

# Mind’s Eye: A Benchmark of Visual Abstraction, Transformation and Composition for Multimodal LLMs

**Rohit Sinha\***

CSE Dept., IIT Hyderabad  
Hyderabad, India

rohit.sinha@prjt.cse.iith.ac.in

**Aditya Kanade**

Microsoft Research  
Bengaluru, India

kanade850@gmail.com

**Sai Srinivas Kancheti\***

CSE Dept., IIT Hyderabad  
Hyderabad, India

cs21resch01004@iith.ac.in

**Vineeth N Balasubramanian†**

Microsoft Research  
Bengaluru, India

vineeth.nb@microsoft.com

**Tanuja Ganu†**

Microsoft Research  
Bengaluru, India

tanuja.ganu@microsoft.com

## Abstract

Multimodal large language models (MLLMs) have achieved impressive progress on vision language benchmarks, yet their capacity for visual cognitive and visuospatial reasoning remains less understood. We introduce MIND’S EYE, a multiple-choice benchmark of eight visuo-cognitive tasks inspired by classic human intelligence tests and organized under a novel **A–R–T** taxonomy: *Abstraction*, *Relation*, and *Transformation*. The tasks probe core processes of fluid intelligence such as pattern induction, analogical relation mapping, and mental transformation. We evaluate a diverse suite of closed-source and open-source MLLMs and compare their performance with human participants. Humans achieve **80%** accuracy, while top performing MLLMs remain below **50%**. Error analysis reveals failures in: (i) visual attention allocation, (ii) internal perceptual manipulation, and (iii) weak abstraction of underlying visual concepts. Our findings suggest that current MLLMs exhibit limited visuospatial reasoning capabilities, when compared with human participants, highlighting the need for more cognitively grounded evaluation frameworks<sup>1</sup>.

\*Work done at Microsoft Research India

†Corresponding author

<sup>1</sup>Code and Benchmark are available at: <https://github.com/microsoft/Mind-s-Eye>

## 1 Introduction

Multimodal Large Language Models (MLLMs) have demonstrated compelling visual understanding in recent years: identifying objects, reading text, or describing spatial relationships in presented scenes (Li et al., 2023c). These tasks primarily test whether models can encode visual inputs and map them to linguistic outputs. As MLLMs become stronger, it becomes imperative to study their performance on increasingly complex visual tasks that are natural to humans. To this end, visuospatial transformation tasks require such models to generate novel visual states not present in the input such as mentally rotating a 3D object, predicting how a surface unfolds, or visualizing shape compositions. Such transformations form the core of spatial reasoning but demand a capability beyond perceptual encoding: the construction and manipulation of implicit spatial representations (Vandenberg and Kuse, 1978a). The capability of MLLMs to possess such generative spatial understanding remains an open empirical question.

Existing evaluations of MLLMs can be broadly categorized into three kinds: broad evaluations that test surface perception and prioritize scale (Liu et al., 2024a; Yue et al., 2024), synthetic diagnostics that probe compositional reasoning through pattern matching (rather than mental simulation) (Zhang et al., 2019), and cognitive bench-

marks that test spatial rule learning (Chollet et al., 2025). We point out two key gaps across these studies: (1) Firstly, existing evaluations do not isolate and study visuospatial transformation, the capability (that comes to humans naturally) to mentally rotate, fold, or recompose shapes (Shepard and Metzler, 1971; Fleuret et al., 2011); and (2) Most existing studies often conflate visual evidence with linguistic priors, leaving it unclear on whether models reason from images or exploit language shortcuts (Suhr et al., 2019).

Keeping these in mind, we herein introduce MIND’S EYE, a cognitively grounded benchmark derived from classic cognitive psychology tests such as mental rotation and paper folding (Ekstrom et al., 1976a). Eight visuo-cognitive tasks are organized under an Abstraction-Relation-Transformation (ART) taxonomy: Abstraction tests study pattern induction, Relation tests study analogical mapping, Transformation tests study mental manipulation of shapes. Our generation process allows us to isolate visuospatial reasoning from world knowledge and linguistic priors. Each item includes diagnostic distractors targeting specific error types, enabling fine-grained analysis of where and why models fail. In particular, we organize our investigation around three questions: (1) *How do MLLMs perform and compare with human performance on controlled visuospatial tests?* (2) *Which cognitive factors drive the largest deficits in the performance of MLLMs on these tasks?* (3) *Can prompting interventions help improve the performance of MLLMs on the considered tasks, or do failures reflect general model limitations?*

Evaluation of 18 MLLMs on MIND’S EYE reveals significant underperformance relative to humans in visuospatial reasoning. Humans average 80% accuracy across tasks, while the best models remain below 50%. The largest deficits appear on Transformation and Abstraction tasks, both of which require mental simulation rather than surface pattern matching. Notably, while human accuracy degrades from easy to hard instances, MLLM performance remains flat across difficulty levels, indicating the absence of foundational visuo-cognitive operations rather than mere struggles with complexity. Prompting interventions yield dimension-dependent effects: structured scaffolding benefits Abstraction tasks but consistently impairs Transformation performance, suggesting that prompting facilitates rule deriva-

tion yet fails to elicit procedural visuospatial operations. Attention analysis further reveals that while models can localize relevant answer regions, they fail to reason reliably over this information—they identify where to look but not how to reason over what they see.

Our work makes the following contributions:

- A new benchmark, MIND’S EYE, for visuo-cognitive understanding of MLLMs, grounded in cognitive constructs of Abstraction-Relation-Transformation (ART) which includes diagnostic distractors.
- Evaluation of 18 MLLMs on the benchmark and comparison with a human baseline; the study includes prompting strategies, as well as fine-tuning and reinforcement learning-based alignment on a strong open-source model.
- Diagnostic analyses showing attention misalignment, difficulty-invariant failure patterns, and reasoning trace errors.

## 2 Related Work

### Multimodal and visual reasoning benchmarks.

General-purpose benchmarks such as MMBench, SEED-Bench, MathVista, and MMMU (Liu et al., 2024b; Li et al., 2024, 2023a; Lu et al., 2024a; Yue et al., 2024; Yu et al., 2024) measure breadth across Visual QA, OCR, and mathematical reasoning, but lack parametric control for studying visuospatial understanding. Compositional reasoning benchmarks including CLEVR, RAVEN, and CV-Bench (Johnson et al., 2017; Zhang et al., 2019; Fan et al., 2025; Hudson and Manning, 2019; Suhr et al., 2019) target attribute binding and relational comparison over single-frame perception, yet offer limited control over geometric transformations (e.g., rotation angle, fold parity) or mental manipulation.

### Cognitive and analogical reasoning benchmarks.

Recent efforts have moved toward cognitive testing: *Mind the Gap* (Stogiannidis et al., 2024) for spatial completion, *Bongard-LOGO* (Nie et al., 2020a) and *VisuLogic* (Xu et al., 2025a) for rule induction, and *Do You See Me* (Kanade and Ganu, 2025) for perception grounding. However, benchmarks like ARC, Bongard-LOGO, and SVRT (Chollet, 2019; Fleuret et al., 2011; Nie

Table 1: **Closest benchmarks vs. MIND’S EYE along diagnostic axes:** A comparative evaluation of MIND’S EYE against other benchmarks on key diagnostic criteria. The table highlights the unique features that make MIND’S EYE a more controlled and cognitively grounded diagnostic tool for assessing fluid visuospatial intelligence. ✓=explicit support; ◻=partial; ✗=absent.

Dataset	Formal Psychometric Taxonomy	Psychometric Task Derivation	Distractors Keyed to Confounds	No Knowledge Reliance	Parametric Control	Scalability
RAVEN (Zhang et al., 2019)	✗	✓	✗	✓	✗	✓
Bongard-LOGO (Nie et al., 2020a)	✗	✗	✗	✓	✗	✓
CLEVR (Johnson et al., 2017)	✗	✗	✗	✗	✓	✓
VGRP-Bench (Ren et al., 2025)	✗	✗	✗	◻	✓	✓
VisualPuzzles (Song et al., 2025)	✗	✗	✗	✓	✓	✗
AlgotpuzzleVQA (Ghosal et al., 2025)	✗	✗	✗	◻	✗	✓
VisFactor (Huang et al., 2025a)	✗	✓	✗	✓	✗	✗
IQBench (Pham et al., 2025)	✗	✗	✗	✓	✗	✗
NTSEBench (Pandya et al., 2025)	✗	✗	✗	◻	✗	✗
SpatialVisBench (Wang et al., 2025)	✗	✗	✗	✓	✓	✗
<b>Mind’s Eye (ours)</b>	✓	✓	✓	✓	✓	✓

et al., 2020a) emphasize rule discovery and analogy over stepwise geometric simulation. VisFactor (Huang et al., 2025a) evaluates basic perceptual factors by digitizing FRCT-style tests, while IQBench (Pham et al., 2025) assesses broader IQ-style reasoning including RPMs and analogies.

**Cognitive science foundations.** Our benchmark draws from classical studies on mental rotation (Shepard and Metzler, 1971), the Vandenberg & Kuse MRT (Vandenberg and Kuse, 1978b), and CogAT paper-folding tests (Publishing, 2009), as well as Hofstadter’s theory of analogy (Hofstadter, 1979) and Newell’s unified cognition framework (Newell, 1994). While existing multimodal benchmarks emphasize either *recognition* or *abstraction*, they under-specify *internal simulation*. MIND’S EYE bridges this gap through programmatic generation of tasks probing whether MLLMs can perform internal transformations—rotation, folding, composition, symmetry recognition—central to human visuospatial intelligence.

**Positioning Mind’s Eye.** Table 1 compares benchmarks along six diagnostic axes. *Formal Psychometric Taxonomy* asks whether tasks are organized under a structured cognitive framework with explicit construct-coverage mappings (e.g., a q-matrix), not merely motivated by cognitive science. *Task Derivation from Established Assessments* indicates whether tasks are adapted from validated psychometric instruments. *Distractors Keyed to Confounds* captures whether wrong-answer options are designed to diagnose specific reasoning errors rather than sampled randomly. *No Knowledge Reliance* marks benchmarks solv-

able without domain knowledge or linguistic priors. *Parametric Control* indicates whether independent generation parameters enable systematic difficulty manipulation, and *Scalability* indicates whether new items can be produced programmatically at negligible cost. To our knowledge, Mind’s Eye is the first benchmark in this space to satisfy all six criteria simultaneously.

### 3 MIND’S EYE: The Benchmark

Going beyond assessing *what* models perceive (such as in object recognition or scene description of images), our proposed benchmark seeks to study the capabilities of models when one focuses on *how* models reason over visual input. Core capacities of human visual intelligence, such as mentally rotating objects, tracking structure through spatial transformations, or inducing abstract rules from visual patterns together constitute *visuocognitive reasoning*, i.e. cognitive operations performed over visual representations, encompassing not only spatial manipulation but also pattern abstraction and relational inference. MIND’S EYE, our proposed benchmark, is grounded in an **Abstraction–Relation–Transformation (ART)** taxonomy that isolates these visuocognitive processes. The taxonomy draws on Carroll’s construct of fluid intelligence (Carroll, 1993) to decompose visual reasoning into three complementary facets: inferring abstract patterns, mapping relational correspondences, and mentally manipulating spatial structure. We now detail the conceptual foundations of ART (§3.1), as well as the benchmark’s design principles and task suite (§3.2).

### 3.1 The ART Taxonomy

MIND’S EYE organizes visuocognitive reasoning along three dimensions: **Abstraction**, **Relation**, and **Transformation** (ART). These are complementary facets of fluid visual reasoning, viz. the capacity to solve novel problems through deliberate, knowledge-independent thought (Carroll, 1993; Schneider and McGrew, 2018). Each dimension isolates a distinct cognitive operation over visual input:

- **Abstraction** requires *inducing latent structure from surface variation*. The solver must identify an underlying rule, pattern, or category that unifies disparate visual instances; for example, recognizing that two differently oriented configurations share identical hierarchical organization. This corresponds to inductive reasoning in psychometric models of fluid intelligence (McGrew, 2005).
- **Relation** requires *mapping correspondences across visual structures*. The solver must detect how elements in one configuration align with elements in another, supporting analogical transfer and structural comparison. This corresponds to the relational reasoning central to analogy-making and fluid intelligence (Halford et al., 2010; Nie et al., 2020a).
- **Transformation** requires *mentally simulating spatial operations*. The solver must internally rotate, fold, compose, or otherwise manipulate visual representations to predict outcomes—engaging spatial working memory and figural reasoning (Ekstrom et al., 1976b; Vandenberg and Kuse, 1978b).

This framework draws on Carroll’s Three Stratum Theory, which situates fluid intelligence (Gf) as a broad factor underlying performance on novel reasoning tasks (Carroll, 1993). Crucially, fluid intelligence manifests not only in verbal or symbolic reasoning but also in figural and spatial domains; tasks such as Raven’s Progressive Matrices are canonical measures precisely because they require abstract rule induction over visual patterns (Raven, 2000). The ART taxonomy makes explicit the component processes conflating such tasks, enabling targeted diagnosis of where models succeed or fail.

The ART taxonomy provides the theoretical scaffold; the benchmark instantiates it through

tasks that satisfy three design principles. (1) *Cognitive isolation*: tasks require reasoning over visual structure, not retrieval of world knowledge, ensuring that performance reflects visuocognitive capacity rather than domain familiarity. (2) *Diagnostic precision*: each item includes carefully constructed distractors tied to specific reasoning errors (e.g., mirrored transformations, parity mistakes), enabling fine-grained failure analysis. *Psychometric rigor*: stimulus generation follows factorial designs with calibrated difficulty, and all items use standardized multiple-choice format to permit reliable comparison across models and against human baselines. The following subsection details how these principles are realized in the benchmark’s eight tasks.

### 3.2 Benchmark Design

**Task Suite.** Our benchmark comprises eight tasks, distributed across the ART dimensions (Figure 1). *Abstraction* is probed through Visual Relation Abstraction (VRA) and Hierarchical Pattern Equivalence (HPE), which require inducing latent rules or detecting recursive structure from visual exemplars. *Relation* is probed through Dynamic Structural Correspondence (DSC), Visual Conceptual Slippage (VCS), and Symmetric Structures (SS), which require mapping correspondences across configurations or detecting violations of relational invariants. *Transformation* is probed through Mental Transformation (MT), Paper Folding (PF), and Mental Composition (MC), classic tests of spatial manipulation adapted from the psychometric literature (Vandenberg and Kuse, 1978b; Publishing, 2009). Each task is operationalized as a multiple-choice problem with four or six options. The formal task–construct mapping is specified via a q-matrix (Table 12 in the Appendix), following psychometric design standards for construct coverage (Embretson and Reise, 2013).

**Stimulus Generation.** All stimuli are programmatically generated as scalable vector graphics, enabling control over geometric parameters and ensuring perceptual uniformity across items. Generation follows a factorial design: structural parameters that determine task difficulty (e.g., rotation magnitude, fold count, hierarchy depth) are varied independently of nuisance parameters (e.g., color, spatial layout, surface texture) that should be task-irrelevant. This separation serves two pur-

Abstraction	Relation		
Visual Relation Abstraction	Dynamic Structural Isomorphism	Visual Conceptual Slippage	Symmetric Structures
<p>Question: Identify which figure does not adhere to the underlying visual concept</p>	<p>Question: Follow the transformations and identify the fifth image at <math>t = 1.0</math></p>	<p>Question: Which figure does not adhere to the underlying visual concept that is shared by the majority of these figures</p>	<p>Question: Identify the odd figure in the grid.</p>
Hierarchical Pattern Equivalence	Transformation		
<p>Question: Identify the image that violates the pattern</p>	Mental Composition	Mental Transformation	Paper Folding
	<p>Question: When the 2d figure is folded into a 3d shape, which of the following is the correct shape?</p>	<p>Question: Which option is the correct transformation for the original shape?</p>	<p>Question: If the paper is folded as shown, which options correct after unfolding?</p>

Figure 1: **Overview of the eight tasks in the proposed MIND’S EYE benchmark:** Each panel shows an example image-question pair of the benchmark

poses. Firstly, it permits a priori difficulty calibration based on structural complexity (Embretson, 1983; Ekstrom et al., 1976b). Secondly, it mitigates shortcut learning: models cannot exploit incidental correlations between surface features and correct answers. Full generation specifications for each task are provided in Appendix C.

### Rationale for Synthetic Stimulus Generation:

The use of synthetic, programmatically generated SVG images in our benchmark follows established practices in cognitive psychology. Synthetic stimulus generation enables precise control over confounding variables while isolating specific cognitive abilities, an approach employed by foundational assessments that remain the gold standard for measuring human cognition, including the Kit of Factor-Referenced Cognitive Tests (Ekstrom et al., 1976a; Vandenberg and Kuse, 1978c; Thurstone, 1938; Guilford and Zimmerman, 1948; McFall et al., 1993). Critically, evidence demonstrates that performance on synthetic reasoning tasks correlates with general visual cognition and real-world capabilities across domains (Burton, 2003; Moen et al., 2020; Kunda et al., 2012). This approach also aligns with how the community has adopted benchmarks like ARC-AGI (Chollet et al., 2025) as measures of progress toward general intelligence. By grounding each visuospatial reasoning task in established cognitive science assessments (Section 3.1), our benchmark provides diagnostic insights into whether failures in visual understanding of models stem from fundamental cognitive limitations versus superficial perception

gaps, a distinction crucial for understanding and improving MLLM capabilities. Similar contemporary work has validated this approach for evaluating visual reasoning and abstract reasoning (Xu et al., 2025a; Stogiannidis et al., 2025b) in vision-language models.

**Diagnostic Distractors.** Each item includes distractor choices in the answer options designed to capture the granularity of a model’s reasoning error. For Transformation tasks, distractors include reflections mistaken for rotations, incorrect fold parity, and off-by- $\theta$  rotation errors. For Relation tasks, distractors swap corresponding elements or preserve surface similarity while violating structural correspondence. For Abstraction tasks, distractors match superficial features (e.g., shape, color) but violate the latent rule. This design helps analyze errors beyond a binary (correct/incorrect) observation: the distractor chosen reveals the model’s understanding and approach to the solution, enabling a relatively more fine-grained comparison across models. Detailed distractor generations are discussed in Appendix D.

**Benchmark Scale.** Our evaluation set comprises 800 items: 100 per task, balanced across the three ART dimensions. The size of our benchmark follows the scale of existing synthetic benchmarks for targeted capability assessment (Xu et al., 2025b; Stogiannidis et al., 2025b). We however note that while our objective in this work is diagnostic assessment of MLLMs, since our stimuli are programmatically generated, our dataset

	Accuracy $\uparrow$							
	Abstraction		Relation			Transformation		
	VRA	HPE	DSC	VCS	SS	MT	PF	MC
Random Choice	16.0	25.0	25.0	16.0	25.0	25.0	25.0	25.0
Human	68.0	88.0	81.2	87.0	78.0	81.0	80.1	82.0
<b>Open-source multimodal LLMs</b>								
Idefics - 8B	24.0 $\pm$ 0.02	25.1 $\pm$ 0.08	32.3 $\pm$ 0.08	12.2 $\pm$ 0.01	25.0 $\pm$ 0.01	32.0 $\pm$ 0.09	<b>41.5<math>\pm</math>0.00</b>	20.0 $\pm$ 0.02
InternVL3 - 8B	22.0 $\pm$ 0.01	<b>29.1<math>\pm</math>0.21</b>	31.0 $\pm$ 0.03	23.7 $\pm$ 0.06	29.1 $\pm$ 0.01	29.1 $\pm$ 0.05	24.6 $\pm$ 0.09	28.0 $\pm$ 0.05
LLaMa-3.2 - 11B	22.0 $\pm$ 0.06	29.0 $\pm$ 0.02	31.2 $\pm$ 0.05	23.1 $\pm$ 0.02	29.3 $\pm$ 0.03	29.2 $\pm$ 0.08	24.5 $\pm$ 0.02	28.0 $\pm$ 0.02
Llava-1.6-Mistral - 7B	16.0 $\pm$ 0.01	23.7 $\pm$ 0.02	<b>32.4<math>\pm</math>0.04</b>	30.6 $\pm$ 0.01	24.5 $\pm$ 0.02	<b>35.8<math>\pm</math>0.04</b>	24.1 $\pm$ 0.04	29.1 $\pm$ 0.02
Phi3.5-vision-instruct - 8B	22.0 $\pm$ 0.01	<b>29.1<math>\pm</math>0.02</b>	31.0 $\pm$ 0.04	23.5 $\pm$ 0.01	29.7 $\pm$ 0.01	29.3 $\pm$ 0.03	24.3 $\pm$ 0.01	28.7 $\pm$ 0.02
Qwen-2.5-VL - 3B	20.0 $\pm$ 0.00	26.2 $\pm$ 0.02	31.0 $\pm$ 0.09	21.0 $\pm$ 0.01	21.2 $\pm$ 0.02	22.4 $\pm$ 0.01	25.0 $\pm$ 0.01	27.9 $\pm$ 0.01
Qwen-2.5-VL - 7B	19.1 $\pm$ 0.01	24.2 $\pm$ 0.01	30.4 $\pm$ 0.01	22.7 $\pm$ 0.01	20.2 $\pm$ 0.04	25.7 $\pm$ 0.02	25.1 $\pm$ 0.02	36.4 $\pm$ 0.01
Qwen-2.5-VL - 32B	25.1 $\pm$ 0.01	18.3 $\pm$ 0.01	22.6 $\pm$ 0.04	30.2 $\pm$ 0.07	26.3 $\pm$ 0.02	27.6 $\pm$ 0.01	32.0 $\pm$ 0.02	49.5 $\pm$ 0.02
Blip - 2.7B	11.2 $\pm$ 0.07	22.7 $\pm$ 0.02	18.3 $\pm$ 0.02	09.1 $\pm$ 0.04	17.0 $\pm$ 0.01	10.1 $\pm$ 0.05	21.4 $\pm$ 0.02	24.0 $\pm$ 0.02
InstructBlip - 4B	16.3 $\pm$ 0.01	26.4 $\pm$ 0.02	19.1 $\pm$ 0.02	12.3 $\pm$ 0.04	15.0 $\pm$ 0.05	28.1 $\pm$ 0.02	11.3 $\pm$ 0.01	13.0 $\pm$ 0.07
Paligemma - 3B	12.5 $\pm$ 0.02	17.2 $\pm$ 0.02	12.7 $\pm$ 0.03	34.7 $\pm$ 0.03	13.1 $\pm$ 0.01	14.0 $\pm$ 0.02	26.4 $\pm$ 0.02	29.6 $\pm$ 0.03
Smol - 2.2B	11.3 $\pm$ 0.03	21.2 $\pm$ 0.03	19.2 $\pm$ 0.06	21.2 $\pm$ 0.21	15.1 $\pm$ 0.01	23.5 $\pm$ 0.01	26.0 $\pm$ 0.02	28.2 $\pm$ 0.31
<b>Open-source multimodal LRMs</b>								
Vision-G1 - 7B	22.3 $\pm$ 0.05	24.2 $\pm$ 0.17	29.7 $\pm$ 0.11	24.1 $\pm$ 0.16	23.6 $\pm$ 0.07	25.1 $\pm$ 0.01	28.1 $\pm$ 0.17	38.1 $\pm$ 0.12
GT-Thinker - 7B	23.1 $\pm$ 0.11	25.5 $\pm$ 0.09	28.1 $\pm$ 0.04	25.7 $\pm$ 0.07	24.1 $\pm$ 0.14	26.7 $\pm$ 0.01	27.9 $\pm$ 0.04	39.6 $\pm$ 0.17
V-Thinker - 8B	21.5 $\pm$ 0.15	22.6 $\pm$ 0.08	27.1 $\pm$ 0.61	25.5 $\pm$ 0.71	22.9 $\pm$ 0.16	24.5 $\pm$ 0.10	29.2 $\pm$ 0.18	32.2 $\pm$ 0.21
<b>API-based models</b>								
GPT-o3	21.1	22.4	11.2	23.7	22.3	25.1	25.6	43.1
GPT-4o	28.4	26.6	30.3	25.2	19.1	32.7	29.0	35.0
Gemini-2.5	<b>29.0</b>	20.2	30.0	<b>35.3</b>	<b>31.4</b>	35.6	31.1	<b>51.8</b>

Table 2: **Task-wise results of MLLMs:** *Abstraction:* VRA (Visual Relation Abstraction), HPE (Hierarchical Pattern Equivalence). *Relation:* DSC (Dynamic Structural Correspondence), VCS (Visual Conceptual Slippage), SS (Symmetric Structures). *Transformation:* MT (Mental Transformation), PF (Paper Folding), MC (Mental Composition).

can scale without additional annotation cost, if required for training purposes. We provide an extended set of 2,500 items per task (20,000 total), generated with identical templates and difficulty controls, for fine-tuning or representation learning. Both the diagnostic and extended versions of our benchmark will be made public on acceptance. The diagnostic and extended partitions do not have data overlap, and hence are maintained separately to support both fair comparison. Full generation procedures, per-task specifications, and dataset statistics are detailed in Appendix C.

**Human Evaluation.** To study how humans perform on the benchmark, we recruited 30 participants of the age group ranging from 20 to 40, with a gender distribution of 19 males and 11 females. Each participant was presented with 5 questions from each task, sampled via inverse-frequency weighting from a pool of 20 questions per task (total of 40 questions across tasks). To minimize bias and ensure consistency, all participants first completed an identical calibration phase consisting of 8 examples spanning all tasks. Human accuracy

for each subtask was measured by comparing participant responses against the ground truth. For more details on the human evaluation protocol, see Appendix F.

## 4 Experiments and Results

**Experimental Setup.** We evaluate MIND’S EYE on a wide range of recent MLLMs, including GPT-4o, GPT-o3 and Gemini-2.5 pro which are accessed via their respective proprietary APIs, as well as open-source models: LLaVA-1.6-7B, Llama-3.2-11B-Vision, phi-4-multimodal-instruct-5.7B, Qwen2.5-VL-Instruct (3B, 7B and 32B) and InternVL3.5-8B. To ensure fair comparison, all models are evaluated on identical visual inputs and standardized textual prompts. Since modern MLLMs often produce long, free-form outputs, rule-based answer extraction can be unreliable (Duan et al., 2024; Fu et al., 2024). Following recent practice (Lu et al., 2024b; Zhang et al., 2024), we adopt an *expert LLM evaluation pipeline* comprising three stages: (1) candidate model receives the image and question in a fixed template; (2) an answer extractor, Gemma-

3 (Team et al., 2025), parses the raw output into a concise response; and (3) the parsed response is mapped to standardized task-specific labels for accuracy computation across all eight tasks. This approach leverages robust semantic extraction via a large model, while maintaining fully automated, reproducible scoring (details in Appendix E). To prevent positional bias, correct answer options were randomly rotated across positions following standard MCQ evaluation practice.

**Prompting strategies.** Since multimodal reasoning can be sensitive to prompt phrasing (Wei et al., 2022; Kojima et al., 2022), we evaluate four structured prompting paradigms: Chain-of-Thought (CoT), Meta-Task Framing, Step-by-Step Instruction (SBS), and Hint-based prompting. Full prompt templates and examples for each strategy are provided in Appendix E and H. Results for CoT-based prompts are in Table 2; results of other prompting strategies are in Appendix G.

**Main Results.** Our primary results are reported in Table 2. The results reveal a general weakness of MLLMs on the considered visuospatial reasoning tasks, especially when compared to human performance. Although these models can often identify 3D arrangements or object correspondences, they struggle to integrate this perception into consistent reasoning, frequently selecting implausible answers (Fig 21 in Appendix). Models particularly struggle with tasks requiring interpreting temporal sequences and tracking visual elements across transformations. For example, *Dynamic Structural Correspondence* tests unidirectional tracking of changes across a sequence, while *Paper Folding* requires not only forward tracking but also mentally reversing the process; in both cases, models frequently misinterpret the visual dynamics. Scaling analysis (Fig 4) shows that performance generally increases with model size, reinforcing the role of model parameter scale; however, performance also increases when moving from abstraction-heavy tasks to transformation-oriented ones, underscoring persistent weaknesses in reasoning abilities based on mental manipulation. A likely root cause of these failures may be the models’ limited capability to *cognitively group figures and concepts* into coherent representations. We also observe a *strong dependence on surface perception*: in *Mental Composition*, models succeed when the unfolded net visually resembles a cube but fail when correct in-

ference requires mentally folding a shape into a nontrivial 3D structure (Fig 22 in Appendix). We note that while the specific tasks are different, related previous efforts (Huang et al., 2025b; Stogiannidis et al., 2025a; Urgan and Ari, 2025) also report similar performance numbers. Our studies with prompt variations corroborate our above observations (see Appendix D).

## 5 Analysis and Discussion

**Attention Alignment and Accuracy.** We analyze whether option-directed attention predicts reasoning success. For each item, we compute an *Option-Specific Attention Score* ( $OAS_{\text{correct}}$ ): the mean normalized attention mass directed toward the correct option’s spatial region during reasoning-token generation. Across 200 items (25 per task, stratified by difficulty),  $OAS_{\text{correct}}$  correlates positively with accuracy (point-biserial  $r_{pb} = 0.34, p < 0.001$ ). Yet even in the highest-attention quartile, accuracy remains well below human performance ( $>80\%$ ), indicating that attention alignment is necessary but not sufficient for correct reasoning. This dissociation is further supported by a paired analysis: on correct predictions, attention to the correct option exceeds attention to distractors (0.24 vs. 0.16;  $t(87) = 4.32, p < 0.001$ ), whereas incorrect predictions show no such preference (0.18 vs. 0.17;  $t(112) = 0.84, p = 0.40$ ). Thus, while models are able to localize the relevant information but fail to reason over it reliably (see Appendix B.2).

**Robustness Under Relative Attention Normalization.** A natural concern with the preceding analysis is that raw softmax attention may be confounded by register-token artifacts (Darcet et al., 2024), which can inflate diffuse background attention and obscure genuine spatial focus. To address this, we follow Zhang et al. (2025) and recompute all attention metrics using *relative attention*.

Reassuringly, the findings reported above not only hold but become somewhat sharper under this normalization. The point-biserial correlation between  $OAS_{\text{correct}}$  and accuracy increases from  $r_{pb} = 0.34$  to  $r_{pb} = 0.41$  ( $p < 0.001$ ), suggesting that relative attention provides a cleaner predictor once register noise is factored out. On correct predictions, the effect size for attention preference toward the correct option over distractors grows from Cohen’s  $d = 1.15$  to  $d = 1.41$  ( $t(86) = 5.89, p < 0.001$ ). On incorrect predictions, relative attention to the selected wrong

option remains statistically indistinguishable from attention to the correct option ( $p = 0.31$ ), confirming that the dissociation is not an artifact of noisy attention extraction. Mean Region-Aligned Attention (RAA) increases modestly from 0.18 to 0.21, yet even this improved grounding remains far below what would be needed for reliable reasoning and well below human performance ( $>80\%$ ).

This reinforces our earlier conclusion: the bottleneck appears to lie not in *where* models attend, but in their limited ability to perform the cognitive operations required to reason over correctly localized information.

Attention to relevant regions correlates with accuracy, yet localization alone proves insufficient: *models can identify where to look but not how to reason over what they see.*

### Reasoning Stability under Prompt Variations.

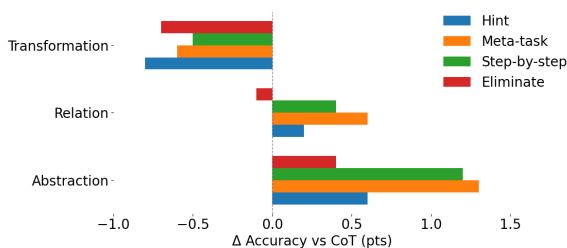


Figure 2: Change in accuracy ( $\Delta$  Accuracy) of different prompt variations w.r.t. CoT performance across ART dimensions

Figure 2 reports the differential effects of prompt variations on model performance when compared with baseline CoT performance. The results reveal that prompting effects are dimension-dependent rather than uniformly beneficial. Transformation tasks exhibit consistent performance degradation across all alternative prompting strategies, with Hint prompting showing the largest drop (approximately  $-0.9$  pts), suggesting that tasks requiring internal simulation are particularly sensitive to instruction framing. In contrast, Abstraction tasks benefit from structured guidance, with Meta-task and Step-by-step prompting yielding gains of approximately  $+1.3$  pts, indicating that explicit scaffolding facilitates latent rule derivation. Relation tasks show intermediate behavior, with modest improvements under Meta-task prompting but near-baseline performance otherwise. These asymmetric effects suggest that while

prompting can enhance pattern recognition and abstraction, it fails to induce the procedural operations underlying robust transformation reasoning (see Appendix B.4).

Prompting benefits abstraction but impairs transformation performance: *scaffolding facilitates rule derivation yet fails to elicit procedural visuocognitive operations.*

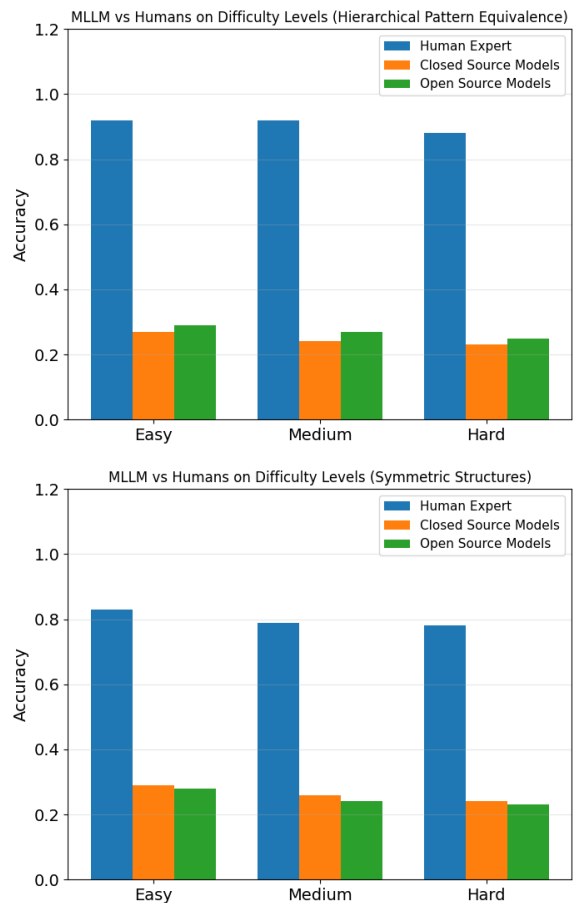


Figure 3: Human-model performance gap across ART taxonomy dimensions stratified by difficulty. Each bar represents the macro-average accuracy for a task across all models in that category (see Table 2).

**Performance Across ART dimension by Difficulty.** To examine how model capabilities vary across the three core dimensions of our ART taxonomy, we analyzed performance stratified by difficulty level (Easy, Medium, Hard). Figure 3 reveals a consistent pattern: human accuracy degrades predictably with difficulty (from  $> 0.80$  on Easy to  $0.25$  on Hard), while both open and closed-source models exhibit flat performance curves ( $0.20$ – $0.45$ ) regardless of difficulty. On cer-

tain hard instances, particularly Mental Transformation and Visual Relation Abstraction, human accuracy drops to model-level performance. However, this convergence is asymmetric: humans fail because the task is hard; models fail because they cannot perform the underlying operation at any difficulty level. This difficulty-invariant failure across all three dimensions suggests that current MLLMs lack foundational visual-cognitive operations beyond merely struggling with complex instances (refer Appendix K).

MLLMs fail uniformly across difficulty levels: *they lack foundational visuo-cognitive operations rather than merely struggling with complexity.*

## 6 Conclusions

We present MIND’S EYE, a visuo-cognitive benchmark for evaluating MLLMs on visual intelligence tasks, organized along three axes inspired by Carroll’s three-stratum theory: Abstraction, Relation, and Transformation. Our evaluation reveals a persistent human-model gap: non-expert humans achieve 80% mean accuracy while top MLLMs remain below 50%. Prompting strategies yield task-dependent but modest improvements without altering error profiles. Key findings include: (i) MLLMs rely heavily on perceptual cues with limited coupling between textual reasoning and visual evidence; (ii) scaling improves surface-matching tasks more than those requiring internal simulation; and (iii) our ART-aligned, parametric design exposes specific failure modes, suggesting directions for advances in grounded attention, spatial working memory, and transformation-aware representations. Future work includes open-ended responses, 3D perception tasks, and human studies across expertise levels.

## Limitations

As stated earlier, our Mind’s Eye benchmark focuses on using a *multiple-choice* scoring for reliability and objectivity of comparison; however, open-ended generation may bring about its own unique set of insights. Secondly, our tasks herein center on 2D renderings with controlled 3D implications; fully 3D inputs and interactions remain a focus of future work. Thirdly, our human baseline uses non-expert adults in a single language setting;

cross-lingual and expert cohorts may shift absolute levels of performance (we hypothesize though that relative gaps are likely to remain, based on our observations in this work).

**Threats to Validity.** *Construct validity.* While tasks target Abstraction/Relation/Transformation, they are proxies for broader visuo-cognition; we limit language priors but cannot eliminate all heuristics. *External validity.* Findings on synthetic, controlled items may not transfer to natural images; we release generators to enable domain shifts. *Reliability.* Item difficulty and distractor quality are controlled parametrically; bootstrap CI’s and mixed-effects models quantify uncertainty.

**Risks of Anthropomorphism.** Cognitive-style performance can invite anthropomorphic interpretations, for e.g., ascribing ‘mental rotation,’ ‘working memory,’ or ‘attention’ in the human sense to models. This risks conflating *functional* success on a narrowly specified task with *mechanistic* equivalence to human cognition. Over-interpretation can also invert causality: improvements from prompt engineering or data exposure may be mistaken for emergent cognitive faculties. To mitigate this, we treat model outputs as *behavioral signatures* under controlled stimuli, avoid mentalistic language, and separate construct-level claims (what is measured) from implementation claims (how models compute).

**Broader Impact/Ethics.** Our benchmark uses synthetic, knowledge-minimal stimuli designed to reduce privacy, content, and demographic risks; nevertheless, we report a few broader limitations that hold for almost all benchmarks. Firstly, publishing leaderboards may encourage narrow optimization, masking real-world limitations in safety-critical contexts (education, assessment, medical imaging). Secondly, human baselines reflect a specific population (age, language, interface); results should not be used to rank individuals or groups. Thirdly, cognitive-style tests could be misapplied as gatekeeping tools in hiring or education; our license and documentation explicitly prohibit human evaluation or selection use. We release generators, seeds, and scoring code to enable transparent replication and stress-testing, and we encourage researchers to report uncertainty, disclose inference settings, and evaluate interventions (e.g., grounded-attention or working-

memory modules) for safety as well as performance.

## References

- Eshaan Agarwal, Joykirat Singh, Vivek Dani, Raghav Magazine, Tanuja Ganu, and Akshay Nambi. 2024. [Promptwizard: Task-aware prompt optimization framework](#). *arXiv preprint arXiv:2405.18369*.
- Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. 2025. [Thought anchors: Which llm reasoning steps matter?](#) *Preprint*, arXiv:2506.19143.
- Lois Burton. 2003. [Examining the relation between visual imagery and spatial ability tests](#). *International Journal of Testing*, 3(3):277–291.
- John B. Carroll. 1993. *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.
- Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. 2025. [Arc-agi-2: A new challenge for frontier ai reasoning systems](#). *Preprint*, arXiv:2505.11831.
- François Chollet. 2019. [On the measure of intelligence](#). *arXiv preprint arXiv:1911.01547*.
- Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. 2024. [Vision transformers need registers](#). In *The Twelfth International Conference on Learning Representations*.
- Paul De Boeck and Mark Wilson. 2003. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer.
- Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. 2024. [Vlmevalkit: An open-source toolkit for evaluating large multi-modality models](#). In *Proceedings of the 32nd ACM International Conference on Multimedia, MM '24*, page 11198–11201, New York, NY, USA. Association for Computing Machinery.
- Ruth B. Ekstrom, John W. French, and Harry H. Harman. 1976a. [Manual for kit of factor-referenced cognitive tests](#).
- Ruth B. Ekstrom, John W. French, Harry H. Harman, and Diran Dermen. 1976b. *Kit of Factor-Referenced Cognitive Tests*. Educational Testing Service, Princeton, NJ.
- Susan E. Embretson. 1983. [Construct validity: Construct representation versus nomothetic span](#). *Psychological Bulletin*, 93(1):179–197.
- Susan E. Embretson and Steven P. Reise. 2013. *Item Response Theory for Psychologists*. Psychology Press.
- Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. 2025. [Grit: Teaching mllms to think with images](#). *arXiv preprint arXiv:2505.15879*.

- François Fleuret, Tingting Li, Charles Dubout, Eric K. Wampller, Steven Yantis, and Donald Geman. 2011. Comparing machines and humans on a visual categorization test. *PNAS*, 108(43):17621–17625.
- Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. [Blink: Multimodal large language models can see but not perceive](#). In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXIII*, page 148–166, Berlin, Heidelberg. Springer-Verlag.
- Deepanway Ghosal, Vernon Toh, Yew Ken Chia, and Soujanya Poria. 2025. [AlgoPuzzleVQA: Diagnosing multimodal reasoning challenges of language models with algorithmic multimodal puzzles](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9615–9632, Albuquerque, New Mexico. Association for Computational Linguistics.
- J. P. Guilford and Wayne S. Zimmerman. 1948. [The guilford-zimmerman aptitude survey](#). *Journal of Applied Psychology*, 32:24–34.
- Graeme S. Halford, William H. Wilson, and Steven Phillips. 2010. Relational complexity and reasoning. *Cognitive Science*, 34(8):1451–1476.
- Douglas R Hofstadter. 1979. *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books.
- Jen-Tse Huang, Dasen Dai, Jen-Yuan Huang, Youliang Yuan, Xiaoyuan Liu, Wenxuan Wang, Wenxiang Jiao, Pinjia He, Zhaopeng Tu, and Haodong Duan. 2025a. [Human cognitive benchmarks reveal foundational visual gaps in mllms](#). *Preprint*, arXiv:2502.16435.
- Jen-Tse Huang, Dasen Dai, Jen-Yuan Huang, Youliang Yuan, Xiaoyuan Liu, Wenxuan Wang, Wenxiang Jiao, Pinjia He, Zhaopeng Tu, and Haodong Duan. 2025b. [Human cognitive benchmarks reveal foundational visual gaps in mllms](#). *Preprint*, arXiv:2502.16435.
- Drew A Hudson and Christopher D Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*.
- Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*.
- Aditya Kanade and Tanuja Ganu. 2025. [Do you see me : A multidimensional benchmark for evaluating visual perception in multimodal llms](#). *Preprint*, arXiv:2506.02022.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA. Curran Associates Inc.
- Maithilee Kunda, Keith McGregor, and Ashok Goel. 2012. Reasoning on the raven’s advanced progressive matrices test with iconic visual representations. In *Proceedings of the Cognitive Science Society*, Sapporo, Japan.
- Michael Lan, Philip Torr, and Fazl Barez. 2024. [Towards interpretable sequence continuation: Analyzing shared circuits in large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12576–12601, Miami, Florida, USA. Association for Computational Linguistics.
- J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2023a. [Seed-bench-2: Benchmarking multimodal large language models](#).
- Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13299–13308.
- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023b. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven CH Hoi. 2023c. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024a. [Mmbench: Is your multi-modal model an all-around player?](#) In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VI*, page 216–233, Berlin, Heidelberg. Springer-Verlag.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024b. [Mmbench: Is your multi-modal model an all-around player?](#) In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VI*, page 216–233, Berlin, Heidelberg. Springer-Verlag.

- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024a. *Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts*. In *International Conference on Learning Representations (ICLR)*.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024b. *Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts*. *Preprint*, arXiv:2310.02255.
- Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. *Learn to explain: Multimodal reasoning via thought chains for science question answering*. *Preprint*, arXiv:2209.09513.
- Suzanne A. McFall, Jeanette C. Deitz, and Terry K. Crowe. 1993. *Test-retest reliability of the test of visual perceptual skills with children with learning disabilities*. *American Journal of Occupational Therapy*, 47(9):819–824.
- Kevin S McGrew. 2005. *The Cattell-Horn-Carroll theory of cognitive abilities*. Springer.
- Kevin C. Moen, Michael R. Beck, Samuel M. Saltzman, and 1 others. 2020. *Strengthening spatial reasoning: elucidating the attentional and neural mechanisms associated with mental rotation skill development*. *Cognitive Research: Principles and Implications*, 5(1):20.
- Allen Newell. 1994. *Unified theories of cognition*. Harvard University Press.
- Weili Nie, Zhiding Yu, Lei Mao, Ankit B. Patel, Yuke Zhu, and Animashree Anandkumar. 2020a. *Bongard-logo: a new benchmark for human-level concept learning and reasoning*. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Red Hook, NY, USA. Curran Associates Inc.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020b. *Adversarial nli: A new benchmark for natural language understanding*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901.
- Pranshu Pandya, Vatsal Gupta, Agney S Talwarr, Tushar Kataria, Dan Roth, and Vivek Gupta. 2025. *NTSEBENCH: Cognitive reasoning benchmark for vision language models*. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3680–3708, Albuquerque, New Mexico. Association for Computational Linguistics.
- Tan-Hanh Pham, Phu-Vinh Nguyen, Dang The Hung, Bui Trong Duong, Vu Nguyen Thanh, Chris Ngo, Tri Quang Truong, and Truong-Son Hy. 2025. *Iqbench: How "smart" are vision-language models? a study with human iq tests*. *Preprint*, arXiv:2505.12000.
- Riverside Publishing. 2009. *Cognitive Abilities Test (CogAT) Form 7, Paper Folding Items*. Riverside. Psychometric test battery.
- Achyuta Rajaram, Neil Chowdhury, Antonio Torralba, Jacob Andreas, and Sarah Schwettmann. 2024. *Automatic discovery of visual circuits*.
- John Raven. 2000. *Raven's progressive matrices*. *Handbook of Nonverbal Assessment*, pages 223–237.
- Yufan Ren, Konstantinos Tertikas, Shalini Maiti, Junlin Han, Tong Zhang, Sabine Süsstrunk, and Filippos Kokkinos. 2025. *Vgrp-bench: Visual grid reasoning puzzle benchmark for large vision-language models*. *Preprint*, arXiv:2503.23064.
- W. Joel Schneider and Kevin S. McGrew. 2018. *Intelligence in education: Cattell-horn-carroll theory and assessment*. *Psychology in the Schools*, 55(1):7–43.
- Alessandro Serra, Francesco Ortu, Emanuele Panizon, Lucrezia Valeriani, Lorenzo Basile, Alessio Ansuini, Diego Doimo, and Alberto Cazzaniga. 2025. *The narrow gate: Localized image-text communication in native multimodal models*. *Preprint*, arXiv:2412.06646.
- Roger N Shepard and Jacqueline Metzler. 1971. *Mental rotation of three-dimensional objects*. *Science*.
- Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. 2025. *Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge*. *Preprint*, arXiv:2504.10342.
- Ilias Stogiannidis, Steven McDonagh, and Sotirios A Tsaftaris. 2024. *Mind the Gap: Benchmarking spatial reasoning in vision-language models*. *arXiv preprint arXiv:2403.19707*.
- Ilias Stogiannidis, Steven McDonagh, and Sotirios A. Tsaftaris. 2025a. *Mind the gap: Benchmarking spatial reasoning in vision-language models*. *Preprint*, arXiv:2503.19707.
- Ilias Stogiannidis, Steven McDonagh, and Sotirios A. Tsaftaris. 2025b. *Mind the gap: Benchmarking spatial reasoning in vision-language models*. *arXiv preprint arXiv:2503.19707*.
- Alane Suhr, Adam Trischler, Jackie Chi Kit Cheung, and Yoav Artzi. 2019. *A corpus for reasoning about natural language grounded in photographs*. In *ACL*.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others.

2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Louis Leon Thurstone. 1938. *Primary Mental Abilities*. University of Chicago Press, Chicago.
- Sinan Urgan and Seçkin Ari. 2025. [An analysis of architectural impact on llm-based abstract visual reasoning: A systematic benchmark on raven-fair](#). *Preprint*, arXiv:2511.11916.
- Steven G. Vandenberg and Allan R. Kuse. 1978a. [Mental rotations, a group test of three-dimensional spatial visualization](#). *Perceptual and Motor Skills*, 47(2):599–604.
- Steven G Vandenberg and Allan R Kuse. 1978b. [Mental rotations, a group test of three-dimensional spatial visualization](#). *Perceptual and Motor Skills*, 47(2):599–604.
- Steven G. Vandenberg and Allan R. Kuse. 1978c. [Mental rotations, a group test of three-dimensional spatial visualization](#). *Perceptual and Motor Skills*, 47(2):599–604.
- Siting Wang, Minnan Pei, Luoyang Sun, Cheng Deng, Kun Shao, Zheng Tian, Haifeng Zhang, and Jun Wang. 2025. [Spatialviz-bench: An mllm benchmark for spatial visualization](#). *Preprint*, arXiv:2507.07610.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA. Curran Associates Inc.
- Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhai Wang, Jifeng Dai, and Jinguo Zhu. 2025a. [Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models](#). *Preprint*, arXiv:2504.15279.
- Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhai Wang, Jifeng Dai, and Jinguo Zhu. 2025b. [Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models](#).
- Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. [Mm-vet: evaluating large multimodal models for integrated capabilities](#). In *Proceedings of the 41st International Conference on Machine Learning, ICML'24*. JMLR.org.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. [Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi](#). In *Proceedings of CVPR*.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [Hellaswag: Can a machine really finish your sentence?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.
- Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. 2019. [Raven: A dataset for relational and analogical visual reasoning](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. 2025. [MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs](#). In *The Thirteenth International Conference on Learning Representations*.
- Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. 2024. [Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?](#) *Preprint*, arXiv:2403.14624.

## Appendix

We present the following additional results and discussions which we could not include in the main paper owing to space constraints:

- A Additional Results
  - B Extended Analysis
  - C More about MIND’S EYE
  - D Detailed Benchmark Design
  - E Evaluation Setup Details
  - F Human Evaluation Protocol
  - G Prompting Strategies and Styles
  - H Full Prompt Templates
  - I Analysis of CoT Reasoning Quality
  - J Carroll’s Three-Stratum Theory of Fluid Intelligence
  - J Performance comparison of humans and models across cognitive subtasks by difficulty level.
- ### A Additional Results

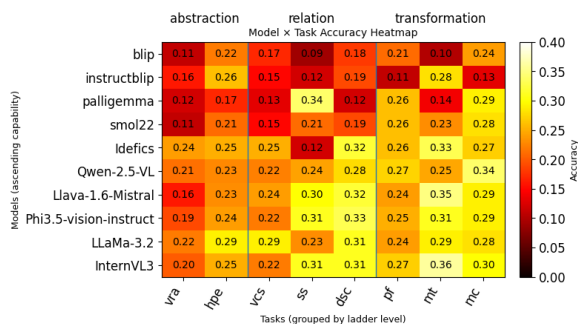


Figure 4: **Model performance across all tasks:** Heatmap of model performance across tasks, with rows denoting models and columns denoting tasks (color intensity represents accuracy). Models ordered by increasing capability (top to bottom), and tasks grouped by ART, revealing that even top-tier models have significant/varied weaknesses.

**Benchmark Distribution** Our benchmark consists of eight tasks: Dynamic Structural Correspondence, Hierarchical Pattern Equivalence, Mental Composition, Mental Transformation, Paper Folding, Visual Conceptual Slippage, Symmetric Structures, and Visual Relation Abstraction. These tasks are grouped into three categories

that align with core dimensions of fluid intelligence: Pattern Abstraction (Visual Relation Abstraction, Hierarchical Pattern Equivalence), Relation (Dynamic Structural Correspondence, Visual Conceptual Slippage, Symmetric Structures), and Transformation (Mental Transformation, Paper Folding, Mental Composition). Each task is programmatically generated <sup>2</sup>, allowing precise control over difficulty by varying task specific parameters. Figure 6 illustrates the dataset distribution employed for probing and evaluating model performance.

We compare overall benchmark performance against model size (in billions of parameters) across a diverse set of multimodal models in Figure 7b. Interestingly, performance does not scale monotonically with size: some medium scale models (e.g., InternVL3, LLaMA-3.2, Phi-3.5) achieve competitive or even superior performance relative to much larger counterparts, while smaller models (e.g., BLIP, InstructBLIP, PaliGemma) consistently underperform. Notably, Qwen-2.5-VL exhibits strong performance at both small and large scales, suggesting architectural and training choices play a larger role than raw parameter count. A correlation analysis confirms this observation, with Pearson’s  $r \approx 0.62$ , indicating only a moderately positive relationship between model size and benchmark performance. Taken together, these results highlight that *scaling* yields non uniform gains across our tasks, suggesting that parameter growth alone may not suffice under this benchmark, and that improved training and architecture could be equally important.

**Prompting strategies performance deltas :** We compare the effect of four prompting strategies (*eliminate*, *hint*, *meta-task*, *step-by-step*) against chain-of-thought (CoT) across the eight benchmark tasks in Figure 8. Each subplot shows the relative score change, where values to the left of the vertical dashed line indicate deterioration and values to the right indicate improvement. The results reveal a heterogeneous landscape:

- **Consistent improvements:** Tasks such as *Hierarchical Pattern Equivalence* and *Visual Relation* benefit from meta-task, eliminate and step-by-step prompts, which appear to help models engage in multistep reasoning.
- **Mixed or task dependent effects:** Tasks like

<sup>2</sup>Code: [https://anonymous.4open.science/r/Minds\\_Eye-0801/](https://anonymous.4open.science/r/Minds_Eye-0801/)

*Dynamic Structural Correspondence*, *Visual Conceptual Slippage* and *Paper Folding* show both gains and regressions depending on the prompting strategy.

- **Clear deterioration:** Cognition heavy tasks such as *Mental Transformation* and *Mental Composition* exhibit consistent performance drops across most prompting strategies relative to CoT.
- **Instability of eliminate and hint:** These strategies occasionally yield benefits, but more often result in deterioration across tasks.

Overall, the figure highlights that while prompting can produce gains in reasoning intensive tasks, it can also worsen performance in many tasks, underscoring the lack of a universally beneficial prompting strategy.

Figure 7a shows the performance of Qwen-2.5-VL models of different sizes (3B, 7B, 32B) on the ART benchmark under chain-of-thought (CoT) prompting. Several patterns emerge. First, scaling does not yield uniform improvements across tasks: while the 32B variant outperforms the smaller models on conceptual relation and transformation heavy tasks such as *Mental Transformation*, *Mental Composition* and *Paper Folding*, the smaller 3B and 7B variants remain competitive or superior on temporal relation and abstraction oriented tasks like *Dynamic Structural Correspondence* and *Hierarchical Pattern Equivalence*. This reinforces that scale alone is insufficient to overcome reasoning deficits, and that certain tasks demand structured cognitive mechanisms rather than brute force capacity. Second, *Mental Composition* is a particularly challenging task, where both 7B and 32B improve substantially over 3B, yet overall accuracy remains low, reflecting the persistent difficulty of compositional reasoning. Finally, we note a strong dependency on perceptual similarity in some tasks: while larger models exploit surface level cues more effectively (in *Paper Folding*, *Symmetric Structures*, *Visual Conceptual Slippage*, *Visual Relational Abstraction*), they continue to fail when success requires internal simulation of transformations.

Taken together, these results highlight that while larger models can achieve gains in perception heavy reasoning tasks, smaller models sometimes generalize better in abstraction oriented set-

tings. This suggests that scaling amplifies perceptual alignment but does not induce the higher level grouping or cognitive mechanisms required for robust visuo-cognitive reasoning.

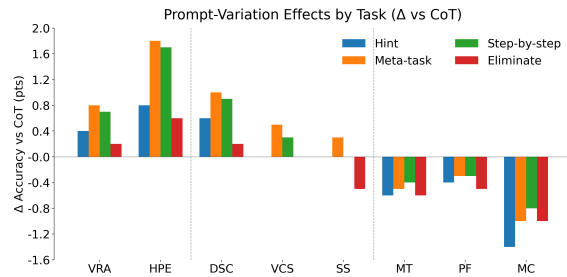


Figure 9: **Differential Effects of prompting across ART dimensions:** Average accuracy of models relative to CoT performance aggregated by Abstraction, Relation, and Transformation.

Figure 9 presents a task-level analysis of prompt variation effects relative to chain-of-thought (CoT), revealing substantial heterogeneity within each ART dimension. Within Abstraction, *Hierarchical Pattern Equivalence* (HPE) is the most prompt-sensitive task, with meta-task and step-by-step prompting yielding gains of approximately +1.8 pts and +1.7 pts, respectively, while *Visual Relation Abstraction* (VRA) shows smaller improvements (+0.4–0.8 pts) and remains substantially less sensitive to prompt framing. Relation tasks exhibit the greatest internal variability: *Dynamic Structural Correspondence* (DSC) benefits from structured prompting, achieving gains of up to +1.0 pt under meta-task prompting, whereas *Visual Conceptual Slippage* (VCS) remains near-invariant ( $\leq +0.3$  pts across all prompts), and *Symmetric Structures* (SS) shows mixed behavior, with a modest gain under meta-task prompting ( $\approx +0.3$  pts) but degradation under elimination-based prompting ( $\approx -0.5$  pts). In contrast, Transformation tasks are uniformly brittle to prompt variation. *Mental Composition* (MC) exhibits the largest and most consistent drops across all prompting strategies (-0.8 to -1.4 pts), followed by *Mental Transformation* (MT) ( $\approx -0.4$  to -0.6 pts), while *Paper Folding* (PF) shows comparatively smaller declines ( $\approx -0.3$  to -0.5 pts). Importantly, no transformation task shows systematic improvement under any alternative prompting strategy. These task-level dissociations indicate that prompting primarily benefits tasks requiring explicit symbolic rule induction (e.g., HPE, DSC), while consistently

disrupting tasks that depend on multi-step internal visual simulation (e.g., MC, MT), reinforcing that prompt engineering modulates surface reasoning behavior but does not address the underlying transformation bottleneck identified by the ART framework.

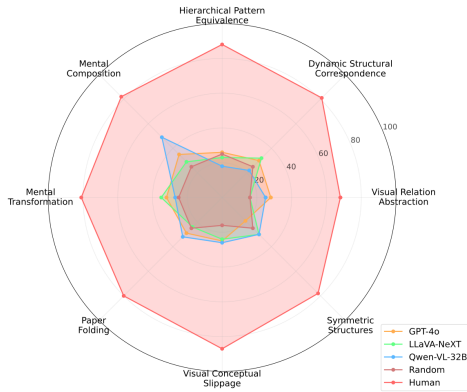


Figure 10: Accuracies of multimodal LLMs on MIND’S EYE Benchmark. Please refer to 2 for more results and discussions.

**Effect of Image Resolution.** A natural question is whether the performance gap we observe could be attributed to visual quality rather than reasoning limitations. To test this, we varied image resolution between 100 DPI ( $600 \times 800$  px) and 300 DPI ( $1024 \times 1024$  px) and evaluated Qwen-2.5-VL-7B across all eight tasks. As shown in Table 3, no statistically significant difference was observed at either resolution, suggesting that image quality is not a bottleneck for model performance on our benchmark. We note that all stimuli provided to models during evaluation are rendered as SVGs and exported at  $1024 \times 1024$  px at 300 DPI, ensuring that option labels and geometric details are fully legible at inference time.

## B Extended Analysis

### B.1 Attention Maps

**Attention Heatmap Analysis.** To further probe the internal reasoning failures, we analyzed the attention heatmaps. Surprisingly, when tokens explicitly referred to specific figures or options (e.g., “option A”, “shape”, “comparing”), the model’s attention was not concentrated on the corresponding visual regions. Instead, attention was diffusely spread across background areas or unrelated parts of the image. Across sampled instances ( $N = 20$  for each tasks, where  $N$  is the sampled CoT Traces), less than 20% of the model’s normalized

attention mass was directed toward the objects explicitly referenced in the reasoning trace. This failure mode highlights a limitation: although the model produces fluent chain-of-thought reasoning, the underlying attention does not ground the reasoning process in the visual input. In other words, query specific tokens fail to anchor attention to the corresponding figures, undermining the fidelity of the reasoning process. These findings reinforce our broader conclusion that current models may rely more on linguistic priors than grounded visual attention when tackling cognitive reasoning tasks.

**Grounded Attention Metric.** We quantify grounding with *Region Aligned Attention* (RAA):  $RAA = \frac{1}{T} \sum_{t=1}^T \sum_{p \in \mathcal{R}(t)} a_t(p)$ , where  $a_t(p)$  is the normalized attention over pixels (or patches) at token  $t$ , and  $\mathcal{R}(t)$  is the union of regions referenced by token  $t$  (e.g., “option A”, “shape”, “compare”). We evaluate RAA on  $N=20$  tokens for which perception dependence is high (e.g. Option A, shape, color) across 20 items sampled stratified by task; mean RAA is **0.18**, corroborating the qualitative observation that the attention of query specific tokens often fails to align precisely with the corresponding figures.

### B.2 Attention Performance Correlation Analysis

**Option-Specific Attention Score (OAS)** For each item  $i$  with  $K$  options  $\{O_1, \dots, O_K\}$ , we define the *Option-Specific Attention Score* (OAS) as the average normalized attention mass allocated to an option’s spatial region during reasoning token generation. Let  $a_t(p)$  denote normalized attention at reasoning token  $t$  to visual patch  $p$ , with  $\sum_p a_t(p) = 1$ . Let  $\mathcal{T}_{\text{reason}}$  denote reasoning tokens and  $\mathcal{R}_k$  the patch set corresponding to option  $O_k$ . Then:

$$OAS_k(i) = \frac{1}{|\mathcal{T}_{\text{reason}}|} \sum_{t \in \mathcal{T}_{\text{reason}}} \sum_{p \in \mathcal{R}_k} a_t(p). \quad (1)$$

This metric quantifies the average proportion of attention allocated to option  $k$  across all reasoning tokens. For our analysis, we compute three variants: We analyze  $OAS_{\text{correct}}$ ,  $OAS_{\text{selected}}$ , and mean  $OAS_{\text{distractors}}$ .

**Correlation and Trend Analysis** Let  $y_i \in \{0, 1\}$  denote correctness. We compute the point-biserial correlation between  $OAS_{\text{correct}}$  and  $y_i$ :

$$r_{pb} = \frac{\bar{x}_1 - \bar{x}_0}{s_x} \sqrt{\frac{n_1 n_0}{n(n-1)}}, \quad (2)$$

Table 3: Resolution ablation on Qwen-2.5-VL-7B. No statistically significant difference is observed across the two settings.

Resolution	VRA	HPE	DSC	VCS	SS	MT	PF	MC
100 DPI / 600×800	18.7±0.02	24.4±0.05	30.1±0.08	22.1±0.02	20.7±0.05	25.2±0.30	24.8±0.10	36.1±0.40
300 DPI / 1024×1024	19.1±0.01	24.2±0.01	30.4±0.01	22.7±0.01	20.2±0.04	25.7±0.02	25.1±0.02	36.4±0.01

where:

- $\bar{x}_1 = \frac{1}{n_1} \sum_{i:y_i=1} \text{OAS}_{\text{correct}}(i)$  is the mean OAS for correct predictions
- $\bar{x}_0 = \frac{1}{n_0} \sum_{i:y_i=0} \text{OAS}_{\text{correct}}(i)$  is the mean OAS for incorrect predictions
- $s_x$  is the standard deviation of OAS across all items
- $n_1 = \sum_i y_i$  and  $n_0 = n - n_1$  are the counts of correct and incorrect predictions
- $n$  is the total number of items analyzed

**Paired Attention Comparisons** For correct predictions, we test whether attention favors correct options over distractors:

$$H_0 : \mathbb{E}[\text{OAS}_{\text{correct}} - \text{OAS}_{\text{distractors}}] = 0, \quad (3)$$

using paired  $t$ -tests. For incorrect predictions, we compare  $\text{OAS}_{\text{selected}}$  against  $\text{OAS}_{\text{correct}}$  to assess attention misallocation.

**Implementation and Sampling** Cross-attention weights are extracted from the final decoder layer and averaged across heads. Option regions are defined via fixed or programmatic bounding boxes depending on task layout. Analyses are conducted on  $N = 200$  items (25 per task).

Analysis	Statistic	Result
Point-biserial correlation	$r_{pb}$	0.34 ( $p < 0.001$ )
<i>Paired Comparisons (Correct Predictions, <math>n = 87</math>)</i>		
$\text{OAS}_{\text{correct}}$	Mean	0.24±0.08
$\text{OAS}_{\text{distractors}}$	Mean	0.16±0.06
Paired t-test	$t(86)$	4.32 ( $p < 0.001$ )
<i>Paired Comparisons (Incorrect Predictions, <math>n = 113</math>)</i>		
$\text{OAS}_{\text{selected}}$	Mean	0.18±0.07
$\text{OAS}_{\text{correct}}$	Mean	0.17±0.07
Paired t-test	$t(112)$	0.84 ( $p = 0.40$ )

Table 4: Statistical results from attention-performance correlation analysis. All p-values are two-tailed except where noted.

### B.3 Interpretation

The positive point-biserial correlation ( $r_{pb} = 0.34$ ) provide convergent evidence that attention alignment to correct options is a significant predictor of task performance. However, the modest effect size and low absolute accuracy even in the highest attention quartile (35.7% vs. > 80% human performance) indicate that attention is *necessary but insufficient* for correct reasoning.

The paired comparison results reveal a critical asymmetry: when models answer correctly, they allocate significantly more attention to correct options than distractors (Cohen’s  $d = 1.15$ , large effect). However, when models err, their attention to the selected (incorrect) option is statistically indistinguishable from attention to the correct option ( $p = 0.40$ ), suggesting that errors arise from *attention misallocation* rather than systematic biases away from correct answers. This pattern is consistent with a model that lacks robust visual grounding: it attends to plausible options without the cognitive mechanisms to reliably distinguish correctness from perceptual similarity.

## B.4 Pairwise Analysis of CoT vs Non CoT of Same Model Family

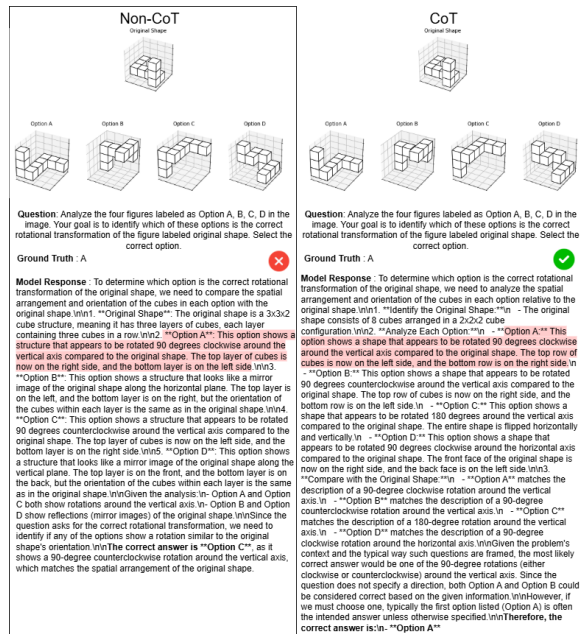


Figure 12: Reasoning with prompt variations: (Left) Response with Non-CoT Prompt, (Right) Response with CoT Prompt, for the same image-question pair. The highlighted section of the reasoning traces shows the superficial shift in reasoning of an option, without any principled justification.

Our analysis of reasoning traces (Figure 12) highlights two critical failure modes. First, we observe a clear perception error: the models often misinterpret the underlying structure of the figure. For instance, when reasoning about a shape, the model incorrectly encodes it as a  $3 \times 2 \times 2$  cube, indicating persistent misperception of visual structure. Second, we find a systematic instability in reasoning: altering the prompt does not induce substantive changes in the underlying reasoning process, but instead produces superficial shifts in response orientation. The traces provide no principled justification for why the answer changes, suggesting limited visuo-cognitive grounding and inconsistent reasoning explanations.

## B.5 Similar Answer Selection Propensity

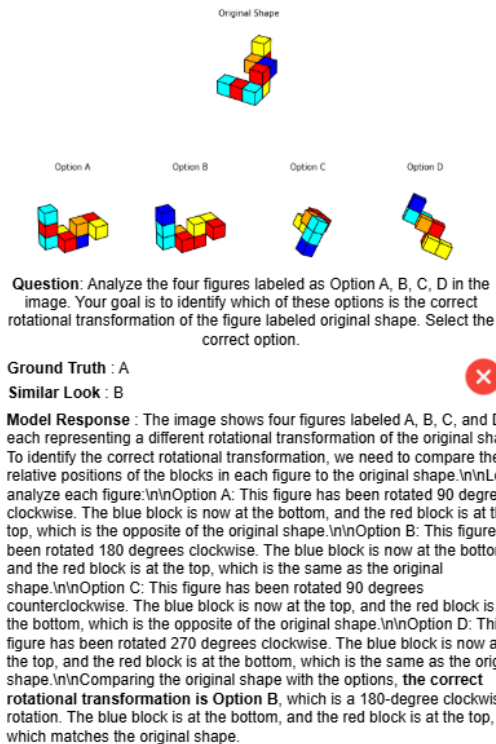


Figure 13: Similar Answer Selection reasoning : The figure illustrates cases where a distractor option closely resembles the correct answer. The model doesn't perform the necessary multistep reasoning and final disambiguation. The bold text shows the final answer selection

Models	Mental Transformation			Paper Folding		
	Correct	Similar	Incorrect	Correct	Similar	Incorrect
InternVL2.5	0.20	0.32	0.48	0.21	0.24	0.50
InternVL3.5	0.34	0.30	0.36	0.18	0.34	0.48
Qwen2.5-VL-7B	0.34	0.18	0.48	0.18	0.24	0.58
Qwen2.5-VL-3B	0.28	0.24	0.48	0.25	0.25	0.50
Qwen2.5-VL-32B	0.36	0.24	0.40	0.31	0.22	0.47
LLaVa	0.26	0.20	0.54	0.15	0.20	0.65
Idefics	0.28	0.22	0.50	0.30	0.25	0.45

Table 5: Propensity for Similar Answer Selection : The table reports the proportion of times the model selects the correct, incorrect, and visually similar (distractor) options for each task.

In tasks where one of the distractor options closely resembles the correct answer, successful solving requires multi-step reasoning to disambiguate between the two. As shown in the figure 13, the model seldom engages in such multi-step reasoning and final disambiguation step and instead falls to the wrong option uniformly as shown in Section B.8, ultimately leading to systematic errors.

## B.6 Domain Knowledge Dependence

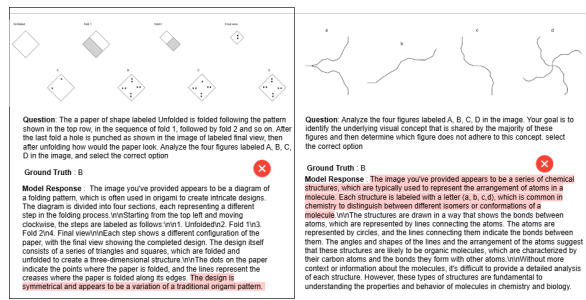


Figure 14: **Concept Misunderstanding** : This shows two examples of model failure due to applying inappropriate domain knowledge. The model misinterprets abstract symmetric structures as "chains of molecules" instead of reasoning about their geometric properties and misinterprets paper folding as a task related to Origami, leading to an incorrect conclusion.

For many tasks, the models try to retrieve an answer from its domain of knowledge, which leads to error in understanding the underlying concepts of perception. Like for symmetric structures, it infers them as chain of molecules rather than trying to understand their underlying concepts, Fig 14. Models shows poor performance in understanding the underlying concept due to their heavy reliance on domain knowledge based interpretation. Like for symmetric figures, it thinks these are chains of molecules.

## B.7 Causal Intervention

Causal intervention for circuit discovery in MLLM entails selectively ablating model components—such as attention heads, residual streams, or activations—to assess their functional role in a target task (Serra et al., 2025; Rajaram et al., 2024; Lan et al., 2024). The primary objective of knockout based intervention is to causally identify subnetworks or circuits within the language tower that are responsible for specific behaviors or multimodal communication, by observing the disruption of model outputs when these components are ablated (Rajaram et al., 2024; Lan et al., 2024; Serra et al., 2025).

Recent works have demonstrated the utility of knockout interventions for mechanistic discovery in MLLM. For instance, Serra et al. (2025) employs attention knockout to localize circuits mediating image-to-text transfer. Similarly, Rajaram et al. (2024) utilizes cross layer attribution followed by activation knockout to validate discovered circuits, and Lan et al. (2024) investigates the

causal impact of ablation on shared subnetworks. These methodologies collectively establish knockout intervention as a central paradigm for causal interpretability in MLLM.

Motivated by these techniques, we performed knockout interventions aligned with current methodology to determine whether circuits exist for a specific task within Qwen-7B. In contrast to prior findings, our knockout experiments did not reveal any functional circuit whose ablation affected model performance on the tested task. To verify this observation, we performed similar intervention across intra-family (Qwen-3B, Qwen-7B) and across models (Qwen-7b and LLaVa-7B) for all the tasks. This negative result suggests that, for the task investigated, no distinct causal circuits could be isolated within the model using this approach.

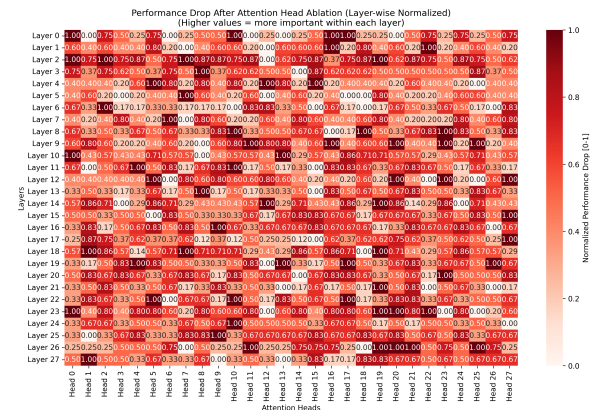


Figure 15: **Performance Variation from Attention Head Knockout**: A heatmap from a causal intervention experiment on the Qwen-7B model indicates that disabling individual attention heads did not cause a significant performance drop, suggesting the model lacks a specific, localized circuit for the Mental Composition task.

## B.8 Model Consensus: Role of Distractors

A key concern in the evaluation of these tasks is whether models are genuinely failing to reason about the underlying transformation, or merely being misled by distractors that resemble the correct answer.

distribution of incorrect responses with respect to the designated distractor option. For the **Mental Transformation (MT)** subset, The chi-squared test yielded  $\chi^2 = 0.59$ ,  $p = 0.44$ . For the **Paper Folding (PF)** subset, The chi-squared statistic was  $\chi^2 = 0.42$ ,  $p = 0.52$ .

**Human Performance.** To validate the distractor effects on human performance, we performed the same  $\chi^2$  goodness-of-fit test on the distribution of incorrect responses with respect to the designated distractor option. For the **Mental Rotation Test (MRT)** subset, The chi-squared test yielded  $\chi^2 = 3.21$ ,  $p = 0.073$ . For the **Paper Folding (PF)** subset, The chi-squared statistic was  $\chi^2 = 6.12$ ,  $p = 0.014$ . This indicates a trend toward distractor concentration among the wrong options.

In both cases, the  $p$ -values are far above the conventional significance threshold ( $\alpha = 0.05$ ), meaning we *fail to reject the null hypothesis*. This indicates that models are not disproportionately attracted to the annotated distractors. Instead, their errors appear uniformly spread across all incorrect alternatives.

**(ii) Per-item exact tests.** To check whether any individual question disproportionately attracted errors to its distractor, we ran an *exact binomial test* per item. This test compares the observed fraction of distractor errors against the null 1/3 baseline. A small  $p$ -value would indicate that, for that item, models were systematically biased toward (or away from) the distractor.

In both tasks, while a few items reached uncorrected  $p < 0.05$ , none survived Holm-Bonferroni correction for multiple comparisons (PF: 5/50 uncorrected, 0/50 corrected; MRT: 3/25 uncorrected, 0/25 corrected). Thus, no item showed a reliable per-item distractor effect.

**(iii) Mixed-effects logistic regression with clustered inference.** Because responses to the same item are not independent, we fit a logistic generalized estimating equation (GEE) with clustering by question. This model tests whether the log-odds of choosing the distractor differ from the null logit (1/3) baseline, while accounting for within question correlation.

If the intercept were significantly different from zero, it would indicate a systematic shift toward (or away from) distractors across all items. In practice, both tasks showed non significant results, confirming the absence of such a bias.

Task	$\hat{\beta}$	SE	OR vs. null	$p$
MRT	0.156	0.266	1.17	0.556
PF	0.088	0.170	1.09	0.605

Both tasks show intercepts not different from zero; the estimated odds ratios relative to the null (1.00) are close to 1 and non significant.

Across pooled  $\chi^2$  tests, per-item exact tests with multiplicity control, and a mixed effects logistic model that accounts for within item dependence, we find **no statistical evidence** that models are disproportionately choosing the annotated distractors. Their wrong answers are distributed roughly uniformly across all incorrect options. These results suggest that models are not simply confused by visually similar foils, but may instead reflect a broader challenge in integration cognition with perception.

## B.9 Thought Anchors CoT Annotation Analysis

Adopting the [Bogdan et al. \(2025\)](#) framework, we categorized sentences in chain-of-thought traces into six reasoning stages. Our fine-grained analysis of Qwen-7B on the Mental Transformation task highlights that current MLLM lack the mechanisms required for genuine spatial reasoning. Despite generating detailed chain-of-thought (CoT) traces, the model consistently failed to align its reasoning with perceptual evidence. To systematically analyze the internal reasoning structure of the model, we adopt the framework in [Bogdan et al. \(2025\)](#) to automatically labeled each sentence in the chain-of-thought (CoT) trace into one of six categories using an LLM based auto labeling procedure:

- 1. problem\_setup:** Parsing or rephrasing the problem statement, often reflecting initial comprehension.
- 2. plan\_generation:** Stating or deciding on a plan of action (e.g., outlining steps of reasoning).
- 3. option\_analysis:** Analyzing a specific option (A, B, C, or D) in detail with supporting reasoning.
- 4. final\_answer\_emission:** Explicitly providing the predicted answer or sentences directly leading to the answer.

5. **self\_checking**: Verifying prior steps, double checking logic, or expressing re-confirmation of reasoning.
6. **unknown**: Reserved for sentences that do not fit any of the above, including purely stylistic or filler expressions.

This categorization enables us to disentangle where failures occur within the reasoning process, whether at the stage of problem comprehension, option level analysis, or final answer selection.

Each example in our dataset is annotated with the ground-truth transformation and the correct option. This allows us to probe not only whether the model predicted the correct option, but also whether its reasoning steps were aligned with the ground truth transformation.

**Near chance performance.** The model achieved an overall accuracy of 32.2%, only marginally above random guessing among four options. In contrast, human participants on comparable tasks reliably achieve accuracies above 80%. This gap underscores a fundamental inability to simulate mental transformations.

**Axis-specific biases.** Performance varied strongly by the axis of rotation: 28.2% for Y-axis, 32.4% for X-axis, and 35.7% for Z-axis. The samples generated were uniformly sampled across all the axis of rotations. Such anisotropy is inconsistent with human visuo-spatial reasoning, where performance is relatively robust across axes. This suggests that the model relies on superficial 2D heuristics rather than constructing flexible 3D representations.

**Mis-binding of reasoning and answers.** In 61.1% of cases, the model’s intermediate reasoning correctly described the ground truth transformation, but failed on two accounts **(1)** the final predicted option was incorrect. **(2)** The rotation angle across the axes were incorrect or misaligned. This indicates a systematic *mis-binding error*: the model can verbalize the correct transformation but fails to ground it in the corresponding visual candidate, indicating a loose coupling between linguistic reasoning and visual perception.

**Summary.** Together, these analyses suggest that current MLLMs exhibit limited evidence of embodied visuo-cognitive processes required for

Table 6: Overall performance and reasoning–answer consistency on Mental Transformation (Qwen-7B).

Metric	Value
Overall Accuracy	32.2%
Mis-binding	61.1%

Table 7: Accuracy by dominant rotation axis.

Dominant axis	# Items	Accuracy
X	68	32.4%
Y	39	28.2%
Z	42	35.7%

these tasks. Rather than performing internal perceptual transformations, they rely on shallow symbolic heuristics, leading to systematic and structured errors in mental rotation and transformation tasks. For Qwen-7B, overall accuracy was 32.2%, with 61.1% of reasoning steps correctly describing the transformation but yielding incorrect final answers: a *mis-binding* failure between verbal reasoning and visual grounding. Axis-specific results reveal anisotropy across rotation axes, reinforcing the lack of cognitive representation of perception.

## B.10 Effect of Prompt Optimization on Performance

We further examined whether performance limitations could be attributed to prompt ambiguity or poor phrasing by applying the framework in (Agarwal et al., 2024), which iteratively refines instructions and examples through a feedback driven critique and synthesis process. For this experiment, we generated three optimized prompt variations for the Qwen2.5-VL model and evaluated them across representative tasks from our benchmark. Table 9 reports the comparison between the baseline prompt and optimized variations.

While the optimized prompts yielded small but consistent gains across tasks (e.g., +0.08 on Visual Conceptual Slippage, +0.08 on Dynamic Structural Correspondence, +0.07 on Mental Transformation, and +0.06 on Hierarchical Pattern Equivalence), the overall improvements remained modest. These differences, though positive, do not substantially alter the performance profile of the model.

This suggests that the observed errors cannot be explained away as artifacts of ambiguous prompt

Table 8: Accuracy by number of active axes in the ground truth transformation.

# Active axes	# Items	Accuracy
1 axis	87	29.9%
2 axes	26	30.8%
3 axes	36	38.9%

Table 9: **Prompt Optimization**. This table compares the performance of Qwen 2.5-VL:7B with a baseline prompt versus an optimized version on four tasks. The modest improvements demonstrate that while better phrasing helps, it does not fix the core limitations of the model. All deltas <0.10 absolute.

Prompt	VCS	DCS	MRT	HPE
Baseline	0.30	0.32	0.35	0.24
Variation	0.38	0.40	0.42	0.30

wording. Instead, the persistence of core error patterns across both baseline and optimized prompts indicates that the primary bottleneck lies in the model’s inherent reasoning limitations rather than surface level prompt design. Thus, prompt optimization serves to confirm that the challenges exposed by our benchmark are fundamentally model driven rather than prompt driven.

This observation reinforces our conclusion that the benchmark exposes genuine deficiencies in visuo-cognitive reasoning, rather than artifacts of prompt design.

## C More about MIND’S EYE

We construct the MIND’S EYE benchmark by procedurally generating eight families of visuospatial reasoning tasks. Each family implements a well-defined cognitive operation (rotation, folding, composition, abstraction, etc.) and produces itemized question–answer pairs with explicit metadata (answer key, violation type, difficulty). Table 11 summarizes the controllable parameters, visual layout, and answer annotations for all tasks.

**Visual Conceptual Slippage.** We adapt classical “odd-one-out” paradigms to probe sensitivity to abstract visual relations. Each item draws six panels arranged in a  $2 \times 3$  grid. Five panels conform to a chosen concept (e.g., equidistant spacing, global symmetry, enclosure of one shape by another). Exactly one panel is designated as vio-

lating the concept. For word-symmetry items, a random uppercase string is rendered and mirrored to induce or break bilateral symmetry. Controlled parameters include concept type, variation seeds, and (for word-symmetry) word length. Random seeds are set to ensure reproducibility. The metadata records the violating option, the concept type, and, in word trials, the sampled word.

**Visual Relation Abstraction.** Visual Relation Abstraction items follow the Bongard problem style as in (Nie et al., 2020a). Using curated shape attributes (e.g., convexity, line crossings, polygonal regularity), we generate six figures: five positives sharing an attribute and one negative. Images are arranged in a  $2 \times 3$  grid with randomized positions. The annotation records the negative label and a decomposed textual reason string (e.g., “others are convex closed shapes; this one is not”).

**Mental Transformation.** Mental Transformation tests are generated from polycube assemblies. Building on the mental rotation subtask introduced in (Stogiannidis et al., 2025b), we extend it along an additional reasoning dimension to evaluate the model’s capacity for multistep reasoning as well. Each item shows a 3D “Original Shape” above four candidate rotations. Controlled factors are (i) shape identity, (ii) difficulty (single-axis vs. multi-axis rotation), and (iii) cube coloring (monochrome vs. varied). The correct answer is the candidate that matches the rotated original; metadata includes applied angles and difficulty level.

**Mental Composition.** This task probes net-to-solid reasoning. A 2D net (cube, prism, pyramid, cone, etc.) is rendered alongside four 3D candidate solids. Nets are chosen from a mapping (cube, cuboid, prism, pyramid, cone; harder items also include octahedron, dodecahedron, icosahedron). For easy items, nets are restricted to simple solids with uniform coloring; for hard items, complex polyhedra with confounding colorings are used. The net is drawn in the top-left of a  $2 \times 4$  grid, and candidate solids are rendered in the bottom row with distinct colors. The correct candidate is the folded realization of the net. Annotations store net identity, correct solid, distractors, color assignments, and difficulty.

**Paper Folding.** We simulate folding and hole punching on polygonal sheets. The sheet is a square or hexagon, represented as a polygon with

Table 10: **Closest benchmarks vs. MIND’S EYE along diagnostic axes** : A comparative evaluation of MIND’S EYE against other benchmarks on key diagnostic criteria such as Parametric Control, Distractor Quality, and the presence of a Human Baseline. It highlights the unique features that make MIND’S EYE a more controlled and diagnostic tool for assessing fluid intelligence. ✓=explicit support; ◻=partial; ✗=absent.

Dataset	Parametric Control	Cognitive Factor (ART)	Distractors Keyed to Confounds	No Knowledge Reliance	Format (MCQ/Open)	Multi-Pass Human Evaluation
CLEVR-like	✓	◻ (Abstraction)	◻	✓	Open	✗
Bongard-LOGO	✓	✓ (Abstraction/Relation)	◻	✓	MCQ	✗
RPM (RAVEN/I-RAVEN)	◻	✓ (Abstraction/Relation)	◻	✓	MCQ	✗
Mega-bench (MMMU/SEED/...)	✗	◻ (Mixed)	✗	✗	Mixed	✗
<b>Mind’s Eye (ours)</b>	✓	✓ (A/R/T)	✓	✓	MCQ	✓

vertices. A sequence of two folds is sampled either from vertical/horizontal reflections or from diagonal reflections. After folding, a single hole is punched at a random valid coordinate inside the polygon. The algorithm recursively unfolds the sheet and computes the mirrored hole positions. The final composite image shows: (i) initial unfolded sheet, (ii) two intermediate folds, (iii) the final folded sheet with hole, and (iv) four candidate unfolded sheets (A–D), one correct and three foils generated by removing, mirroring, or randomizing holes. The annotation records the fold group and the correct label. The task is to infer the unfolded hole pattern. Images show the fold sequence (top row) and four candidate unfolded sheets (bottom row, A–D). The correct option reproduces the true unfolded hole distribution.

**Dynamic Structural Correspondence.** Dynamic isomorphism tasks evaluate extrapolation of geometric motion. Two shapes undergo distinct continuous transformations (e.g., rotate-back-and-forth, bounce, wiggle, pulsate, swirl, shear, compress-and-stretch). The top row shows their trajectories at  $t \in \{0.0, 0.25, 0.5, 0.75\}$ . The bottom row contains four candidate continuations for  $t = 1.0$ , with one true continuation and three distractors (e.g., using mismatched functions or perturbed times). Parameters control shape identities, transformation pair, and time discretization. Annotations specify the correct continuation and textual explanation of which transformation applied to each shape.

**Symmetric Structures.** This task probes symmetry detection in line drawings. Each item shows four connected-path drawings: three exhibiting a chosen symmetry (vertical, horizontal, or rotational) and one lacking it. We generate random line paths by chaining ten short segments with random turns. Symmetry is imposed by re-

flection (vertical/horizontal) or rotation of order  $k \in \{2, 4\}$ . The layout is a  $1 \times 4$  grid (A–D), and the answer is the asymmetric panel.

**Hierarchical Pattern Equivalence.** Hierarchical reasoning is tested using recursively defined drawings (nested circles, concentric hexagons, fractal trees, L-systems, Sierpinski gaskets, Pythagoras trees, etc.). Each  $2 \times 2$  grid shows three valid hierarchical constructions and one violation consisting of random disconnected strokes. A random seed per panel ensures reproducible but varied instantiations. Parameters include which hierarchical generator is sampled and the seed for randomness. The correct answer is the violating panel.

## D Benchmark Design

The stimuli in MIND’S EYE are generated using scalable vector graphics (SVG) to keep a tight control over the geometric properties of the generated figures. Our design follows established principles from cognitive psychometrics (Embretson and Reise, 2013; De Boeck and Wilson, 2003) and recent best practices in multimodal evaluation (Li et al., 2023b; Liu et al., 2024a).

**Content blueprint and construct coverage.** Each task targets a distinct visuospatial construct, e.g. axis-aware 3D rotation (MRT) or relational structure preservation (Analogies). We developed a  $q$ -matrix mapping items to latent skills, ensuring coverage across multiple reasoning domains while avoiding construct underrepresentation (Embretson and Reise, 2013). This blueprint is intended to ensure broad cognitive coverage rather than overfitting to a narrow skill domain.

**Factorial item generation.** To minimize annotation artifacts and superficial shortcuts, we implemented *parametric, factorial generators* for all tasks. Each generator independently randomizes

Task	Controlled Parameters	Image Layout	Answer Key / Annotation
Visual Conceptual Slippage	Concept type (spacing, alignment, number, enclosure, symmetry, topology, border, hollowness, word symmetry), variation seeds	$2 \times 3$ grid of six labeled panels (A–F), one violates the rule	Index of violating panel; concept label; violation description
Visual Relation Abstraction	Shape attributes (convexity, symmetry, straight lines, angles, closure, regularity), positive vs. negative sets	$2 \times 3$ grid of six figures (A–F), one negative embedded among positives	Label of negative sample; reasoning string decomposed from attributes
Mental Transformation	Shape identity, difficulty level (single vs. multi-axis rotation), cube colors	Top row: “Original” 3D polycube; bottom row: four rotated candidates (A–D)	Correct option among A–D, rotation angles, difficulty tag
Mental Composition	Net type (cube, tetrahedron, prism, pyramid, cone, etc.), color assignments, difficulty (easy vs. hard nets)	Left: 2D net; right: four 3D candidate solids (A–D)	Correct option matching the folded solid; net/shape pair metadata
Paper Folding	Paper polygon size, fold sequence (V/H, diagonal), number and placement of punched holes	Top row: folding sequence; bottom row: four unfolded candidates (A–D)	Correct option label; fold sequence metadata
Dynamic Structural Correspondence	Shape type (triangle, square, pentagon, hexagon, diamond), transformation pair from library (rotate, shear, pulsate, bounce, etc.), time steps	$2 \times 4$ grid: first row shows transformation sequence; second row shows four candidate continuations (A–D)	Correct continuation label (fifth frame) with transformation description
Symmetric Structures	Symmetry type (vertical, horizontal, rotational), path complexity	$1 \times 4$ grid of line drawings (A–D), three symmetric and one asymmetric	Asymmetric label; annotation “lacks symmetry”
Hierarchical Pattern Equivalence	Structure type (nested circles, fractal trees, Sierpinski, L-system, etc.), violation injection	$2 \times 2$ grid of hierarchical drawings (A–D), one random violation	Label of violating structure; hierarchical function metadata

Table 11: **Overview of task generation** : The technical blueprint for the benchmark, detailing the specific controlled parameters, image layout, and answer annotations for each of the eight procedurally generated tasks.

*structural factors* (e.g. fold sequence length in Paper Folding, transformation chain length in Dynamic Isomorphism) and *nuisance factors* (rendering styles, color schemes, layout jitter). Orthogonal variation across these factors ensures item variety while balancing distractor plausibility. This design philosophy draws inspiration from adversarial benchmark construction in NLP and vision (Nie et al., 2020b; Zellers et al., 2019).

**Difficulty calibration.** Difficulty levels were calibrated both *a priori*, by manipulating structural complexity (e.g. rotation angle magnitude, hierarchy depth, color confounders), consistent with psychometric principles of item design where structural manipulations systematically af-

fect item difficulty (Embretson, 1983; Embretson and Reise, 2013; Ekstrom et al., 1976b; Vandenberg and Kuse, 1978a).

**Difficulty Calibration via Human Consensus:**

To establish empirically grounded difficulty levels for each item, we leverage human performance data collected during our evaluation study (Appendix F). Each item was independently evaluated by exactly five randomly sampled participants from our cohort of 30 non-expert adults. We operationalize difficulty through a consensus based framework rooted in inter-annotator agreement principles: items are classified as Easy if all five annotators provide the correct response (perfect agreement,  $\kappa = 1.0$  for that item), Hard if all five annotators respond incorrectly (perfect agreement

Task	S1:		S2:		S3:		S4:		S5:		S6:	
	Mental Rotation	Folding/ Topology	Relational Mapping	Symmetry/ Group Actions	Composition/ Decomposition	Slippage/ Robustness						
MT	✓						✓					
PF		✓						✓				
DSCS	✓									✓		
HPE			✓									
VRA			✓				✓					
SS							✓					
MC									✓			
VCS			✓									✓

Table 12: **Q-matrix blueprint.** Each task is mapped to a vector of latent visuo-cognitive skills. This matrix operationalizes the benchmark’s construct coverage and supports multi-trait psychometric modeling (De Boeck and Wilson, 2003; Embretson and Reise, 2013).

Table 14: Abstraction: VRA (Visual Relation Abstraction), HPE (Hierarchical Pattern Equivalence). Relation: DSC (Dynamic Structural Correspondence), VCS (Visual Conceptual Slippage), SS (Symmetric Structures). Transformation: MT (Mental Transformation), PF (Paper Folding), MC (Mental Composition).

on failure,  $\kappa = 1.0$ ) or any one answers correctly, and Medium if annotators exhibit split judgments with 2-3 correct responses (partial agreement,  $0.33 \leq \kappa \leq 0.67$ ). This approach aligns with established psychometric practice where item difficulty is calibrated against empirical response patterns rather than a priori structural complexity alone (Landis and Koch, 1977). Formally, for item  $i$  with human responses  $\{r_1, r_2, r_3, r_4, r_5\} \in \{0, 1\}$ , we assign difficulty  $d(i)$  as: Easy if  $\sum r_j = 5$ ; Hard if  $\sum r_j \leq 1$ ; Medium otherwise. To quantify the reliability of these assignments, we computed Fleiss’ kappa across all items within each task family, yielding moderate to substantial agreement ( $\kappa = 0.71$  across tasks), confirming that our difficulty categories capture stable individual differences rather than measurement noise. This consensus-driven calibration ensures that difficulty labels reflect actual human performance distributions and provides a principled basis for stratified analysis of model performance across varying levels of cognitive demand. The distribution of difficulty levels across our benchmark is: Easy (32%), Medium (45%), Hard (23%), ensuring sufficient representation of all difficulty strata for robust evaluation.

**Distractor taxonomy.** We construct options via confound keyed templates: *Transformation* — For MT the confounding dimensions were mirrored objects, varying color sequence of blocks. For MC the distractors were mirrored folds, parity

Subtasks / Dimensions	Ours	Mind the Gap	Bongard Logo	Visulogic	Bongard Hoi
<b>Abstraction</b>					
VRA	✓	✓			
HPE	✓		✓		✓
<b>Relation</b>					
DSC	✓		✓		
VCS	✓				
SS	✓				
<b>Transformation</b>					
MT	✓				
PF	✓			✓	
MC	✓	✓			

Table 13: **Comparison of subtask coverage across benchmarks:** Comparison of MIND’S EYE to other cognitive reasoning benchmarks. under the Abstraction-Relation-Transformation (A-R-T) framework.

punches, total punch holes reduced by 1. For MC colors and similar number of faces of the 3D object were the distractors. *Relation* — For DSC similar transformations applied to opposite shape and similar shapes with different transformation applied were used. For VCS shapes, colours and counts were used for the distractors. *Abstraction* — superficial feature match of the figures and motif substitution were the distractor’s features. Each item’s options include exactly one ground truth and one confounds sampled from distinct templates to avoid ambiguity and remaining wrong options. Table 11 shows the control parameters for each task more in detail.

**Answer encoding and randomization.** Each item is a 1+4 panel (query + options A–D) except the tasks VCS and VRA. These tasks are 1+6 (query + options A-F). Option order is uniformly randomized; keys are uniformly distributed across all options. Unless stated, each subtask contains 100 items (balanced across difficulty bins), yielding 800 items total for the main suite.

**Item specification.** We author in SVG and export to PNG at  $1024 \times 1024$  px (300 DPI) with fixed stroke widths and sans-serif labels; background is uniform. Panels use consistent gutters and margins to minimize layout cues. A schematic is shown in Fig. 1.

**Benchmark properties.** The Q-matrix in Table 12 specifies the mapping between each bench-

mark task and the three core processes of fluid reasoning : **Abstraction, Relation, and Transformation**. In psychometric terms, the Q-matrix operationalizes our construct blueprint (De Boeck and Wilson, 2003; Embretson and Reise, 2013), serving as an explicit hypothesis about the latent skills each item requires. For example, *Visual Relation Abstraction* is coded purely under Abstraction, while *Hierarchical Pattern Equivalence* loads on both Abstraction and Relation, since it demands generalization of hierarchical patterns and recognition of their structural equivalence. Similarly, *Dynamic Structural Correspondence* and *Symmetric Structures* are placed at the Relation-Transformation intersection, as they require both analogical mapping and mental manipulation of visual forms. Pure Transformation tasks such as *Mental Rotation*, *Paper Folding*, and *Mental Composition* emphasize dynamic visuospatial manipulation without strong abstraction demands.

This structured mapping justifies that the benchmark covers a broad range of reasoning processes identified as central to fluid intelligence (Carroll, 1993; Schneider and McGrew, 2018; McGrew, 2005). Moreover, it allows us to move beyond raw accuracy by fitting multi trait IRT or cognitive diagnostic models, thereby diagnosing which cognitive processes (A, R, T) different models succeed or fail on. In effect, the Q-matrix both grounds our task design theoretically and provides the statistical scaffold for psychometric calibration and analysis.

## E Evaluation Setup

To ensure a fair comparison across models, all systems are evaluated with identical visual inputs and standardized textual prompts. Since modern MLLMs often generate extended free form outputs, conventional rule based extraction is brittle and prone to errors (Duan et al., 2024; Fu et al., 2024; Lu et al., 2022). Following recent practice (Lu et al., 2024b; Zhang et al., 2024), we adopt an expert LLM-based evaluation protocol.

The procedure consists of three stages.

1. **Input Presentation:** The model under evaluation receives both the image and textual question in a fixed prompt template designed to minimize variation across models.
2. **Answer Extraction:** We employ Gemma-3 (Team et al., 2025) as the judging model

to parse free form answers into concise responses. This method builds on prior work showing that large LLMs can perform semantic normalization of outputs with high reliability (Liu et al., 2024b; Fu et al., 2024).

3. **Label Standardization:** Extracted responses are mapped to task specific discrete labels (e.g., multiple choice option identifiers or numeric values). Accuracy is then computed against the ground truth key for each of the eight subtasks.

**Prompting strategies.** Since multimodal reasoning is highly sensitive to prompt design (Wei et al., 2022; Kojima et al., 2022), we explore four prompting strategies designed to elicit cognitive reasoning rather than shallow pattern matching:

- **Meta-task Framing.** Before presenting a question, the prompt explicitly describes the type of reasoning required. For example: “*This is a mental transformation test. You need to imagine folding or rotating the shape in 3D.*” “*This is a paper folding puzzle. At the end, identify which option shows the holes in the unfolded paper.*” Such framing aligns the model’s reasoning pathway with the intended cognitive faculty, similar to task oriented prompting used in prior cognitive benchmarks (Nie et al., 2020a; Zhang et al., 2019).
- **Step-by-Step Instruction Prompts.** Models are encouraged to reason structurally by decomposing problems: “*First, describe the shapes. Then, identify the transformation (rotation, reflection, folding, symmetry). Finally, choose the answer.*” This mirrors structured reasoning templates shown effective in prior chain-of-thought prompting work (Wei et al., 2022).
- **Hints via Concept Tags.** To reduce ambiguity about the task type, we prepend task specific tags. For example: “[*Task: Mental Transformation*] *Which option matches the rotated version of the shape?*” “[*Task: Symmetry Detection*] *Which figure preserves the symmetry of the original?*” Such concept scaffolding helps models focus on execution rather than inferring task intent, following recent evaluations of role tagged prompting in

multimodal reasoning (Liu et al., 2024b; Xu et al., 2025a).

- **Chain-of-Thought Anchors.** Instead of generic “think step by step,” we provide explicit anchors to guide reasoning stages: “*Step 1: Identify the primitive shapes. Step 2: Detect how they move or fold. Step 3: Eliminate mismatched answers.*” This builds on structured CoT prompting approaches (Wei et al., 2022; Kojima et al., 2022) and ensures models engage in interpretable intermediate reasoning rather than shortcutting to an answer.

**Hardware** : The computational experiments presented in this paper were executed using a server equipped with four NVIDIA RTX A6000 graphics processing units, each providing 48 GB of dedicated memory to support the inferential and analytical demands of the evaluated models.

**Closed source** We use the following setup of OpenAI API for evaluation:

```
OpenAI model name: o3-2025-04-16
response = client.responses.create(
    model="gpt-o3",
    reasoning={"effort": "medium"},
    input=[
        {
            "role": "user",
            "content": prompt
        }
    ],
    max_output_tokens=500,
)
```

## F Human Evaluation Protocol

To establish a meaningful baseline and to ground our benchmark in psychometric validity, we conducted a controlled human evaluation study. A total of  $N = 30$  participants were recruited through university mailing lists and professional networks. All participants were adults between the ages of 20 and 40 years, ensuring that the sample represents a cognitively mature population while minimizing potential confounds associated with either adolescent development or age related decline in visuospatial processing. The cohort comprised 17 male and 13 female participants. None of the participants reported any prior expertise with the specific tasks used in our benchmark, and all provided informed consent.

Table 15: Human participant statistics.

Metric	Value
Sample size ( $N$ )	30
Age range (years)	20–40
Mean age (years)	25.3
Male : Female	17 : 13
Prior task expertise	None (self reported)
Recruitment	University lists, professional networks
Consent	Informed consent obtained

Each participant completed the full battery of eight task families. The tasks were presented in randomized order to control for ordering effects, and items within each task family were also randomized to reduce learning or memorization effects. The evaluation was administered under standardized conditions: participants were given detailed instructions at the beginning of each task type. Responses were collected digitally through a custom interface that mirrors the image based multiple choice format used for multimodal language models.

The total testing time for each participant was approximately 60 minutes, which included both task instructions and the full set of items across all task families. This duration was sufficient to collect reliable performance data while avoiding participant fatigue. The resulting dataset of human responses provides not only an upper bound reference for model comparison but also enables us to quantify item difficulty and discrimination ability through psychometric analysis.

## G Prompt Style Performance

The performance under four different prompting strategies was evaluated to understand model sensitivity to instructions. Detailed results are presented for Hint-Based prompting (Table 16), Elimination-Based prompting (Table 19), Meta-Task prompting (Table 17), and Step-by-Step prompting (Table 18).

## H Prompts

Following the overview of evaluation strategies in Appendix E, this section presents the specific prompt templates used in our experiments. Figure 17 illustrates the general question structure applied to each task. The detailed templates for our three primary prompting methods—Elimination, Hint-Based, and Meta-Task—are provided in Figures 18, 19, and 20, respectively.

	Abstraction		Relation			Transformation		
	VRA (100)	HPE (100)	DSC (100)	VCS (100)	SS (100)	MT (100)	PF (100)	MC (100)
Idefics - 8B	0.24	0.27	0.35	0.26	0.16	0.34	0.27	0.27
InternVL3 - 8B	0.22	0.28	0.31	0.25	0.30	0.37	0.28	0.28
LLaMa-3.2 - 11B	0.22	0.28	0.31	0.26	0.23	0.30	0.25	0.28
Llava-1.6-Mistral - 7B	0.18	0.24	0.33	0.25	0.31	0.36	0.25	0.30
Phi3.5-vision-instruct - 8B	0.21	0.26	0.32	0.24	0.30	0.33	0.26	0.30
Qwen-2.5-VL - 7B	0.25	0.26	0.30	0.26	0.15	0.39	0.21	0.22

Table 16: **Performance on the task splits using Hint prompts.** Abstraction: VRA (Visual Relation Abstraction), HPE (Hierarchical Pattern Equivalence). Relation: DSC (Dynamic Structural Correspondence), VCS (Visual Conceptual Slippage), SS (Symmetric Structures). Transformation: MT (Mental Transformation), PF (Paper Folding), MC (Mental Composition).

	Abstraction		Relation			Transformation		
	VRA (100)	HPE (100)	DSC (100)	VCS (100)	SS (100)	MT (100)	PF (100)	MC (100)
Idefics - 8B	0.23	0.28	0.34	0.27	0.14	0.34	0.27	0.28
InternVL3 - 8B	0.23	0.29	0.32	0.25	0.30	0.37	0.28	0.29
LLaMa-3.2 - 11B	0.23	0.29	0.32	0.26	0.23	0.31	0.25	0.29
Llava-1.6-Mistral - 7B	0.18	0.25	0.33	0.25	0.31	0.36	0.25	0.30
Phi3.5-vision-instruct - 8B	0.22	0.27	0.33	0.25	0.30	0.33	0.26	0.30
Qwen-2.5-VL - 7B	0.27	0.28	0.31	0.26	0.22	0.25	0.32	0.37

Table 17: **Performance on the task splits using Meta Task prompts.** Abstraction: VRA (Visual Relation Abstraction), HPE (Hierarchical Pattern Equivalence). Relation: DSC (Dynamic Structural Correspondence), VCS (Visual Conceptual Slippage), SS (Symmetric Structures). Transformation: MT (Mental Transformation), PF (Paper Folding), MC (Mental Composition).

## I Qualitative CoT Analysis

To further understand the internal reasoning behavior of multimodal large language models, we qualitatively analyzed the reasoning traces produced by GPT-4o across representative tasks from the MIND’S EYE benchmark. Across these tasks, the reasoning traces reveal a consistent pattern: although the models often provide syntactically coherent explanations and occasionally arrive at the correct answer, their reasoning is largely surface level and perceptually driven. Rather than performing the required cognitive operations of perception, such as mental transformation, folding/unfolding, or abstraction of relational structure, the model tends to depend on low level visual heuristics (e.g., color matching, spatial alignment, or visual distinctiveness among options).

Together, these analyses indicate that model reasoning traces often rely on heuristic visual cues rather than systematic cognitive reasoning.

## J Carroll’s Fluid Intelligence to ART Framework

Figure 25 illustrates the alignment between the constructs of Fluid Intelligence from the Cattell-

Horn-Carroll (CHC) framework and our proposed Abstraction, Relation, and Transformation (A-R-T) taxonomy.

## K Difficulty Analysis

The performance patterns across the eight cognitive subtasks reveal several critical insights into the visual reasoning capabilities of current MLLMs compared to human performance.

**Human Performance Sensitivity to Difficulty:** Human participants demonstrate the expected sensitivity to difficulty calibration, with performance systematically declining across difficulty levels. On Easy items, humans achieve accuracies of 0.85-0.95, consistent with the calibration criterion (all 5 annotators correct). Performance drops to 0.55-0.65 on Medium items (2-3 annotators correct), and further declines to 0.10-0.25 on Hard items (0-1 annotators correct). This graded degradation validates our difficulty manipulation and demonstrates that humans engage genuinely with increasing cognitive demands. The decline is particularly pronounced in Transformation tasks (Mental Composition: 0.94  $\rightarrow$  0.14) and Abstraction tasks (Hierarchical Pattern Equivalence: 0.92  $\rightarrow$  0.18), where spatial manipulation and pattern

	<b>Abstraction</b>		<b>Relation</b>			<b>Transformation</b>		
	VRA (100)	HPE (100)	DSC (100)	VCS (100)	SS (100)	MT (100)	PF (100)	MC (100)
Idefics - 8B	0.24	0.28	0.35	0.27	0.15	0.35	0.27	0.28
InternVL3 - 8B	0.23	0.29	0.32	0.26	0.31	0.38	0.28	0.29
LLaMa-3.2 - 11B	0.23	0.29	0.32	0.26	0.23	0.31	0.25	0.29
Llava-1.6-Mistral - 7B	0.19	0.25	0.33	0.26	0.31	0.37	0.25	0.30
Phi3.5-vision-instruct - 8B	0.22	0.27	0.33	0.26	0.31	0.34	0.26	0.30
Qwen-2.5-VL - 7B	0.25	0.28	0.30	0.26	0.23	0.26	0.26	0.32

Table 18: **Performance on the task splits using SBS prompts.** Abstraction : VRA (Visual Relation Abstraction), HPE (Hierarchical Pattern Equivalence). Relation: DSC (Dynamic Structural Correspondence), VCS (Visual Conceptual Slippage), SS (Symmetric Structures). Transformation: MT (Mental Transformation), PF (Paper Folding), MC (Mental Composition).

	<b>Abstraction</b>		<b>Relation</b>			<b>Transformation</b>		
	VRA (100)	HPE (100)	DSC (100)	VCS (100)	SS (100)	MT (100)	PF (100)	MC (100)
Idefics - 8B	0.23	0.27	0.34	0.26	0.15	0.34	0.27	0.28
InternVL3 - 8B	0.22	0.28	0.32	0.25	0.31	0.37	0.28	0.28
LLaMa-3.2 - 11B	0.22	0.28	0.32	0.25	0.23	0.30	0.25	0.29
Llava-1.6-Mistral - 7B	0.18	0.24	0.33	0.25	0.31	0.36	0.25	0.30
Phi3.5-vision-instruct - 8B	0.21	0.26	0.33	0.25	0.31	0.33	0.26	0.29
Qwen-2.5-VL - 7B	0.22	0.30	0.32	0.22	0.21	0.34	0.24	0.29

Table 19: **Performance on the task splits using Eliminate prompts.** Abstraction: VRA (Visual Relation Abstraction), HPE (Hierarchical Pattern Equivalence). Relation: DSC (Dynamic Structural Correspondence), VCS (Visual Conceptual Slippage), SS (Symmetric Structures). Transformation: MT (Mental Transformation), PF (Paper Folding), MC (Mental Composition).

abstraction become progressively more demanding.

**The Model Performance Gap:** Both closed-source and open-source models exhibit substantially lower performance compared to humans, with accuracy typically ranging between 0.2-0.5 across tasks. This performance gap is consistent across all eight subtasks, indicating systematic limitations in visual-cognitive reasoning rather than isolated weaknesses. The gap is particularly pronounced in Transformation (Mental Composition, Paper Folding, Mental Transformation) and Abstraction (Visual Relation Abstraction, Hierarchical Pattern Equivalence) dimensions of the ART framework.

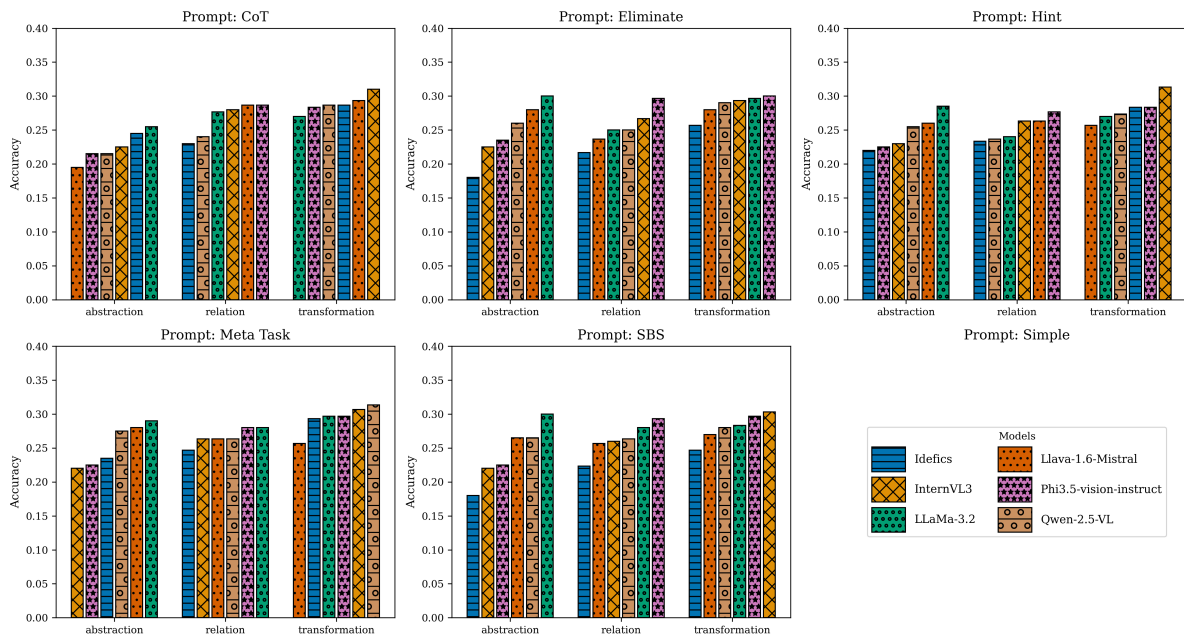
**Flat Model Difficulty Curves: A Critical Divergence** In stark contrast to humans, both model categories show minimal sensitivity to task difficulty, with performance remaining relatively

flat (typically varying by only 0.02-0.08 accuracy points) across Easy, Medium, and Hard conditions. This flat difficulty curve reveals a fundamental limitation: while humans struggle progressively more with harder instantiations of genuine spatial reasoning, models appear unable to perform the core cognitive operations at any difficulty level. The models’ consistent low performance (0.20-0.45) regardless of difficulty suggests they lack the foundational visual-cognitive mechanisms required for these tasks.

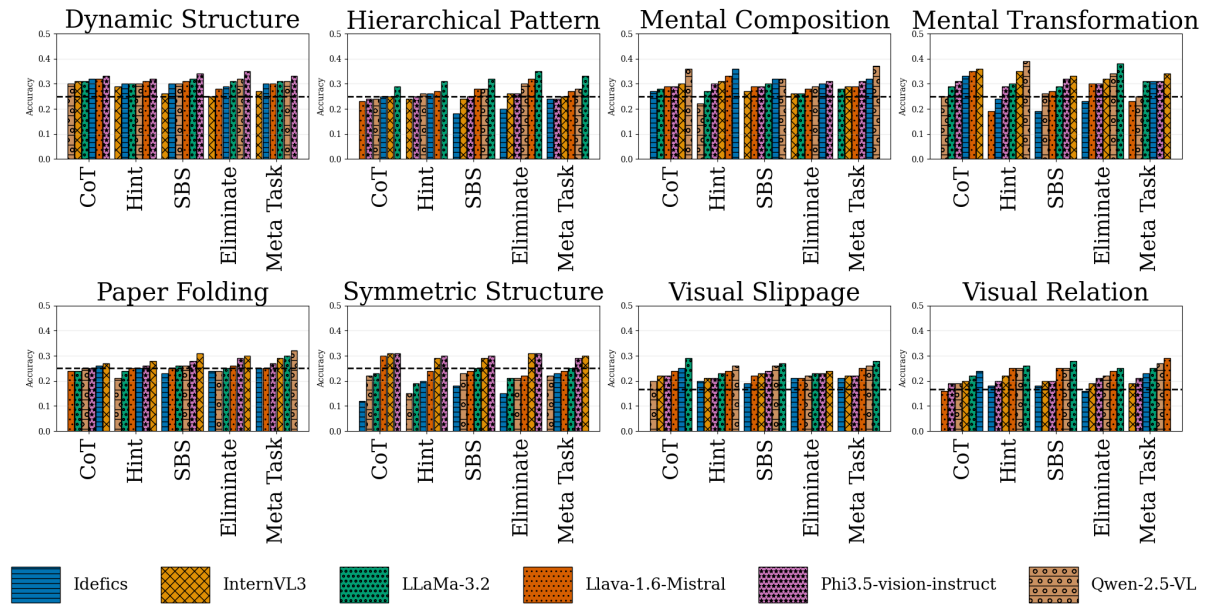
**Closed-Source vs. Open-Source Models** Closed-source models consistently outperform open-source models across all tasks and difficulty levels, though both remain substantially below human performance and exhibit similarly flat difficulty curves. The performance advantage of closed-source models is most pronounced in Transformation tasks, where they achieve 0.30-

0.45 accuracy compared to 0.25-0.35 for open-source models. However, this advantage narrows in Relation tasks (Symmetric Structures, Visual Conceptual Slippage) and Abstraction tasks (Hierarchical Pattern Equivalence, Dynamic Structural Correspondence), suggesting that certain types of visual reasoning present fundamental challenges even for state-of-the-art proprietary models. Critically, neither model category shows the systematic performance degradation across difficulty levels that characterizes human performance.

These results suggest that current MLLMs may lack fundamental visual-cognitive capabilities that humans deploy effortlessly. The divergence between human difficulty sensitivity and flat model performance curves provides compelling evidence that models are not merely worse at these tasks—they are solving them through fundamentally different (and inadequate) mechanisms. While humans engage in genuine visuospatial reasoning that scales with task complexity, models appear to rely on shallow heuristics that fail uniformly across difficulty levels. This suggests that bridging the human-model gap will require architectural innovations that enable true perceptual transformation and cognitive simulation, rather than simply scaling existing approaches.



(a) Skill-level bar chart showing average model performance across the three cognitive levels (Transformation, Relation, Abstraction). Tasks are grouped by level, and bars indicate per model averages.



(b) Effect of different prompting styles (CoT, Hint, Meta) on per task performance. The dotted line denotes the random choice baseline for each task.

Figure 5: **Model Performance and Prompting Effects** : (a) Average model performance across cognitive skill levels. (b) Prompting style effects on task wise performance. Together, these plots illustrate the inconsistent and often modest impact of different prompting styles (CoT, Hint, Meta) on per task performance

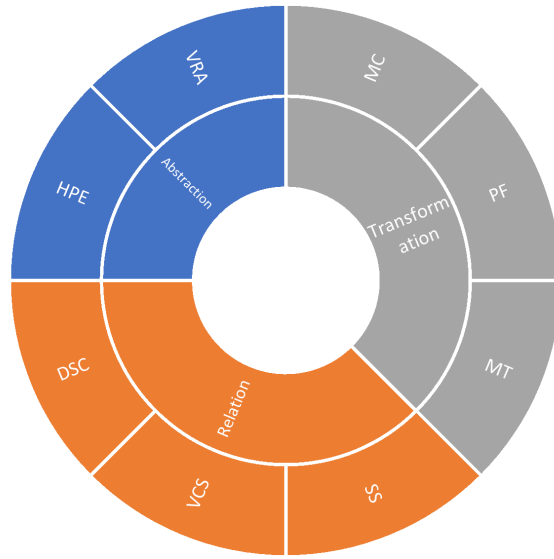
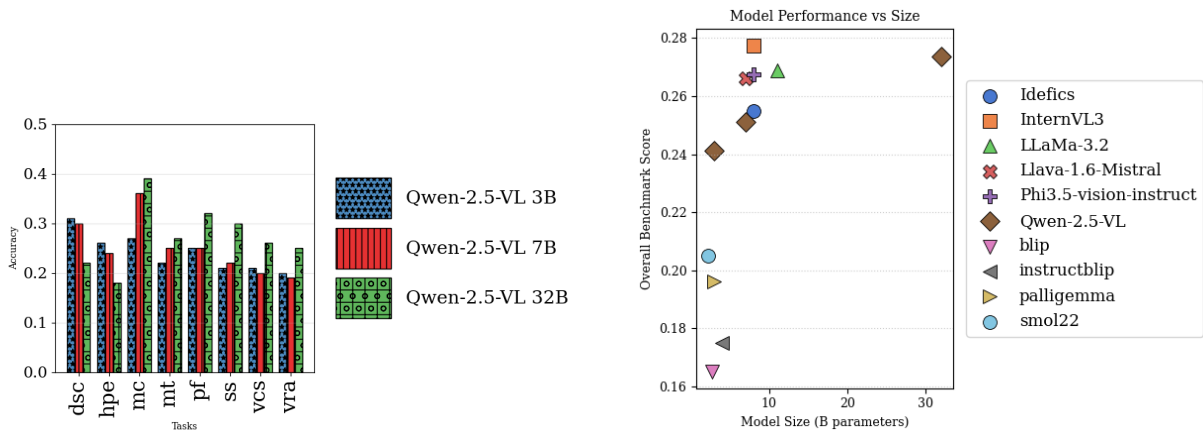


Figure 6: **Dataset Distribution as per the ART Framework:** The inner ring represents the three A-R-T cognitive categories, while the outer ring shows the eight specific tasks and their alignment within this framework.



(a) **Accuracy of Qwen-2.5-VL models (3B, 7B, 32B) across ART tasks :** The performance comparison of Qwen-2.5-VL models of three different sizes (3B, 7B, 32B) across the eight tasks. It highlights that scaling provides non uniform gains, with larger models improving on some tasks but not uniformly across all the tasks, reinforcing that scale alone is insufficient to overcome the reasoning deficits.

(b) **Model performance versus size on our benchmark.** While larger models (e.g., Qwen-2.5-VL-32B) achieve strong results, several medium sized models (InternVL3, LLaMA-3.2, Phi-3.5) match or exceed them, indicating that scaling alone is insufficient and that training design and architecture critically influence cognitive reasoning performance.

Figure 7: **Impact of Scale :** (a) and (b) provides a compelling evidence that scaling alone is not sufficient enough to improve performance on this benchmark

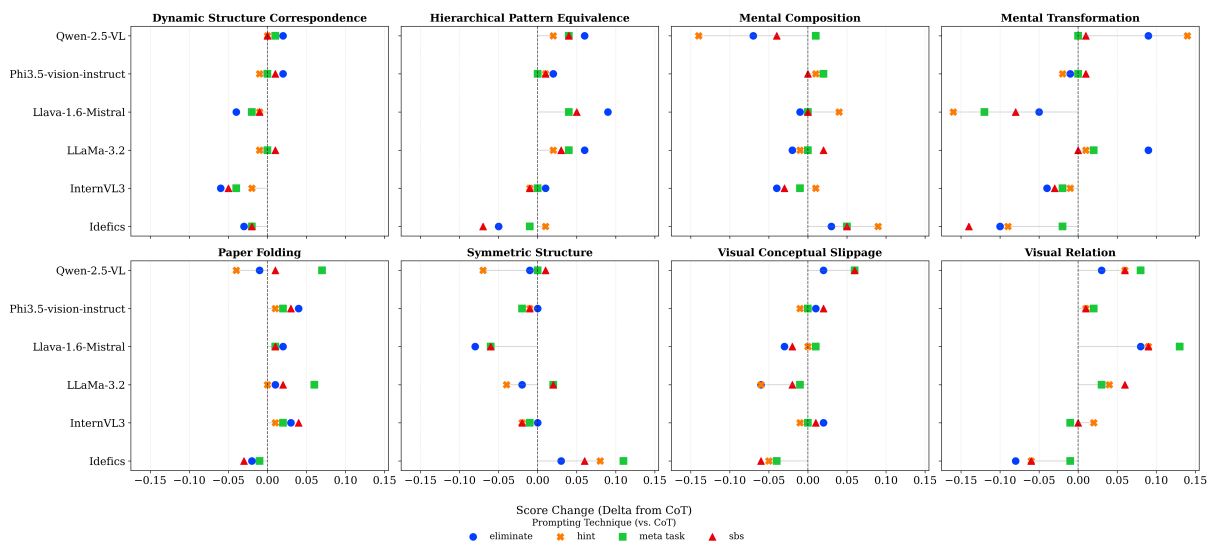


Figure 8: **Relative effect of prompting strategies versus chain-of-thought (CoT) across tasks.** Points to the left of the dashed line indicate performance deterioration, while those to the right indicate improvement. Meta-task and step-by-step (SBS) prompts often improve tasks like *Hierarchical Pattern Equivalence*, *Visual Relation*, *Paper Folding*, but abstraction tasks like *Symmetric Structure*, *Mental Composition*, *Mental Transformation*, *Visual Conceptual Slippage* show consistent declines. Prompting strategies therefore exert strongly task dependent effects, with no universally reliable method for improving performance.

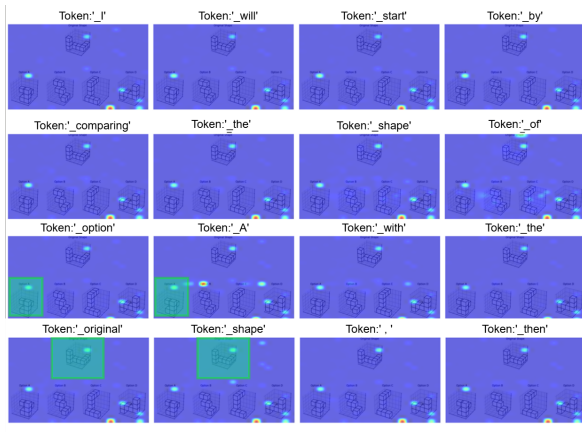


Figure 11: **Misplaced model attention** : Attention map of the LLaVa-7B model for the Mental Transformation Task. The green boxes shows the expected regions of attention.

**Judge LLM prompt for answer extraction**

Review the following text which contains a detailed explanation and a final answer choice for a multiple-choice question. Your task is to extract ONLY the final letter of the selected option (e.g., A, B, C, D). Do not provide any explanation or extra text. If there is not clear final answer choice give option X  
 Text: "{llm\_response\_text}"  
 Extracted Option:

**Example Prompts for Mind's Eye Tasks**

**Visual Conceptual Slippage** : Analyze the six figures labeled A, B, C, D, E, and F in the image. Your goal is to identify the underlying visual concept that is shared by the majority of these figures and then determine which figure does not adhere to this concept. Choose from the options A,B,C,D,E and F

**Visual Relation Abstraction**: Analyze the six figures labeled A, B, C, D, E, and F in the image. Your goal is to identify the underlying visual concept that is shared by the majority of these figures and then determine which figure does not adhere to this concept. Choose from the options A,B,C,D,E and F

**Mental Transformation** : Analyze the four figures labeled A, B, C, D in the image. Your goal is to identify which of these options is the correct rotational transformation of the figure labeled original shape. Choose from the options A,B,C and D

**Mental Composition** : Analyze the four figures labeled A, B, C, D in the image. These are 3D Figures. Your goal is to identify which of these options can be constructed by folding the top 2D figure labeled as Original Shape. Choose from the options A,B,C and D

**Paper Folding** : The a paper of shape labeled Unfolded is folded following the pattern shown in the top row, in the sequence of fold 1, followed by fold 2 and so on. After the last fold a hole is punched as shown in the image of labeled final view, then after unfolding how would the paper look. Analyze the four figures labeled A, B, C, D in the image, how they are folded and answer. Choose from the options A,B,C and D

**Dynamic Structural Correspondence** : Follow the transformation of the shapes in the top row, starting from  $t=0.0$  to  $t=0.75$ . Analyze the four figures labeled A, B, C, D in the image. Your goal is to identify which of these options is the fifth image at  $t = 1.0$ . Choose from the options A,B,C and D.

**Symmetric Structures** : Analyze the four figures labeled A, B, C, D in the image. Your goal is to identify the underlying visual concept that is shared by the majority of these figures and then determine which figure does not adhere to this concept. Choose from the options A,B,C and D

**Hierarchical Pattern Equivalence** : Analyze the four figures labeled A, B, C and D in the image. Your goal is to identify the underlying visual concept that is shared by the majority of these figures and then determine which figure does not adhere to this concept. Choose from the options A,B,C and D

Figure 17: **Prompts**: (Top) The judge LLM prompt is used for extracting selected options form the free following answers of the model. (Bottom) The question prompt for each task of the benchmark

<p style="text-align: center;"><b>Visual Conceptual Slippage</b></p> <p>Analyze the six figures labeled A, B, C, D, E, and F in the image. Your goal is to identify the underlying visual concept that is shared by the majority of these figures and then determine which figure does not adhere to this concept. Use elimination reasoning following these steps:</p> <ol style="list-style-type: none"> <li>1. For each figure (A-F), identify its key visual property (e.g., symmetry, rotation, structure).</li> <li>2. Determine the property shared by the majority.</li> <li>3. Eliminate figures that do not fit this majority property, stating the mismatch briefly.</li> </ol> <p>The single figure that fails the majority rule is the odd one out. The final answer is the odd one out from one of the six options (A-F). Choose that option as the final answer</p>	<p style="text-align: center;"><b>Mental Transformation</b></p> <p>Analyze the four figures labeled A, B, C, D in the image. Your goal is to identify which of these options is the correct rotational transformation of the figure labeled original shape. Use elimination reasoning following these steps:</p> <ol style="list-style-type: none"> <li>1. For each option (A-D), check if it can be obtained by a pure rotation of the original shape (no reflection or distortion).</li> <li>2. Eliminate options that require flipping, mirroring, or that misplace parts after rotation.</li> </ol> <p>The remaining option is the correct rotational transformation. Give the final answer from one of the four options (A-D).</p>
<p style="text-align: center;"><b>Paper Folding</b></p> <p>The a paper of shape labeled Unfolded is folded following the pattern shown in the top row, in the sequence of fold 1, followed by fold 2 and so on. After the last fold a hole is punched as shown in the image of labeled final view, then after unfolding how would the paper look. Use elimination reasoning following these steps:</p> <p>For each option (A-D):</p> <ol style="list-style-type: none"> <li>1. Check hole count matches expected reflections from the fold sequence.</li> <li>2. Check symmetry about each fold line is preserved.</li> <li>3. Check relative positions and offsets of holes from folds/edges are consistent.</li> </ol> <p>Reject any option failing one of these checks. The remaining option is the correct answer. Give the final answer from one of the four options (A-D).</p>	<p style="text-align: center;"><b>Dynamic Structural Correspondence</b></p> <p>Follow the transformation of the shapes in the top row, starting from <math>t=0.0</math> to <math>t=0.75</math>. Analyze the four figures labeled A, B, C, D in the image. Your goal is to identify which of these options is the fifth image at <math>t = 1.0</math>. Use elimination reasoning:</p> <ol style="list-style-type: none"> <li>1. Identify the transformation rule from <math>t=0.0 \rightarrow t=0.75</math> (e.g., rotation angle, size change, element shift).</li> <li>2. Predict the state at <math>t=1.0</math> by applying the same rule once more.</li> <li>3. For each option (A-D), check: <ul style="list-style-type: none"> <li>- Does it continue the transformation consistently?</li> <li>- Does it violate the rule (wrong angle, size, count, or orientation)?</li> </ul> </li> </ol> <p>Reject options that break the rule, and explain the mismatch briefly. The remaining option is the correct answer. Give the final answer from one of the four options (A-D).</p>
<p style="text-align: center;"><b>Mental Composition</b></p> <p>Analyze the six figures labeled A, B, C, D in the image. Your goal is to identify which of these options can be constructed by folding the top 2D figure labeled as Original Shape. Use elimination reasoning following these steps:</p> <ol style="list-style-type: none"> <li>1. For each option (A-D), check if its face adjacencies and orientations can result from folding the given 2D net labeled as original shape.</li> <li>2. Eliminate any option with impossible adjacencies, overlaps, or mirrored faces.</li> </ol> <p>The remaining option is the correct 3D construction. Give the final answer from one of the four options (A-D).</p>	<p style="text-align: center;"><b>Symmetric Structures</b></p> <p>Analyze the four figures labeled A, B, C, D in the image. Your goal is to identify the underlying visual concept that is shared by the majority of these figures and then determine which figure does not adhere to this concept. Use elimination reasoning:</p> <ol style="list-style-type: none"> <li>1. Identify the symmetry property of each figure (mirror axes, rotational order, or none).</li> <li>2. Determine the property shared by the majority.</li> <li>3. For each option (A-D), check if it satisfies that property. <ul style="list-style-type: none"> <li>- If it breaks the symmetry rule, reject it.</li> </ul> </li> </ol> <p>The single figure that fails the majority rule is the odd one out. The final answer is the odd one out from one of the four options (A-D).</p>
<p style="text-align: center;"><b>Hierarchical Pattern Equivalence</b></p> <p>Analyze the four figures labeled A, B, C and D in the image. Your goal is to identify the underlying visual concept that is shared by the majority of these figures and then determine which figure does not adhere to this concept. Use elimination reasoning:</p> <ol style="list-style-type: none"> <li>1. For each figure (A-D), check if it shows recursive structure (self-similarity across multiple scales).</li> <li>2. Identify the majority property: "recursive" vs. "non-recursive."</li> <li>3. Eliminate figures that do not fit the majority property, stating the mismatch (e.g., lacks smaller copies, no multi-scale repetition).</li> </ol> <p>The figure that fails the majority rule is the odd one out. The final answer is the odd one out from one of the four options (A-D).</p>	<p style="text-align: center;"><b>Visual Relation Abstraction</b></p> <p>Analyze the six figures labeled A, B, C, D, E, and F in the image. Your goal is to identify the underlying visual concept that is shared by the majority of these figures and then determine which figure does not adhere to this concept. Use elimination reasoning following these steps:</p> <ol style="list-style-type: none"> <li>1. For each figure (A-F), identify its key visual property (e.g., symmetry, rotation, structure).</li> <li>2. Determine the property shared by the majority.</li> <li>3. Eliminate figures that do not fit this majority property, stating the mismatch briefly.</li> </ol> <p>The single figure that fails the majority rule is the odd one out. The final answer is the odd one out from one of the six options (A-F). Choose that option as the final answer</p>

Figure 18: Elimination Prompt for all the tasks.

<p style="text-align: center;"><b>Visual Conceptual Slippage</b></p> <p>Analyze the six figures labeled A, B, C, D, E, and F in the image. Your goal is to identify the underlying visual concept that is shared by the majority of these figures and then determine which figure does not adhere to this concept. To solve this question observe how the spaces ,alignment, number of each object in each figure change. The final answer is the odd one out from one of the six options. Give the final answer from one of these six options (A-F)</p>	<p style="text-align: center;"><b>Mental Transformation</b></p> <p>Analyze the four figures labeled A, B, C, D in the image. Your goal is to identify which of these options is the correct rotational transformation of the figure labeled original shape. Observe the given figures and pay attention to how the original shape can be rotated to match the given options.. Give the final answer from one of the four options (A-D).</p>
<p style="text-align: center;"><b>Paper Folding</b></p> <p>The a paper of shape labeled Unfolded is folded following the pattern shown in the top row, in the sequence of fold 1, followed by fold 2 and so on. The papers are folded in a manner that the grey part of the paper goes on top after folding. After the hole is placed in the figure label Final view, follow the symmetry of the folds to unfold the paper and identify where the hole position is in that unfolded step. Keep on unfolding and tracking the holes until the paper is completely unfolded. Give the final answer from one of the four options ( A-D ).</p>	<p style="text-align: center;"><b>Dynamic Structural Correspondence</b></p> <p>Follow the transformation of the shapes in the top row, starting from <math>t=0.0</math> to <math>t=0.75</math>. Analyze the four figures labeled A, B, C, D in the image. Your goal is to identify which of these options is the fifth image at <math>t = 1.0</math>. Consider how the elements of the top figure transforms over time. Following that transformation identify which of these options is the fifth image at <math>t = 1.0</math> Give the final answer from one of the four options (A-D).</p>
<p style="text-align: center;"><b>Mental Composition</b></p> <p>Analyze the six figures labeled A, B, C, D n the image. Your goal is to identify which of these options can be constructed by folding the top 2D figure labeled as Original Shape. Observe the given figures and pay attention to how the original shape can be folded or composed to match the given options Give the final answer from one of the four options (A-D).</p>	<p style="text-align: center;"><b>Symmetric Structures</b></p> <p>Analyze the four figures labeled A, B, C, D in the image. Your goal is to identify the underlying visual concept that is shared by the majority of these figures and then determine which figure does not adhere to this concept. Observe the given figures and pay attention to their symmetry properties. Give the final answer from one of the four options (A-D).</p>
<p style="text-align: center;"><b>Hierarchical Pattern Equivalence</b></p> <p>Analyze the four figures labeled A, B, C and D in the image. Your goal is to identify the underlying visual concept that is shared by the majority of these figures and then determine which figure does not adhere to this concept. Observe the given figures and pay attention to their recursive structure properties. Give the final answer from one of the four options (A-D).</p>	<p style="text-align: center;"><b>Visual Relation Abstraction</b></p> <p>Analyze the six figures labeled A, B, C, D, E, and F in the image. Your goal is to identify the underlying visual concept that is shared by the majority of these figures and then determine which figure does not adhere to this concept. Observe carefully as five of the size figures would have one of these properties convex has curves, has straight line, symmetric, has three straight lines, has four straight lines, has five straight lines, has six straight lines, has seven straight lines, has eight straight lines, thin shape, closed shape, has acute angle, has obtuse angle, has angle, has line crossing, has two parts, exists triangle, exists quadrangle, exist sectors. Give the final answer from one of the six options ( A - F )</p>

Figure 19: **Hint Prompt** for all the tasks.

<p style="text-align: center;"><b>Visual Conceptual Slippage</b></p> <p>Analyze the six figures labeled A, B, C, D, E, and F in the image. Your goal is to identify the underlying visual concept that is shared by the majority of these figures and then determine which figure does not adhere to this concept. This is a conceptual slippage + odd-one-out question. Your task is not to guess randomly but to reason about transformations.</p> <p><b>Goal:</b></p> <ul style="list-style-type: none"> <li>- First, analyze the six figures (A, B, C, D, E, F).</li> <li>- Identify the shared underlying visual concept that appears in the majority of figures.</li> <li>- Then, determine which single figure does NOT follow this concept.</li> </ul> <p><b>Important:</b></p> <ul style="list-style-type: none"> <li>- Compare using any one of these concepts spacing, alignment, number, enclosure, symmetry, hollowness, topology or border thickness.</li> <li>- Focus on structural differences, not surface-level noise.</li> <li>- Eliminate options systematically before deciding.</li> </ul> <p><b>Finally:</b></p> <ul style="list-style-type: none"> <li>- Clearly state the shared concept.</li> <li>- Identify the odd-one-out.</li> <li>- Choose one of the option (A-F) which is the odd one out</li> </ul>	<p style="text-align: center;"><b>Mental Transformation</b></p> <p>Analyze the four figures labeled A, B, C, D in the image. Your goal is to identify which of these options is the correct rotational transformation of the figure labeled original shape. This is a MENTAL ROTATION puzzle. Your task is not to guess but to apply spatial reasoning.</p> <p><b>Goal:</b></p> <ul style="list-style-type: none"> <li>- Analyze the figure labeled "Original Shape."</li> <li>- Compare it with the four options (A, B, C, D).</li> <li>- Identify which option is the correct rotational transformation of the original figure.</li> </ul> <p><b>Important:</b></p> <ul style="list-style-type: none"> <li>- Use mental rotation reasoning; imagine rotating the original figure in your mind until it aligns with one of the options.</li> <li>- Focus on orientation, edges, colors, and relative placement of components.</li> <li>- Do not confuse rotation with reflection or mirroring.</li> <li>- Eliminate options that involve flipping or distortion instead of pure rotation.</li> </ul> <p><b>Finally:</b></p> <ul style="list-style-type: none"> <li>- State the correct option.</li> <li>- Give the final answer from one of the four options (A-D).</li> </ul>
<p style="text-align: center;"><b>Paper Folding</b></p> <p>The a paper of shape labeled Unfolded is folded following the pattern shown in the top row, in the sequence of fold 1, followed by fold 2 and so on. After the last fold a hole is punched as shown in the image of labeled final view, then after unfolding how would the paper look. This is a PAPER FOLDING WITH HOLE PUNCHING puzzle. Your task is not to guess but to apply step-by-step spatial reasoning.</p> <p><b>Goal:</b></p> <ul style="list-style-type: none"> <li>- Start with the paper labeled "Unfolded."</li> <li>- Follow the sequence of folds shown in the top row (Fold 1 then Fold 2 then so on).</li> <li>- After the final fold, a single hole is punched (as shown in the "Final View").</li> <li>- Predict how the paper will look once it is completely unfolded.</li> <li>- Compare your prediction with the four options (A, B, C, D) and identify the correct one.</li> </ul> <p><b>Important:</b></p> <ul style="list-style-type: none"> <li>- Use paper folding reasoning: each fold creates a symmetry line that replicates the hole across multiple regions when unfolded.</li> <li>- Track carefully how the hole's position maps onto the unfolded paper after each fold.</li> <li>- Do not confuse folding with rotation or reflection: the key is symmetry about fold lines.</li> <li>- Eliminate options that show the wrong number of holes, incorrect symmetry, or misplaced positions.</li> </ul> <p><b>Finally:</b></p> <ul style="list-style-type: none"> <li>- State the correct option.</li> <li>- Give the final answer from one of the four options (A-D).</li> </ul>	<p style="text-align: center;"><b>Dynamic Structural Correspondence</b></p> <p>Follow the transformation of the shapes in the top row, starting from <math>t=0.0</math> to <math>t=0.75</math>. Analyze the four figures labeled A, B, C, D in the image. Your goal is to identify which of these options is the fifth image at <math>t=1.0</math>. This is a DYNAMIC TRANSFORMATION puzzle. Your task is not to guess but to reason about how shapes evolve over time.</p> <p><b>Goal:</b></p> <ul style="list-style-type: none"> <li>- Observe the sequence of shapes in the top row, starting from <math>t=0.0</math> up to <math>t=0.75</math>.</li> <li>- Identify the transformation pattern that occurs across these time steps for both the figures (e.g., rotation, expansion, contraction, bouncing).</li> <li>- Predict what the shape will look like at the next time step (<math>t=1.0</math>).</li> <li>- Compare your prediction with the four options (A, B, C, D).</li> <li>- Select the correct option.</li> </ul> <p><b>Important:</b></p> <ul style="list-style-type: none"> <li>- Use temporal reasoning: track how elements change gradually from one frame to the next.</li> <li>- Consider orientation, position, size, and structural transformations of each component.</li> <li>- Focus on consistency: the next state must logically continue the progression, not reset or break it.</li> <li>- Eliminate options that distort the sequence, skip a step, or contradict the observed trend.</li> </ul> <p><b>Finally:</b></p> <ul style="list-style-type: none"> <li>- State the correct option.</li> <li>- Give the final answer from one of the four options (A-D).</li> </ul>
<p style="text-align: center;"><b>Mental Composition</b></p> <p>Analyze the six figures labeled A, B, C, D in the image. Your goal is to identify which of these options can be constructed by folding the top 2D figure labeled as Original Shape. This is a 3D UNFOLDING / PAPER FOLDING puzzle. Your task is not to guess but to apply spatial reasoning.</p> <p><b>Goal:</b></p> <ul style="list-style-type: none"> <li>- Analyze the "Original Shape," which is a 2D net of a 3D figure.</li> <li>- Compare it with the four 3D options (A, B, C, D).</li> <li>- Identify which option can be constructed by folding the 2D net into a 3D solid.</li> </ul> <p><b>Important:</b></p> <ul style="list-style-type: none"> <li>- Use 3D mental folding reasoning: imagine folding the flat net step by step to form the 3D figure.</li> <li>- Pay attention to which faces touch each other when folded.</li> <li>- Track the relative positions of markings, edges, and orientations.</li> <li>- Eliminate options that are impossible (e.g., faces that would overlap incorrectly, reversed orientations, or extra/missing adjacencies).</li> </ul> <p><b>Finally:</b></p> <ul style="list-style-type: none"> <li>- State the correct option.</li> <li>- Give the final answer from one of the four options (A-D).</li> </ul>	<p style="text-align: center;"><b>Symmetric Structures</b></p> <p>Analyze the four figures labeled A, B, C, D in the image. Your goal is to identify the underlying visual concept that is shared by the majority of these figures and then determine which figure does not adhere to this concept. This is a SYMMETRY DETECTION + ODD-ONE-OUT puzzle. Your task is not to guess but to apply structured visual reasoning.</p> <p><b>Goal:</b></p> <ul style="list-style-type: none"> <li>- Analyze the four figures (A, B, C, D).</li> <li>- Identify the underlying symmetry property (e.g., reflectional symmetry, rotational symmetry, or asymmetry) that is shared by the majority of figures.</li> <li>- Then, determine which single figure does not follow this property.</li> </ul> <p><b>Important:</b></p> <ul style="list-style-type: none"> <li>- Focus on the type of symmetry: <ul style="list-style-type: none"> <li>- Reflection (mirror symmetry across a vertical, horizontal, or diagonal axis)</li> <li>- Rotational (the figure looks the same after a certain degree of rotation)</li> </ul> </li> <li>- Absence of symmetry (asymmetric shapes).</li> <li>- Compare each figure systematically.</li> <li>- Eliminate the one that violates the majority symmetry rule.</li> </ul> <p><b>Finally:</b></p> <ul style="list-style-type: none"> <li>- Clearly state the shared symmetry property.</li> <li>- Identify the odd-one-out figure.</li> <li>- The final answer is the odd one out from one of the four options (A-D).</li> </ul>
<p style="text-align: center;"><b>Hierarchical Pattern Equivalence</b></p> <p>Analyze the four figures labeled A, B, C and D in the image. Your goal is to identify the underlying visual concept that is shared by the majority of these figures and then determine which figure does not adhere to this concept. This is a RECURSIVE STRUCTURE + ODD-ONE-OUT puzzle. Your task is not to guess but to apply structured pattern reasoning.</p> <p><b>Goal:</b></p> <ul style="list-style-type: none"> <li>- Analyze the four figures (A, B, C, D).</li> <li>- Identify the underlying recursive or hierarchical structure property that is shared by the majority of the figures.</li> <li>- Then, determine which single figure does not follow this recursive property.</li> </ul> <p><b>Important:</b></p> <ul style="list-style-type: none"> <li>- Focus on recursion in the design: shapes that are built from smaller copies of themselves or that repeat a structural rule at multiple levels.</li> <li>- Look for hierarchical patterns: a large shape composed of smaller versions, nested repetition, or self-similarity across scales.</li> <li>- Compare systematically to see which figure breaks the recursive structure rule.</li> <li>- Eliminate the option that does not match the recursive or hierarchical pattern.</li> </ul> <p><b>Finally:</b></p> <ul style="list-style-type: none"> <li>- The final answer is the odd one out from one of the four options (A-D).</li> </ul>	<p style="text-align: center;"><b>Visual Relation Abstraction</b></p> <p>Analyze the six figures labeled A, B, C, D, E, and F in the image. Your goal is to identify the underlying visual concept that is shared by the majority of these figures and then determine which figure does not adhere to this concept. This is a puzzle of ABSTRACT SHAPES. Your task is not to guess but to apply structured pattern reasoning.</p> <p><b>Goal:</b></p> <ul style="list-style-type: none"> <li>- Analyze the six figures (A, B, C, D, E, F).</li> <li>- Identify the underlying abstract property by the majority of the figures.</li> <li>- Then, determine which single figure does not follow this recursive property.</li> </ul> <p><b>Important:</b></p> <ul style="list-style-type: none"> <li>- Compare systematically to see which figure breaks the abstract structure rule.</li> <li>- Eliminate the option that does not match the recursive or hierarchical pattern.</li> </ul> <p><b>Finally:</b></p> <ul style="list-style-type: none"> <li>- The final answer is the odd one out from one of the six options (A-F). Choose that option as the final answer</li> </ul>

Figure 20: Meta Task Prompt for all the tasks.

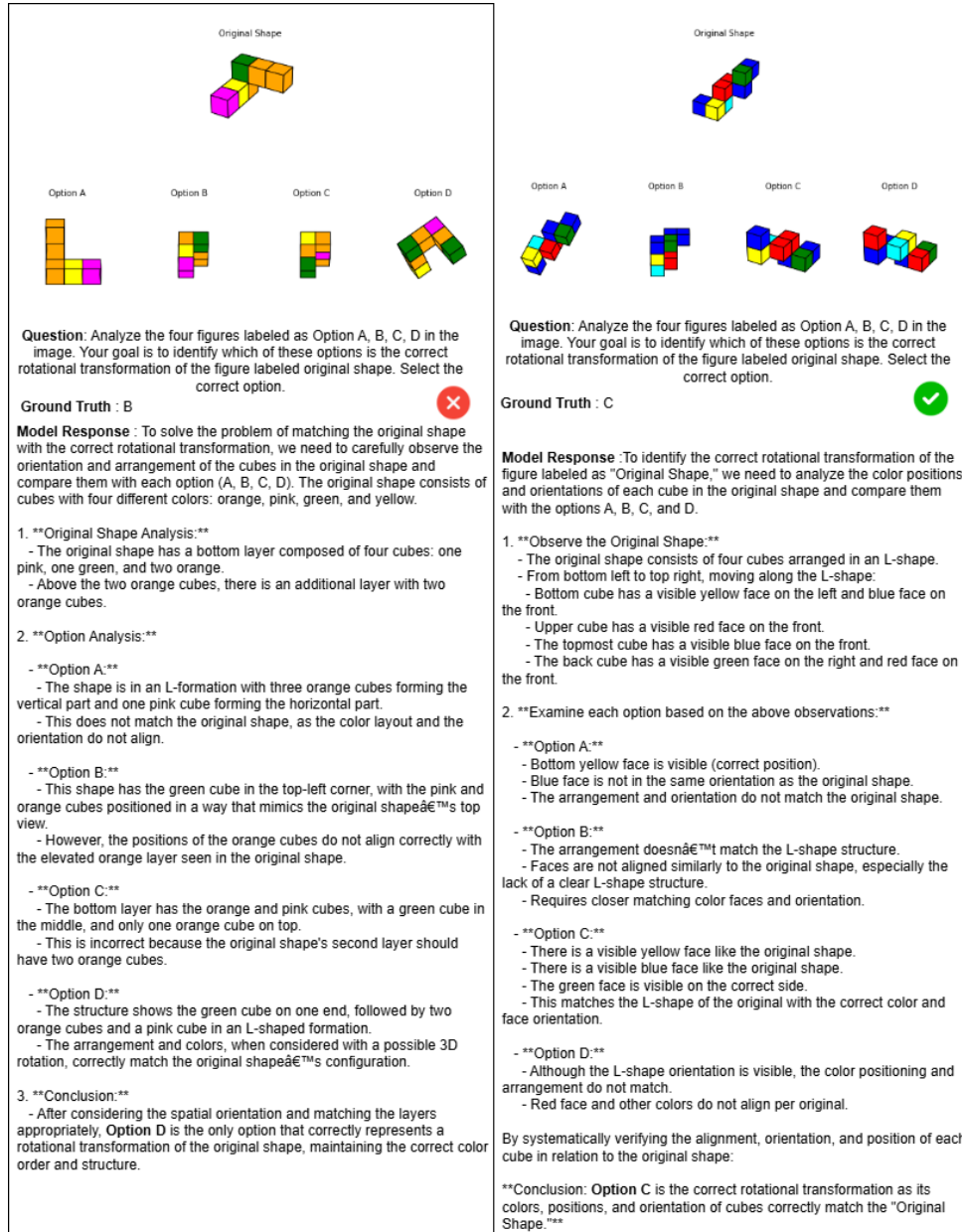
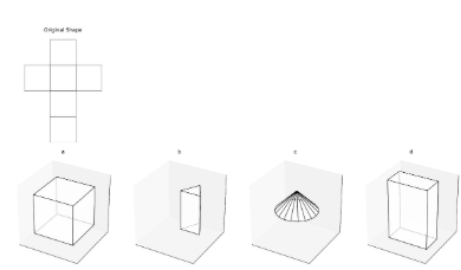


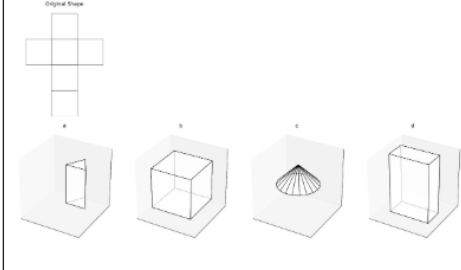
Figure 21: Reasoning Trace Analysis for GPT-4o on Mental Transformation Task : (Left) Incorrect Answer, (Right) Correct Answer. In the *conclusion* of the reasoning traces, the final answer selection is done. Analyzing the reasoning traces for GPT-4o for the Mental Transformation Task (MT) shows that the models are relying on color as heuristic to try to match the options with the original shape. This reasoning traces suggests that the model's functional accuracy may not be consistent with the mechanistic equivalent of the capabilities required to reason about these solutions and reach the correct answer



**Question:** Analyze the four figures labeled A, B, C, D in the image. These are 3D Figures. Your goal is to identify which of these options can be constructed by folding the top 2D figure labeled as Original Shape. Select the correct option.

**Ground Truth :** A ✘

**Model Response :** The original shape at the top is a net of a cube. When folded, it forms a cube, which matches with image "d" in the row of shapes shown below. The net consists of six square faces, arranged in a pattern that allows them to be folded into the three-dimensional cube shown in "d".



**Question:** Analyze the four figures labeled A, B, C, D in the image. These are 3D Figures. Your goal is to identify which of these options can be constructed by folding the top 2D figure labeled as Original Shape. Select the correct option.

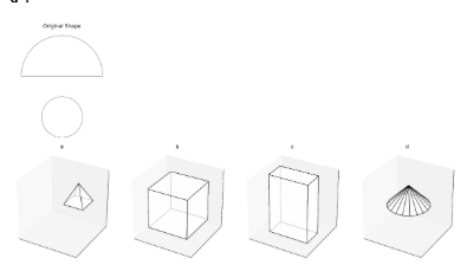
**Ground Truth :** B ✔

**Model Response :** The image shows four different 3D shapes that could potentially be constructed from the given 2D net labeled "Original Shape." The 2D net is a cross-shaped configuration of squares, which resembles the net of a cube.

Let's evaluate each option:

- Option (a) is a triangular prism, which cannot be formed from the net provided, as it requires triangles, not squares.
- Option (b) is a cube, which can be formed from the given net of squares.
- Option (c) is a cone, which cannot be constructed from a net of squares, as it requires a circular base.
- Option (d) is a rectangular prism (or cuboid), but the net matches exactly with a cube, as all squares are the same size.

Therefore, **option (b)** is the correct 3D shape that can be constructed from the given net of a cube.



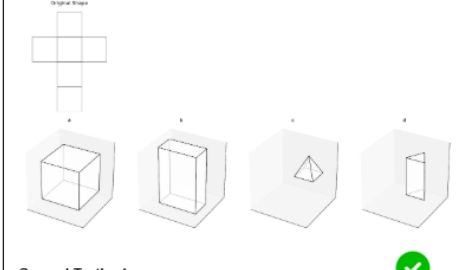
**Question:** Analyze the four figures labeled A, B, C, D in the image. These are 3D Figures. Your goal is to identify which of these options can be constructed by folding the top 2D figure labeled as Original Shape.

**Ground Truth :** D ✘

**Model Response :** The original shapes at the top are a semicircle and a full circle. The images labeled a, b, c, and d show different 3D objects.

- Image A shows a cone, which can be formed by rotating a semicircle around its straight edge.
- Image B is a cube, which is not directly related to either the semicircle or circle shown above.
- Image C shows a cylinder, which can be formed by rotating a rectangle (not a semicircle or circle) around one of its edges.
- Image D shows a hemisphere, which is formed by rotating a semicircle around its diameter.

Therefore, images **A and D** are the correct 3D interpretations through rotation of the original semicircle shape.



**Ground Truth :** A ✔

**Model Response :** The image shows a shape net labeled "Original Shape" at the top, and below, there are four separate 3D shapes labeled a, b, c, and d.

- The "Original Shape" is a net for a cube, consisting of six squares arranged in a cross pattern.
- **Shape a** is a cube, which matches the original shape net.

Figure 22: Reasoning Trace Analysis for GPT-4o on Mental Composition Task: (Left) Incorrect Answer, (Right) Correct Answer. The reasoning trace shows that when GPT-4o correctly identified the unfolded figure as the cube's net, it was able to infer the correct folded shape and select the right answer. However, in cases where it failed to recognize the net structure, the model could not mentally simulate the folding operation, leading to incorrect predictions.

**Question:** The a paper of shape labeled Unfolded is folded following the pattern shown in the top row, in the sequence of fold 1, followed by fold 2 and so on. After the last fold a hole is punched as shown in the image of labeled final view, then after unfolding how would the paper look. Analyze the four figures labeled A, B, C, D in the image, and select the correct option

**Ground Truth :** B

**Model Response :** The image shows a sequence of folds made on a square sheet of paper, resulting in a triangular shape, denoted as "Final View." Below the folding sequence are four options: A, B, C, and D, which show possible crease patterns on an unfolded square with dots that represent holes after punching through the paper.

To find the correct option that corresponds to the final triangular shape with holes in it, the folding steps need to be analyzed:

- Fold 1:** involves folding the top half of the square down to meet the bottom half, creating a smaller triangle.
- Fold 2:** involves folding one side of the triangle inward, further aligning the creases, resulting in an even smaller triangle.

The "Final View" presents a triangular shape with two holes positioned in such a way that they match the folding sequence.

Among the options A, B, C, and D:

- Option A: Shows four dots near the corners.
- Option B: Shows four dots near the center.
- Option C: Shows two dots near two corners and two in the middle.
- Option D: Shows two dots near the bottom and two dots near the top.

The correct option is "D": It matches the expected positions of holes after the folding sequence and provides the same orientation seen in the final triangular shape, with dots positioned in alignment with the folds and punching locations.

**Question:** The a paper of shape labeled Unfolded is folded following the pattern shown in the top row, in the sequence of fold 1, followed by fold 2 and so on. After the last fold a hole is punched as shown in the image of labeled final view, then after unfolding how would the paper look. Analyze the four figures labeled A, B, C, D in the image, and select the correct option

**Ground Truth :** D

**Model Response :** The image depicts a sequence of folding operations on a hexagon, with a final view showing the result, and four options (A, B, C, D) of where dots appear on the hexagon after folding.

To determine the correct option:

- Unfolded Hexagon:** Start with a plain hexagon.
- Fold 1:** The hexagon appears to be folded such that the top portion is brought down horizontally.
- Fold 2:** The resulting shape from Fold 1 is further folded vertically.

The "Final View" indicates that only a triangular section of the hexagon is visible after these folds, with dots marked on the corners.

To match this with one of the options (A, B, C, D), consider the visibility and position of corners after the described folds.

Analyzing each option:

- Option A: shows dots on three consecutive corners at the bottom.
- Option B: places dots on opposite corners.
- Option C: has dots on the upper and middle parts, suggesting a different fold scenario.
- Option D: shows dots in positions consistent with sections that overlap in a triangular fold.

Since the task is to determine which option (A, B, C, or D) matches the final view after the given folds, the answer is "D". It best represents the dot placement as seen in the "Final view" after considering the fold operations described.

Figure 23: Reasoning Trace Analysis for GPT-4o on Paper Folding Task: (Left) Incorrect Answer, (Right) Correct Answer. Analysis of the reasoning traces shows that while the model correctly identifies how the paper is folded, its option analysis and final answer selection provide no evidence of tracking the holes through the unfolding process. Instead, the model appears to rely on superficial spatial matching between hole positions in the folded and unfolded states, rather than mentally simulating the unfolding operations to derive the correct answer.

**Question:** "Analyze the six figures labeled A, B, C, D, E, and F in the image. Your goal is to identify the underlying visual concept that is shared by the majority of these figures and then determine which figure does not adhere to this concept. Select the Correct Option

**Ground Truth :** E

**Model Response :** The image presents a pattern recognition problem where different shapes are arranged in groups. Here's an analysis of each group in the image:

- A row of five black square shapes.
- A row of five black diamond shapes.
- A row of five black square shapes.
- A row of five black star shapes.
- Three black shapes: A square, a diamond, and a circle arranged close to each other, forming a triangular pattern.
- A row of three black circular shapes.

The puzzle seems to involve different shapes arranged in different patterns or sequences. The task is likely to identify the rule or pattern and determine the correct continuation or identify any outliers. Based on the arrangement and repetition of the shapes, one could deduce that rows (a), (b), (c), and (d) contain consistent shapes while (e) and (f) disrupt this pattern by having different shapes or fewer shapes in a row.

**Question:** "Analyze the six figures labeled A, B, C, D, E, and F in the image. Your goal is to identify the underlying visual concept that is shared by the majority of these figures and then determine which figure does not adhere to this concept. Select the Correct Option

**Ground Truth :** D

**Model Response :** The image shows a sequence of patterns labeled (A) to (F), each consisting of different shapes arranged in a line. Here's a breakdown of each pattern:

- (A) consists of four circles.
- (B) consists of four squares.
- (C) consists of four circles.
- (D) consists of one circle, one diamond, one cross, and one square.
- (E) consists of five circles.
- (F) consists of six stars.

The patterns in (A), (B), (C), (E), and (F) consist of identical shapes repeated a certain number of times, while (D) consists of four distinct shapes. The number of identical shapes in each pattern is different, suggesting a progression or rule related to the number of shapes used in each example.

Figure 24: Reasoning Trace Analysis for GPT-4o on Visual Conceptual Slippage Task: (Left) Incorrect Answer, (Right) Correct Answer. Analysis of the reasoning trace suggests that the model relies primarily on superficial visual cues and perceptual artifacts when evaluating the options, rather than grasping the underlying abstract relations shared across the figures. The model arrives at the correct answer only because the correct option exhibits a distinct visual difference, not due to genuine conceptual understanding.

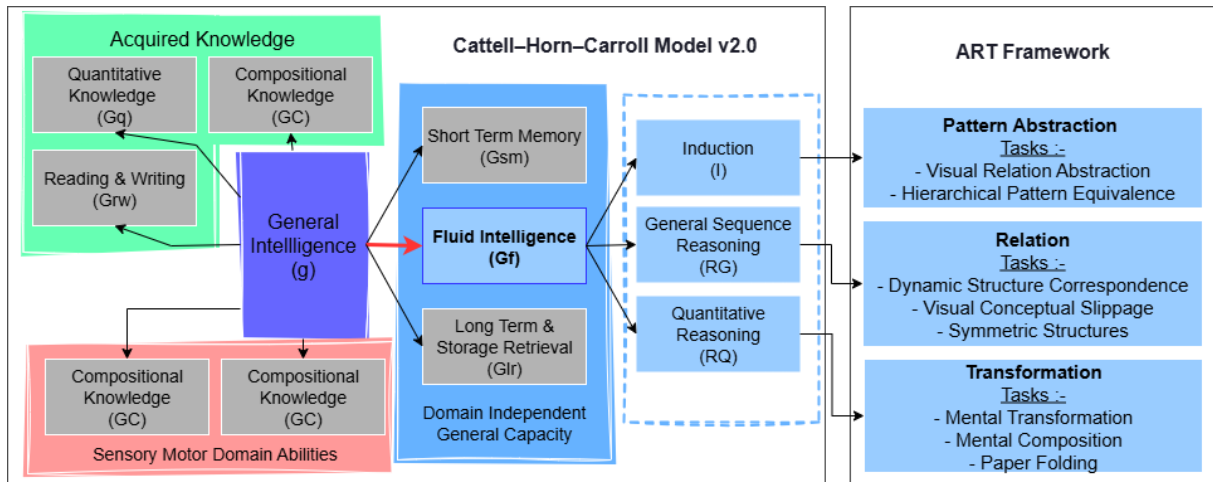


Figure 25: **Mapping Carroll's Three Stratum Theory to the Mind's Eye ART taxonomy:** The figure illustrates how **fluid intelligence (Gf)**: a core construct in Carroll's Three Stratum Theory of cognitive abilities—corresponds to the three visucognitive dimensions evaluated in **Mind's Eye: Abstraction, Relation, and Transformation (ART)**. Arrows denote the conceptual linkage from fluid reasoning to these visual faculties, grounding the benchmark's task design in established psychometric theory.

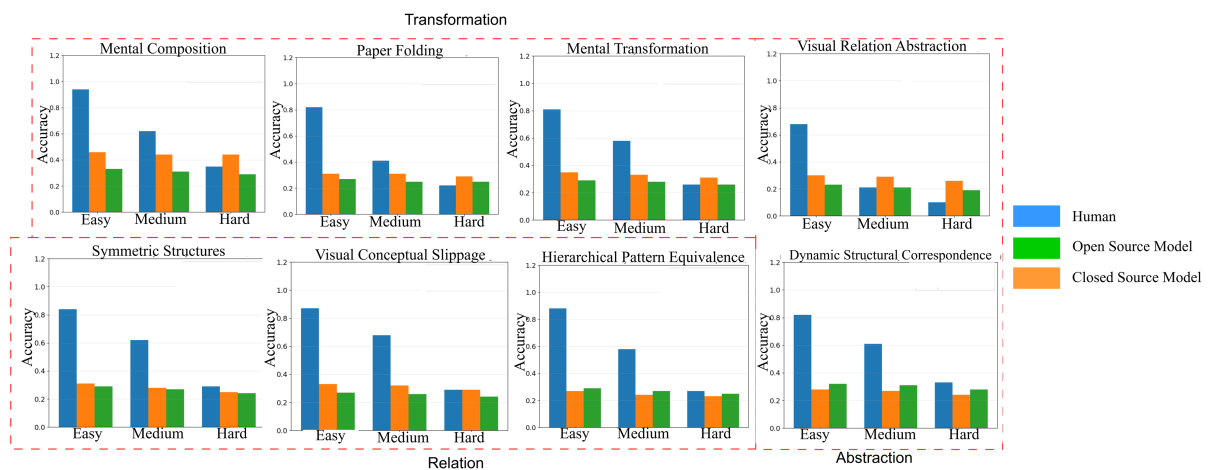


Figure 26: **Performance across ART dimensions by difficulty level.** Both closed-source and open-source models struggle across all dimensions with flat difficulty curves (0.20-0.45 accuracy), while human experts maintain robust performance (>0.80) across easy, medium, and hard tasks. Each bar represents the macro-average accuracy for a task across all models in that category (see Table 2).