of overall ability, the study equates different individual abilities, thus no single ability will have a biased weight due to a larger number of questions.

Using zero-shot or few-shot experiments, we designed the following prompt template for VLMs: "You are taking a spatial ability test. Please output the result in pure text format as: question number, answer." Each question is presented separately, while the question and options are displayed in a single image, since the case study indicates negligible difference of model performance with and without explicit separation of answer choices (Table 4), and the ability to properly recognize option indices are also considered an elementary ability beneath spatial abilities. To ensure comparability with human benchmarks, we retained the instruction phase from the original human experiments.

Additionally, to address potential variations caused by differences in image size and color used in the tests, we conducted supplementary control experiments. Specifically, we resized test images to double and half their original dimensions, and also converted them into grayscale. The results show that image size led to minimal performance differences. However, when the images were converted to grayscale, several models exhibited improved performance. This suggests that color information may sometimes interfere with the models' ability to extract spatial information. Nevertheless, to ensure consistency and comparability with human experimental results, we ultimately used the original color images in our main experiments.

The score for each question is calculated as the number of correct options selected divided by the total number of correct options, and the score is zero if the answer contains false options. Specifically, scores are not counted if a model always guesses the same answer due to its failure to understand the question. The score for each test is the total of all question scores included in that test, converted into a percentage. The score for each BSA is calculated as the average score of all tests under that ability, while the Overall Ability Score is the average of the scores for all BSAs.

4.2 Results

As shown in Table 3, each cell in the table represents a model's score on a specific test. To evaluate the stability of model performance, each test was repeated three times for every model. The numbers in parentheses indicate the standard deviation of these three scores, indicating performance stability. The final rows of the table present the results for all models across each test, along with the corresponding human experiment results.

4.2.1 Human vs VLM

As illustrated in Figure 4, the overall ability scores of the 13 VLMs are relatively close, ranging from 16.31 to 31.22, with an average of 24.95, which is significantly lower than the human average of 68.38. When analyzed across the five BSAs, human performance consistently surpasses that of VLMs. In terms of averages, both human and VLMs show the same performance ranking across the abilities, with spatial orientation being the best and mental rotation the worst.

The red dashed lines in Figure 4 indicate random answer baseline scores. Some VLMs performed below these baselines, indicating that their abilities on certain tasks were genuinely inadequate. It is important to clarify that these below-baseline scores differ from the cases where models provided

[3] Image source: Hegarty et al. (2009).

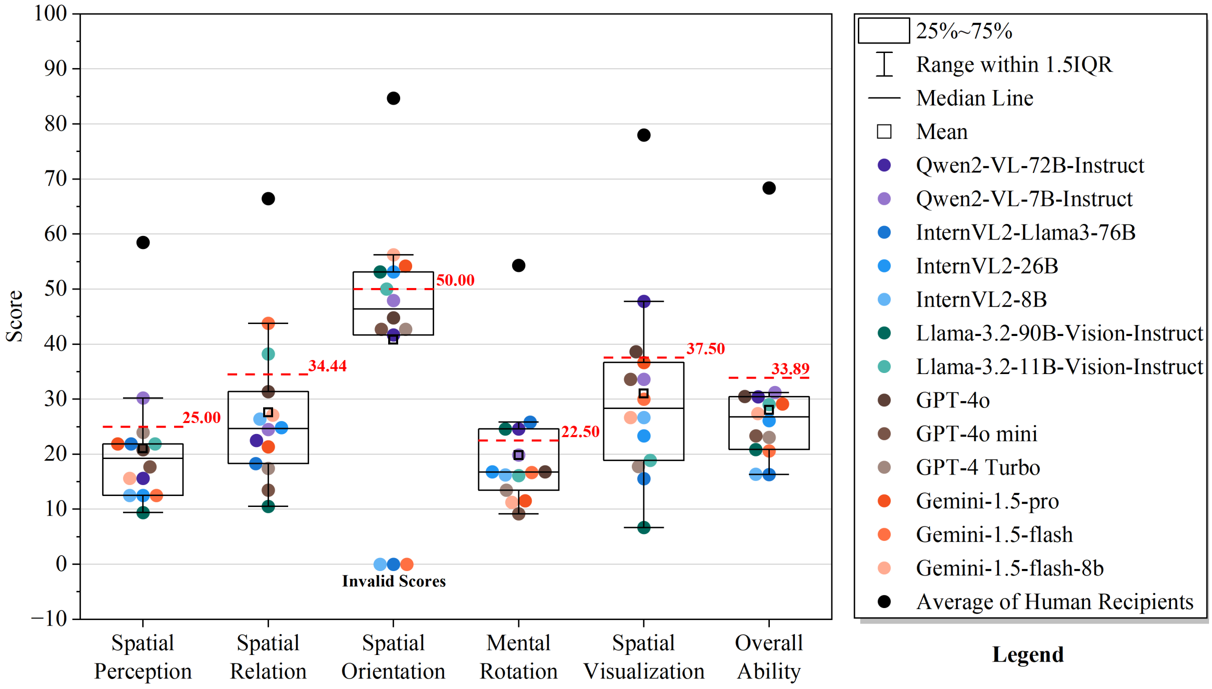


Figure 4: Comparison of VLMs' and humans' five basic spatial abilities and overall ability.

the same answers to all questions and were marked as invalid (which means they were just "guessing").

4.2.2 VLM vs VLM

From Figure 4, it can be observed that the performance of VLMs show consistency in SP and MR. In contrast, SR and SV exhibit more significant variability. Most models perform relatively well in SO, yet the discrepancy is also considerable, with three models (Intern-VL2-76B, Intern-VL2-8B, and Gemini-1.5-flash) failing completely. In terms of overall ability, the performance differences among models are not particularly significant, though some models (e.g. Qwen2-VL-7B and GPT-4o), stand out with relatively strong results.

4.2.3 Correlation of Basic Spatial Abilities

For the test results of the models' BSAs, we performed a correlation analysis between the abilities to examine their associations. We used the Pearson correlation test, as shown in Figure 5. The results show that Pearson's correlation coefficients (r) between any two variables are less than 0.4, indicating that all of the ability combinations show very weak or no correlation. Since the events satisfied the two-dimensional normal distribution, no correlation is equivalent to independence. Thus, it can be considered that the spatial abilities do not show a correlation with each other, meaning that each spatial ability is sufficiently "basic" and proves the credibility of the BSA framework.

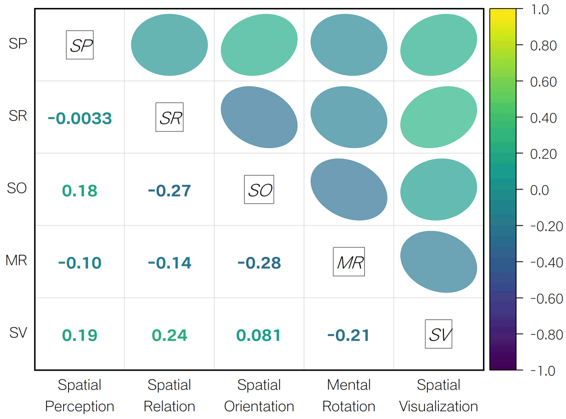


Figure 5: Pearson's correlation coefficients (r) between five basic spatial abilities.

4.3 Discussion

4.3.1 VLMs' Overall Performance

Our evaluation confirms a significant gap between VLMs' BSAs and human benchmarks across all dimensions, aligning with prior studies on individual abilities (e.g., spatial relations (Yamada et al., 2024; Cohn and Hernandez-Orallo, 2023), spatial orientation (Momennejad et al., 2023)). This raises fundamental questions about whether VLMs operate as programmable pattern recognizers or genuinely emulate human-like spatial intelligence.

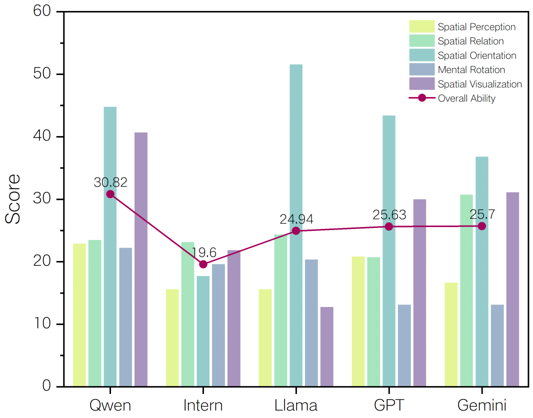


Figure 6: Comparison of different series of VLMs' basic spatial abilities and overall ability.

Notably, VLMs mirror human performance rankings across BSAs: highest in spatial orientation, followed by spatial visualization, and lowest in mental rotation (Figure 4). This correlates with task complexity that 2D tasks require simpler coordinate reasoning, while 3D demands dynamic mental manipulation, suggesting VLMs struggle disproportionately with higher-dimensional transformations just like humans. This performance gradient implies a developmental pathway: strengthening basic 2D spatial reasoning may scaffold advanced 3D capabilities (Tang et al., 2024). Such hierarchical training strategies could bridge current limitations.

VLMs' better performances on Spatial Orientation and Spatial Relation also align with some recent psychological findings which suggest that these abilities may be more fundamental or "basic" (Friedman et al., 2019). However, according to broader psychological research, BSAs are typically regarded as equally foundational components of spatial intelligence (Carroll, 1993). Therefore, further empirical studies from psychological perspectives are required to clarify these relationships at the underlying cognitive level.

4.3.2 Impact of VLM Manufacturer and Size

As shown in Figure 6, models from different manufacturers exhibit noticeable performance differences. The Qwen series demonstrates clear superiority with an overall score of 30.82, outperforming competitors across multiple assessment dimensions. Mid-tier performers including Gemini, GPT, and Llama series cluster around 25, while InternVL2 trails notably at 19.6.

Notably, individual spatial ability performance shows distinct patterns from aggregate scores. For instance, Gemini-1.5-flash achieves peak spatial relation performance (43.78) but demonstrates marked deficiencies in mental rotation (16.67) and complete absence of spatial orientation capability. The disparities likely originate from heterogeneous training data distributions across different BSAs.

Turning to model size analysis (Figure 7), our findings challenge conventional scaling laws. No positive correlation between parameter count and BSA performance is observed. Smaller models (leftward positioned in manufacturer groupings) frequently outperform their larger counterparts, a trend particularly evident in the Qwen and Llama series where around 10B-parameter models surpass larger variants. Similar scaling anomalies are discovered in the Gemini and InternVL2 series.

Smaller models such as Qwen2-VL-7B-Instruct and Llama-3.2-11B-Vision-Instruct exhibit relatively high performance (overall scores of 31.22 and 29.02), which may be attributed to their efficient multimodal alignment mechanisms and targeted instruction tuning. These models are often trained with highly curated datasets and optimized alignment strategies (e.g., supervised fine-tuning or RLHF), enhancing their ability to follow structured tasks like basic spatial ability tests.

Larger models such as GPT-4 Turbo demonstrate lower-than-expected performance (23.07), which may result from overfitting to generic pretraining objectives or reliance on surface-level heuristics rather than abstract spatial reasoning. In highly structured, low-context tasks like those assessing spatial cognition, such tendencies can hinder performance compared to smaller models that generalize in a simpler, more robust fashion.

The spatial ability tasks used in this study are designed to align closely with validated psychometric assessments. Their abstract, rule-based format differs significantly from the open-ended, language-heavy tasks many larger VLMs are tuned for. Smaller models, when fine-tuned specifically for multimodal instruction tasks, may thus exhibit stronger inductive biases or more focused learning beneficial for such structured evaluations.

In several similar studies, the performance of smaller models approaches or surpasses that of larger models in fundamental tasks such as text generation or classification (Matarazzo and Torlone, 2025; Li et al., 2025), where model performance is related to the characteristics and complexity of the

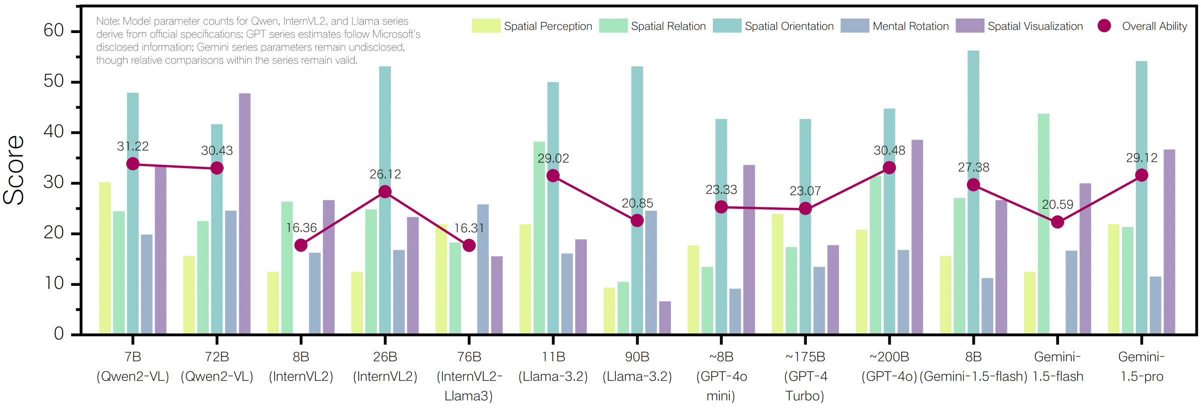

Note: Model parameter counts for Qwen, InternVL2, and Llama series derive from official specifications; GPT series estimates follow Microsoft's disclosed information; Gemini series parameters remain undisclosed, though relative comparisons within the series remain valid.

Figure 7: Comparison of the basic spatial abilities of VLMs with different sizes and manufacturers.

tasks. In tasks requiring logical reasoning, small models may outperform larger models due to more efficient architectures or training strategies.

4.3.3 Improvement by Alternative Reasoning Techniques

To assess intervention strategies, we implemented various experimental settings, including chain-of-thoughts (CoT), CoT self-consistency and tree-of-thoughts (ToT) for spatial reasoning (understanding the 3D shape, analyzing the plane, determining the cross-section, matching cross-section to options, giving the answer, as shown in Figure 9) and conducted few-shot learning on the SBST test (Table 6).

As shown in Figure 8, the accuracy improvement of CoT and few-shot learning is unstable across different VLMs. While examples help VLMs recognize spatial patterns such as prism cross-sections, they fail to address fundamental limitations in dynamic 3D mental simulation, as VLMs overly relied on the general patterns learnt during pre-training process. CoT self-consistency training proved effective on GPT-4o mini and Qwen2-VL-7B-Instruct, while ToT boosted overall baseline accuracy by nearly 0.10, enhancing visual-semantic alignment and multi-step reasoning.

4.3.4 From BSA to Real World Application

To further examine how well the BSA results may generalize to more naturalistic settings, such as robotics, real-time navigation, and interactive environments, we conducted additional experiments using the UrbanVideo-Bench dataset, designed for evaluating embodied visual capabilities (Zhao et al., 2025). The results show that models with high BSA scores (e.g., GPT-4o and GPT-4o mini) also perform better in the embodied tasks (Table 5). However, the relative importance of different BSAs can vary across real-world applications in embodied AI:

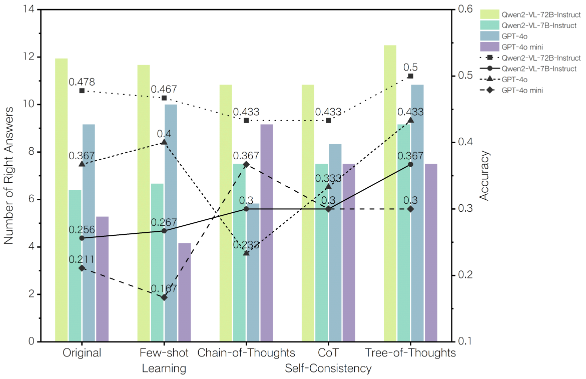


Figure 8: The impact of different experimantal settings on the accuracy of VLMs' answers.

Spatial Perception tends to support low-level sensory integration and scene interpretation, such as understanding occlusion relationships or segmenting geometric structures from vision inputs (Qiu and Di, 2024; Wang and Ke, 2024). It is particularly important in upstream modules of spatial reasoning pipelines, feeding perceptual representations into higher-level spatial processes.

Spatial Relation is highly relevant for fine-grained spatial decision-making, such as interpreting instructions in household robotics, or for understanding the relative positioning of individuals in crowd simulation and surveillance systems (Chen et al., 2024a; Yan et al., 2024).

Spatial Orientation is foundational in navigation-intensive applications such as mobile robotics, self-driving vehicles, and drones, which must maintain consistent heading and pose relative to dynamic and global reference frames. Orientation is also crucial in humanoid robotics, where continuous recalibration of spatial awareness is necessary in environments with shifting perspectives (Gao et al.).

Mental Rotation is critically important for tasks that require perspective-taking and manipulation of objects in 3D space (Wang et al., 2024; Lin et al., 2025). This includes robotic grasping, where a robot must infer how an object appears from a different angle in order to align its end effector correctly. In human-robot collaboration, mental rotation supports shared spatial understanding between agents. This ability is also essential in AR/VR interfaces and autonomous navigation, where systems must match real-world object orientations to internal representations.

Spatial Visualization plays a central role in tasks involving mental simulation of multi-step spatial transformations, such as multi-hop path planning, architectural reasoning, or surgical robotics (Zargarzadeh et al., 2025). This ability supports internal modeling of space over time, which is vital for predicting future environmental configurations.

While our current experimental design isolates each BSA for rigorous evaluation, a key direction for future work is to investigate their combined contributions in complex, real-world scenarios, and how specific applications may prioritize different ability. For example, in urban search and rescue operations, an embodied agent must perceive obstacles (SP), reason about its position relative to victims and hazards (SR), simulate multiple escape routes (SV), and navigate through rotating and cluttered environments (MR and SO) under dynamic conditions (Gao et al.).

4.3.5 Constraints in VLMs' BSAs

Our analysis reveals four fundamental constraints in VLMs' BSAs. Firstly, VLMs struggle to distinguish subtle shape variations such as hexagon/octagon cross-sections and misinterpret spatial relationships such as interior/exterior boundaries, indicating weak metric encoding in visual representations. This limitation may stem from constraints in training data or a lack of prior knowledge such as geometric principles, which example-based training can partially mitigate. Secondly, even with CoT prompting, VLMs exhibit shallow reasoning chains and cannot dynamically simulate 3D transformations such as mental rotation trajectories, contrasting human parietal lobe-driven simulation mechanisms (Jung and Haier, 2007). Thirdly, erratic behaviors like multi-answer selection in single-choice 3D tasks expose poor cross-modal grounding, which is a critical gap between textual instruction parsing and visual feature extraction. Lastly, overreliance on pre-training patterns limits adaptability to novel spatial configurations.

These limitations underscore architectural deficiencies beyond data scarcity. Unlike humans who integrate dorsal and ventral visual streams for spatial processing, VLMs lack dedicated modules for dynamic spatial simulation. Bridging this gap may require hybrid architectures embedding geometric priors and neurosymbolic reasoning.

5 Conclusion

This study establishes a psychometric framework for evaluating five basic spatial abilities (BSAs) in visual language models, benchmarking 13 mainstream models across nine rigorous experiments.

Our findings reveal three key insights: Firstly, VLMs exhibit a significant performance gap compared to humans (average score: 24.95 vs. 68.38) while mirroring human performance hierarchies, excelling in 2D spatial orientation but struggling with 3D mental rotation. Secondly, model performance varies significantly by manufacturer, reflecting architectural priorities. Contrary to scaling laws, compact models frequently outperform larger counterparts. Thirdly, alternative reasoning techniques yield measurable but limited improvements, highlighting architectural constraints beyond training data or experimental setting limitations.

By establishing the first quantitative linkage between psychometric BSAs and VLM capabilities, this work can serve as a benchmark for future developments in spatial reasoning, and can be applied to several areas. For methodology, the BSA framework can help identify which BSAs are underdeveloped in current models, allowing researchers to target model design and training toward these weaker abilities and construct specialized datasets for finetuning or pretraining. For real world application, our findings have direct relevance for embodied AI systems, particularly those rely on manipulation, navigation, and human-robot interaction (Wang et al., 2024; Lin et al., 2025; Zargarzadeh et al., 2025; Chen et al., 2024a; Yan et al., 2024). In conclusion, combining advances in fundamental-level BSAs with higher-level physical learning and tasks offers a promising pathway for robust spatial intelligence. Therefore, future progress calls for closer collaboration between machine learning and cognitive science communities.

Limitations

This study revealed certain limitations. For example, spatial orientation ability was assessed using only one test due to the accessibility to the data from existing studies. Many models liperformed poorly on this test, making it difficult to discern subtle differences between them. Additionally, three tests involved choosing the correct answer from two options, which increased the likelihood of correct guesses. To avoid distortion of results, models that clearly failed to understand the questions or consistently guessed the same answer were assigned a score of zero for these tests. Future research can build upon the ability framework proposed in this study to curate or design new tests for assessing the basic spatial abilities of VLMs, addressing potential reliability issues of individual tests. It is also essential to include human experimental results as a benchmark.

The models tested in this study are all VLMs released in recent months, and the number of multimodal models with visual capabilities remains limited. Future research can use the nine tests employed in this study to evaluate the basic spatial abilities of newly developed VLMs and compare their performance to the models tested here, highlighting progress over time.

Acknowledgement

We thank the anonymous reviewers for their feedback on this work. This research is supported by the National Key Research and Development Program of China (Grant No. 2024YFC3307603) and the National Natural Science Foundation of China (Grant No. 72442026, U23B2030, 62171260).

Ethics Statement

This study utilized the API service provided by OpenAI, DeepInfra, Google, and SiliconFlow in full compliance with their terms of service and usage policies.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A frontier large vision-language model with versatile abilities.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity.

D. Ben-Chaim, G. Lappan, and R. T. Houang. 1986. Development and analysis of a spatial visualization test for middle school boys and girls. *Perceptual and Motor Skills*, 63(2 Pt 1):659–669.

George K. Bennett, Harold G. Seashore, and Alexander G. Wesman. 2012. The differential aptitude test : A review and critique.

George Bodner and Roland Guay. 1997. The purdue visualization of rotations test. *Chemical Educator*, 2:1–17.

Marc H. Bornstein and Howard Gardner. 1986. Frames of mind: The theory of multiple intelligences. *Journal of Aesthetic Education*, 20(2):120.

Jacqueline M. Caemmerer, Timothy Z. Keith, and Matthew R. Reynolds. 2020. Beyond individual intelligence tests: Application of cattell-horn-carroll theory. *Intelligence*, 79:101433.

John B. Carroll. 1993. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge University Press, Cambridge.

Raymond B. Cattell. 1963. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54(1):1–22.

Yaran Chen, Wenbo Cui, Yuanwen Chen, Mining Tan, Xinyao Zhang, Dongbin Zhao, and He Wang. 2024a. RoboGPT: An intelligent agent of making embodied long-term decisions for daily instruction tasks.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024b. How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites.

Anthony G. Cohn and Jose Hernandez-Orallo. 2023. Dialectical language model evaluation: An initial appraisal of the commonsense spatial reasoning abilities of LLMs.

Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. 2022. A survey of embodied AI: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244.

Zane Durante, Bidipta Sarkar, Ran Gong, Rohan Taori, Yusuke Noda, Paul Tang, Ehsan Adeli, Shrinidhi Kowshika Lakshmikanth, Kevin Schulman, Arnold Milstein, Demetri Terzopoulos, Ade

Famoti, Noboru Kuno, Ashley Llorens, Hoi Vo, Katsu Ikeuchi, Li Fei-Fei, Jianfeng Gao, Naoki Wake, and Qiuyuan Huang. 2024. An interactive agent foundation model.

Özlem Erkek, Mine Isiksal, and Erdinc Cakiroglu. 2011. *The Relationship between Preservice Teachers' Spatial Anxiety and Geometry Self-Efficacy in Terms of Gender and Undergraduate Program*.

Benedict Fehringer. 2021. Supplementary materials for: Implementation of the R-cube-vis test in its long and short version in english as well as german.

Benedict C. O. F. Fehringer. 2023. R-Cube-SR Test: A New Test for Spatial Relations Distinguishable From Visualization. *European Journal of Psychological Assessment*, 39(1):37–48.

Alinda Friedman, Bernd Kohler, Peri Gunalp, Alexander P. Boone, and Mary Hegarty. 2019. A computerized spatial orientation test. *Behavior Research Methods*, 52(2):799–812.

Alinda Friedman, Bernd Kohler, Peri Gunalp, Alexander P. Boone, and Mary Hegarty. 2020. A computerized spatial orientation test. *Behavior Research Methods*, 52(2):799–812.

Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. 2024. Scene-LLM: Extending Language Model for 3D Visual Understanding and Reasoning.

Chen Gao, Baining Zhao, Weichen Zhang, Jinzhu Mao, Jun Zhang, Zhiheng Zheng, Fanhang Man, Jianjie Fang, Zile Zhou, Jinqiang Cui, Xinlei Chen, and Yong Li. Embodiedcity: A benchmark platform for em- bodied agent in real-world city environment.

David C. Geary. 2022. Spatial ability as a distinct domain of human cognition: An evolutionary perspective. *Intelligence*, 90:101616.

David C. Geary, Mary K. Hoard, Lara Nugent, and John E. Scofield. 2021. In-class attention, spatial ability, and mathematics anxiety predict across-grade gains in adolescents' mathematics achievement. *Journal of Educational Psychology*, 113(4):754–769.

Gemini Team. 2024. Gemini: A family of highly capable multimodal models.

Diane F. Halpern. 1992. *Sex Differences in Cognitive Abilities, 2nd Ed.* Sex Differences in Cognitive Abilities, 2nd Ed. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US.

Mary Hegarty, Madeleine Keehner, Peter Khooshabeh, and Daniel R. Montello. 2009. How spatial abilities enhance, and are enhanced by, dental education. *Learning and Individual Differences*, 19(1):61–70.

Mary Hegarty, Daniel R. Montello, Anthony E. Richardson, Toru Ishikawa, and Kristin Lovelace. 2006. Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. *Intelligence*, 34(2):151–176.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023. 3D-LLM: Injecting the 3D World into Large Language Models.

John L. Horn. 1968. Organization of abilities and the development of intelligence. *Psychological Review*, 75(3):242–259.

W Johnson and T Bouchardjr. 2005. The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence*, 33(4):393–416.

Wendy Johnson and Thomas J. Bouchard. 2007. Sex differences in mental abilities: g masks the dimensions on which they lie. *Intelligence*, 35(1):23–39.

Rex E. Jung and Richard J. Haier. 2007. The parieto-frontal integration theory (P-FIT) of intelligence: Converging neuroimaging evidence. *Behavioral and Brain Sciences*, 30(2):135–154.

Michael J. Kane and Randall W. Engle. 2002. The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, 9(4):637–671.

Petros Katsioloudis, Vukica Jovanovic, and Mildred Jones. 2014. A comparative analysis of spatial visualization ability and drafting models for industrial and technology education students. *Journal of Technology Education*, 26:88–104.

Sihang Li, Jin Huang, Jiaxi Zhuang, Yaorui Shi, Xiaochen Cai, Mingjun Xu, Xiang Wang, Linfeng Zhang, Guolin Ke, and Hengxing Cai. 2025. SciLitLLM: How to adapt LLMs for scientific literature understanding.

Bingqian Lin, Yunshuang Nie, Ziming Wei, Jiaqi Chen, Shikui Ma, Jianhua Han, Hang Xu, Xiaojun Chang, and Xiaodan Liang. 2025. NavCoT: Boosting LLM-based vision-and-language navigation via learning disentangled reasoning.

Marcia C. Linn and Anne C. Petersen. 1985. Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development*, 56(6):1479–1498.

Llama Team, AI @ Meta. 2024. The llama 3 herd of models.

Eleanor E. Maccoby and Carol N. Jacklin. 1974. *The Psychology of Sex Differences*. The Psychology of Sex Differences. Stanford University Press.

Yukiko Maeda, So Yoon Yoon, K. Kim-Kang, and P.K. Imbrie. 2013. Psychometric properties of the revised PSVT: R for measuring first year engineering students' spatial ability. *International Journal of Engineering Education*, 29.

Peter Herbert Maier. 1996. Spatial geometry and spatial ability–How to make solid geometry solid. In *Selected Papers from the Annual Conference of Didactics of Mathematics*, Osnabrueck, Germany: Gesellschaft für Didaktik der Mathematik (GDM).

Andrea Matarazzo and Riccardo Torlone. 2025. A survey on large language models with some insights on their capabilities and limitations.

Kevin S. McGrew. 2009. CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, 37(1):1–10.

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

Ida Momennejad, Hosein Hasanbeig, Felipe Vieira, Hiteshi Sharma, Robert Osazuwa Ness, Nebojsa Jojic, Hamid Palangi, and Jonathan Larson. 2023. Evaluating cognitive maps and planning in large language models with CogEval. https://arxiv.org/abs/2309.15129v1.

Anita Pawlak-Jakubowska and Ewa Terczyńska. 2023. Evaluation of STEM students' spatial abilities based on a novel net cube imagination test. *Scientific Reports*, 13(1):17296.

James W. Pellegrino, David L. Alderton, and Valerie J. Shute. 1984. Understanding spatial ability. *Educational Psychologist*, 19(4):239–253.

M. Peters, B. Laeng, K. Latham, M. Jackson, R. Zaiyouna, and C. Richardson. 1995. A Redrawn Vandenberg and Kuse Mental Rotations Test - Different Versions and Factors That Affect Performance. *Brain and Cognition*, 28(1):39–58.

Ronen Porat and Ciprian Ceobanu. 2023. Spatial ability: Understanding the past, looking into the future. *European Proceedings of Educational Sciences*, Education, Reflection, Development - ERD 2022.

Wenmo Qiu and Xinhan Di. 2024. OCC-MLLM:empowering multimodal large language model for the understanding of occluded objects.

Manasi Sharma. 2023. Exploring and improving the spatial reasoning abilities of large language models.

C. Spearman. 1904. "general intelligence," objectively determined and measured. *American Journal of Psychology*, 15(2):201.

Yihong Tang, Ao Qu, Zhaokai Wang, Dingyi Zhuang, Zhaofeng Wu, Wei Ma, Shenhao Wang, Yunhan Zheng, Zhan Zhao, and Jinhua Zhao. 2024. *Sparkle: Mastering Basic Spatial Capabilities in Vision Language Models Elicits Generalization to Composite Spatial Reasoning*.

L. L. Thurstone. 1950. Some primary abilities in visual thinking. *Proceedings of The American Philosophical Society*, 94(6):517–521.

Steven G. Vandenberg and Allan R. Kuse. 1978. Mental Rotations, a Group Test of Three-Dimensional Spatial Visualization. *Perceptual and Motor Skills*, 47(2):599–604.

Philip E. Vernon. 1965. Ability factors and environmental influences. *American Psychologist*, 20(9):723–733.

Guy Vingerhoets, Engelien Lannoo, and Sabien Bauwens. 1996. Analysis of the money road-map test performance in normal and brain-damaged subjects. *Archives of Clinical Neuropsychology*, 11(1):1–9.

Daniel Voyer, Susan Voyer, and M. Philip Bryden. 1995. Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, 117(2):250–270.

Junchi Wang and Lei Ke. 2024. LLM-seg: Bridging image segmentation and large language model reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1765–1774.

Tian Wang, Junming Fan, and Pai Zheng. 2024. An LLM-based vision and language cobot navigation approach for Human-centric Smart Manufacturing. *Journal of Manufacturing Systems*, 75:299–305.

Yutaro Yamada, Yihan Bao, Andrew K. Lampinen, Jungo Kasai, and Ilker Yildirim. 2024. Evaluating Spatial Understanding of Large Language Models.

Yuwei Yan, Qingbin Zeng, Zhiheng Zheng, Jingzhe Yuan, Jie Feng, Jun Zhang, Fengli Xu, and Yong Li. 2024. OpenCity: A scalable platform to simulate urban activities with massive LLM agents.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of LMMs: Preliminary explorations with GPT-4V(ision).

Sadra Zargarzadeh, Maryam Mirzaei, Yafei Ou, and Mahdi Tavakoli. 2025. From decision to action in surgical autonomy: Multi-modal large language models for robot-assisted blood suction. *IEEE Robotics and Automation Letters*, 10(3):2598–2605.

Baining Zhao, Jianjie Fang, Zichao Dai, Ziyou Wang, Jirong Zha, Weichen Zhang, Chen Gao, Yue Wang, Jinqiang Cui, Xinlei Chen, and Yong Li. 2025. UrbanVideo-bench: Benchmarking vision-language models on embodied intelligence with video data in urban spaces.

A Data Statistics

As a necessary complement to the results in Section 4.2 and 4.3, this section provides the original statistics of the data from VLMs' BSA evaluation. The complete results are listed in Table 3.

Table 3: Test results of 13 VLMs and human recipients ("/" indicates invalid scores).

	SVT	**NCIT**	**DATSR**	**RCSR**	**MRMT**	**MRT**	**PSVTR**	**SBST**	**RCVis**
Qwen2-VL-72B-Instruct	15.63 (0)	18.75 (0)	10.57 (1.67)	38.19 (0.98)	41.67 (1.47)	29.17 (0)	20.00 (2.72)	47.78 (1.57)	/
Qwen2-VL-7B-Instruct	30.21 (1.47)	37.50 (0)	11.46 (0.75)	/	47.92 (1.47)	20.83 (0)	18.89 (1.57)	25.56 (1.57)	41.67 (0)
InternVL2-Llama3-76B	21.88 (0)	18.75 (0)	17.82 (1.19)	/	/	25.00 (0)	26.67 (0)	15.56 (1.57)	/
InternVL2-26B	12.50 (0)	37.50 (0)	12.15 (1.18)	/	53.13 (0)	12.50 (3.40)	21.11 (1.57)	23.33 (7.20)	/
InternVL2-8B	12.50 (0)	43.75 (0)	9.00 (0)	/	/	12.50 (0)	20.00 (0)	26.67 (2.72)	/
Llama-3.2-90B-Vision-Instruct	9.38 (0)	12.50 (0)	8.50 (0)	/	53.13 (0)	29.17 (0)	20.00 (0)	6.67 (0)	/
Llama-3.2-11B-Vision-Instruct	21.88 (0)	50.00 (0)	26.45 (0)	/	50.00 (0)	22.22 (1.96)	10.00 (0)	18.89 (1.57)	/
GPT-4o	20.83 (5.31)	37.50 (5.10)	10.10 (3.12)	46.53 (1.96)	44.79 (8.20)	12.50 (5.89)	21.11 (5.67)	36.67 (7.20)	40.56 (2.83)
GPT-4o mini	17.71 (7.80)	18.75 (10.21)	8.19 (1.09)	/	42.71 (1.47)	2.78 (1.96)	15.56 (3.14)	21.11 (1.57)	46.11 (1.57)
GPT-4 Turbo	23.96 (3.90)	18.75 (0)	16.08 (3.78)	/	42.71 (1.47)	12.50 (3.40)	14.44 (6.85)	17.78 (4.16)	/
Gemini-1.5-pro	21.88 (0)	18.75 (0)	23.93 (0.98)	/	54.17 (2.95)	4.17 (0)	18.89 (1.57)	36.67 (0)	40.00 (1.36)
Gemini-1.5-flash	12.50 (0)	62.50 (0)	25.05 (0)	/	/	16.67 (0)	16.67 (0)	30.00 (0)	/
Gemini-1.5-flash-8B	15.63 (0)	37.50 (0)	16.72 (0.59)	/	56.25 (0)	12.50 (0)	10.00 (0)	26.67 (0)	/
Average of 13 VLMs	*18.19*	*31.37*	*15.08*	*5.90*	*37.42*	*16.35*	*17.95*	*25.64*	*12.95*
Standard Deviation of 13 VLMs	5.87	15.19	6.54	16.00	21.84	8.49	4.62	10.60	20.26
Average of human recipients	*58.47*	*55.00*	*55.33*	*89.00*	*84.69*	*45.00*	*63.60*	*68.00*	*88.00*
Standard Deviation of human	19.72	19.13	/	9.00	14.41	20.83	20.53	23.00	7.00
Number of human recipients	1007	105	1480	51	61	636	1022	223	52

Table 4: VLMs' performance in Santa Barbara Solids Test with and without explicit separation of answer choices.

Model	**Original Experiment**	**Supplementary Experiment**
Qwen2-VL-72B-Instruct	47.78	23.33
Qwen2-VL-7B-Instruct	25.56	26.67
GPT-4o	36.67	33.33
GPT-4o mini	21.11	26.67
GPT-4 Turbo	17.78	13.33

Table 5: VLMs' performances in UrbanVideo-Bench embodied tasks align with their BSA performances.

	Random Baseline	InternVL2 -8B	InternVL2 -26B	InternVL2 -Llama3-76B	GPT-4o	GPT-4o mini
1. Overall BSA Score	**33.9**	**20.5**	**19.4**	**20.4**	**30.5**	**23.3**
1.1 Spatial Percpetion	25.0	12.5	12.5	21.9	20.8	17.7
1.2 Spatial Relation	34.4	26.4	24.8	18.3	31.4	13.5
1.3 Spatial Orientation	50.0	/	/	/	44.8	42.7
1.4 Mental Rotation	22.5	16.3	16.8	25.8	16.8	9.2
1.5 Spatial Visualization	37.5	26.7	23.3	15.6	38.6	33.6
2. Overall UrbanVideo-Bench Score	**21.3**	**33.3**	**34.3**	**34.8**	**52.2**	**45.5**
2.1 Recall	***18.3***	***31.3***	***36.8***	***33.6***	***60.1***	***50.1***
2.1.1 Trajectory Captioning	18.5	23.4	24.3	19.5	47.6	33.0
2.1.2 Sequence Recall	17.0	23.2	36.6	38.4	58.9	53.6
2.1.3 Object Recall	20.8	35.0	35.0	37.5	65.0	48.3
2.1.4 Scene Recall	13.5	52.3	61.3	54.1	67.6	59.5
2.1.5 Start/End Position	21.8	22.5	26.8	18.3	61.3	56.3
2.2 Perception	***26.6***	***36.3***	***34.4***	***39.7***	***48.5***	***45.6***
2.2.1 Proximity	37.8	58.0	51.2	65.5	63.0	69.7
2.2.2 Duration	35.6	44.7	40.2	48.5	47.7	51.5
2.2.3 Landmark Position	19.7	23.1	19.9	22.9	36.8	33.3
2.2.4 Goal Detection	18.0	27.4	28.1	28.1	42.4	31.3
2.2.5 Cognitive Map	21.9	28.3	32.4	33.6	52.8	42.4
2.3 Reasoning	***20.5***	***35.4***	***34.6***	***33.8***	***52.3***	***45.4***
2.3.1 Causal	18.2	33.6	32.7	30.9	66.4	65.5
2.3.2 Counterfactual	25.0	45.5	44.7	43.2	44.7	47.7
2.3.3 Association	18.3	27.0	26.5	27.4	45.8	22.9
2.4 Navigation	***18.0***	***29.5***	***29.8***	***29.7***	***45.3***	***37.9***
2.4.1 Progress Evaluation	21.8	31.5	28.9	31.3	34.2	30.8
2.4.2 High-level Planning	15.9	35.7	37.6	34.5	67.8	57.5
2.4.3 Action Generation	16.4	21.4	22.8	23.2	33.8	25.4

Table 6: VLMs' performances show a slight improvement in supplementary experimental settings.

Santa Barbara Solids Test Performance Score	**Qwen2-VL-72B-Instruct**	**Qwen2-VL-7B-Instruct**	**GPT-4o**	**GPT-4o mini**
Original Score	47.78	25.56	36.67	21.11
Chain-of-Thoughts	43.33	30.00	36.67	23.33
Few-shot Learning	46.67	26.67	40.00	16.67
CoT Self-Consistency	43.33	30.00	33.33	30.00
Tree-of-Thoughts	50.00	36.67	43.33	30.00

Table 7: Image size has little effect on model performance, while image color may introduce notable variation.

Santa Barbara Solids Test Performance Score	**Qwen2-VL-72B-Instruct**	**Qwen2-VL-7B-Instruct**	**GPT-4o**	**GPT-4o mini**
Original Image	47.78	25.56	36.67	21.11
Image Size x2	45.56	33.33	30.00	16.67
Image Size x(1/2)	43.33	33.33	30.00	16.67
Image Color to Greyscale	53.33	46.67	30.00	23.33

Table 8: Increasing the temperature parameter typically leads to decreased VLM performances.

Santa Barbara Solids Test Performance Score	**Qwen2-VL-72B-Instruct**	**Qwen2-VL-7B-Instruct**	**GPT-4o**	**GPT-4o mini**
Original Score (Temperature=0)	47.78	25.56	36.67	21.11
Temperature=0.2	40.00	43.33	30.00	16.67
Temperature=0.5	33.33	43.33	40.00	16.67
Temperature=1.0	30.00	16.67	20.00	30.00

B Prompts for VLMs

In this section, we provide the designed prompts for the nine BSA tests to receive results from VLMs.

B.1 MGMP Spatial Visualization Test (SVT)

Prompt for Extracting LLMs' Results from MGMP Spatial Visualization Test (SVT)

[Instruction]
You are taking a spatial ability test. The test consists of 32 questions. Before we start, you need to look at two example questions.

[Answer the Example Question]
The image above presents a TOP view of a structure composed of stacked cubes, with the numbers indicating the quantity of cubes in each stack. From which direction is this structure observed to create the 3D view below? (The orientation is already marked on the TOP view above)

[Provide answer after response]
The correct answer is D. Next, we will start 32 formal tests.

[Formal test]
This is Question {i}. Please output the result in pure text format as: question number,answers. For example: 1,B. Please answer.

B.2 Net Cube Imagination Test (NCIT)

Prompt for Extracting LLMs' Results from Net Cube Imagination Test (NCIT)

[Instruction]
You are taking a spatial ability test. This test consists of two sections.

Section A: On the basis of the spatial element and the entered cut line, please indicate the correct solution of development (a flat net). The cut line determines how to observe the element and determines the front wall of the spatial element. The front wall is always located as the first on the left side in the net. Stages of unfolding a spatial element into the form of a flat net is illustrated in the examples below.

Section B: Indicate, based on the net and the introduced cut line, the correct solution for the spatial element. The cut line determines the front wall for the spatial element. The examples below is illustrated the stages of net folding. The examples for section B are below.

[Formal test]
This test consists of 8 questions in section A and 8 questions in section B. Question 1~8 are in section A, and question 9~16 are in section B.
This is Question {i}. Please output the result in pure text format as: question number,answers. For example: 1,B. Please answer.

B.3 Differential Aptitude Test: Space Relation (DAT:SR)

Prompt for Extracting LLMs' Results from Differential Aptitude Test: Space Relation (DAT:SR)

[Instruction]
You are taking a spatial ability test. This test consists of 40 patterns which can be folded into figures. To the right of each pattern there are five figures. You are to decide which of these figures can be made from the pattern shown. The pattern always shows the outside of the figure.

[Answer the Example Question]
Here is an example: Which of these five figures—A, B, C, D, E—can be made from the pattern at the left?

[Provide answer after response]
In Example X, A and B certainly cannot be made; they are not the right shape. C and D is correct both in shape and size. You cannot make E from this pattern. So the right answers are CD.
Remember:
1. In this test there will always be a row of five figures following each pattern.
2. In every row there is at least one correct figure.
3. Usually more than one is correct. In facts, in some cases, all five may be correct.
Next, we will start 40 formal tests.

[Formal test]
This is Question {i}. Please output the result in pure text format as: question number,answers. For example: 1,BDE. Please answer.

B.4 R-Cube-Spatial Relation Test (R-Cube-SR)

Prompt for Extracting LLMs' Results from R-Cube-Spatial Relation Test (R-Cube-SR)

[Instruction]
You are taking a spatial ability test. You will be shown several pairs of cubes. These cubes have six different colors, but you will see only three of them. In each task, you have to decide whether the left cube can be transferred to the right one by turning or tilting it. In doing so, a new color then becomes visible.
The test is divided into two blocks. In each block, 24 tasks will be presented in succession.

[Answer the Example Question]
Now we will start the Block 1. Before the formal test, the training phase begins. You have the opportunity to see if you have understood everything in four tasks. After each task you will see whether your answer was correct or incorrect.
This is Sample Task 1. Please answer whether the left cube can be transferred to the right one by turning or tilting it.

[Provide answer after response]
The correct answer is False. It is NOT possible to transfer the left cube into the right cube. Next, we will start 24 formal tests.

[Formal test]
This is Question {i}. Please output the result in pure text format as: question number,answer ("T" for True or "F" for False). For example: 1,T. Please answer.

B.5 Money Road-Map Test (MRMT)

Prompt for Extracting LLMs' Results from Money Road-Map Test (MRMT)

[Instruction]
You are taking a spatial ability test. This picture shows a stylized city map. Imagine the dashed line is the route you walk through the city. As indicated by the number along the route, you take a total of 32 turns. You need to answer the directions you take at each turn (left or right).

Please output the result in an order from the start to the end with pure text format, with "L" for left, "R" for right, and each answer separated with a comma ",".

B.6 Mental Rotation Test (MRT)

Prompt for Extracting LLMs' Results from Mental Rotation Test (MRT)

[Instruction]
You are taking a spatial ability test. This is an image containing five smaller pictures. The top picture is the target, displaying an object. Your task is to identify which two of the four pictures below show objects that are identical to the target, but just rotated versions of it. Among the four pictures below, two are correct answers, and two are incorrect.

This is Question {i}. Please output the result in pure text format as: question number,answer. For example: 1,BC.

B.7 Purdue Spatial Visualization Tests: Visualization of Rotations (PSVT:R)

Prompt for Extracting LLMs' Results from Purdue Spatial Visualization Tests: Visualization of Rotations (PSVT:R)

[Instruction]
You are taking a spatial ability test. This test consists of 30 questions designed to see how well you can visualize the rotation of three-dimensional objects.

[Answer the Example Question]
Shown below is an example of the type of question included in the second section.
You are to:
1. study how the object in the top line of the question is rotated;
2. picture in your mind what the object shown in the middle line or the question looks like when rotated in exactly the same manner;
3. select from among the five drawings (A, B, C, D, or E) given in the bottom line of the question the one that looks like the object rotated in the correct position.
Please answer: What is the correct answer to the example shown above?

[Provide answer after response]
Answers A, B, C, and E are wrong. Only drawing D looks like the object rotated according to the given rotation, Remember that each question has only one correct answer.
Next, we will start 30 formal tests.

[Formal test]
This is Question {i}. Please output the result in pure text format as: question number,answer. For example: 1,B. Please answer.

B.8 Santa Barbara Solids Test (SBST)

Prompt for Extracting LLMs' Results from Santa Barbara Solids Test (SBST)

[Instruction]

You are taking a spatial ability test. This test consists of 30 questions about cross-sections. A cross-section is the 2D shape that results when a cutting plane intersects an object.
There are many examples of cross-sections in everyday life. For example, when you slice an apple from top to bottom, the resulting cut surface is a cross-section of the apple.
Picture 1 shows an apple with some worms inside. Note that the cross section on the right shows both the apple and the shapes and locations of the sliced worms inside the apple.

In this multiple choice test, you will be asked to identify the cross sections of three types of figures as shown in picture 2.
Here are some important things to remember:
1. All figures are solid (not hollow) objects.
2. The objects are about 6-8 inches tall. Imagine that they are on the table in front of you.
3. Attached figures are "glued together" at their edges.
4. Nested objects consist of one object inside another. In the nested object above, the cylinder extends all the way through the cube. If you sliced this figure, you would see the cylinder inside the cube.
The cutting planes, shown in grey, will have different orientations, as shown in picture 3.

You will see three types of cutting planes: horizontal, vertical, and oblique.
For each type of cutting plane, try to imagine the cross section that would result if you faced the cutting plane head-on, as if you were looking at your reflection in a mirror, as shown in the picture 4.

You should also assume that the objects are 6-8 inches tall, and that they are sitting on the desk in front of you. In the example below, the cutting plane would produce the cross section on the right, as shown in picture 5.

[Answer the Example Question]

Now you need to do a sample question shown in picture 6.
Select from among the four answers (A, B, C, or D) given in the bottom line of the picture the cross-section that you would see when the grey cutting plane slices the object.
Imagine that you are facing the cutting plane head-on, as if you were looking in a mirror. Make your choice based on the shapes of the possible answers, not their sizes. Please answer.

[Provide answer after response]

Answers A, B, and D are wrong. Only C looks like the cross-section that you would see when the grey cutting plane slices the object. Remember that each question has only one correct answer. Next, we will start 30 formal tests.

[Formal test]

This is Question {i}. Please output the result in pure text format as: question number,answer. For example: 1,B. Please answer.

B.9 R-Cube-Visualization Short Test (R-Cube-Vis)

Prompt for Extracting LLMs' Results from R-Cube-Visualization Short Test (R-Cube-Vis)

```
[Instruction]
You are taking a spatial ability test. You will be shown several pairs of cubes. These cubes have
six different colors, but you may see fewer of them. These pairs consist of either 3-cubes or
4-cubes. In each task you always have to decide whether the left cube can be transformed into the
right one by rotating elements. In this process, one or more new colors may also become visible.
There will be 3 blocks with different tasks, and each block is announced in advance. In the first
block, only one element is rotated. In the second block, two elements are rotated in parallel. In
the third block, two elements are rotated that cross each other.
There will be a training phase before each block, In this phase, you can practice the upcoming tasks
and will receive feedback.

[Answer the Example Question]
Now we will start the Block 1. There will be shown 3- and 4- cubes with ONE rotated element.
First, you have the opportunity to see whether you have understood the task with four sample tasks
in a training phase. After each task, you will be shown whether you answered correctly or incorrect-
ly. This is Sample Task 1. Please answer whether the left cube can be transformed into the right one
by rotating ONE element.

[Provide answer after response]
The correct answer is True. If you turn the second right column from left to top, then it is possi-
ble to transform the left cube into the right cube.
Next, we will start 20 formal tests.

[Formal test]
This is Question {i}. Please output the result in pure text format as: question number,answer ("T"
for True or "F" for False). For example: 1,T. Please answer.
```

C CoT Method and Example-based Training Demonstration

In this section, we selected the SBST test, using GPT-4o to evaluate the extent to which the chain-of-thought (CoT) method and example-based training enhance the model's ability to solve spatial problems.

In the first experiment, we designed a structured CoT reasoning process (understanding the 3D shape, analyzing the plane, determining the cross-section, matching cross-section to options, giving the answer) along with a single example question. We then assessed the model's accuracy under four different conditions: (a) directly answering without additional guidance, (b) using only example-based training, (c) employing only the CoT method, and (d) combining both CoT and example-based training, where the example question was analyzed using the CoT approach.

In the second experiment, we examined the impact of varying the number of example questions (1, 3, 5, and 10) on response accuracy, without incorporating the CoT method.

To enhance the reliability of the findings, both experiments were conducted multiple times to take the average result. The experimental framework is illustrated in Figure 9.

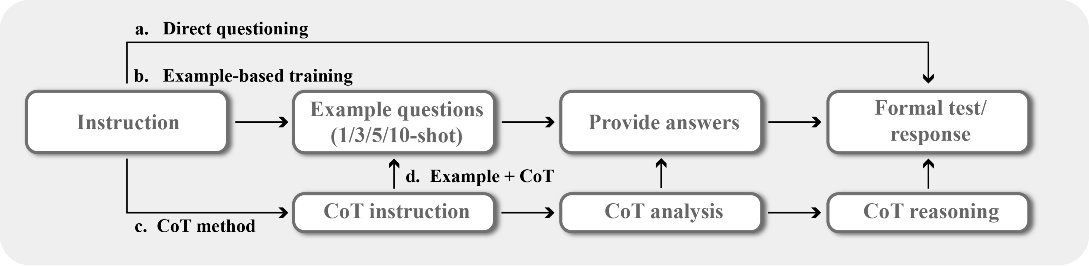
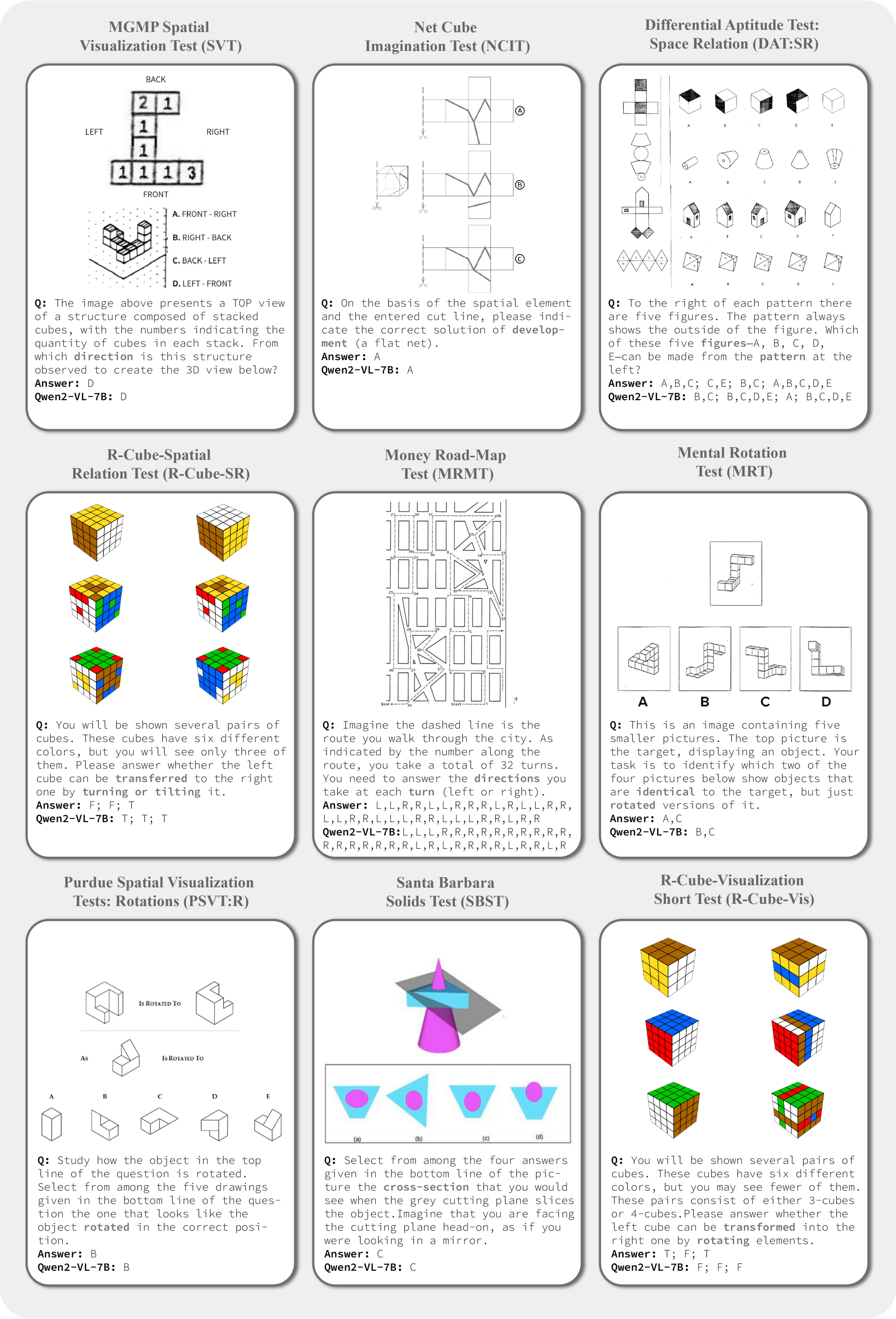


Figure 9: Framework of CoT Method and Example-based Training

D Sample Data Demonstration

In this section, we take Qwen2-VL-7B, the model that acquired the highest score as an example, to demonstrate its response to nine sample questions of different tests, as shown in Figure 10.



MGMP Spatial Visualization Test (SVT)

Q: The image above presents a TOP view of a structure composed of stacked cubes, with the numbers indicating the quantity of cubes in each stack. From which **direction** is this structure observed to create the 3D view below?
Answer: D
Qwen2-VL-7B: D

Net Cube Imagination Test (NCIT)

Q: On the basis of the spatial element and the entered cut line, please indicate the correct solution of **development** (a flat net).
Answer: A
Qwen2-VL-7B: A

Differential Aptitude Test: Space Relation (DAT:SR)

Q: To the right of each pattern there are five figures. The pattern always shows the outside of the figure. Which of these five **figures**—A, B, C, D, E—can be made from the **pattern** at the left?
Answer: A,B,C; C,E; B,C; A,B,C,D,E
Qwen2-VL-7B: B,C; B,C,D,E; A; B,C,D,E

R-Cube-Spatial Relation Test (R-Cube-SR)

Q: You will be shown several pairs of cubes. These cubes have six different colors, but you will see only three of them. Please answer whether the left cube can be **transferred** to the right one by **turning or tilting** it.
Answer: F; F; T
Qwen2-VL-7B: T; T; T

Money Road-Map Test (MRMT)

Q: Imagine the dashed line is the route you walk through the city. As indicated by the number along the route, you take a total of 32 turns. You need to answer the **directions** you take at each **turn** (left or right).
Answer: L,L,R,R,L,L,R,R,R,L,R,L,L,R,R,L,L,R,R,L,L,L,R,R,L,L,L,R,R,L,R,R
Qwen2-VL-7B: L,L,L,R,R,R,R,R,R,R,R,R,R,R,R,R,R,R,R,R,L,R,L,R,R,R,R,L,R,R,L,R

Mental Rotation Test (MRT)

Q: This is an image containing five smaller pictures. The top picture is the target, displaying an object. Your task is to identify which two of the four pictures below show objects that are **identical** to the target, but just **rotated** versions of it.
Answer: A,C
Qwen2-VL-7B: B,C

Purdue Spatial Visualization Tests: Rotations (PSVT:R)

Q: Study how the object in the top line of the question is rotated. Select from among the five drawings given in the bottom line of the question the one that looks like the object **rotated** in the correct position.
Answer: B
Qwen2-VL-7B: B

Santa Barbara Solids Test (SBST)

Q: Select from among the four answers given in the bottom line of the picture the **cross-section** that you would see when the grey cutting plane slices the object. Imagine that you are facing the cutting plane head-on, as if you were looking in a mirror.
Answer: C
Qwen2-VL-7B: C

R-Cube-Visualization Short Test (R-Cube-Vis)

Q: You will be shown several pairs of cubes. These cubes have six different colors, but you may see fewer of them. These pairs consist of either 3-cubes or 4-cubes.Please answer whether the left cube can be **transformed** into the right one by **rotating** elements.
Answer: T; F; T
Qwen2-VL-7B: F; F; F

Figure 10: Data samples from tests for BSA^2.