Do You See Me : A Multidimensional Benchmark for Evaluating Visual Perception in Multimodal LLMs

Aditya Kanade
Microsoft Research India

Tanuja Ganu[*]
Microsoft Research India

Multimodal Large Language Models (MLLMs) show reasoning promise, yet their visual perception is a critical bottleneck. Paradoxically, MLLMs sometimes produce correct answers while misinterpreting crucial visual elements, masking these underlying perception failures. Our preliminary analysis on a joint perception-reasoning dataset revealed that 29% of correct reasoning answers from a leading MLLM contained perception errors. To systematically study visual perception abilities of MLLMs, we introduce **Do You See Me**- a scalable, programmatically generated benchmark with 1758 images and 2612 questions across seven core subtasks spanning 2D and 3D variants (twelve total tasks) providing parametric control over difficulty levels. The benchmark tasks are inspired by human psychology. Our evaluation of eleven leading MLLMs reveals a stark deficit: humans achieve 95.83% accuracy, while top MLLMs average below 50%. This performance gap widens drastically as task complexity increases. Further diagnostics show: (1) supervised finetuning offers only modest gains (11%), (2) models tend to exploit task "shortcuts" like MCQ formats over detailed visual analysis, and (3) Chain-of-Thought prompting can degrade complex visual tasks by verbalizing images into lossy text. These findings expose the foundational perception limits in current MLLMs and highlight the need for robust visual perception improvements in MLLMs. The benchmark dataset, source code and evaluation scripts are available at[*].

1 Introduction

Multimodal Large Language Models (MLLMs) demonstrate moderate reasoning capabilities, yet a striking paradox remains: models can produce correct answers while fundamentally misinterpreting crucial visual elements. This phenomenon masks underlying perceptual failures. For instance, a leading model like Claude Sonnet-3.5 [Anthropic, 2024] can arrive at the correct final answer for a logical pattern-completion task despite misinterpreting the visual components of the puzzle (see Figure 1). Such cases, where correct reasoning outcomes mask foundational perception errors, raise a critical question: do MLLMs truly *see* what they reason about?

This perceptual brittleness is a critical issue, as robust visual perception forms the foundation for trustworthy higher-order reasoning [Chalfant and Scheffelin, 1969]. The problem has persisted largely because current evaluation paradigms are not equipped to diagnose it. While benchmarks such as MMVP [Tong et al., 2024b], CV-Bench [Tong et al., 2024a], and MVP-Bench [Li et al., 2024a] provide valuable assessments of general multimodal capabilities, their methodologies are not primarily designed for the isolated and fine-grained analysis of perception. Specifically, their reliance on manual annotation limits scalability, they lack systematic control over task difficulty, and many sub-tasks within these benchmarks often treat perception as a secondary objective. This clear gap in current evaluation methodology motivates our investigation, which is guided by three key research questions: **RQ1:** *To what extent do correct reasoning outcomes mask underlying perceptual failures?*; **RQ2:** *How well do MLLMs perform on core visual perception skills across controlled difficulty levels compared to humans ?*; and **RQ3:** *What structural limitations and failure modes underlie MLLMs' visual perception deficits ?*

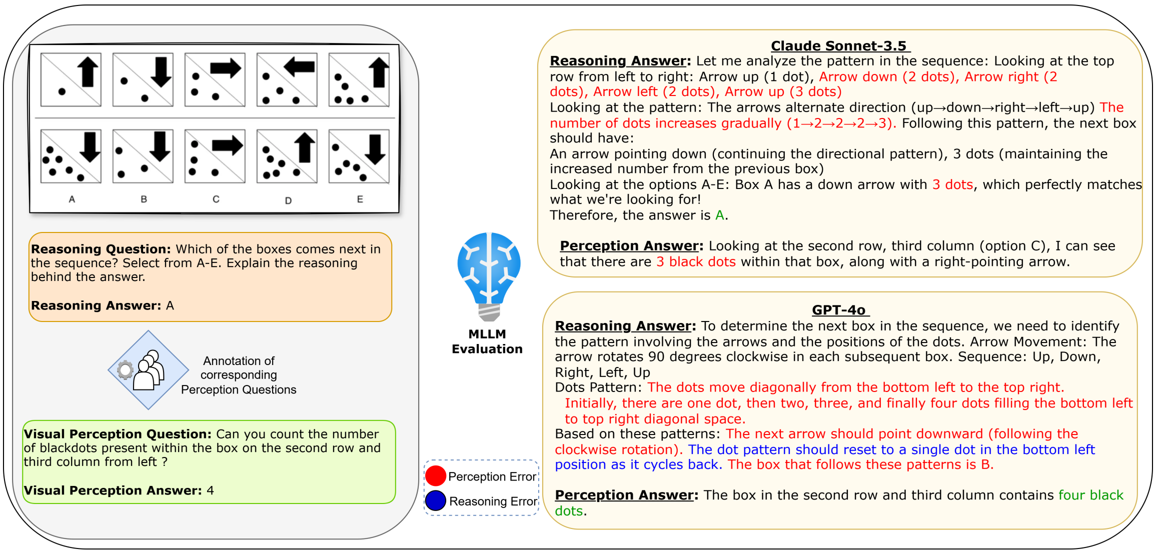
Our investigation begins by probing the premise of RQ1 with a preliminary study. We curated a joint perception-reasoning dataset of 150 logic-based IQ test problems drawn from MathVista [Lu et al., 2023] and LogicVista [Xiao et al., 2024]. Our analysis provided initial, striking evidence: for Claude-Sonnet-3.5, 29% of its correctly answered reason-

---

[*] Correspondence to: Tanuja Ganu, `tanuja.ganu@microsoft.com`

[*] `https://github.com/microsoft/Do-You-See-Me`

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7285–7326
March 24-29, 2026 ©2026 Association for Computational Linguistics



Reasoning Question: Which of the boxes comes next in the sequence? Select from A-E. Explain the reasoning behind the answer.

Reasoning Answer: A

Visual Perception Question: Can you count the number of blackdots present within the box on the second row and third column from left ?

Visual Perception Answer: 4

Claude Sonnet-3.5

Reasoning Answer: Let me analyze the pattern in the sequence: Looking at the top row from left to right: Arrow up (1 dot), Arrow down (2 dots), Arrow right (2 dots), Arrow left (2 dots), Arrow up (3 dots)
Looking at the pattern: The arrows alternate direction (up→down→right→left→up) The number of dots increases gradually (1→2→2→2→3). Following this pattern, the next box should have:
An arrow pointing down (continuing the directional pattern), 3 dots (maintaining the increased number from the previous box)
Looking at the options A-E: Box A has a down arrow with 3 dots, which perfectly matches what we're looking for!
Therefore, the answer is A.

Perception Answer: Looking at the second row, third column (option C), I can see that there are 3 black dots within that box, along with a right-pointing arrow.

GPT-4o

Reasoning Answer: To determine the next box in the sequence, we need to identify the pattern involving the arrows and the positions of the dots. Arrow Movement: The arrow rotates 90 degrees clockwise in each subsequent box. Sequence: Up, Down, Right, Left, Up
Dots Pattern: The dots move diagonally from the bottom left to the top right. Initially, there are one dot, then two, three, and finally four dots filling the bottom left to top right diagonal space.
Based on these patterns: The next arrow should point downward (following the clockwise rotation). The dot pattern should reset to a single dot in the bottom left position as it cycles back. The box that follows these patterns is B.

Perception Answer: The box in the second row and third column contains four black dots.

Figure 1: *Visual Misinterpretations in Popular Multimodal LLMs*

ing questions still contained fundamental visual perception errors. This finding empirically affirms that correct reasoning outcomes can indeed mask perception failures, underscoring the need for a more rigorous and systematic evaluation of visual perception in MLLMs.

Towards this end, we introduce **Do You See Me**, a benchmark designed to systematically evaluate core visual perception skills in MLLMs. Grounded in established frameworks from human psychology [Chalfant and Scheffelin, 1969], the benchmark isolates fundamental abilities such as visual discrimination, figure-ground perception, and spatial relations. To overcome the limitations of prior work, **Do You See Me** is programmatically generated, ensuring scalability and enabling fine-grained analysis through parametric control over task difficulty. The benchmark comprises 1758 images and 2612 questions across seven core subtasks, presented in both 2D and 3D photorealistic scenes variants (twelve total tasks) to assess performance across varied visual contexts. Our evaluations on this benchmark reveal a striking deficit: while human participants achieve 95.83%[†] accuracy, the best-performing MLLMs average below 50%, with this performance chasm widening drastically as task complexity increases.

Beyond this performance gap, the benchmark's controlled design enables a deeper diagnostic analysis of MLLM failure modes. For instance, by reformulating an open-ended task into a multiple-choice question, we observed an accuracy jump from 23.19% to 41.80%, revealing a reliance on task cues over genuine visual analysis. Our analysis shows that Chain-of-Thought (CoT) prompting can hurt performance on tasks requiring holistic visual understanding, suggesting that translating rich visual input into step-by-step text introduces information loss. Finally, a large-scale supervised finetuning (SFT) experiment yielded only modest gains (approx. 11%), confirming that these perceptual limitations are foundational and not easily resolved by data scaling alone.

In summary, this work makes three primary contributions: (i) we introduce a scalable, programmatically generated benchmark, **Do You See Me**, designed to systematically evaluate core visual perception skills; (ii) we present a curated joint dataset to disentangle reasoning success from underlying perceptual accuracy, which provides the empirical motivation for our work; and (iii) we provide a comprehensive evaluation of eleven leading MLLMs that quantifies their perceptual deficits and provides a diagnostic analysis of their failure modes.

2 Related Work

While many benchmarks evaluate the high-level reasoning of Multimodal Large Language Models (MLLMs) [Lu et al., 2023, Zhang et al., 2024, Lu et al., 2022, Saikh et al., 2022, Li et al., 2024b, Yue et al., 2023], a growing body of work reveals that their foundational visual perception is a critical bottleneck [Zhang et al., 2024, Wu et al., 2024]. For instance, recent studies demonstrate that even state-of-the-art models fail at trivial visual tasks like counting overlapping shapes or identifying a circled letter [Rahmanzadehgervi et al., 2025].

[†] Macro Average of 2D and 3D performance

Such failures have been theoretically linked to the cognitive science concept of the binding problem, with MLLMs showing human-like performance degradation in tasks requiring feature binding, like conjunctive visual search [Campbell et al., 2025]. This highlights a clear gap between the models' touted complex reasoning abilities and their actual performance on basic perceptual tasks.

Table 1: Comparison of multimodal benchmarks across key dimensions.

Benchmark	**Automated Data Collection**	**Difficulty Levels**	**Categories**	**Size**	**Human Annotation**
MVP-Bench [Li et al., 2024a]	✗	✗	13 high-level 5 low-level	605	✓
MME [Fu et al., 2023]	✗	✗	14	1,147	✗
CV-Bench [Tong et al., 2024a]	✗	✗	4	2,638	✗
MMVP [Tong et al., 2024b]	✗	✗	9	300	✓
Do You See Me (Ours)	✓	✓	7	**2,612***	✓

*Can be extended in an automated manner.

Existing benchmarks that focus on visual perception, such as MMVP [Tong et al., 2024b], CV-Bench [Tong et al., 2024a], MME [Fu et al., 2023], and MVP-bench [Li et al., 2024a], present several key methodological limitations. As summarized in Table 1, they are often derived from common datasets like ImageNet [Russakovsky et al., 2014] and COCO [Lin et al., 2014], posing a risk of data contamination since these are likely part of MLLM training corpora. Furthermore, their reliance on manual annotation limits scalability, and they generally lack mechanisms for parametrically controlling task difficulty, which hinders a more fine-grained analysis of model capabilities.

Drawing inspiration from human psychology for a more principled evaluation, our work is grounded in established motor-free assessments such as the Test of Visual Perceptual Skills (TVPS) [Gardner, 1988] and the Motor-Free Visual Perception Test (MVPT) [Colarusso, 2003]. These tests are designed to systematically isolate and assess core perceptual dimensions [Chalfant and Scheffelin, 1969] while avoiding confounding factors like motor skills [Colarusso, 2003, Gardner, 1988, Hammill et al., 2016]. Accordingly, we introduce the Do You See Me benchmark, which leverages these established perceptual categories and a programmatic generation pipeline to offer a scalable and difficulty-controlled evaluation of MLLM visual skills, directly addressing the limitations of prior benchmarks.

3 Preliminary Study - Joint Visual Perception and Reasoning Dataset

Most of the existing benchmarks assess MLLM's visual reasoning capabilities by solely relying on the final answer based accuracy. However, this approach can obscure the exact source of errors. In particular, three primary sources of error can arise: 1) Visual Perception- inaccuracies in identifying or interpreting elements in the provided image, 2) Reasoning- errors in the logical or conceptual steps used to arrive at the final answer, or 3) Arithmetic- mistakes in performing numerical or algebraic calculations. To accurately distinguish between different error sources, it is essential to analyze not only final answers but also the reasoning chains. We introduce a *joint perception-reasoning* dataset specifically designed to separate visual perception errors from higher-level reasoning failures.

Why IQ-Type Questions? IQ-style diagrammatic questions primarily feature basic geometric shapes and patterns, minimizing reliance on domain-specific knowledge. This allows for a focused evaluation of visual perception and reasoning skills without introducing extraneous complexity.

3.1 Data Collection

Our dataset is drawn from two established visual reasoning benchmarks: MathVista [Lu et al., 2023] and LogicVista [Xiao et al., 2024]. We selected logic-based tasks centered around geometric shapes and pattern recognition from:

- *IQtest* subset of MathVista (focusing on spatial and pattern-based problems).
- *Diagrams* subset of LogicVista (pattern completion tasks).

These subsets feature universally understood shapes in controlled layouts, allowing systematic evaluation of perception and reasoning. We curated 15 problems from MathVista's *IQtest* and 135 from LogicVista's *Diagrams*, yielding a total of 150 examples in our final dataset.

3.2 Data Annotation

We extend each original problem (I, R, A_R)—where I is the image, R is the reasoning question, and A_R is the corresponding ground-truth answer—by adding a *visual perception* question P with its ground-truth A_P. The extended sample is thus: $\big(I,\ (R, A_R),\ (P, A_P)\big)$. The perception questions are manually devised such that they are directly relevant to each reasoning question (e.g., "How many triangles are in the figure?”). More details on dataset creation and distribution are added in Appendix C.

4 Do You See Me

Human psychology systematically categorizes human visual perception as a combination of five core abilities [Chalfant and Scheffelin, 1969]: visual discrimination, the ability to recognize dominant features (e.g., position, shape, form, color); visual figure-ground, the ability to distinguish the main object from its background; visual memory, the ability to remember sequences of presented images; visual closure, the ability to complete partially obscured shapes; and visual spatial, the ability to perceive positions of objects relative to oneself and to other objects. Assessments such as the *Test of Visual Perception Skills (TVPS)* [Gardner, 1988] and *Motor-Free Visual Perception Test (MVPT)* [Colarusso, 2003] systematically evaluate these abilities through structured visual tasks and associated questions. While there are fundamental differences between MLLM and human visual processing, the perceptual categories defined in human psychology offer an established framework for structured evaluation of analogous capabilities in MLLMs. Building on these principles, we introduce the **Do You See Me** benchmark, a fully automated test suite designed to evaluate MLLM visual perception across dimensions analogous to human tests.

4.1 General Benchmark Design

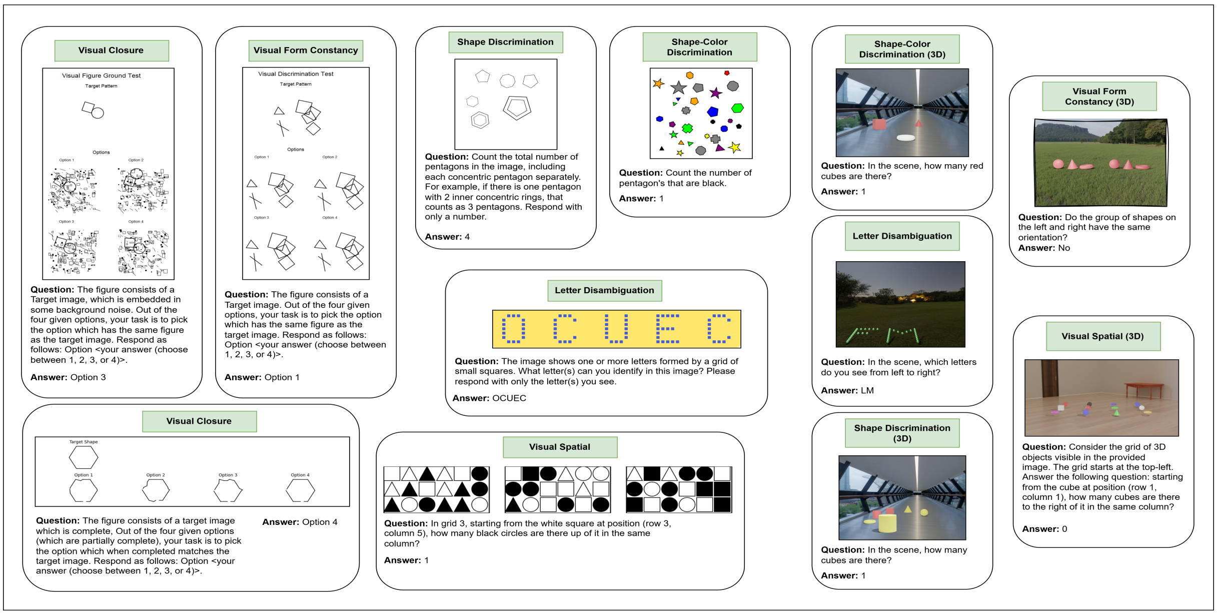
Consistent with many human visual perception tests, stimuli in the **Do You See Me** benchmark are primarily generated in a 2D setting using Scalable Vector Graphics (SVG), allowing for precise control over geometric properties. To enable a broader analysis of MLLM capabilities on more complex scenes, the benchmark also incorporates photorealistic 3D rendered settings for several subtasks. These 3D scenes are produced using Blender [Community, 2018], drawing on a prior work for scene setup [Abbavaram Gowtham Reddy and Balasubramanian, 2022]. For 3D tasks, standard shapes include five basic forms: sphere, cube, cone, cylinder, and torus. Correspondingly, 2D tasks utilize fundamental SVG-generated geometric figures such as capsules, stars, hexagons, circles, pentagons, rectangles, and triangles, Table 2 lists the distribution of 2D and 3D subtasks.

Table 2: Dataset statistics for 2D and 3D visual perception tasks in Do You See Me benchmark.

Set.	**Task**	**Imgs**	**Ques.**	**Type**
2D	Shape Discrimination	241	241	Int
	Joint Shape-Color	90	408	Int
	Letter Discrimination	135	135	Text
	Form Constancy	270	270	MCQ
	Spatial Grids	270	806	Int
	Visual Figure-Ground	90	90	MCQ
	Visual Closure	166	166	MCQ
3D	Shape Discrimination	120	120	Int
	Joint Shape-Color	120	120	Int
	Letter Discrimination	96	96	Text
	Form Constancy	80	80	MCQ
	Spatial Grids	80	80	Int
Total (2D)		**1,262**	**2,116**	
Total (3D)		**496**	**496**	
Overall		**1,758**	**2,612**	

The **Do You See Me** benchmark probes distinct facets of visual perception (in 2D and photorealistic 3D) in MLLMs through seven core subtasks. *Shape Discrimination* involves identifying or counting specified shapes, often in cluttered or occluded scenes, and requires a numeric answer. *Joint Shape-Color Discrimination* requires linking shapes with their designated colors, also necessitating a numeric response. *Letter Disambiguation* focuses on recognizing textual characters despite varied presentation conditions, with answers provided as text. The *Form Constancy* task challenges MLLMs to identify the correct matching group from four options; the other three present slight transformations (e.g., size, rotation). *Spatial Grids* assesses comprehension of object arrangements and relationships within grid structures. In *Visual Figure-Ground*, the aim is to select a target pattern amidst a noisy background. *Visual Closure* requires choosing the option that correctly completes a partially presented shape. A distinct 3D version of *Visual Figure-Ground* is not included, as our photorealistic 3D rendered scenes inherently test this ability. Furthermore, a 3D equivalent for *Visual Closure* is omitted due to ambiguities in judging 3D shape closure from a single camera perspective. Table 6 and Table 7 defines parameters for fine-grained control over image generation. To establish difficulty-modulating parameter ranges for each subtask, a preliminary evaluation was conducted. These limits were determined by observing when the GPT-4o [OpenAI et al., 2024a] model started to exhibit a high degree of failure.

A detailed description of the benchmark’s construction, including the specific control parameters for modulating difficulty in each subtask, can be



Question: The figure consists of a Target image, which is embedded in some background noise. Out of the four given options, your task is to pick the option which has the same figure as the target image. Respond as follows: Option <your answer (choose between 1, 2, 3, or 4)>.

Answer: Option 3

Question: The figure consists of a Target image. Out of the four given options, your task is to pick the option which has the same figure as the target image. Respond as follows: Option <your answer (choose between 1, 2, 3, or 4)>.

Answer: Option 1

Question: Count the total number of pentagons in the image, including each concentric pentagon separately. For example, if there is one pentagon with 2 inner concentric rings, that counts as 3 pentagons. Respond with only a number.

Answer: 4

Question: Count the number of pentagon's that are black.

Answer: 1

Question: In the scene, how many red cubes are there?

Answer: 1

Question: Do the group of shapes on the left and right have the same orientation?

Answer: No

Question: The image shows one or more letters formed by a grid of small squares. What letter(s) can you identify in this image? Please respond with only the letter(s) you see.

Answer: OCUEC

Question: In the scene, which letters do you see from left to right?

Answer: LM

Question: Consider the grid of 3D objects visible in the provided image. The grid starts at the top-left. Answer the following question: starting from the cube at position (row 1, column 1), how many cubes are there to the right of it in the same column?

Answer: 0

Question: The figure consists of a target image which is complete, Out of the four given options (which are partially complete), your task is to pick the option which when completed matches the target image. Respond as follows: Option <your answer (choose between 1, 2, 3, or 4)>.

Answer: Option 4

Question: In grid 3, starting from the white square at position (row 3, column 5), how many black circles are there up of it in the same column?

Answer: 1

Question: In the scene, how many cubes are there?

Answer: 1

Figure 2: ***Do You See Me*** *benchmark visual perception dimensions*

found in Appendix D. We open-source the synthetic data generation code and the dataset at [‡].

4.2 Synthetic Data Generation and Human Performance Benchmarking

To benchmark human performance against current MLLMs, we conducted a study with fifteen human subjects using this dataset. For each subtask, participants were presented with two randomly selected visual perception questions for each combination of control parameters. They typed their answers and rated each question's difficulty (Easy, Moderate, or Hard). To minimize bias and ensure consistency, all participants underwent a calibration phase before each subtask, involving seven distinct examples spanning the subtask's difficulty range, and everyone received the same calibration and main study examples. Human accuracy for each subtask was determined by comparing their answers to the ground truth. Further details on the human benchmarking protocol are available in Appendix F.

5 Experiments

5.1 Experimental Setup and Evaluation

We evaluate the models listed in Table 3 on Do You See Me. To ensure fair comparison, all models receive identical visual content and uniform textual prompts. Performance of models is measured in terms of final answer accuracy. We closely follow evaluation protocols laid out by [Lu et al., 2023, Xiao et al., 2024]. Please refer Appendix B for a detailed evaluation protocol.

Table 3: Overview of the MLLMs evaluated in our study.

Closed-Source (Proprietary API)	Open-Source (Local Inference)
GEMINI-1.5 FLASH [Team et al., 2024]	LLAMA-3.2-11B-VISION [Grattafiori et al., 2024]
GPT-4O [OpenAI et al., 2024a]	GEMMA-3-12B-INSTRUCT [Team et al., 2025]
CLAUDE-SONNET-3.5 [Anthropic, 2024]	PHI-4-MULTIMODAL-INSTRUCT-5.7B [Abdin et al., 2024]
GEMINI-2.0-FLASH [Google DeepMind and Google, 2024]	QWEN2.5-VL-7B-INSTRUCT [Wang et al., 2024]
GEMINI-2.5-FLASH [Comanici et al., 2025]	INTERNVL2.5-8B [Chen et al., 2025]
OPENAI-O3 [OpenAI et al., 2024b]	

5.2 Experimental Results

5.2.1 Joint Perception-Reasoning Dataset

To probe the relationship between reasoning success and perceptual accuracy, we evaluate MLLMs on a **joint perception–reasoning dataset**. Each sample in this dataset contains an image (I), a perception question (P), and a reasoning question (R). Each model is separately prompted to answer both questions with explicit instructions to provide a detailed chain-of-thought. We employ an LLM-based grader with expert human oversight (detailed in Section 5.1) to score the correctness of the answers.

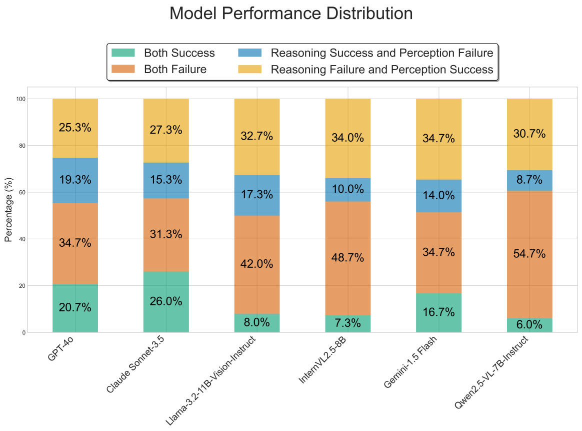
Perception vs. Reasoning Performance: Figure 3a presents the distribution of MLLM perfor-

---

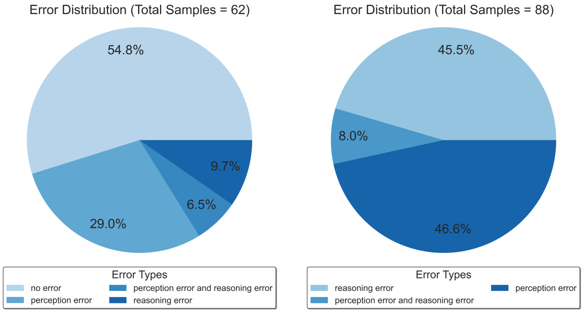
[‡] https://anonymous.4open.science/r/DoYouSeeMe-F52E/README.md

Table 4: Performance of various models on 2D and 3D visual perception tasks. Values are reported as mean accuracy$_{\text{std. dev.}}$. Row colors distinguish between Human performance (yellow), closed-source models (blue), and open-source models (gray).

Model	2D Visual Perception Tasks								3D Visual Perception Tasks					
	Figure Ground	Visual Spatial	Color Disamb.	Shape Disamb.	Letter Disamb.	Form Const.	Visual Closure	Avg. Acc.	Color Disamb.	Letter Disamb.	Shape Disamb.	Form Const.	Visual Spatial	Avg. Acc.
Human	$98.52_{3.8}$	$93.33_{3.1}$	$100.00_{0.0}$	$89.72_{8.17}$	$84.94_{6.12}$	$98.15_{1.35}$	$94.72_{1.84}$	$\mathbf{94.19}_{3.48}$	$96.11_{1.45}$	$99.58_{1.13}$	$96.44_{0.83}$	$97_{2.71}$	$98.21_{1.08}$	$\mathbf{97.46}_{1.44}$
Claude Sonnet-3.5	$41.48_{2.8}$	$35.28_{0.83}$	$76.8_{0.14}$	$33.75_{1.1}$	$15.56_{0.74}$	$90.74_{0.37}$	$57.07_{2.49}$	$50.09_{0.95}$	$66.67_{1.18}$	$20.83_{2.95}$	$73.33_{0.0}$	$18.12_{0.88}$	$34.38_{0.88}$	$42.66_{1.17}$
Gemini-1.5-Flash	$37.04_{0.64}$	$13.19_{0.38}$	$46.98_{0.37}$	$15.14_{1.27}$	$27.16_{0.86}$	$80.37_{0.0}$	$49.6_{0.91}$	$38.49_{0.39}$	$78.89_{0.48}$	$5.9_{0.6}$	$46.67_{2.2}$	$54.17_{1.44}$	$31.25_{3.31}$	$43.37_{1.60}$
GPT-4o	$29.26_{4.49}$	$27.42_{1.88}$	$74.51_{0.65}$	$23.19_{0.64}$	$32.35_{0.86}$	$74.94_{1.19}$	$56.94_{2.48}$	$45.51_{1.28}$	$96.67_{1.44}$	$18.4_{4.21}$	$80.83_{0.83}$	$42.92_{5.64}$	$37.08_{8.04}$	$55.18_{4.03}$
O3 (2025-04-16)	$25.19_{3.39}$	$35.16_{0.43}$	$81.11_{0.25}$	$39.7_{1.94}$	$53.71_{1.11}$	$66.34_{1.67}$	$49.51_{4.73}$	$50.10_{1.50}$	$86.39_{0.96}$	$22.57_{3.35}$	$80.83_{1.67}$	$29.17_{1.44}$	$41.67_{3.15}$	$52.12_{2.11}$
Gemini 2.5 Flash	$8.15_{1.7}$	$54.28_{1.18}$	$88.4_{0.37}$	$53.89_{1.92}$	$30.62_{1.13}$	$51.85_{1.34}$	$51.59_{1.91}$	$48.39_{0.51}$	$71.39_{1.27}$	$4.17_{2.08}$	$77.78_{2.1}$	$36.67_{4.02}$	$9.17_{4.73}$	$39.83_{2.84}$
Gemini 2.0 Flash	$34.44_{1.11}$	$39.66_{1.65}$	$88.73_{0.42}$	$40.83_{0.72}$	$35.8_{1.13}$	$79.63_{1.34}$	$72.82_{2.09}$	$55.98_{0.51}$	$89.44_{1.27}$	$17.71_{1.04}$	$85.28_{0.48}$	$25.83_{1.44}$	$47.08_{4.39}$	$53.06_{1.72}$
Qwen2.5-VL-7B-Instruct	$29.63_{4.49}$	$25.35_{0.72}$	$61.6_{0.71}$	$17.22_{0.24}$	$4.69_{0.43}$	$56.79_{1.07}$	$47.62_{3.09}$	$34.70_{1.49}$	$78.89_{0.48}$	$0.69_{0.6}$	$79.72_{0.96}$	$79.58_{1.44}$	$36.67_{7.22}$	$55.11_{2.14}$
Gemma-3-12B-instruct	$31.11_{1.92}$	$10.22_{0.14}$	$64.13_{1.84}$	$12.92_{1.44}$	$8.15_{0.0}$	$44.94_{2.14}$	$28.97_{1.72}$	$28.63_{0.81}$	$81.11_{0.96}$	$3.82_{0.6}$	$53.89_{0.48}$	$45.0_{0.0}$	$32.5_{0.0}$	$43.26_{0.40}$
Phi-4-Multimodal-Instruct-5.7B	$23.7_{3.21}$	$18.44_{1.37}$	$24.67_{1.84}$	$7.78_{1.2}$	$6.42_{0.43}$	$23.58_{1.5}$	$18.85_{0.34}$	$17.63_{0.89}$	$65.0_{2.2}$	$0.69_{0.6}$	$47.5_{1.44}$	$17.5_{2.17}$	$30.42_{9.04}$	$32.22_{3.09}$
InternVL2.5-8B	$30.00_{6.42}$	$16.00_{0.36}$	$55.64_{0.25}$	$19.17_{1.68}$	$0.74_{0.43}$	$34.81_{3.32}$	$29.76_{3.78}$	$26.58_{2.41}$	$82.50_{0.36}$	0.00_{0}	$66.67_{1.34}$	$53.75_{2.23}$	$33.75_{4.43}$	$47.33_{1.67}$
Llama3.2-11B-Vision-Instruct	$26.67_{2.34}$	$6.08_{1.23}$	$25.74_{0.97}$	$3.33_{1.43}$	$11.85_{0.72}$	$22.22_{1.31}$	$22.02_{2.21}$	$16.84_{1.46}$	$78.33_{0.72}$	$2.08_{0.23}$	$69.17_{1.74}$	$90.00_{1.89}$	$10.00_{2.21}$	$49.91_{1.35}$



(a) MLLM performance on joint perception-reasoning questions.



(b) Error distribution for correct final answers.

(c) Error distribution for incorrect final answers.

Figure 3: *Comparison of MLLM visual reasoning performance (a) and error breakdowns (b, c) for correct and incorrect final answers respectively (Claude Sonnet-3.5).*

mance on the paired perception and reasoning questions. While all evaluated models demonstrate varying degrees of failure across these tasks, a noteworthy and recurring pattern emerges: in several instances, models correctly answer the reasoning questions while simultaneously failing on the corresponding perception questions designed to assess their understanding of crucial visual elements.

Visual Perception Errors in MLLM Reasoning Response: We manually annotated responses from the highest-performing model-*Claude Sonnet-3.5*, categorizing each error as visual perception, reasoning, or arithmetic (as defined in Section 3). Figure 3 presents the error distribution for Claude Sonnet-3.5. For incorrect final answers, we found both visual perception and reasoning errors present in the response chain. Notably, even when models produced correct final answers, we identified visual perception errors in 29.0% of these responses, with only 54.8% of correct responses being free of all error types. These findings suggest that models may arrive at correct answers despite misperceiving visual elements, highlighting the importance of comprehensive evaluation of visual perception capabilities independent of reasoning performance.

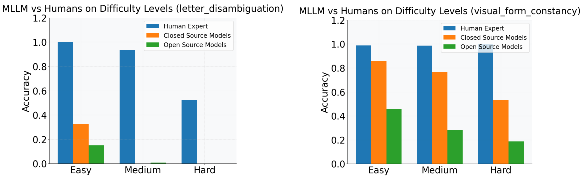
Insight: Relying on final-answer accuracy to judge MLLMs' visual reasoning can create a false sense of success, masking persistent perception errors beneath correct outcomes.

5.2.2 Do You See Me

In this section, we present a comprehensive evaluation of popular MLLMs on **Do You See Me**. Further, we also extensively analyze possible root cause of failure.

Overall Performance: Table 4 illustrate MLLM accuracy across the twelve visual subtasks in their respective 2D and 3D settings. A consistent and primary finding across both 2D and 3D tasks is that current MLLMs fall significantly short of human performance. On average, humans achieve an accuracy of 95.83% (see Appendix O for detailed per-task scores), whereas the best-performing MLLMs operate at considerably lower accuracy levels across nearly all subtasks in both dimensional settings. Further, we observe that overall human subjects exhibit very relatively significantly low-variance across participants.

MLLM Performance Across Human-Perceived Difficulty Levels: Using perceived difficulty ratings (Easy, Medium, and Hard) collected from human subjects across all seven visual perception tasks, we compared MLLM performance (grouped as open- vs. closed-source) against human performance. Figure 4 reveals striking patterns: in visual form constancy, the human performance gap for closed-source models widens from 12% (Easy) to 45% (Hard), while in letter disambiguation, both model types fail completely at Medium difficulty and beyond as humans maintain high accuracy. Similar degradation appears across all subtasks (see Appendix O).



(a) Letter Disambiguation (b) Visual Form Constancy

Figure 4: *Comparison of MLLM and human performance across controlled difficulty levels: (a) on Letter Disambiguation task, and (b) on Visual Form Constancy.*

MLLM Performance Over Increasing Difficulty Control Parameters: As described in §4, we established control parameters for each of the seven subtasks to modulate visual stimuli difficulty in **Do You See Me**. We observe that all MLLMs consistetly exhibit a decline in visual perception performance as the subtasks become harder, culminating in near-zero accuracy at the most challenging settings (refer Appendix O for plots).

Insight: MLLMs show a marked visual perception deficit relative to humans on the **Do You See Me** benchmark. The gap widens sharply as task complexity increases—whether human-rated or parametrically controlled—while human accuracy remains stable.

6 Discussion

Evaluations on **Do You See Me** expose key weaknesses in current MLLMs: minimal gains from fine-tuning, sensitivity to visual complexity, and instability in visual reconstruction.

Comparing 2D and 3D Perceptual Performance MLLMs' performance varies unpredictably between 2D and 3D tasks. Stronger performance in one setting does not guarantee transfer to the other (Figure 19 and Figure 20);—likely reflecting pretraining bias toward real-world (3D) imagery. For example, factors like occlusion and noise in 2D stimuli typically reduces MLLM accuracy in Shape Discrimination and Figure-Ground tasks, but exhibits stronger performance on similar 3D setting.

Limited Gains from Supervised Finetuning: To assess if supervised finetuning (SFT) addresses core visual perception limits, we finetuned Qwen2.5-VL-7B-Instruct—with approximately 67,000 new benchmark-conformant image-text pairs (methodology in Appendix J). While this SFT improved the model's average accuracy by approximately 11% (from 40.91% to 51.75%), its performance remained significantly below human accuracy of 95.83%. Therefore, merely scaling SFT with more benchmark-like data appears insufficient to overcome fundamental visual perception limitations in MLLMs, highlighting the need for alternative approaches.

Impact of Task Format on Final Answer Correctness We analyze the possibility that high final-answer accuracy is artificially inflated by the subtask format. For example, even for humans, it is much more likely to land on the correct answer using the process of elimination when answering Multiple-Choice Questions (MCQ) versus providing open-ended answers. To test whether MLLMs also exploit task format cues to achieve higher accuracy, we performed a supplementary experiment converting an open-ended task into an MCQ format. Specifically, we adapted the open-ended Shape Discrimination (2D) subtask. The effect was significant: GPT-4o's accuracy *jumped* from 23.19% ± 0.64% on the original task to 41.80% ± 2.54% on the MCQ version. This performance gain indicates that the model is not performing 'from-scratch' visual perception but is instead exploiting non-visual cues within the provided options. This finding demonstrates that high final-answer accuracy can mask a lack of true multimodal understanding by allowing the model to exploit task format cues.

Probing Visual Stimuli Reconstruction: To probe how MLLMs internally represent visual inputs, we prompted models to first *reconstruct 2D stimuli as SVG code* before answering related questions. This served as an explicit test of their low-level visual perception and structural understanding. However, reconstruction-based prompting generally *failed to improve*—and often *degraded*—task accuracy compared to direct visual questioning (quantitative results in Appendix K). Qualitative inspection revealed frequent reconstruction errors in shape, orientation, and spatial alignment, underscoring a key perceptual bottleneck. Interestingly, the quality of SVG reconstructions *correlates with downstream task performance*: models that generated more faithful stimuli (e.g., *Claude* on color–shape disambiguation) also performed better on corresponding tasks. This suggests that visual reconstruction can serve as a useful **probing exercise**—revealing not only whether MLLMs can "see" correctly, but also how perception quality constrains reasoning success.

The Role of Verbalization in Multimodal Reasoning. To probe how language-based reasoning interacts with visual perception, we conducted an ablation comparing *direct-answer* and *Chain-of-Thought (CoT)* prompting. In the CoT setup, models were instructed to reason step-by-step before giving a final answer.

We find that a task's **verbalizability**—*the degree to which its visual content can be accurately and completely expressed in language*—is a key determinant of CoT effectiveness. High-verbalizability tasks, such as *Visual Spatial* or *Letter Disambiguation*, benefit from CoT because describing spatial arrangements or symbolic details provides a structured reasoning scaffold. In contrast, low-verbalizability tasks, such as *Form Constancy* or *Visual Closure*, depend on holistic visual understanding that cannot be faithfully captured in words. Here, CoT prompting often *degrades* performance by imposing a **lossy verbalization** of visual information, causing models to reason over their own flawed verbal reconstructions rather than the raw perceptual signal.

This trend is mirrored in model comparisons: *O3*, optimized for structured internal reasoning, significantly outperforms GPT-4o on high-verbalizability tasks, but not on low-verbalizability ones. Together, these findings suggest that verbalizability fundamentally governs when language-based reasoning helps or hinders multimodal understanding.

> **Insight:** MLLMs can reason, but they still struggle to see. Their perception gaps persist beyond finetuning and worsen when vision is forced into words— underscoring the need for better perceptual grounding.

Control Parameter Sensitivity vs. Performance: We employed Kruskal-Wallis test [kru, 2008] to identify control parameters that significantly impact MLLM performance across various visual tasks. These analyses reveal that parameter significance varies by model and task, with fine-tuned and closed-source models often showing sensitivity to a broader range of factors, possibly indicating lesser randomness in answering visual perception questions, and thus, better perception. (refer Appendix I).

Misallocated Visual Attention as a Perceptual Bottleneck: Our analysis of patch-level attention maps pinpoints misallocated visual attention as a critical perceptual bottleneck: MLLMs frequently direct only sparse attention—around 10%—towards query-relevant object regions. This fundamental failure to engage with crucial visual details severely limits their ability to ground language queries in visual evidence. Consequently, advancing MLLM visual perception urgently necessitates more effective query-driven attention mechanisms (detailed in Appendix L).

Instability of Encoder Representations at Fine-Grained Resolutions: Our investigation into how MLLMs process fine-grained visual details revealed critical limitations linked to the visual encoder's patch resolution (typically 14x14 pixels). We evaluated performance on a form constancy task involving objects with subtle geometric distortions, systematically varying the stimuli size from several times larger than the patch size down to its approximate dimensions (methodology in Appendix N). A significant decline in task accuracy was observed as object sizes approached or fell below this patch resolution, contrasting sharply with more reliable perception of larger stimuli. This pronounced difficulty in handling low-granularity visual information strongly suggests a fundamental constraint in current encoders, limiting their ability to robustly interpret small objects or intricate details essential for accurate visual perception.

7 Conclusion

In this work, we first motivated the need for a strong visual perception dataset by documenting flawed visual perception in seemingly correct reasoning answers by current MLLMs; our preliminary analysis found that 29% of correct reasoning answers from a leading MLLM still contained perception errors. We proposed Do You See Me as a diagnostic tool, a programmatically generated and scalable benchmark with over 2,600 questions designed to evaluate core visual skills. Our evaluations showed a stark performance deficit: on our human-psychology-inspired benchmark, human performance was approximately 95.83%, while most MLLMs performed below 50%. We also found that while humans maintain a high degree of accuracy on harder samples, MLLM performance drops drastically as task complexity increases.

Further diagnostic analyses revealed these deficits to be foundational. Supervised finetuning offered only modest gains of around 11% , and we found that models often bypass genuine perception by exploiting task "shortcuts," with performance on one task jumping from 23.19% to 41.80% in an MCQ format. Moreover, the utility of Chain-of-Thought (CoT) prompting is contingent on a task's "verbalizability," often degrading performance on holistic visual tasks by forcing a lossy translation of the image into text. These findings underscore that robust visual perception remains a critical bottleneck, suggesting that current models can reason over what they are told, but still struggle to truly see.

8 Limitations

Our work has a few limitations that we acknowledge and plan to address in future research. First, the size of our joint perception-reasoning dataset is relatively small. However, we have made every effort to include all possible samples where it was feasible to generate non-ambiguous and correlated visual perception questions. To address this limitation, we plan to employ LLM + Image-Diffusion techniques in the future to generate a more diverse and controlled format of perception+reasoning questions, thus expanding our dataset. Next, in the current setup, we have restricted our visual perception prompts to the English language only, including the letter disambiguation task. This decision was made in the interest of managing the overall cost of benchmarking closed-source MLLMs. However, we recognize the importance of language diversity and plan to expand our coverage to other non-English languages in future iterations of our work. Overall, we believe that our work provides a valuable contribution to the understanding of MLLM capabilities in visual perception tasks and lays the foundation for future research in this area.

References

2008. *Kruskal-Wallis Test*, pages 288–290. Springer New York, New York, NY.

Benin Godfrey L Abbavaram Gowtham Reddy and Vineeth N Balasubramanian. 2022. On causally disentangled representations. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Anthropic. 2024. Claude 3.5 sonnet. `https://www.anthropic.com`. Large language model, released October 2024.

Declan Campbell, Sunayana Rane, Tyler Giallanza, Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M. Frankland, Thomas L. Griffiths, Jonathan D. Cohen, and Taylor W. Webb. 2025. Understanding the limits of vision language models through the lens of the binding problem. *Preprint*, arXiv:2411.00238.

James C Chalfant and Margaret A Scheffelin. 1969. Central processing dysfunctions in children: A review of research.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Ronald P. Colarusso. 2003. Mvpt-3: Motor-free visual perception test.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Blender Online Community. 2018. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv*, abs/2306.13394.

Morrison F. Gardner. 1988. Tvps, test of visual-perceptual skills (non-motor).

Google DeepMind and Google. 2024. Introducing Gemini 2.0: our new AI model for the agentic era. `https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024`. Blog post on The Keyword. Accessed: 3 Oct. 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

D.D. Hammill, J.K. Voress, and N.A. Pearson. 2016. *DTVP-3*. Manual Moderno.

Guanzhen Li, Yuxi Xie, and Min-Yen Kan. 2024a. Mvp-bench: Can large vision-language models conduct multi-level visual perception like humans? *ArXiv*, abs/2410.04345.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024b. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *Annual Meeting of the Association for Computational Linguistics*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyue Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and A. Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *ArXiv*, abs/2209.09513.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024b. Openai o1 system card. *Preprint*, arXiv:2412.16720.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024c. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. 2025. Vision language models are blind: Failing to translate detailed visual features into words. *Preprint*, arXiv:2407.06581.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. 2014. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211 – 252.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: a novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23:289 – 301.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

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.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024a. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Preprint*, arXiv:2406.16860.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024b. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *Preprint*, arXiv:2401.06209.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. 2024. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In *The Twelfth International Conference on Learning Representations*.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *ArXiv*, abs/2407.04973.

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. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.

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? In *European Conference on Computer Vision*.

A Table of Contents

1. Experimental Details
    - Closed-Source Model Configuration
    - Open-Source Model Configuration
    - Evaluation Protocol
2. Prelimnary Dataset Statistics
3. Do You See Me - Additional Details
4. Do You See Me - Control Parameter Details
5. Detailed Results
6. Human Performance Benchmarking
7. Joint Perception-Reasoning Dataset - Qualitative Analysis
    - Correct Reasoning Incorrect Perception
    - Qualitative Analysis of Reasoning Chain Errors
8. MLLM Prompts
9. Parameter Importance
10. Finetuning Details
11. Visual Stimuli Reconstruction
12. Sparse Attention to Query Relevant Regions
13. Limits of Visual Perception
    - Visual Form Constancy: Sensitivity to Rotation at Varying Scales
    - Visual Discrimination: Shape Counting Accuracy at Varying Scales

B Experimental Details

We evaluated a range of MLLMs, categorized as follows:

- **Closed-Source Models:**
  - GEMINI-1.5 FLASH
  - GEMINI-2.0 FLASH
  - GEMINI-2.5 FLASH
  - GPT-4o (`gpt-4o_2024-08-06`)
  - CLAUDE-SONNET-3.5 (`claude-3-5-sonnet-2024102`)
  - O3 (`o3_2025-04-16`)
- **Open-Source Models:**
  - LLAMA-3.2-11B-VISION
  - GEMMA-3-12B-INSTRUCT
  - PHI-4-MULTIMODAL-INSTRUCT-5.7B
  - QWEN2.5-VL-7B-INSTRUCT
  - INTERNVL2.5-8B

Closed-Source Model Configuration

The closed-source models were accessed via their respective proprietary APIs (Google AI API for gemini-1.5-flash, OpenAI API for gpt-4o_2024-08-06, and Anthropic API for claude-3-5-sonnet-20241022). For all closed-source model evaluations, consistent generation parameters were used to ensure fair comparison:

- Temperature: 1.0
- Top P (nucleus sampling): 0.95
- Maximum new tokens: 200

Open-Source Model Configuration

The open-source MLLMs were run locally on a single NVIDIA A100 80GB GPU. To ensure fair comparison and reproducibility, the same hyperparameter settings used for the closed-source models were also applied to the open-source models for evaluation purposes:

- Temperature: 1.0
- Top P (nucleus sampling): 0.95
- Maximum new tokens: 200

Evaluation Protocol

All models were provided with identical visual content and uniform textual prompts for each task. Recent Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) are increasingly instructed to produce extended textual outputs rather than concise responses, making earlier rule-based or template-matching methods [Lu et al., 2022] difficult to apply. Inspired by recent benchmarks for MLLMs [Lu et al., 2023, Zhang et al., 2024], we employ an expert LLM to evaluate answers. Our framework proceeds in three stages. In the first stage, a MLLM generates a detailed response according to a predefined template (see Appendix H), which includes the task description, the question, and possible choices. Next, an answer extractor (Appendix H), based on GPT-4o [OpenAI et al., 2024c], parses these extended outputs to yield a concise answer. Prior work has shown that such an expert LLM can extract the correct answer with near 100% accuracy [Lu et al., 2023]. Finally, the extracted text is standardized (e.g., reduced to multiple-choice labels or numeric values), and performance metrics are computed. Since the **Do You See Me** dataset contains both multiple-choice (textual) and free-response (numeric) questions, accuracy is used as a measure of performance.

C Joint-Perception and Reasoning Dataset Statistics

Table 5: Distribution of examples across MathVista and LogicVista. "IQ/Logic Qs." refers to pattern-based or spatial reasoning questions.

Dataset	Original Size	IQ/Logic Qs	Selected
$MathVista_{mini}$ (IQtest)	1000	37	15
$LogicVista$ (Diagrams)	448	223	135

D Do You See Me - Additional Details

The **Do You See Me** benchmark is structured into seven distinct subtasks, each meticulously designed to evaluate specific facets of visual perception in Multimodal Large Language Models (MLLMs), as detailed in Section 4.1 of the main paper. For subtasks presented in a 3D setting, a consistent set of five basic geometric shapes is utilized: *sphere, cube, cone, cylinder,* and *torus*. A critical parameter for modulating difficulty in 3D tasks where object occlusion is relevant—specifically Shape Discrimination and Joint Shape-Color Discrimination—is the visibility factor, $\beta_{occ} \in [0, 1]$. This factor dictates the target percentage of an object's surface area that must be visible from the camera, enabling systematic control over task complexity, ranging from fully visible objects ($\beta_{occ} = 1$) to various degrees of occlusion (e.g., $\beta_{occ} = 0.7$ signifies 70% visibility). For other 3D subtasks, namely Visual Form Constancy (3D), Letter Discrimination (3D), and Visual Spatial (3D), occlusion is not a variable factor, and β_{occ} is effectively maintained at 1.

Shape Discrimination (2D and 3D): This subtask evaluates an MLLM's proficiency in identifying and counting specific shapes within a composite visual scene.

2D Setting: Seven fundamental geometric shapes are employed: *rectangle, triangle, circle, pentagon, hexagon, octagon,* and *star*, each rendered with solid black borders and transparent interiors. The complexity is systematically varied using three control parameters: the number of unique shape types present (S), the maximum number of instances permitted for each shape type (S_I), and a separation factor (d_{sep}). To ensure distinct object boundaries for certain conditions, non-overlapping placements ($d_{sep} \geq 0$) are achieved using the Separating Axis Theorem (SAT), which maintains a minimum distance d_{sep} between any two shapes. Conversely, negative values for the separation factor ($d_{sep} < 0$) allow for controlled degrees of overlap, where shapes can interpenetrate by up to $|d_{sep}|$ units. Each generated 2D image is accompanied by a counting question (e.g., "How many circles are in the image?") for which a programmatic ground truth answer is available.

3D Setting: This version uses the standard set of five 3D shapes previously mentioned. Task difficulty is modulated by the number of unique 3D shape types (S), the maximum instances per shape type (S_I), and the visibility parameter (β_{occ}) which controls occlusion. Questions are structurally similar to those in the 2D setting (e.g., "How many spheres are in the scene?").

Joint Shape-Color Discrimination (2D and 3D): This subtask assesses the MLLM's ability to handle compositional queries that require the simultaneous identification of object shape and color through counting tasks.

2D Setting: Six distinct 2D shapes (*star, triangle, pentagon, hexagon, octagon, cross*) are used in conjunction with eight standard colors (*red, green, blue, orange, purple, black, gray, yellow*). To prevent ambiguity in shape-color binding, all shapes are rendered without overlap. The difficulty level is primarily controlled by two parameters: the number of unique shape types (S) and the number of unique colors (C) present in the image. An example query is: "Count all red triangles".

3D Setting: The standard five 3D shapes are utilized, rendered with the same eight colors available in the 2D version. Difficulty is adjusted through three main parameters: the number of unique 3D shape types (S_{3D}), the maximum number of instances allowed for any unique shape-color pair (I_{sc}), and the visibility parameter (β_{occ}) controlling occlusion. Questions require joint discrimination, such as: "How many red spheres are present in the scene?".

Letter Discrimination (2D and 3D): This subtask is designed to test an MLLM's capability to recognize textual characters presented under various conditions.

2D Setting: Characters are rendered as patterns within a 5×7 grid of LED-style blocks. The difficulty of letter identification is manipulated by three parameters: the block spacing factor (β), which adjusts the distance between constituent blocks of a letter; the color contrast (ΔC) between the letter blocks and the image background; and the total number of letters (N) present in the stimulus.

3D Setting: Letters are visually constructed by strategically arranging a collection of simpler 3D primitive shapes—specifically *spheres, cubes,* or *cylinders*—such that they collectively form the appearance of a target letter when viewed from the camera's perspective. The modulation of difficulty in this setting is achieved by varying the type of 3D primitive used for construction, the size of these primitives, the spacing between them, and the total number of distinct letters (N) displayed. A typical question for this task is: “What letter(s) can you identify in this scene?”.

Visual Form Constancy (2D and 3D): This subtask challenges the MLLM to recognize a target pattern even after it has undergone geometric transformations or substitutions, requiring the model to identify the correct match from several options.

2D Setting: A target pattern is first constructed using simple 2D primitives such as a *circle, square, line,* or *triangle*. Alongside the target, three distractor variants are generated by applying transformations controlled by a shape substitution factor (ssf), a scaling factor (α), and a rotation factor (θ_r). The MLLM’s task is to select the one option from four (target + three variants) that perfectly matches the original target’s arrangement.

3D Setting: This version evaluates the MLLM’s ability to discern and compare the orientations of 3D objects. Two groups of objects are presented: a "left group" (target) and a "right group" (comparison). Both groups are initially identical and composed of the standard 3D shapes, with the right group spatially offset from the left. In each instance, a rotation (by angle θ_r) may or may not be randomly applied to the entire right group. If the right group is rotated, its constituent shapes’ orientations will not match those of the left group (leading to a "no" ground truth answer); otherwise, they will match ("yes" answer). The MLLM must respond to the question: “Do all the shapes in the left group and right group have the same orientations?”.

Visual Spatial (2D and 3D): This task evaluates the MLLM’s understanding of object positions and their spatial relationships within structured layouts.

2D Setting: The stimuli consist of one or more $H \times W$ grids. Each cell within these grids contains one of three basic shapes—*circle, square,* or *triangle*—which can be rendered either as solid black or merely outlined. Key parameters controlling the scene complexity include the grid dimensions (D_{grid}) and the number of grids (G) presented. Queries require the MLLM to locate and count shapes based on their spatial position relative to a reference coordinate provided in the question (e.g., “How many solid circles are above the triangle in row 3, column 2?”).

3D Setting: In this version, a grid-like structure is populated using the standard five 3D shapes. Analogous to the 2D setting, the MLLM is required to count 3D shapes based on their positions relative to a specified reference coordinate within the grid. An example query illustrates this: “The image consists of a grid like layout with multiple 3D shapes, starting from the cylinder at position (row 2, column 1), how many tori are there to the right of it in the same row?”.

Visual Figure-Ground (2D): This subtask builds upon the visual form constancy framework by incorporating distracting background elements, thereby challenging the MLLM to distinguish a target pattern from its surroundings. The complexity is primarily controlled by two parameters: the number of shapes (N) composing the target pattern and any distractor patterns, and the background density factor (bdf), which dictates the quantity of visual noise introduced into the scene. The MLLM’s objective is to identify the target pattern among several candidates, despite the presence of this visual noise. As noted in the main paper (Section 4.1), a distinct 3D version for Visual Figure-Ground is not included because the photorealistic 3D rendered scenes used in other 3D subtasks inherently assess this perceptual ability.

Visual Closure (2D): This subtask assesses an MLLM’s ability to mentally complete a partially obscured or incomplete shape and match it to its corresponding complete form. Seven basic 2D shapes are used as targets: *capsule, star, hexagon, circle, pentagon, rectangle,* and *triangle*. For each trial, one shape is selected as the complete target. An incomplete version of this target is created by removing some of its edges. Additionally, three "noisy" distractor options are generated by taking the incomplete target and applying distortions to its vertex positions. The MLLM is presented with the complete target and four options (the correctly incomplete shape and the three distractors) and must identify which of the incomplete options would correctly form the target if its missing parts were filled in. This subtask is exclusively 2D; a 3D equivalent is not provided due to the inherent difficulties and ambiguities in judging 3D shape closure from a single, static camera viewpoint, as mentioned in Section 4.1 of the main paper.

E Do You See Me - Control Parameter Details

Table 6: Control parameters and question types for subtasks[†].

Division	**Subdivision**	**Setting(s)**	**2D Control Parameters**	**Dataset Statistics (2D)**
Visual Discrimination	Shape Discrimination	2D & 3D	Number of Shapes: $S \in [3, 7]$ Instances per Shape: $S_I \in [3, 6, 10]$ Overlap Factor: $\alpha \in [-40, -30, -20, 10]$	Unique Images: 241 Questions: 241 (Integer)
	Joint Shape-Color	2D & 3D	Number of Shapes: $S \in [2, 4, 6]$ Number of Unique Colors: $C \in [2, 4, 6]$	Unique Images: 90 Questions: 408 (Integer)
	Letter Discrimination	2D & 3D	Number of Letters: $N \in [1, 5, 9]$ Foreground-Background Contrast: $\Delta C \in [1, 2, 3]$ Block Size: [0.04, 0.08, 0.1]	Unique Images: 135 Questions: 135 (Text)
	Form Constancy	2D & 3D	Shape Substitution Factor : $ssf \in [0, 1]$ Scaling Factor: $\alpha \in [0.8, 1.1, 1.4]$ Rotation Factor: $\theta_r \in [5, 25, 50]$ Aspect Ratio: $\beta \in [0.8, 1.1, 1.4]$	Unique Images: 270 Questions: 270 (MCQ)
Visual Spatial	Spatial Grids	2D & 3D	Grid Dimension: $D \subseteq [3, 6, 9] \times [3, 6, 9]$ Number of Grids: $G \in [1, 3, 5]$	Unique Images: 270 Questions: 806 (Integer)
Visual Figure-Ground	N.A.	2D	Number of Shapes: $N \in [2, 6, 10]$ Background Density Factor: *bdf* $\in [0.1, 0.3, 0.5]$	Unique Images: 90 Questions: 90 (MCQ)
Visual Closure	N.A.	2D	Number of Full Edges to Remove: $k \in [1, 3]$ Number of Partial Edges to Remove: $l \in [1, 3]$ Number of Edges to Distort: $m \in [1, 3]$ Distortion Factor: $\delta \in [0.1, 0.12, 0.14]$	Unique Images: 166 Questions: 166 (MCQ)

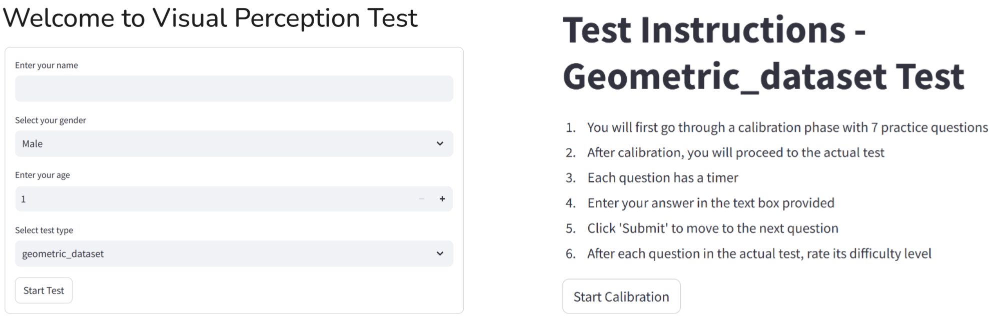
[†] For tasks with 3D versions, 3D-specific counts and parameters are in Appendix E.

Table 7: Control parameters and question types for 3D subtasks in Do You See Me.

Division	**Subdivision**	**Setting(s)**	**3D Control Parameters**	**Dataset Statistics (2D)**
Visual Discrimination	Shape Discrimination	2D & 3D	Number of Shapes: $S \in [1, 2, 3, 4, 5]$ Instances per Shape: $S_I \in [1, 2, 3]$ Occlusion factor: $\beta_{occ} \in [0.7, 0.8, 0.9, 0.99]$	Unique Images: 120 Questions: 120 (Integer)
	Joint Shape-Color	2D & 3D	Number of Shapes: $S \in [1, 2, 3, 4, 5]$ Instances per Shape: $S_I \in [1, 2, 3]$ Occlusion factor: $\beta_{occ} \in [0.7, 0.8, 0.9, 0.99]$	Unique Images: 120 Questions: 120 (Integer)
	Letter Discrimination	2D & 3D	Number of Letters: $N \in [1, 2]$ Shape Size: $\mathcal{S} \in [0.05, 0.08, 0.11, 0.14]$ Inter-Shape Spacing:$\mathcal{S}_{spacing} \in [0.4, 0.5]$	Unique Images: 96 Questions: 96 (Text)
	Form Constancy	2D & 3D	Rotation : $\theta_r \in [5, 7.5, 10, 12.5, 15]$ Number of Shapes $S \in [1, 2, 3, 4]$	Unique Images: 80 Questions: 80 (MCQ)
Visual Spatial	Spatial Grids	2D & 3D	Grid Dimension: $D \subseteq [2, 3, 4, 5] \times [2, 3, 4, 5]$ Number of Grids: $G \in [1]$	Unique Images: 80 Questions: 80 (Integer)

F Human Performance Benchmarking

We recruited 15 participants (11 men, 4 women) for the human performance benchmarking study. The primary selection criterion was that participants had no history of vision-related disorders. All participants were between 25 and 35 years of age. Each of the twelve subtask evaluations took approximately 10 minutes per participant to complete, resulting in a total active testing time of about 120 minutes per participant. To mitigate mental fatigue, these evaluations were conducted over two days. As a token of appreciation for their participation, subjects received a food voucher equivalent to the local minimum wage for two hours. To assess inter-rater reliability, we calculated Fleiss's Kappa. The agreement on task correctness was 0.796, indicating substantial agreement among participants. Additionally, the low standard deviations for human performance, when contrasted with the large gap to mean MLLM performance, demonstrate a clear and significant difference.



Welcome to Visual Perception Test

Enter your name

Select your gender: Male

Enter your age: 1

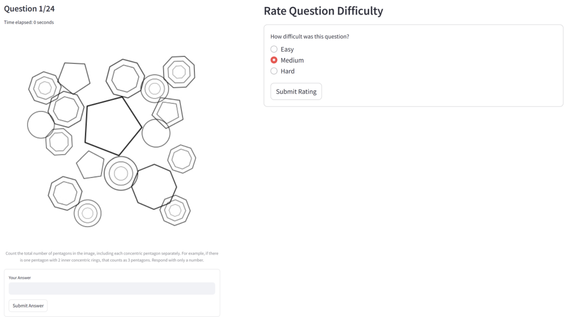
Select test type: geometric_dataset

Test Instructions - Geometric_dataset Test

1. You will first go through a calibration phase with 7 practice questions
2. After calibration, you will proceed to the actual test
3. Each question has a timer
4. Enter your answer in the text box provided
5. Click 'Submit' to move to the next question
6. After each question in the actual test, rate its difficulty level

(a) Information Collected in the Test (b) Calibration Instructions

Figure 5: Human Performance Benchmarking



Question 1/24

Time elapsed: 0 seconds

Count the total number of pentagons in the image, including each concentric pentagon separately. For example, if there is one pentagon with 2 inner concentric rings, that counts as 3 pentagons. Respond with only a number.

Your Answer

Rate Question Difficulty

How difficult was this question?

○ Easy
● Medium
○ Hard

Figure 6: Actual Test and Human Difficulty Rating Collection

G Joint Perception-Reasoning Dataset

G.1 Correct Reasoning Incorrect Perception

Example 1

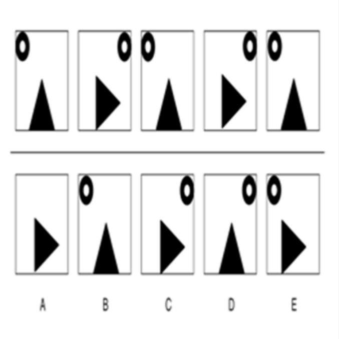


Figure 7

Reasoning Question: Which of the boxes comes next in the sequence? Select answers from A-E. Explain the reasoning behind the answer.
Reasoning Answer (GT): C
Reasoning Explanation: Triangle alternates from pointing up, to pointing right with each turn. Doughnut shape alternates from left corner to right corner with each turn.
Claude Sonnet-3.5 Answer: Let's solve this step by step:
Looking at the top sequence, I notice a pattern of shapes and dots:
Single triangle, single dot above Single triangle, single dot above Single triangle, two dots above Single triangle, single dot above Single triangle, single dot above Also notice that the triangles alternate between pointing left () and right ()
The pattern appears to be:
The triangles alternate direction The dots follow a pattern of: 1, 1, 2, 1, 1, 2, ... Following this pattern, the next box should have:
A triangle pointing right () Two dots above (because we're at the point in the sequence where two dots should appear) Looking at the options A through E: Option C shows a triangle pointing right with two dots above it.
Therefore, the answer is C.
The reasoning is based on both the alternating direction of the triangles and the repeating pattern of dots (1, 1, 2).
VLM Final Answer: YES
Visual Perception Question: Using visual inspection, count the number of black dots on the top-left side of the box where the triangle present in the box is resting on its side and not the vertex.
Visual Perception Answer (GT): 4
Claude Sonnet-3.5 Answer: Let me help you count the black dots as requested.
Looking at the boxes where the triangle is resting on its side (not on its vertex), we need to look at the top-left portion of those boxes.
The triangles resting on their sides (not on vertex) appear in positions B, D, and E in the top row and positions A and E in the bottom row.
Among these:
Top row: B has 2 dots, D has 2 dots, E has 1 dot Bottom row: A has 0 dots, E has 2 dots (**incorrect perception**) Therefore, there are a total of 7 black dots when counting only the top-left portions of boxes where the triangle is resting on its side.
Analysis: The visual perception question shows how model fails answering the question on number of dots on top-left side. The same failure occurs in answering the visual reasoning question.

Example 2

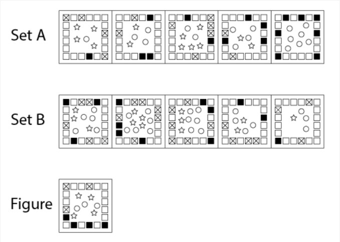


Figure 8

Reasoning Question: Which set does the Figure belong to? Select from A, B, and C. (A) Set A (B) Set B (C) Neither set A nor set B. Explain the reasoning behind the answer.
Reasoning Answer (GT): B
Reasoning Explanation: Set A: The total number of crossed boxes is equal to the total number of stars. The total number of shaded boxes is equal to the total number of circles. Set B: The total number of crossed boxes is equal to the total number of circles. The total number of shaded boxes is equal to the total number of starts. Figure: There are four shaded boxes and four stars. There are three crossed boxes and three circles.
Claude Sonnet-3.5 Answer: The Figure belongs to (B) Set B.
Reasoning:
Looking at the pattern in both sets, we need to focus on the arrangement of black squares and circles in the cards.
Key observations about the Figure:
It has circles arranged in a specific pattern It has three black squares at the bottom The overall layout is similar to other cards in both sets When comparing with Set A and Set B: Set A's cards generally have black squares distributed throughout different positions Set B's cards consistently show black squares grouped at the bottom of each card, similar to the Figure The circle patterns in Set B also appear more similar to the Figure's arrangement The decisive factor is the positioning of the black squares: The Figure shows three black squares aligned at the bottom This arrangement is consistent with the pattern seen in Set B, where black squares are consistently placed at the bottom of the cards This pattern is not characteristic of Set A Therefore, based on the consistent pattern of black square placement at the bottom and similar circle arrangements, the Figure belongs to Set B.
MLLM Final Answer: Correct
Visual Perception Question: In Set B of the diagram, looking at the second square from the left, how many stars (?) are inside that square?
Visual Perception Answer (GT): 4
Claude Sonnet-3.5 Answer: Looking at Set B, in the second square from the left, I can count 2 stars within that square.
MLLM Final Answer: Incorrect
Analysis: The that number of stars within the box are linked to correctly the reasoning question and Claude Sonnet-3.5 fails to answer the count correctly. Further, from the reasoning chain, it is clear that it makes a reasoning error that black squares are grouped/ungrouped in Set A and B, and answers the reasoning question on this basis. This shows a reasoning error in its response, however, the final answer given is correct. Additionally, there is visual hallucination about distribution of black sqaures as well.

G.2 Qualitative Analysis of Reasoning Chain Errors

Example 1

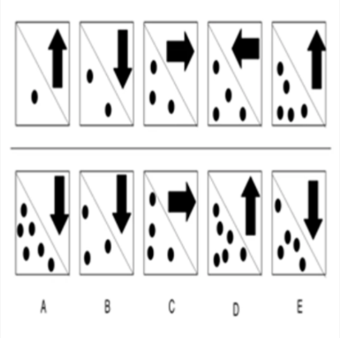


Figure 9

Reasoning Question: Which of the boxes comes next in the sequence? Select from A-E
Reasoning Answer (GT): A
Reasoning Explanation:
GPT4o Answer: To determine which box comes next in the sequence, observe the direction of the arrows and the distribution of dots in each box. 1. The direction of arrows alternates from pointing up to down in every subsequent box. The sequence starts with an arrow pointing up. 2. The number of dots increases incrementally by one in each subsequent box. The first box has one dot, the second has two, the third has three, and so on. Following this pattern:

- The first box in the sequence has one dot and an arrow pointing up.
- The second box has two dots and an arrow pointing down.
- The third box has three dots and an arrow pointing right.
- The fourth box has four dots and an arrow pointing down. **(perception error)**
- The fifth box has five dots and an arrow pointing up.

Given these patterns, the next box should have an arrow pointing down and six dots. Among the choices A to E:

- Choice A fits with an arrow pointing down and six dots.

Therefore, the correct answer is box A.
VLM Final Answer: Correct.
Key Observation: Correct Final Answer But presence of Incorrect Visual Perception in Reasoning Chain.
Analysis: GPT4o makes a visual perception error on the fourth box, the pattern was up, down, right, left, up. Whereas, GPT-4o determines that the pattern is up, down, right, down, up. *This is an example of visual perception error in reasoning chain even while the model gives a correct final answer.*

Example 2

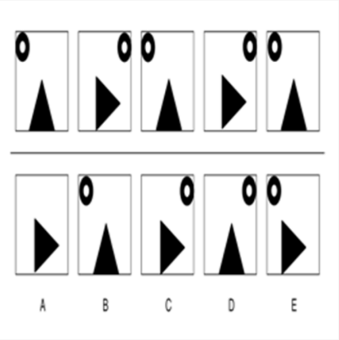


Figure 10

Reasoning Question: Which of the boxes comes next in the sequence? Select from A-E

Reasoning Answer (GT): C

Reasoning Explanation:

GPT4o Answer: To determine which box comes next in the sequence, we need to observe the patterns of the shapes and their rotations in the given boxes. 1. The position of the circle:

- The circle moves through four quadrants: top-right, top-left, bottom-left, and bottom-right. (**visual perception error: the circle never goes to bottom left or right**)
- Based on this pattern, after the bottom-right position in the fifth box, the next position should be top-right.

2. The rotation of the triangle:

- The triangle rotates clockwise: pointing up, right, down, and then left.
- In the fifth box, the triangle is pointing left. Therefore, in the next box, it should point up again.

Given these patterns, we are looking for a box with the circle in the top-right quadrant and the triangle pointing up.

Option **C** shows a circle in the top-right quadrant and a triangle pointing up. This matches the patterns observed, so option C is the correct answer for the next box in the sequence.

VLM Final Answer: Correct.

Key Observation: Correct final answer but *presence of incorrect visual perception* in Reasoning Chain.

Analysis: The circle goes between top-right/left. *Example of visual perception error*

Example 3

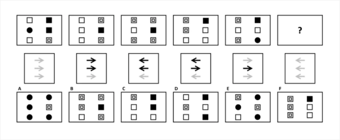


Figure 11

Reasoning Question: The bottom boxes create a rule that has to be applied in the box directly above them. Select which of options A to F corresponds to the rule below the box with the question mark.. Explain the reasoning behind the answer.

Reasoning Answer (GT): E

Reasoning Explanation: The simple rule in this question is that the number of grey arrows in the box below must equal the number of black shaded shapes in the box above. There are three grey arrows below the missing box, so the correct answer must contain three black shapes.

GPT4o Answer: To determine which option (A to F) corresponds to the rule that should be applied above the question mark, we need to analyze the pattern in each column and how the bottom boxes affect the top boxes.

Column Analysis:

Column 1: The bottom box has arrows pointing right. The top box changes the orientation of the shapes accordingly.

Column 2: The bottom box has arrows pointing left. The top box changes by mirroring or flipping the orientation.

Column 3: The bottom box has inward-facing arrows, and the shapes in the top box are transformed or interchanged.

Column 4: The bottom box has outward-facing arrows. It seems to involve swapping the positions of certain elements in the top box.

Column 5 (Question Mark): The bottom box has left-facing arrows, implying a mirroring or flipping operation similar to column 2.

Rule Application:

Given the bottom box under the question mark has left-facing arrows, we should select an option that mirrors or flips the elements horizontally compared to the option from Column 4.

Options Evaluation:

Option E is a horizontally flipped version of the top box in Column 4, making it the correct answer. Therefore, the correct option corresponding to the rule below the box with the question mark is Option E.

VLM Final Answer: Correct

Key Observation: Correct answer, incorrect reasoning.

Analysis: The simple rule here is the number of grey arrows in the box below must equal the number of black shaded shapes in the box above. *However, the model reasons about changing orientations based on the direction of arrows.* However, even with wrong reasoning GPT-4o reaches the correct answer.

Example 4

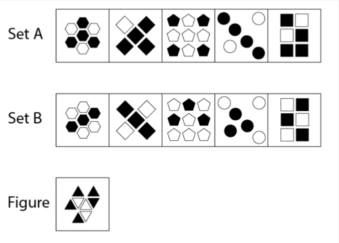


Figure 12

Reasoning Question: Which set does the Figure belong to? Select from A, B, and C. (A) Set A (B) Set B (C) Neither set A nor set B. Explain the reasoning behind the answer.

Reasoning Answer (GT): C

Reasoning Explanation: Set A: Each box contains four shaded shapes. Set B: Each box contains three shaded shapes. Figure: The Figure contains five shaded shapes.

GPT4o Answer: The figure belongs to (C) Neither set A nor set B. Reasoning:

Set A:

In Set A, the figures alternate between filled and empty shapes, or there is a symmetry of arrangement. Set B:

In Set B, there is a symmetrical arrangement as well, but the colors (filled and empty) alternate in a different manner than Set A. Figure Analysis:

The given figure is composed of triangles, both filled and empty, arranged in a different pattern from any of the figures in Set A or Set B. It does not follow the alternation pattern or symmetry rules observed in either set. Therefore, the figure does not belong to Set A or Set B.

VLM Final Answer: Correct

Key Observation: Correct answer, incorrect reasoning.

Analysis: GPT4o reasons that there is a different alternate pattern to fill shapes in Set A and B, and the new shape has both these patterns, hence cannot be alloted to any one set. However, the simple rule here is that 3 shapes are shaded in set B, and 4 shapes are shaded in shape A. Thus, while the model reasons incorrectly, it arrives at the right option.

H MLLM Prompts

Expert LLM Evaluator

You are given a question, a long-form answer generated by a vision-language model, and the ground-truth answer. The model's answer may contain reasoning or explanation, but always ends with a specific count or choice.
Your task is to extract **only the final integer answer** from the model's response and output it in the format: <answer>X</answer>
If no clear answer is found, respond with: <answer>ERROR</answer>
The extracted answer should match the type of the ground truth answer — whether it is a count or an option number. Note, do not directly copy over the ground-truth answer in your response. The answer should only be extracted from the long-form model response.
Examples
Question: In grid 2, starting from the white circle at position (row 1, column 8), how many squares are there left of it in the same row? Ground Truth Answer: 0 VLM Answer: In grid 2, starting from the white circle at position (row 1, column 8), there is 1 square to the left of it in the same row. <answer>1</answer>
Question: The figure consists of a Target image... Option <your answer (choose between 1, 2, 3, or 4)>. Ground Truth Answer: 4 VLM Answer: After analyzing the images, the third option appears most similar to the target image. Option 3. <answer>3</answer>
Question: In grid 1, starting from the white square at position (row 2, column 1), how many black objects are there up of it in the same column? Ground Truth Answer: 0 VLM Answer: There are no black objects above it. <answer>0</answer>
Question: How many circles are present in the image? Ground Truth Answer: 2 VLM Answer: I see that there are 2 circles and 4 triangles. Therefore, there are 2 circles in the image. <answer>2</answer>
Now process the following:
Question: {question}
Ground Truth Answer: {gt_answer}
VLM Answer: {vlm_answer}

Example Prompts For Subtasks in Do You See Me

Visual Figure Ground: The figure consists of a Target image, which is embedded in some background noise. Out of the four given options, your task is to pick the option which has the same figure as the target image. Respond as follows: Option your answer (choose between 1, 2, 3, or 4).

Letter Disambiguation: The image shows one or more letters formed by a grid of small squares. What letter(s) can you identify in this image? Please respond with only the letter(s) you see.

Visual Form Constancy: The figure consists of a Target image. Out of the four given options, your task is to pick the option which has the same figure as the target image. Respond as follows: Option your answer (choose between 1, 2, 3, or 4).

Visual Closure: The figure consists of a target image which is complete, Out of the four given options (which are partially complete), your task is to pick the option which when completed matches the target image. Respond as follows: Option your answer (choose between 1, 2, 3, or 4).

Visual Spatial: In grid 5, starting from the white square at position (row 1, column 5), how many circles are there down of it in the same column?

Color Disambiguation: Count the number of cross's that are purple.

Shape Discrimination: Count the total number of stars in the image, including each concentric star separately. For example, if there is one star with 2 inner concentric rings, that counts as 3 stars. Respond with only a number.

I Parameter Importance

To gain a deeper understanding of the factors driving performance variations among Multimodal Large Language Models (MLLMs) on the **2D tasks** within the "Do You See Me" benchmark, we conducted a detailed parameter importance study. To ascertain statistical significance, we employed Kruskal-Wallis tests, a non-parametric form of one-way ANOVA, considering a parameter's impact significant if the p-value fell below 0.05.

This analysis revealed that most of the 8 MLLMs had at least one significant factor affecting their performance on any given 2D task, suggesting that models exhibiting no sensitivity to parameter changes might be resorting to random guessing. Several consistent patterns emerged. For instance, the `number_of_letters` was a critical determinant in `letter_disambiguation` for **6 out of the 8 MLLMs**. A similar proportion of these models (**5 out of 8**) found `aspect_ratio` to be a significant factor in `visual_form_constancy`. In `joint_shape_color_discrimination` tasks, the `number_of_shapes` (significant for **4 out of 8 models**) and the `number_of_unique_colors` (significant for **4 out of 8 models**) were frequently influential. An interesting distinction arose when comparing the open-source and closed-source MLLMs within this 8-model subset. On average, the closed-source models tended to have a higher number of significant variables affecting their performance on these 2D tasks. This suggests that while these models might achieve higher overall accuracy, their performance is also discernibly modulated by a broader range of specific input complexities within 2D contexts.

Furthermore, the study highlighted the impact of supervised fine-tuning (SFT) by comparing the QWEN2.5-VL-7B-INSTRUCT model with its fine-tuned version, QWEN2.5-VL-7B-INSTRUCT-SFT, on these 2D tasks. The SFT model demonstrated a marked increase in the number of significant variables to which it was sensitive, from four in the base model to nine in the fine-tuned version. This increased sensitivity was particularly evident in 2D tasks related to spatial reasoning, shape discrimination, and letter discrimination. For example, after fine-tuning, `instances_per_shape` and `number_of_shapes` became significant for `shape_discrimination`, and `block_spacing` became an additional factor for `letter_disambiguation`. This observation points that higher number of significant parameters indicates lower randomness, and thus better accuracy on the Do You See Me benchmark.

J Finetuning Details

To assess the impact of supervised finetuning (SFT) on the visual perception capabilities of MLLMs, specifically concerning the tasks presented in the *DoYouSeeMe* benchmark, we finetuned the Qwen2.5-VL-7B-Instruct model. The finetuning process was conducted using the Llama Factory framework on the *DoYouSeeMe-Train* dataset, which comprises approximately 67,000 benchmark-conformant image-text pairs.

We employed LoRA (Low-Rank Adaptation) for efficient finetuning, targeting the vision encoder and the LLM with a LoRA rank of 8. The training was conducted for 5 epochs with a per-device training batch size of 8 and 8 gradient accumulation steps, resulting in an effective batch size of 64. A learning rate of $1.0e^{-4}$ was used with a cosine learning rate scheduler and a warmup ratio of 0.1. The training utilized bf16 precision. Evaluation was performed every 500 steps on a validation set comprising 10% of the training data.

The impact of SFT on the Qwen2.5-VL-7B-Instruct model's performance across various visual perception tasks is presented in Table 8.

Table 8: Performance of Qwen2.5-VL-7B-Instruct Before and After SFT [cite: 601]

Model	Average Acc. (%)	Visual Figure Ground	Visual Spatial	Color Disamb.	Shape Disamb.	Letter Disamb.	Visual Form Const.	Visual Closure
Qwen2.5-VL-7B-Instruct	40.91	27.78	40.69	81.86	19.58	2.96	50.37	63.10
Qwen2.5-VL-7B-Instruct-SFT	51.75	41.11	46.15	60.78	49.58	13.33	92.96	75.89
Human	94.33	100.0	92.59	100.0	100.0	77.77	98.14	91.66

As shown in Table 8, supervised finetuning led to an improvement in the average accuracy of the Qwen2.5-VL-7B-Instruct model from 40.91% to 51.75%. Notable gains were observed in tasks such as Shape Disambiguation (from 19.58% to 49.58%) and Visual Form Constancy (from 50.37% to 92.96%). However, despite these improvements, the finetuned model's performance still remained significantly below human accuracy levels (average 94.33%) across all tasks. For instance, in tasks like Letter Disambiguation, the finetuned model achieved 13.33% compared to human performance of 77.77%. Furthermore, in Color Disambiguation, the performance slightly decreased post-SFT (from 81.86% to 60.78%). This outcome suggests that while SFT on benchmark-conformant data can yield some gains, it does not drastically overcome the fundamental visual perception limitations observed in MLLMs, highlighting the need for alternative approaches or more fundamental architectural changes to enhance MLLM visual perception.

K Visual Stimuli Reconstruction

To further probe the internal visual understanding of Multimodal Large Language Models (MLLMs) and to explore alternative modalities for task resolution, we investigated an intermediate step of visual stimuli reconstruction. In this approach, for each 2D visual task within the *DoYouSeeMe* benchmark, the MLLM was first prompted to convert the given visual stimulus into its constituent Scalable Vector Graphics (SVG) representation. In a subsequent, distinct step, the MLLM was then asked to answer the associated perception question, relying solely on the SVG code it had previously generated. The primary motivations for this exploration were twofold: first, to transform the vision-text task into a text-only one by leveraging the LLM component's text-based reasoning on the structured SVG format; and second, to perform a qualitative perceptual analysis, as the generated SVG itself serves as a tangible artifact of the MLLM's interpretation of the visual scene, offering direct insights into "what the MLLM sees". *Note: Due to the high token output and cost of SVG generation, we sampled two images per parameter sweep for this reconstruction analysis.*

The evaluation of MLLM performance via SVG reconstruction involved a two-step process for each 2D image-question pair. **Step 1: SVG Generation** involved providing the MLLM with the visual stimulus and a prompt such as: "`Given the image, generate the complete SVG code that accurately reconstructs its visual content.`". **Step 2: Question Answering based on Generated SVG** followed, where the MLLM, after generating the SVG, was presented with the original perception question and explicitly instructed to answer based *only* on the SVG code it had generated, using a prompt like: "`Using only the following SVG code: <SVG-Code>, answer the question: <Visual-Perception-Question>`". This two-step methodology ensures that the question-answering performance in the SVG-mediated condition is directly dependent on the quality and accuracy of the MLLM's own SVG reconstruction.

As detailed in the main paper's Discussion section, this strategy of intermediate SVG reconstruction generally **failed to improve, and often notably degraded, task performance** compared to direct visual questioning. This quantitative outcome suggests that generating accurate and detailed SVG representations is a significant challenge for current MLLMs, with errors in SVG generation likely cascading to the question-answering phase. Table 9 presents a performance comparison for GPT-4o and Claude-3.5. "-Image" denotes performance when answering directly from the visual stimulus, while "-SVG" denotes performance based on the MLLM's own generated SVG. All scores are accuracy percentages (%).

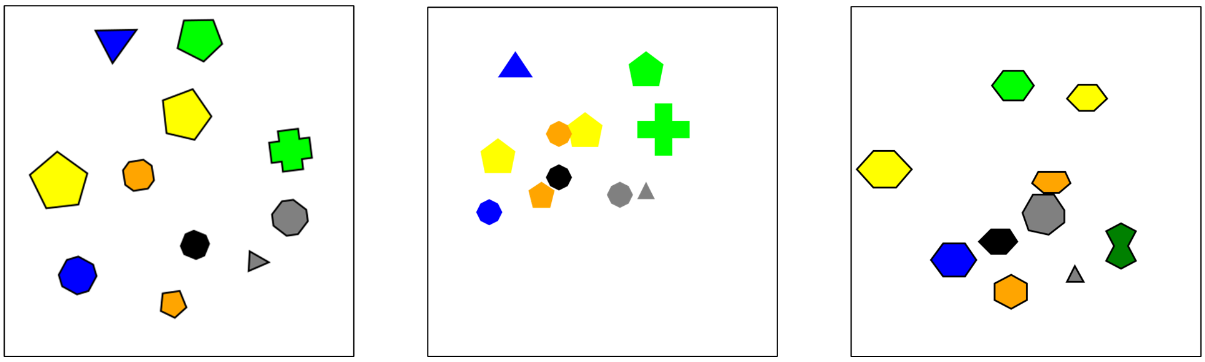
Table 9: MLLM Performance: Image vs. SVG-mediated. All scores are accuracy percentages (%).

Model	Shape Disamb.	Color Disamb.	Letter Disamb.	Visual Spatial	Visual Closure	Figure-Ground	Form Constancy
GPT4o – Image	12.50	77.77	25.92	25.92	58.33	11.11	83.33
GPT4o – SVG	29.16	44.44	11.11	18.51	41.66	0.00	33.33
Claude-3.5 – Image	41.66	77.77	7.40	25.92	54.16	55.55	94.40
Claude-3.5 – SVG	33.33	77.77	11.11	11.11	50.00	11.11	40.70

Observations from these benchmark results indicate a general performance degradation when using generated SVGs. For instance, GPT4o's Form Constancy drops from 83.33% (Image) to 33.33% (SVG), and Claude-3.5's Figure-Ground performance falls from 55.55% (Image) to 11.11% (SVG). The extent of degradation varies across tasks; Color Disambiguation for Claude-3.5 shows identical performance (77.77%) for both Image and SVG, while for GPT4o, it drops significantly. Notably, Figure-Ground for GPT4o drops to 0% with SVG. An interesting outlier is GPT4o's Shape Disambiguation, which increased from 12.50% (Image) to 29.16% (SVG). These results underscore the difficulty MLLMs face in accurately translating visual information into SVG.

A detailed qualitative analysis of the SVGs generated by the MLLMs revealed several common categories of perceptual errors, providing insights into the visual attributes MLLMs struggle with:

1. **Shape Inaccuracies:** Including misidentification (e.g., circle as oval) and geometric distortion (e.g.,

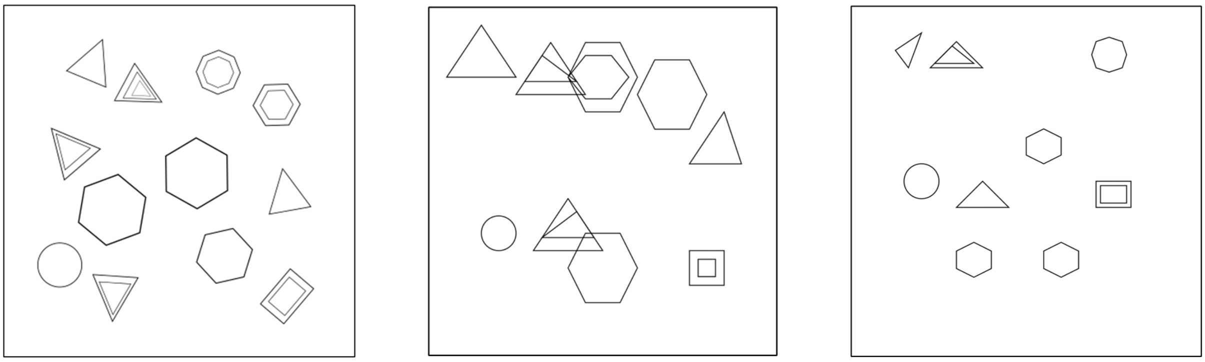


(a) Original Stimulus (b) Claude-3.5 Generated (c) GPT-4o Generated

Figure 13: Comparison of an original visual stimulus with MLLM-generated SVG representations (rendered as images). Such comparisons highlight perceptual errors made by the models.

equilateral triangle as scalene).

2. **Orientation and Positional Errors:** Incorrect rotation and inaccurate relative or absolute positioning of shapes.
3. **Color and Fill/Stroke Errors:** Misidentified colors or incorrect representation of fill and stroke properties.
4. **Count and Completeness Errors:** Omission of existing elements, inclusion of spurious elements, or incorrect counts of repeated shapes.
5. **Grouping and Hierarchical Structure:** General failure to capture complex grouping or hierarchical relationships (concentric shapes).



(a) Original Stimulus (b) Claude-3.5 Generated (c) GPT-4o Generated

Figure 14: Further comparison of an original visual stimulus with MLLM-generated SVG representations, illustrating common perceptual inaccuracies.

These qualitative errors in SVG generation directly reflect underlying perceptual limitations. For example, the inability to correctly render the number of sides of a pentagon indicates a failure in fine-grained shape discrimination.

In conclusion, using SVG reconstruction as an intermediate step proved challenging and often led to a decline in performance. However, this challenge served as a valuable diagnostic tool. The qualitative inaccuracies in MLLM-generated SVGs offer a more direct window into their perceptual processing (or misprocessing) than final task answers alone. Examining these structured visual outputs allows for more effective pinpointing of specific perceptual flaws, such as difficulties with shape constancy, spatial relationships, or attribute binding. This understanding is crucial for guiding future research towards developing MLLMs with more robust and accurate visual perception capabilities.

L Sparse Attention to Query Relevant Regions

This study investigates how Multimodal Large Language Models (MLLMs) allocate visual attention when performing **shape discrimination** (counting) tasks, analyzing the attention maps generated by the Vision Transformer (ViT) component of the MLLM to quantify the attention directed towards image regions crucial for accurately answering visual perception questions related to shape counts. A typical MLLM architecture comprises a vision encoder (V), often a ViT, an adapter module (A), and a Large Language Model (L). An input image I is processed by V into a sequence of visual embeddings or tokens, T_I. The textual prompt P (e.g., "How many squares are in the image?") is tokenized into T_P. The adapter A serves to align these visual embeddings with the language embedding space, transforming T_I into T'_I. This alignment can involve simple linear projections or more complex mechanisms like cross-modal attention layers, ensuring that visual and textual information can be jointly processed. The resulting visual tokens T'_I and prompt tokens T_P are then combined (e.g., concatenated or interleaved) to form a unified multimodal sequence, $\{T'_I, T_P\}$, which is subsequently processed by L to autoregressively generate the textual answer.

The ViT within V processes the input image by dividing it into a grid of patches. Self-attention mechanisms are applied across these patches through multiple layers, allowing the model to weigh the importance of different patches when constructing the representation for each. These attention weights can be aggregated across attention heads and layers to produce an overall attention map that highlights regions of the image the ViT focused on.

In this study, we focus on the **shape discrimination** task, where the MLLM is asked to count instances of a specific shape. To analyze attention, "query-relevant regions" or "important areas" are defined as the pixels occupied by the target shapes in the input image. This typically involves using ground-truth bounding boxes or segmentation masks for each instance of the shape to be counted. Attention maps are then extracted from the ViT layers. By overlaying these attention maps with the masks of the query-relevant regions, the average attention score directed towards these crucial areas is calculated. This quantitative measure reflects the extent to which the vision encoder focuses on the visual elements essential for correctly performing the counting task. In our experiment with the Phi-4-Multimodal-Instruct model, we find that for the **shape discrimination** task, the average attention paid to query-relevant regions within the images was a low **9.79%**. This observation underscores the importance of future work focused on enhancing MLLM mechanisms for identifying and attending to query-relevant visual information more effectively.

M Do Models exploit shortcuts to answer viusal perception questions?

Table 10 indicates that dimensions where verbalization of the image is difficult such as visual closure, visual form constancy, there is a clear drop in performance when model is prompted with CoT style prompt. Whereas, tasks such as visual spatial, and letter disambiguation show a drastic performance improvement when prompted with CoT. Note: We generated the CoT based responses by adding a simple "think step-by-step" prompt to the original question.

Dimension	**CoT Prompt**	**Regular Prompt**
Letter Disambiguation	47.4	31.85
Form Constancy	55.19	74.07
Visual Closure	48.21	57.74
Visual Spatial	35.86	28.91
3D Visual Spatial	38.75	31.25
3D Color Disambiguation	79.17	95.83
3D Letter Disambiguation	32.29	22.92
3D Shape Disambiguation	80.0	81.67
3D Form Constancy	30.0	48.75

Table 10: Performance comparison between CoT Prompt and Regular Prompt across different visual dimensions

N Limits of Visual Perception

N.1 Visual Form Constancy: Sensitivity to Rotation at Varying Scales

(a) $Square\ Size(D = 28); rotation = 2$

(b) $Square\ Size(D = 56); rotation = 2$

Figure 15: Form Constancy

This experiment investigates the MLLM's ability to maintain visual form constancy by detecting subtle rotations of an object, with the primary goal of determining how this perceptual capability is influenced by the absolute size of the visual stimuli, especially in relation to P, the patch size of the vision-encoder (assumed to be 14 pixels for this study). Stimuli consisted of images containing two squares placed side-by-side. The left square served as a static, unrotated reference, while the right square was the target, potentially rotated by a specific angle. An example of such a stimulus is depicted in Figure 15a. The MLLM was prompted with the question: "Is the right square rotated with respect to the left one? Answer Yes or No." Key parameters varied included the square size (D), ranging from approximately $0.5P$ to $8P$, and the rotation angle (θ_r), from 0 to 4. Accuracy was evaluated based on the model's correct identification of rotation presence or absence.

The performance of Claude-Sonnet-3.5 in detecting rotations is illustrated in Figure 15. With no rotation (0), the model achieved 100% accuracy for both 28px ($2P$) and 84px ($6P$) squares specifically highlighted in initial tests, and generally across other tested sizes as shown in the broader heatmap analysis. However, as seen in Figure 16, a subtle 1 rotation caused accuracy to drop to 0% for sizes up to $2P$, highlighting a significant challenge in perceiving minimal rotational changes, even when the object size substantially exceeded the patch size P.

A clear trend emerges regarding the interplay between object size and the ability to detect rotations. When no rotation was applied (0), the model consistently achieved perfect accuracy (1.00) across all tested square sizes, correctly identifying the absence of transformation. The challenge arises with the introduction of even minimal rotation. For a 1 rotation:

- With sub-patch size squares ($D = 7$px, $\approx 0.5P$), accuracy was 0.00.
- With patch-sized squares ($D = 14$px, $\approx 1P$), accuracy remained at 0.00.
- For squares twice the patch size ($D = 28$px, $\approx 2P$), accuracy was still 0.00.
- However, for larger squares ($D = 56$px, $\approx 4P$, and $D = 112$px, $\approx 8P$), the model achieved perfect accuracy (1.00), successfully detecting this subtle 1 rotation.

This pattern indicates a critical size threshold for perceiving minimal rotations. Objects at or below $2P$ were insufficient for the model to discern a 1 change, but larger objects ($\geq 4P$) provided enough visual information.

As the rotation angle increased, performance improved for smaller square sizes:

- For $D = 7$px (sub-patch), the model failed to detect rotations up to 4.
- For $D = 14$px (patch-size), accuracy remained at 0.00 for rotations up to 3, with a partial recovery to 0.67 at 4. This suggests that even when an object is nominally the size of a patch, its internal features might not be sufficiently resolved to detect small angular changes until the rotation becomes more pronounced.
- For $D = 28$px ($2P$), accuracy was 0.00 at 1, recovered to 0.67 at 2, and reached 1.00 for 3 and 4 rotations.
- For $D = 56$px ($4P$) and $D = 112$px ($8P$), the model maintained 1.00 accuracy across all tested rotation angles from 1 to 4.

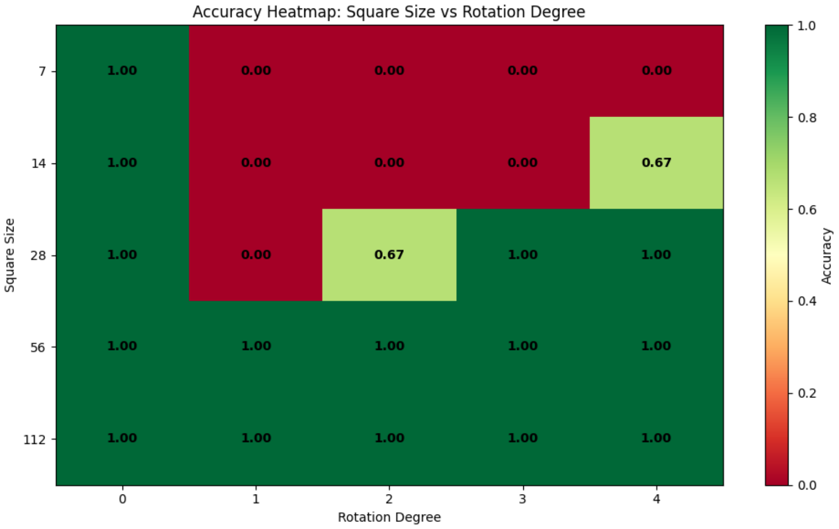
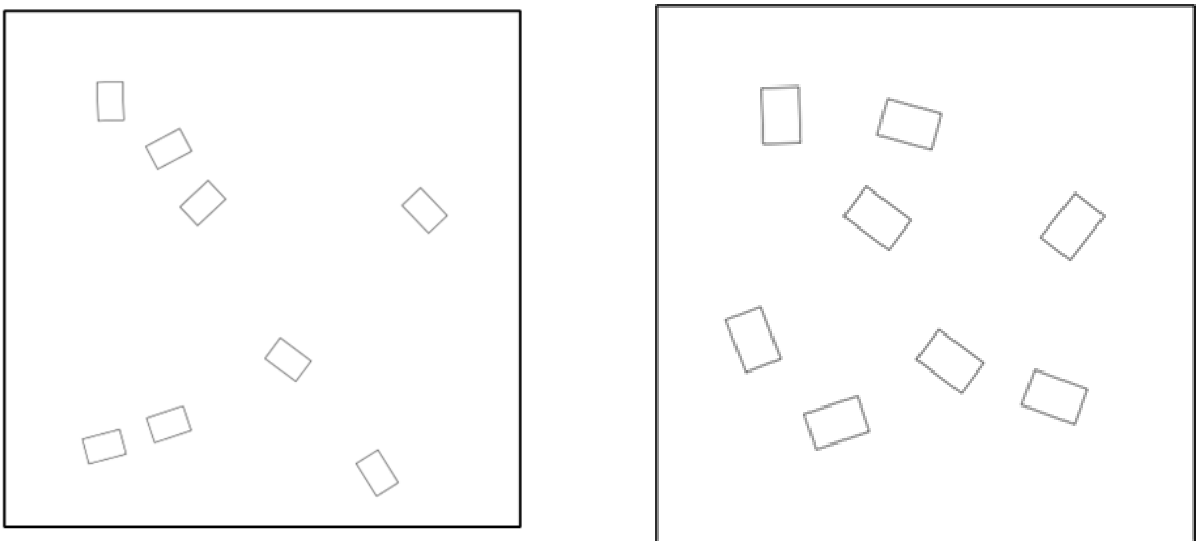


Figure 16: Claude-Sonnet-3.5 performance on rotation detection for various square sizes (scaled relative to patch size P) across different rotation angles ($0 - 4$).

These results refine our understanding of MLLM perceptual limits. Firstly, there is a clear perceptual threshold for detecting subtle rotations, which is heavily dependent on the object's size relative to the encoder's patch resolution. Sub-patch ($0.5P$) and patch-sized ($1P$) objects are particularly challenging for detecting small angular displacements. Secondly, even for objects larger than a single patch (e.g., $2P$), very subtle rotations (1) can remain imperceptible. It appears that a significantly larger effective object size (around $4P$ or more in this experiment) is required for robust detection of minimal (e.g., 1) rotations. This suggests that the model requires a certain aggregation of features over multiple patches, or a higher resolution representation of the object's boundaries, to reliably discern such fine-grained geometric transformations. The findings underscore that robust form constancy is not merely about resolving an object but also about accurately interpreting its subtle geometric attributes, a capability that scales with the object's representation quality, which in turn is linked to its size relative to the visual encoding mechanism.

N.2 Visual Discrimination: Shape Counting Accuracy at Varying Scales

This experiment evaluates the MLLM's visual discrimination capabilities, focusing on its accuracy in counting instances of simple rectangles under varying conditions of object size and number of instances (density). The study investigates how these factors influence enumeration, particularly considering the object size relative to P. Stimuli comprised images with a varying number of non-overlapping rectangles (specifically squares in this instance) of uniform size within each image. An example stimulus is shown in Figure 17. The MLLM was prompted: "How many rectangles are in the image? Answer with a number." Parameters varied were the number of rectangles (2 to 8) and their size. The "Scaling Factor" in the results (Figure 18) corresponds to rectangle sizes (D) from approximately 7px ($P \times 0.5$) to 112px ($P \times 8$). Accuracy was measured by comparing the model's count to the ground truth. *Note that the recatngle size here refers to the longer edge, the shorter edge is fixed at 0.8*rectangle size.*



(a) $Rectangle\ Size(D = 28); Instances = 8$ (b) $Rectangle\ Size(D = 56); Instances = 8$

Figure 17: Shape Discrimination

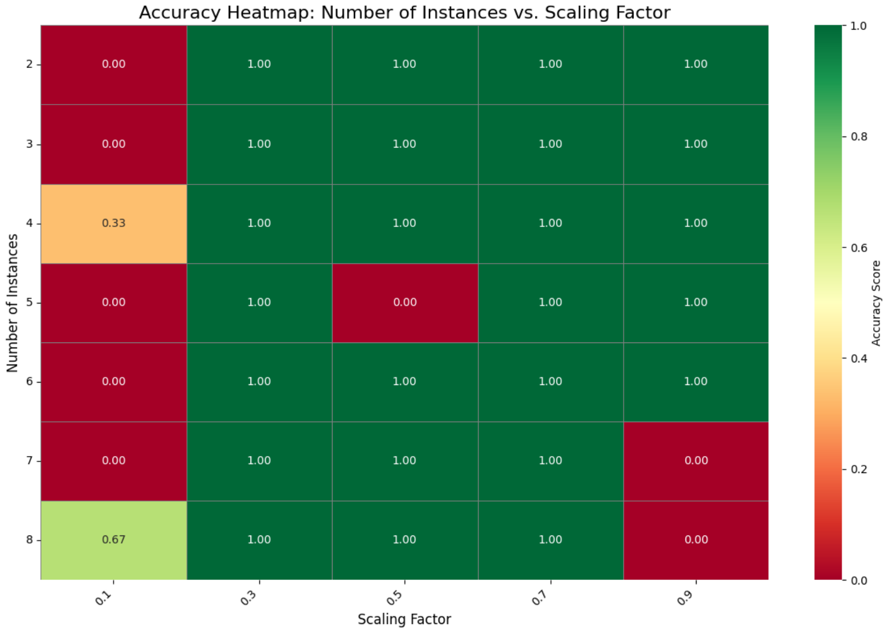


Figure 18: Accuracy heatmap for Claude-Sonnet-3.5 on rectangle (square) counting. Performance varies with the number of instances and scaling factor (size). Darker green indicates higher accuracy.

Figure 18 presents a heatmap of Claude-Sonnet-3.5's counting accuracy. A critical factor was object size: for the smallest rectangles (Scaling Factor 0.1, $D \approx P \times 0.5$), accuracy was consistently low

(0.00 for 2-7 instances, with what appears to be an anomalous 0.67 for 8 instances, possibly a statistical blip). This suggests significant difficulty in resolving or individuating sub-patch-sized objects, likely due to information conflation within single patches. Conversely, for larger rectangles (Scaling Factor ≥ 0.3, $D \geq P \times 2.1$), the model generally achieved perfect or near-perfect accuracy (1.00) across most instance counts. This indicates that if individual objects are clearly resolvable, the basic counting task is manageable. However, anomalous performance drops occurred even for larger rectangles at specific instance counts (e.g., 5 instances at SF 0.5; 7 and 8 instances at SF 0.9, all dropping to 0.00 accuracy). These failures, not attributable to object resolvability by size alone, might stem from attentional lapses, visual crowding effects where objects impair perception of each other, or specific challenging spatial configurations arising from random placement.

In conclusion, MLLM visual discrimination is highly dependent on object size relative to patch resolution. While objects significantly larger than P are generally counted accurately, sub-patch sized objects pose a substantial challenge. Furthermore, unexpected failures with larger objects suggest that factors beyond simple size and count, such as attention mechanisms, resilience to clutter, and interpretation of spatial arrangements, also critically influence perceptual accuracy. These experiments collectively highlight that even fundamental visual tasks can expose nuanced limitations in MLLM perception, emphasizing the need for continued research into enhancing the robustness and fidelity of their visual understanding.

O Detailed Results

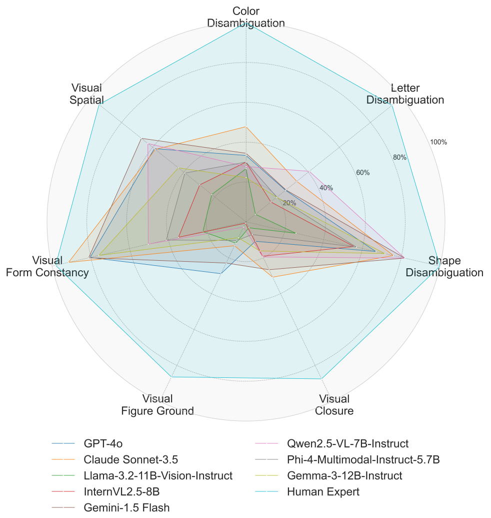


Figure 19: *MLLM performance on the seven subtasks in* ***Do You See Me (2D)*** *benchmark dataset.*

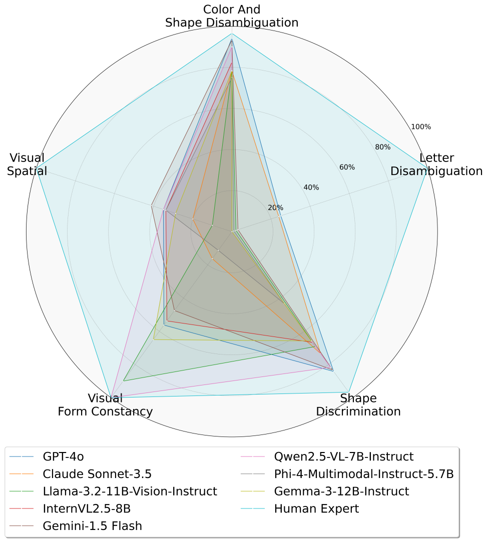


Figure 20: *MLLM performance on the seven subtasks in* ***Do You See Me (3D)*** *benchmark dataset.*

Table 11: Comparison of model performance. Claude Sonnet-3.5 leads in both *reasoning* and *visual perception* questions.

Model	**Reasoning Acc. (%)**	**Perception Acc. (%)**
Claude Sonnet-3.5	**40.95**	**45.21**
GPT-4o	32.97	42.55
Gemini 1.5 Flash	32.97	44.68
Qwen2.5-VL-7B-Instruct	35.63	35.10
Intern2.5-VL-8B	27.66	37.23
Phi-4-Multimodal-Instruct-5.7B	28.72	29.78
Llama3.2-11B-Vision-Instruct	26.06	31.91

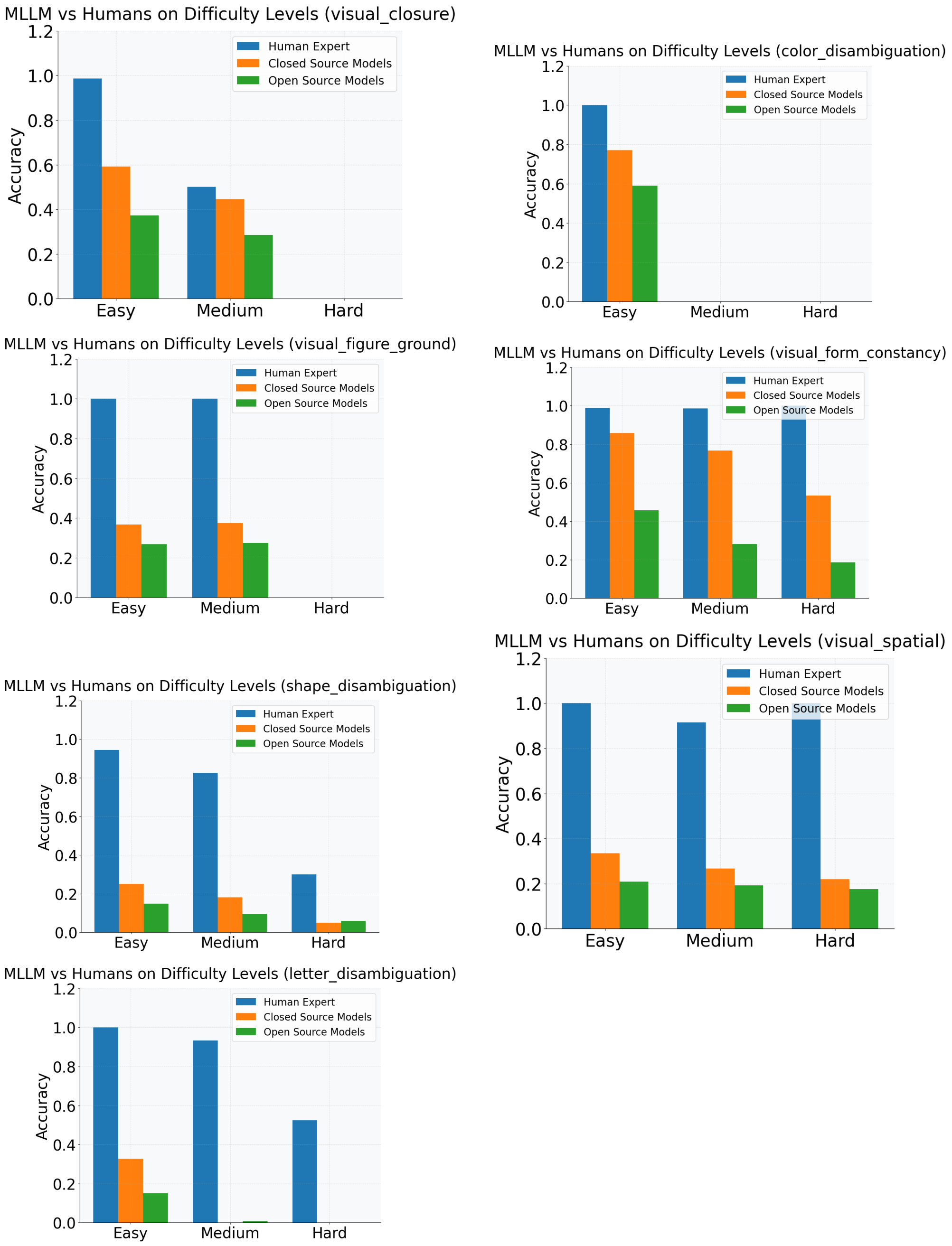


Figure 21: MLLM performance on Human Rated Difficulty Levels. Note: *Empty human bar for a difficulty level indicates that no samples were attributed the corresponding difficulty level.*

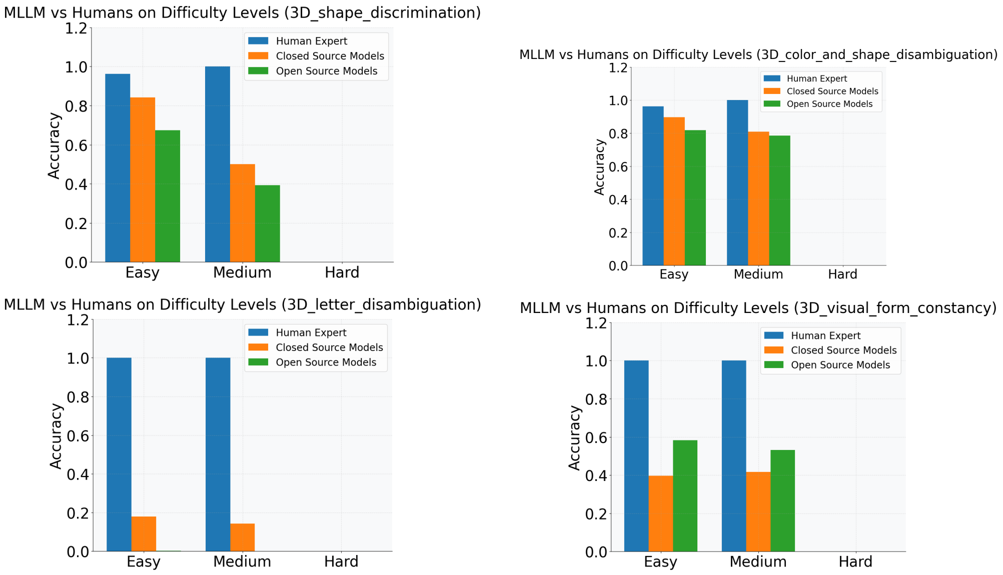


Figure 22: MLLM performance on Human Rated Difficulty Levels. Note: *Empty human bar for a difficulty level indicates that no samples were attributed the corresponding difficulty level*(3D tasks).

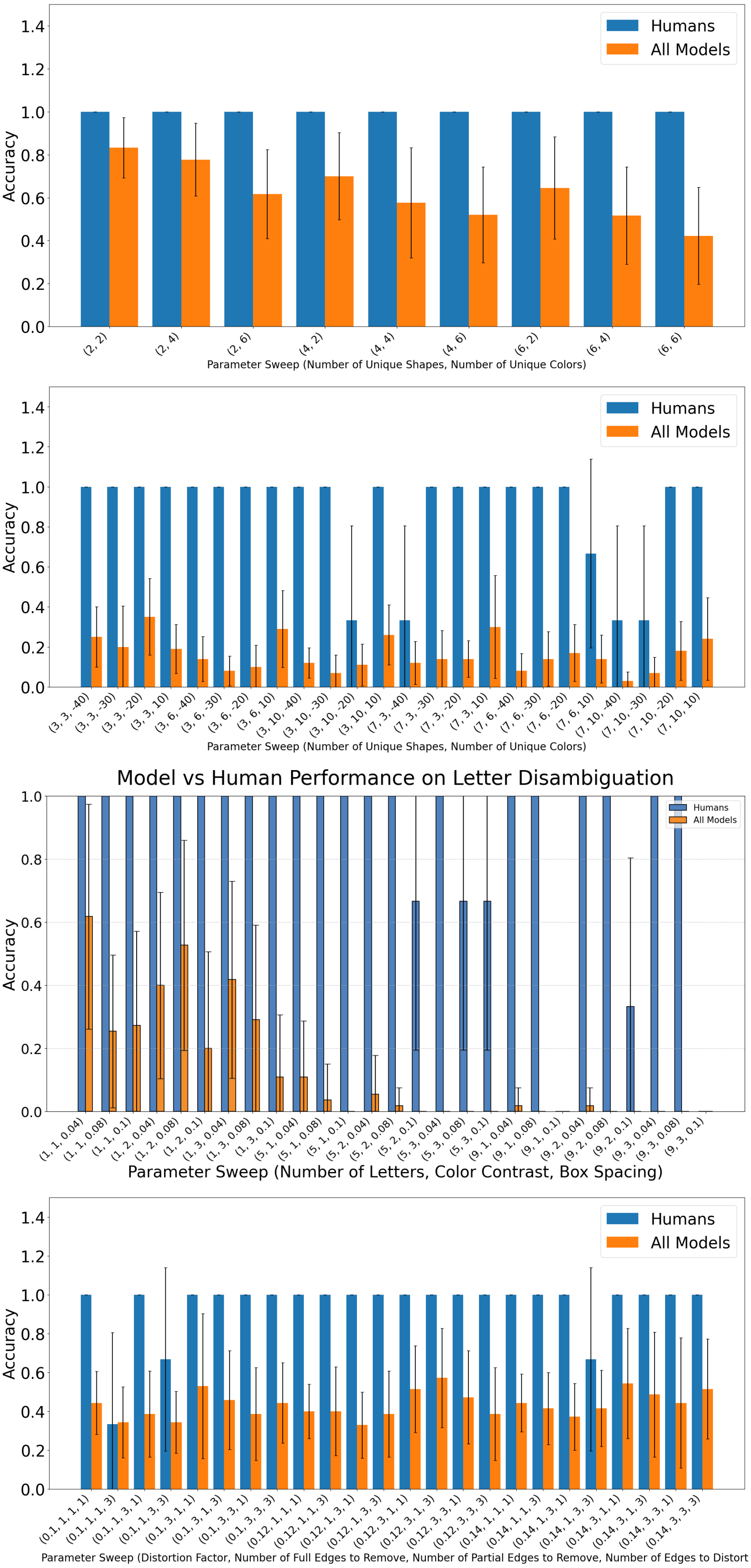


Figure 23: Average MLLM performance over a sweep of combinations of control parameters.

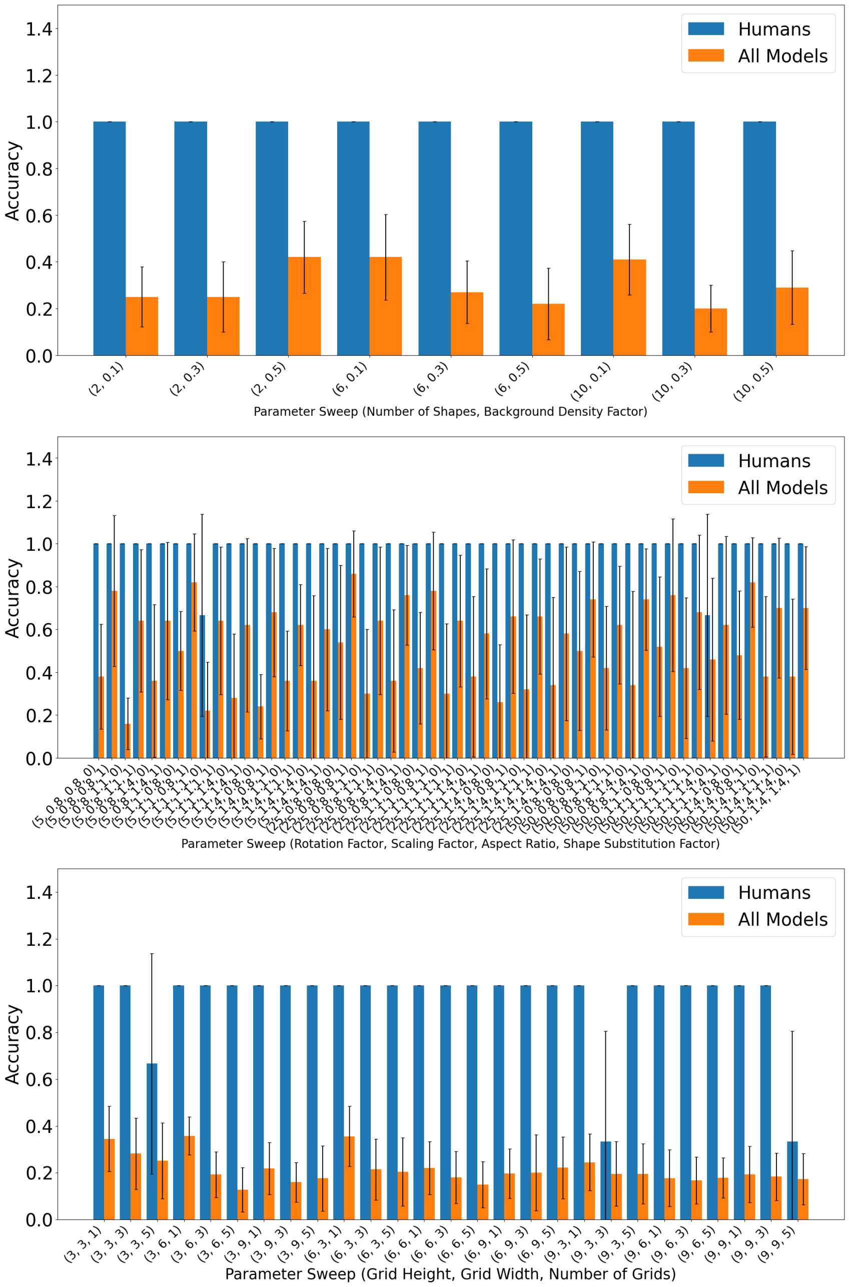


Figure 24: Average MLLM performance over a sweep of combinations of control parameters.

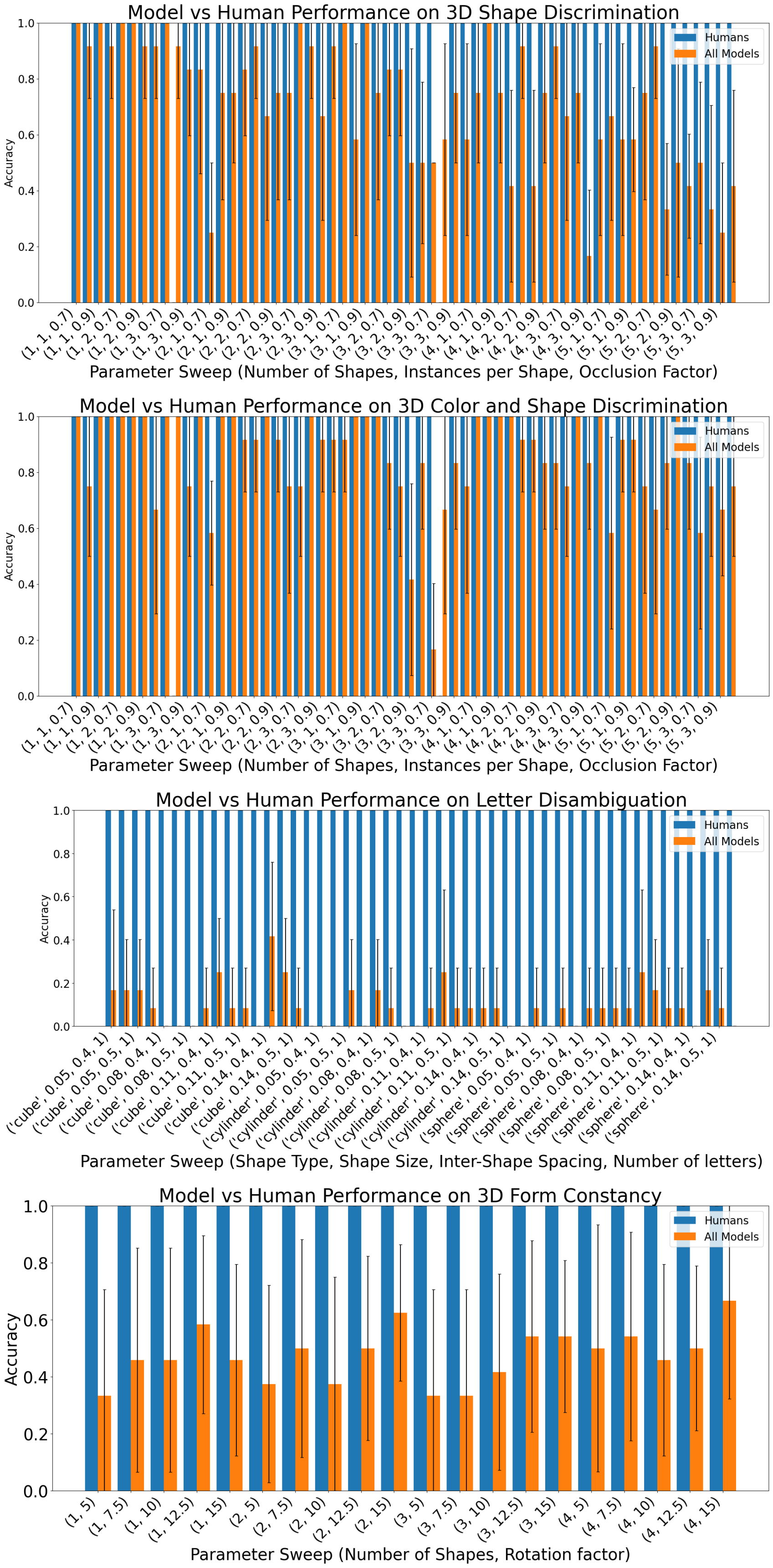


Figure 25: Average MLLM performance over a sweep of combinations of control parameters (3D tasks).