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

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated promising reasoning capabilities in diverse domains, yet their visual perception skills remain a critical bottleneck. In this study, we first investigate the impact of visual perception errors on visual reasoning questions by analyzing the performance of the model on 150 questions. Our findings reveal that incorrect answers often stem from failures in visual perception. In addition, some correct answers arise from hallucinated visual details. Towards this end, we introduce **Do You See Me**, a comprehensive benchmark dataset designed to evaluate visual perception capabilities in MLLMs. Drawing from established principles of human psychology, our dataset comprises 1,262 programmatically generated images that systematically assess different dimensions of visual understanding. Our benchmark consists of seven perception-focused subtasks, each designed with control parameters to modulate task complexity. Additionally, it can be easily extended for new perception tasks and varying complexities. We evaluate multiple state-of-the-art closed source and open source MLLMs and we conduct a comprehensive human study where participants are instructed to complete visual perception tasks and rate the difficulty of each test sample as easy, medium, or hard. Results indicate that MLLMs perform poorly on visual perception tasks, achieving less than 50% accuracy on most subtasks, whereas, humans perform with an average accuracy of around 95%. Furthermore, as task complexity increases, MLLM performance declines drastically, while human performance remains stable. In one of our subtasks, when analyzing performance across human-rated difficulty levels, we observe a progressive widening of gap between human and MLLM accuracy - starting at 12% for easy samples and expanding to 45% for those rated as difficult. This pattern of deteriorating MLLM performance relative to consistent human accuracy is seen across all seven subtasks, highlighting a critical need for enhanced visual perception capabilities in MLLMs.

1 Introduction

Recent advances in Multimodal Large Language Models (MLLMs) have led to impressive performance in diverse areas, including mathematical problem-solving (Lu et al., 2023; Zhang et al., 2024a), physics-based questions (Saikh et al., 2022), and logical reasoning challenges (Xiao et al., 2024). Multiple visual question-answering (VQA) tasks have emerged as a critical benchmark for evaluating artificial intelligence systems, particularly in assessing their ability to combine visual perception with higher-level reasoning (Yue et al., 2023). While the long-term goal is to achieve superhuman reasoning across these tasks, an essential prerequisite is the ability to accurately perceive and interpret visual diagrams or figures—a capability rooted in fundamental spatial and form-recognition skills such as visual discrimination and spatial orientation.

Fig. 1 shows MLLM responses to a logical reasoning question based on pattern completion. We note that both Claude Sonnet-3.5 (Anthropic, 2024) and GPT-4o (OpenAI et al., 2024a) show visual perception errors in their reasoning steps for finding the correct answer with Claude Sonnet-3.5 even reaching the correct final answer despite improper visual grounding. Recent studies have highlighted incorrect or incomplete visual perception as a primary source of error when MLLMs attempt to answer visually grounded reasoning questions (Lu et al., 2023). *This observation underscores the importance of thoroughly understanding the visual-perception capabilities and failures even when a model's overall reasoning might appear correct.* Visual reasoning itself can be viewed as a composition of three distinct processes: *question comprehension (textual or visual), visual perception, and*

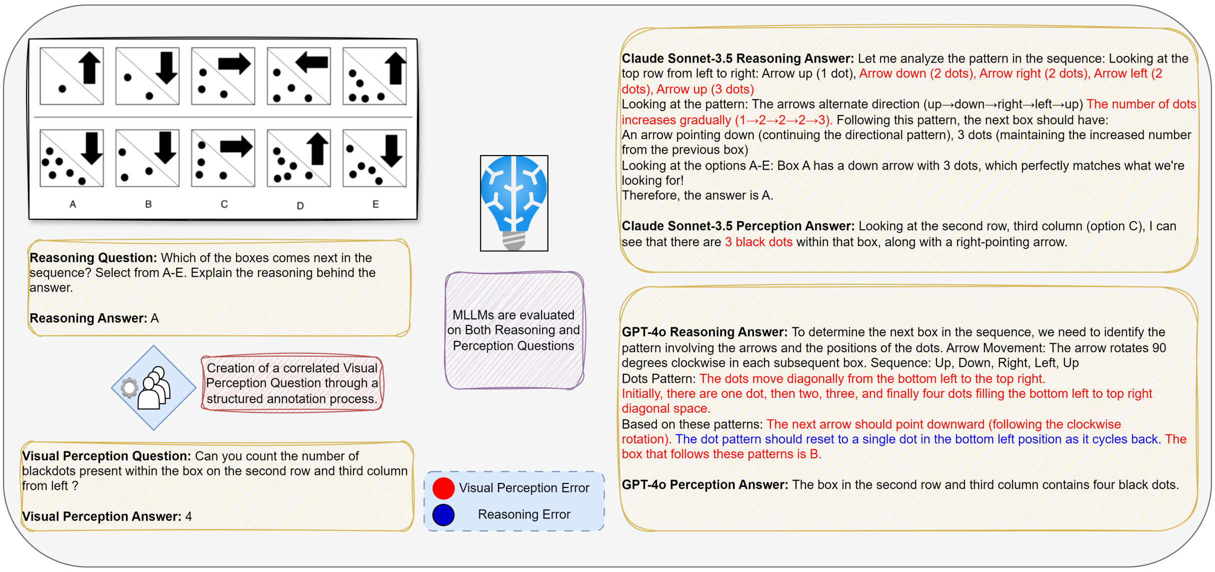


Figure 1: *Visual Misinterpretations in Popular Multimodal LLMs*

reasoning. Because accurate perception is often a prerequisite for coherent reasoning, a model's success in high-level tasks might be misleading if it masks underlying weaknesses in visual skills.

Recent benchmarks such as MMVP (Tong et al., 2024b) and CV-Bench (Tong et al., 2024a) do provide insights into general visual perception abilities in MLLMs, but they depend heavily on detailed human annotations—making them difficult to scale into larger benchmarks. In contrast, tests for human visual perception (e.g., the Test of Visual Perception Skills (Gardner, 1988)) commonly employ structured difficulty levels. For instance, in shape discrimination tasks, the number and arrangement of shapes gradually increase in complexity, and participants must identify the presence of a specific shape. As the test becomes more challenging, performance eventually reaches a ceiling, revealing the upper limit of a subject's visual perception capabilities. A similar approach for MLLMs—where stimuli become progressively harder—would help pinpoint the exact complexity threshold beyond which these models fail to respond correctly, thus providing a more precise measure of their perceptual strengths and weaknesses. This understanding can be crucial in designing more targeted pretraining datasets, ensuring that models receive the necessary variety and difficulty of visual samples. It can also guide architectural refinements, ultimately leading to more robust and accurate multimodal reasoning systems.

Drawing on human psychology (Chalfant and Scheffelin, 1969), visual perception can be categorized into five major interrelated skills: (1) visual discrimination, (2) visual figure-ground, (3) visual spatial, (4) visual closure, and (5) visual sequential memory. These categories provide a more nuanced framework to investigate where exactly models fail. Building on this perspective, our study addresses three primary research questions:

- **RQ1:** *If a MLLM correctly answers a reasoning question based on a visual diagram, does it necessarily succeed on a closely related perception question about the same diagram?,*
- **RQ2:** *How effectively do current MLLMs perform on various dimensions of visual perception skills?* and
- **RQ3:** *What patterns emerge in the successes and failures of MLLMs across different visual perception tasks and their complexities, and how do these patterns compare to human performance?*

To investigate **RQ1** (§2), we curate a joint visual perception-reasoning dataset comprising 150 unique images from logic-based Intelligence Quotient (IQ) tests. We opt for IQ-style questions because they focus on pattern recognition and "pure" reasoning with minimal domain-specific knowledge, making them well-suited for assessing raw reasoning abilities in MLLMs. Our dataset is constructed by merging samples from both MathVista (Lu et al., 2023) and LogicVista (Xiao et al., 2024) that feature logic-VQA tasks. For each image-question pair, we evaluate multiple MLLMs on both perception-oriented and reasoning-oriented questions. We then manually annotate each response, categorizing errors into four types: (1) visual perception error, (2) reasoning error, (3) arithmetic error, and (4) no error. Because a single response can exhibit multiple error types, we assign all applicable labels to each response. This comprehensive annotation process provides a more nuanced understanding of MLLM performance in both perception and reasoning.

Next, to investigate **RQ2** and **RQ3** (§3), we introduce the **Do You See Me** benchmark, a programmatically generated test suite inspired by the five core dimensions of human visual perception. Our benchmark comprises 1,262 unique images designed to test visual perception, accompanied by 2,116 evaluation questions (see Appendix D for details). Our benchmark consists of seven subtasks designed to evaluate different aspects of visual perception. For every subtask, we define a set of control variables (see Appendix Table 4) that can be systematically adjusted, enabling us to increase or decrease task difficulty as needed. Because **Do You See Me** is fully programmatic, we can generate a large number of Image-QA pairs across varied difficulty levels. Our design approach facilitates a fine-grained analysis of MLLMs' perceptual capabilities, making it particularly valuable for pinpointing the specific visual skills at which MLLMs excel—and where they continue to struggle.

Our experiments on the joint perception-reasoning dataset reveal that relying solely on final-answer accuracy may obscure critical shortcomings. For example, Claude Sonnet-3.5, which was *the best-performing model* on reasoning questions with around 41% accuracy, still produced reasoning chains containing visual perception errors. Specifically, around 43% of the correctly answered samples exhibited visual perception errors. While, on the incorrectly answered samples, both reasoning and perception errors were found in equal number of samples.

On the **Do You See Me** benchmark, we observe that no single model outperforms all others across the various dimensions of visual perception. Further, we note that while humans perform well on all the seven subtasks with an average accuracy of 94.31%, the best performing MLLM exhibits an accuracy of 50.05%. Additionally, we observe that as we use increasingly difficult combination of control parameters, MLLMs show a drastic drop in performance (dropping to near zero accuracy), whereas human performance remains largely stable. Results on the proposed **Do You See Me** benchmark clearly indicate that MLLMs fare poorly on visual perception skills, and exhibit a large gap on performance when compared to humans. This suggests an urgent need to improve visual perception capabilities in MLLMs, independently from high-level reasoning.

In summary, our key contributions are:

1. A joint dataset where reasoning questions on logic-based VQA are paired with closely related visual perception questions.
2. An extensive error analysis of SOTA MLLMs for their performance on the joint dataset, including extensive manual annotation of reasoning chains to identify the errors.
3. A *dynamic and synthetic test bench*—**Do You See Me**—that systematically measures fundamental visual-perception skills inspired by established categories in human psychology.
4. A comprehensive evaluation of various closed-source and open-source MLLMs against human performance, uncovering patterns of success and failure, and underscoring the gap between "reasoning" and "seeing" in current MLLMs.

2 Preliminary Study

2.1 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.

2.1.1 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. Table 1 summarizes the distribution.

Table 1: 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

2.1.2 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: $\big(I,\ (R, A_R),\ (P, A_P)\big)$. Five early-stage researchers proposed potential perception questions directly relevant to each reasoning question (e.g., "How many triangles are in the figure?"). Two senior annotators then refined these to ensure each perception question was unambiguous and closely tied to the underlying reasoning skill. Ten ambiguous queries were discarded and replaced, leading to a total of 150 finalized samples.

3 Do You See Me

Human psychology systematically categorizes human visual perception as a combination of five core abilities (Chalfant and Scheffelin, 1969):

- **Visual Discrimination:** Ability to recognize dominant features (e.g., position, shape, form, color).

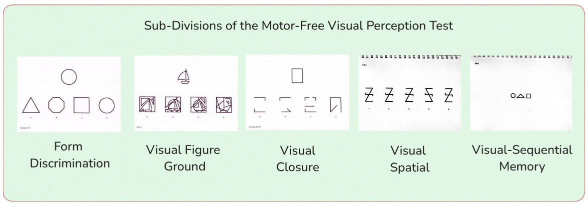


Figure 2: *A collection of samples for each subdivision of the TVPS test (Gardner, 1988)*

- **Visual Figure-Ground:** Ability to distinguish the main object from its background.
- **Visual Memory:** Ability to remember sequences of presented images.
- **Visual Closure:** Ability to complete partially obscured shapes.
- **Visual Spatial:** 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 questionnaires. Although MLLMs differ from human vision, these categories offer a solid framework for evaluating visual perception in MLLMs. Building on these principles, we introduce the **Do You See Me** benchmark, a fully automated test suite that dynamically generates both images and perception-focused questions (VPQA) of incremental difficulty. Our approach allows us to have a graded evaluation of MLLM performance over multiple perception dimensions (refer Appendix D for dataset distribution). Note that visual memory in addition to visual perception is also a test of short term visual memory in humans. Since the MLLMs currently have no memory, except via a textual description of the visual scene, we leave out visual memory from our benchmark. Our code is completely written in Python3 [1], and uses SVG representation to generate the visual images. We open-source the synthetic data generation code, and dataset at [2].

3.1 Visual Discrimination Tasks

We organize visual discrimination into four subtasks.

Shape Discrimination: This subtask assesses the ability to count specific shapes within a composite image. We use seven geometric shapes (*rectan-*

[1]https://www.python.org/downloads/

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

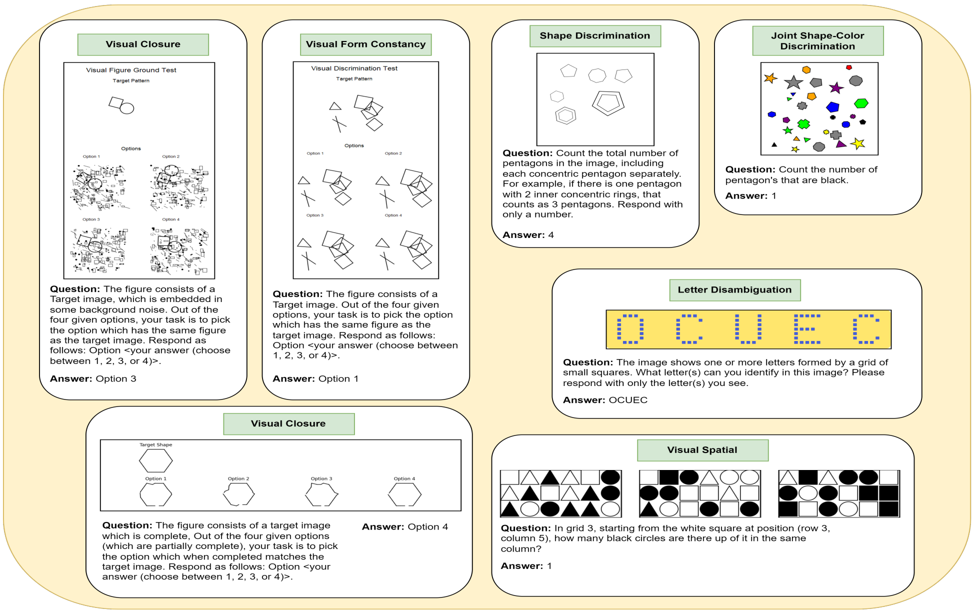


Figure 3: ***Do You See Me*** *benchmark visual perception dimensions*

gle, triangle, circle, pentagon, hexagon, octagon, star), each with a solid black border and transparent interior. Non-overlapping placements are enforced using the Separating Axis Theorem (SAT) (Gottschalk et al., 1996). The difficulty is controlled by three parameters: the number of unique shapes (S), the maximum instances per shape (S_I), and an overlap factor (α). Each generated image is paired with questions (for example, "How many circles are in the image?") and its programmatic ground truth answer.

Joint Shape-Color Discrimination: This subtask focuses on compositional counting queries (e.g., "Count all red triangles"), using six shapes (*star, triangle, pentagon, hexagon, octagon, cross*) and eight colors (*red, green, blue, orange, purple, black, gray, yellow*). We use non-overlapping shapes in the task to avoid ambiguity in shape recognition. Difficulty is governed by the number of unique shapes (S) and the number of unique colors (C).

Letter Discrimination: Letters are constructed from a 5×7 LED-style binary matrix, testing a model's ability to recognize block-based letters. Task difficulty depends on the block spacing factor (β), the color contrast (ΔC) between the letter blocks and the background, and the number of letters (N) present. Higher block spacing (β) or lower color contrast (ΔC) increases difficulty, as does adding more letters in the scene.

Visual Form Constancy: This subtask checks if a model can recognize a target pattern after shape substitutions, rotations, and scaling. The target pattern is constructed from simple primitives (*circle, square, line, triangle*). We create three variants of this pattern by applying a shape substitution factor (ssf), a scaling factor (α), and a rotation factor (θ_r). The MLLM must identify the option that exactly matches the target's arrangement.

3.2 Visual Spatial

This task uses one or more $H \times W$ grids with each cell containing one of three shapes (*circle, square, triangle*) rendered in solid black or outlined form. The parameters are the grid dimension (D) and the number of grids (G). The visual query is contructed instructing the MLLM to locate shapes by row-column coordinates or counting shapes relative to a reference position in the grid (e.g., "How many solid circles are above the triangle in row 3, column 2?").

3.3 Visual Figure-Ground

We extend the visual form constancy framework by introducing distracting background elements. Two parameters control the task complexity: 1) number of shapes (N) and, the background distraction factor (bdf). The *bdf* determines the amount of noise added to the background. The task of the MLLM is to distinguish the target pattern from other candidates under presence of background noise.

3.4 Visual Closure

This subtask evaluates whether the MLLM can match an incomplete shape with the corresponding fully formed target. We define seven basic shapes (*capsule, star, hexagon, circle, pentagon, rectangle, triangle*), remove some edges to create the incomplete target, and then produce three “noisy” distractors by randomly distorting vertex positions. The model must select which incomplete shape correctly closes to the original target.

3.5 Synthetic Data Generation

We specify a range of control parameters for each subtask (see Table 4) and generate all possible parameter combinations. Each combination yields ten images, each paired with multiple questions. The parameter limits are identified by conducting a preliminary study over visual perception accuracy for GPT-4o (OpenAI et al., 2024b). Design of each subtask is elaborated in Appendix B.

3.6 Human Performance Benchmarking

We compare modern MLLMs with seven human participants on the **Do You See Me** dataset. For each subtask, subjects answered two randomly selected questions per parameter combination and rated difficulty. A brief calibration phase preceded each subtask to reduce bias. Human accuracy, computed by matching typed answers to ground truth, provides a baseline for model comparisons (refer Appendix E for more details).

4 Experiments

4.1 Evaluation Protocol

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., 2024a), 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 G), which includes the task description, the question, and possible choices. Next, an answer extractor (Appendix G), based on GPT-4o (OpenAI et al., 2024b), 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.

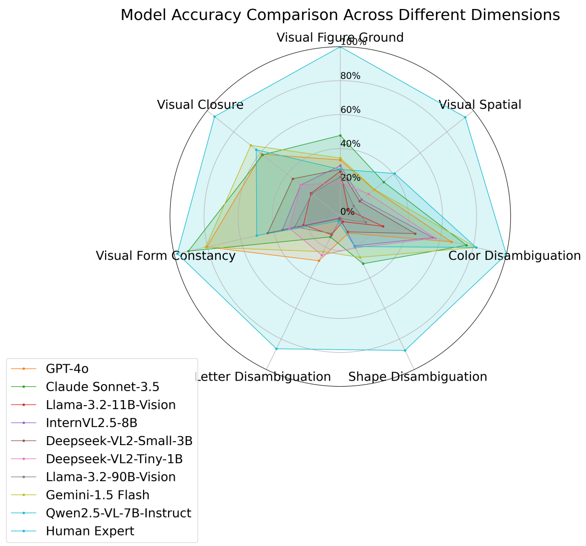


Figure 4: *Radar chart illustrating MLLM performance on the seven subtasks in* ***Do You See Me*** *benchmark dataset.*

4.2 Experimental Setup

We evaluate our *Do You See Me* dataset using both closed-source and open-source Multimodal Large Language Models (MLLMs). For **closed-source models**, we use GEMINI-1.5 FLASH (Team et al., 2024), GPT-4O (OpenAI et al., 2024a), and CLAUDE-SONNET-3.5 (Anthropic, 2024). For **open-source models**, our assessment includes LLAMA-3.2-11B-VISION (Grattafiori et al., 2024), LLAMA-3.2-90B-VISION (Grattafiori et al., 2024), DEEPSEEK-VL2-SMALL-3B (Wu et al., 2024b), DEEPSEEK-VL2-TINY-1B (Wu et al., 2024b), QWEN2.5-VL-7B-INSTRUCT (Wang et al., 2024), and INTERNVL2.5-8B (Chen et al., 2025). Each model is given identical inputs, both the same visual content and uniform textual prompts, to ensure fair comparison. The closed-source MLLMs are accessed through their respective proprietary APIs. In contrast, the open-source MLLMs are run locally with consistent hyperparameter settings, including batch size and decoding parameters.

4.3 Experimental Results

4.3.1 Joint Perception-Reasoning Dataset

In this section, we evaluate MLLMs on a **joint perception–reasoning dataset**, where each sample contains an image (I), a perception question ((P)), and a reasoning question (R). Each model is separately asked to answer both questions with explicit instructions tp respond with a detailed chain-of-thought. We use an LLM-based grader (see Section 4.1) to score the answers on correctness.

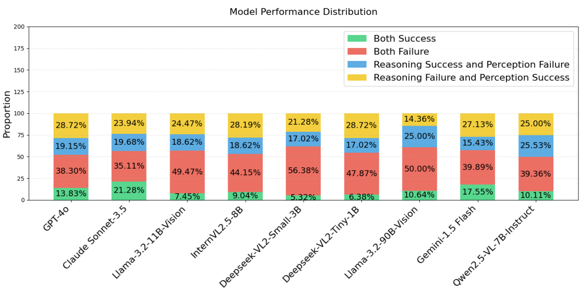
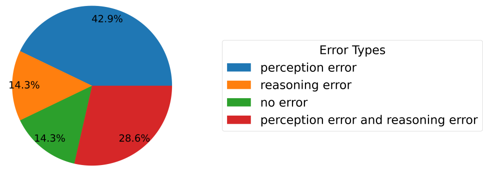


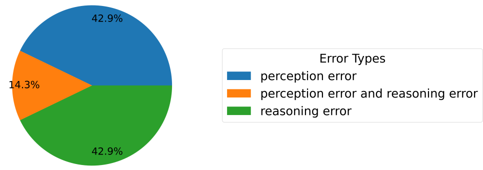
Figure 5: *MLLM performance on joint perception-reasoning questions*

Perception vs. Reasoning Performance. Figure 5 illustrates the distribution of joint performance on the perception and reasoning questions for each image. The stacked bars show four categories: 1) *both perception and reasoning correct*, 2) *both incorrect*, 3) *correct reasoning but incorrect perception*, and 4) *correct perception but incorrect reasoning*. Claude Sonnet-3.5 has the largest fraction of samples with both perception and reasoning correct, whereas DeepSeek-VL2-Small-3B has the highest rate of overall failure. Interestingly, there are some instances where a model provides a correct reasoning answer *despite* failing the corresponding perception question. This can be caused by one of three things: 1) the perception question being only loosely related to the reasoning question, 2) the reasoning question resulting in better interpretation of the figure than a direct visual perception question leading to a correct final answer, or 3) the model arriving at the correct reasoning conclusion through hallucinatory or shortcut strategies, not genuine understanding. Thus, a closer, fine-grained analysis is needed to pinpoint which factor dominates.

Error Breakdown in Reasoning Chains. To gain deeper insights, we manually annotated answers from top-2 models (Claude Sonnet-3.5, GPT-4o), labeling each error as *visual perception*, *reasoning*, or *arithmetic* (as defined in Section 2). Figure 6 presents the error distribution for Claude Sonnet-3.5 (see Appendix C for GPT4o results). We observe that, in the case where the final reasoning answer was determined to be incorrect, both visual perception and reasoning errors were a cause. Whereas, on cases where the final answer was correct, the detailed model response contained around 42.9% visual perception errors, with only 14.3% samples were answered error free. Clearly, these findings underscore the importance of improving visual perception in MLLMs. They also reinforce that success on a final question does not necessarily imply robust grounding in visual details or error free reasoning.



(a) Error distribution for correct final answers.



(b) Error distribution for incorrect final answers.

Figure 6: *Error distributions in reasoning chain for visual reasoning questions (Claude Sonnet-3.5).*

4.3.2 Do You See Me

In this section, we present a comparative evaluation of multiple MLLMs (and human experts) on a set of visual reasoning tasks from the **Do You See Me** benchmark. These tasks assess a broad range of perceptual and interpretive skills, including *visual closure*, *form constancy*, *figure–ground separation*, *spatial reasoning*, *color/shape disambiguation*, and *letter disambiguation*. **Overall Performance Across Dimensions** Figure 4 compares the accuracy of all models across the various subtasks. Each radial spoke corresponds to a distinct visual skill (e.g., visual closure or visual spatial), while the distance from the center (0%–100% accuracy) denotes performance. Consistent with the quantitative results in Table 2, **we observe that current MLLMs still lag behind human performance by a large margin across all subtasks**. Despite this gap, several interesting observations emerge: Most MLLMs are relatively strong visual form constancy. For instance, Claude Sonnet-3.5 achieves an accu-

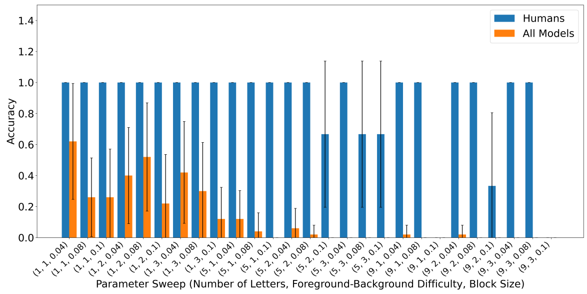
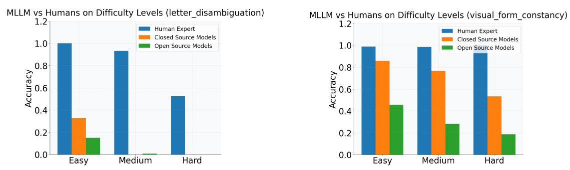


Figure 7: *MLLMs vs Human on Increasingly Difficult Combinations of Control Parameters for Visual Form Constancy Task.*

racy of 91.8%, the highest on form constancy. Further, MLLMs also perform relatively well on joint color-shape disambiguation with Gemini-1.5 Flash and Qwen2.5-VL-7B-Instruct performing the best. Finally, we observe that no single model is the best across all dimensions. Different tasks highlight distinct top performers; for example, Claude Sonnet-3.5 leads on form constancy, whereas Gemini-1.5 Flash and Qwen2.5-VL-7B-Instruct excel at color–shape tasks. Notably, the radar chart also reveals stark performance drops when tasks involve noise, occlusion, or overlap. For example, in visual figure–ground addition of noise in comparison to the form constancy task results in a gap of roughly 50% between top-performers. A similar trend appears between the shape discrimination task (which involves overlapping black outlines) and the joint shape–color task (no overlap). Thus, **it is plausible that MLLM accuracy deteriorates significantly as tasks incorporate occlusion or noise around critical visual areas.**



(a) Letter Disambiguation (b) Visual Form Constancy

Figure 8: *Human vs MLLM Performance on Human Rated Difficulty Levels*

Model Performance Over Increasing Difficulty In §3.6, we described our human study on the **Do You See Me** dataset, where participants provided *perceived difficulty* ratings for each combination of control parameters across all subtasks. We used majority voting to classify each combination as *easy*, *medium*, or *hard*. Next, we partitioned the accuracy results of each MLLM according to these levels. For clearer comparison, we grouped MLLMs into two categories—*closed-source* and *open-source*—and then averaged their performance within each difficulty level, alongside human performance. Figure 8 shows the results for two example tasks: *visual form constancy* and *letter disambiguation*. On visual form constancy—where MLLMs generally perform better than on other subtasks—both open- and closed-source models show progressively lower accuracies as human-rated task difficulty increases. Notably, the performance gap compared to humans widens: for closed-source models, it grows from about 12% at Easy to 20% at Medium and 45% at Hard. Closed-source models also outperform open-source models across all difficulty levels. Meanwhile, on the letter disambiguation task, as soon as the difficulty reaches Medium, both closed- and open-source models effectively fail (performance is near zero), whereas human accuracy remains high. We observed similar trends in the other subtasks as well (see Appendix C).

5 Conclusion

In this study, we examined the visual perception capabilities of MLLMs and their impact on logical reasoning. Our preliminary analysis revealed that many reasoning errors in MLLMs originate from incorrect visual perception, while some correct answers result from hallucinated visual details. These findings highlight the critical role of perception in multimodal reasoning and motivate the need for a systematic evaluation of MLLM visual skills. Motivated by these findings, we introduced the Do You See Me —a programmatic and scalable benchmark with a collection of seven visual perception subtasks inspired by human psychology.

Experimental results demonstrated that MLLMs struggle considerably on these subtasks compared to humans, particularly when complexity of the subtasks increase. While human performance remained robust across varying difficulty levels, MLLMs showed rapid declines in accuracy, reinforcing the need for improved visual processing. Overall, our study highlights the importance of developing more perceptually grounded MLLMs to reduce hallucinations and ensure reliable performance on visual reasoning and related tasks. Our approach of programmatic, scalable and complexity controlled data creation is not only suitable for evaluations, but for synthetic training data creation as well.

6 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. Second, while our dataset considers non-real-world settings for evaluating visual perception, we found this approach extremely useful for generating a vast number of diverse examples with a broad difficulty range. It is worth noting that most human visual perception tests follow a similar format. In future work, we aim to explore the use of 3D tools such as Unity3D [3] and Blender [4] to generate realistic 3D questions based on the same principles presented in our dataset, thus enhancing the ecological validity of our experiments. Third, 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

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

Ted Brown and Lisa Peres. 2018. An overview and critique of the test of visual perception skills – fourth edition (tvps-4). *Hong Kong Journal of Occupational Therapy : HKJOT*, 31:59 – 68.

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, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 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.

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

Stefan Gottschalk, Ming C Lin, and Dinesh Manocha. 1996. Obbtree: a hierarchical structure for rapid interference detection. *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*.

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, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew

[3]https://unity.com/products/unity-engine
[4]https://docs.blender.org/manual

Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 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.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*.

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

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chun yue 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 Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

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, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. 2023. Synthesize, diagnose, and optimize: Towards fine-grained vision-language understanding. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13279–13288.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

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, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kepa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

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, visioncentric 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.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024c. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9568–9578.

Guangzhi Wang, Yixiao Ge, Xiaohan Ding, Mohan S. Kankanhalli, and Ying Shan. 2023. What makes for good visual tokenizers for large language models? *ArXiv*, abs/2305.12223.

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. 2024a. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In *The Twelfth International Conference on Learning Representations*.

Penghao Wu and Saining Xie. 2023. V*: Guided visual search as a core mechanism in multimodal llms. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13084–13094.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. 2024b. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

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, Huan Sun, Yu Su, and Wenhu Chen. 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.

Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. 2023. Contextual object detection with multimodal large language models. *ArXiv*, abs/2305.18279.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *Preprint*, arXiv:2303.15343.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. 2024a. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*.

Zicheng Zhang, Haoning Wu, Erli Zhang, Guangtao Zhai, and Weisi Lin. 2024b. Q-bench+: A benchmark for multi-modal foundation models on low-level vision from single images to pairs. *Preprint*, arXiv:2402.07116.

A Detailed Related Work

Multimodal Large Language Models (MLLMs) have evolved from joint image-language pretraining approaches, exemplified by CLIP (Radford et al., 2021), BLIP (Li et al., 2022, 2023), and SigLIP (Zhai et al., 2023). With the rapid progress of large language models (LLMs), the 'vision as a token' paradigm has emerged, fostering numerous closed and open-source MLLMs that continue to advance through large-scale data and instruction finetuning (OpenAI et al., 2024b; Wu et al., 2024b; Wang et al., 2024). Various domain-specific benchmarks have been introduced in mathematics (Lu et al., 2023; Zhang et al., 2024a), science (Lu et al., 2022; Saikh et al., 2022), scientific documents (Li et al., 2024), and broader AGI contexts (Yue et al., 2023), each requiring intricate reasoning and domain knowledge. However, studies (Zhang et al., 2024a; Wu et al., 2024a) indicate that many VLM failures arise from inadequate visual perception, underscoring the need for robust, perception-focused evaluation.

Several datasets (Peng et al., 2023; Wang et al., 2023; Wu and Xie, 2023; Zang et al., 2023; Tong et al., 2024c; Wu et al., 2024a; Zhang et al., 2024b) do assess visual perception but often incorporate domain-specific reasoning. In contrast, visual benchmarks like MMVP (Tong et al., 2024b) and CV-Bench (Tong et al., 2024a) do provide insights into general visual perception abilities in MLLMs. However, these studies heavily rely on human annotation for dataset construction, which makes it difficult to scale them to larger sizes. Additionally, these datasets lack a mechanism to attribute difficulty levels to the visual perception of their constituent samples, which could be useful in determining fine-grained visual perception differences between MLLMs.

Insights from human psychology reveal that, visual perception encompasses five core dimensions (Chalfant and Scheffelin, 1969). Motor-free tests (Colarusso, 2003; Gardner, 1988), such as the Test of Visual Perceptual Skills (TVPS) (Gardner, 1988), Motor-Free Visual Perception Test (MVPT) (Colarusso, 2003), and Developmental Test of Visual Perception (DVPT) (Hammill et al., 2016) focus exclusively on perception, avoiding motor components. These, tests have evolved to assess children and adults (Brown and Peres, 2018), and have also been adapted for evaluating individuals recovering from visual cortex injuries (Brown and Peres, 2018).

Drawing on the extensive research in human psychology on human visual perception, we introduce the **Do You See Me** benchmark, leveraging established perception categories to evaluate VLMs.

B Do You See Me - Additional Details

B.1 Visual Discrimination

The visual discrimination dimension of our dataset has four subtasks: shape discrimination, joint shape-color discrimination, letter discrimination, and form-constancy. Below, we briefly outline our automated test generation for all the four visual discrimination sub-division.

Shape Discrimination: The difficulty of this task is controlled using: (a) number of unique shapes in the canvas (S), (b) maximum permissible instances per-shape (S_I), and (c) allowable overlap factor (α). The overlap factor α, refers to the separation amount for between shapes. When set to a non-negative value it results into a cluster of non-overlapping shapes, whereas, a negative value allows for some degree of overlap.

Letter Discrimination: The visual perception task's difficulty is modulated by three key factors: block spacing, letter count, and color contrast. The block spacing factor β determines the distance between a letter's constituent blocks, with $\beta = 0.1$ representing large spacing that makes letter identification challenging. As β decreases below 0.1, letter recognition becomes progressively easier. The task difficulty also increases with the number of letters present in the canvas. Color contrast provides another dimension of complexity, defined as $\Delta C = F - B$, where F represents the block color and B represents the background color. We implement three levels of contrast ($\Delta C \in 1, 2, 3$), where 1 indicates high foreground-background contrast and 3 represents low contrast. To generate these color combinations, we begin with seven common background colors in hexadecimal format, converting them to HSV representation. We then create three difficulty levels by manipulating the hue, saturation, and value: for easy combinations (high contrast), we rotate the hue by 180, for medium combinations, we rotate the hue by 90 while maintaining constant saturation and value, for challenging combinations (low contrast), we rotate the hue by 30 and apply a 10% distortion to both saturation and value. These parameters are summarized in Table 4

Visual Form Constancy: It is important to note that low rotation and scaling without substitution make the task more challenging, whereas higher values of rotation, scaling, or substitution simplify it by rendering the variants more visibly distinct.

B.2 Visual Closure

We begin by defining seven basic shapes: `capsule`, `star`, `hexagon`, `circle`, `pentagon`, `rectangle`, and `triangle`. Each shape is represented by a set of vertices connected to form the exact shape. A single shape is randomly selected as the target image. To create an incomplete version of this shape, we first choose k pairs of adjacent vertices and remove the connecting edges entirely. Next, we choose l other pairs of adjacent vertices (distinct from the first k pairs) for partial edge removal. This produces an incomplete shape which, if completed, would match the original target. To generate three distractors (incorrect options), we take the incomplete shape, introduce distortions to m vertices by adding noise specified by the distortion factor δ to the vertex coordinates. Each of the three distractors is created by repeating this process independently, resulting in three unique noisy variants. Finally, the correct incomplete shape (which can be closed to match the target) and the three distractors are shuffled and displayed horizontally, accompanied by the complete target shape on the top (see Fig 3).

C Detailed Results

Table 2: Performance comparison across different visual perception tasks. All values are percentages (%). Models are sorted by average accuracy.

Model	Figure Ground	Visual Spatial	Color Disamb.	Shape Disamb.	Letter Disamb.	Form Const.	Visual Closure	Average Accuracy
Human	100.00	92.59	100.00	100.00	77.77	98.14	91.66	**94.31**
Claude Sonnet-3.5	47.78	32.63	75.98	30.83	13.33	91.48	58.33	50.05
Gemini-1.5	34.44	25.68	81.86	26.67	22.96	81.11	67.26	48.57
GPT-4o	33.33	25.43	66.91	10.83	28.89	80.37	58.93	43.53
Qwen-2.5	27.78	40.69	81.86	19.58	2.96	50.37	63.10	40.91
Deepseek-tiny	20.00	21.34	57.35	20.42	25.19	30.74	29.76	29.26
Intern-VL	30.00	16.00	55.64	19.17	0.74	34.81	29.76	26.59
Deepseek-small	27.78	14.52	45.10	10.00	1.48	43.70	35.71	25.47
LLaMA-11B	26.67	6.08	25.74	3.33	11.85	22.22	22.02	16.84
LLaMA-90B	23.33	10.05	15.44	1.25	11.11	27.41	21.43	15.72

Table 3: 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
LLaMA3.2-Vision-Preview-90B	35.64	25.00
LLaMA3.2-Vision-Preview-11B	26.06	31.91
Deepseek-VL-Small 3B	22.34	26.60
Deepseek-VL-Tiny 1B	23.40	35.01

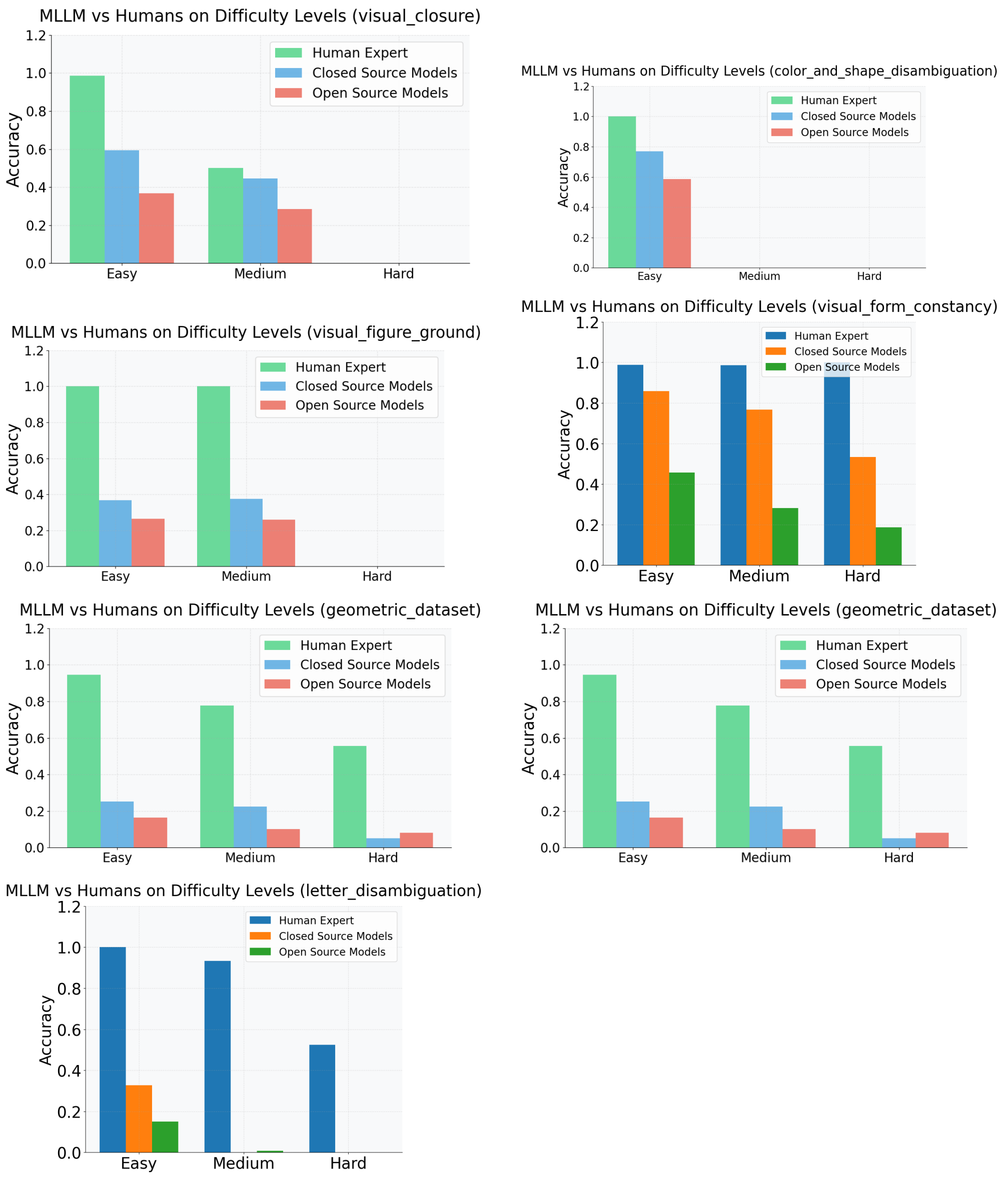


Figure 9: MLLM performance on Human Rated Difficulty Levels. Note: *Empty human bar for a difficulty level indicates that no samples were attributed the corresponding difficulty level.*

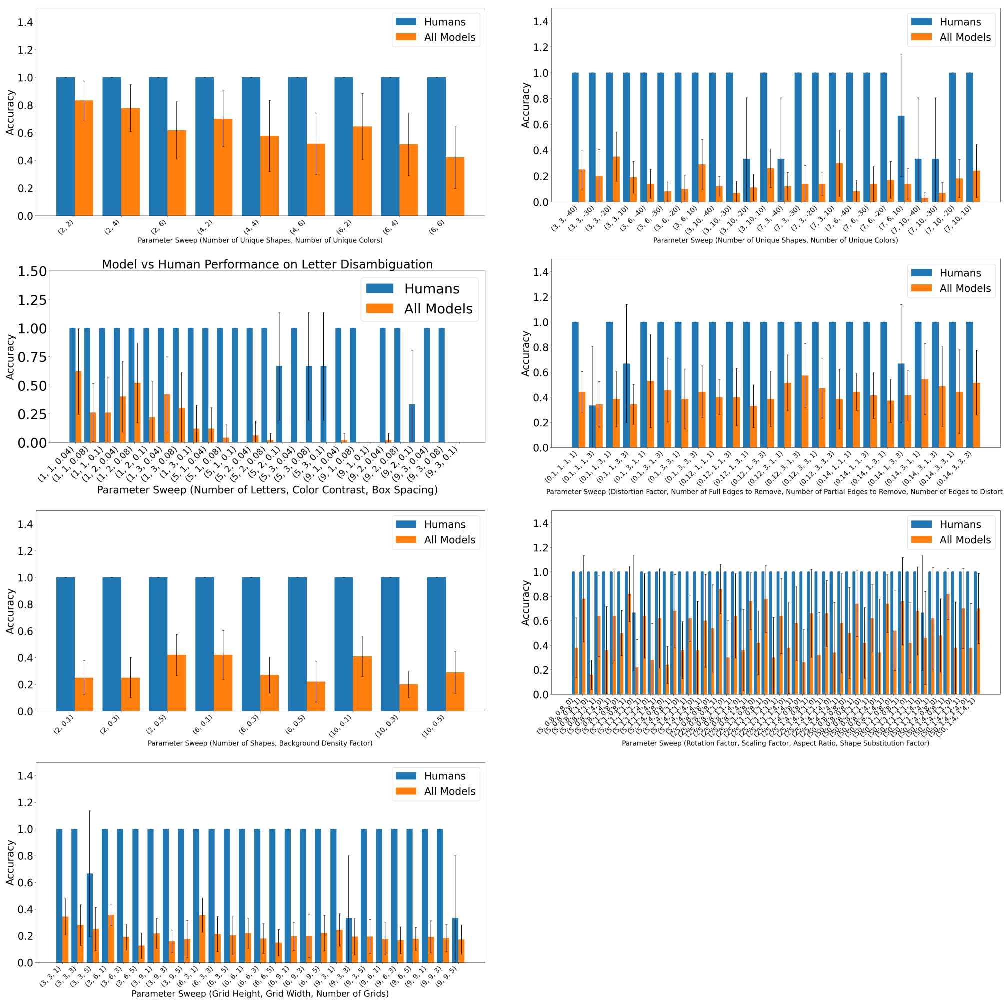


Figure 10: Average MLLM performance over a sweep of combinations of control parameters.

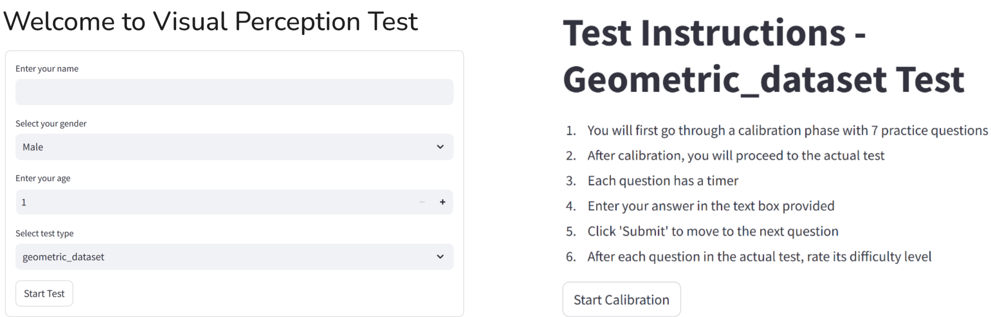
D Do You See Me - Control Parameter Details

Table 4: Control parameters and question types for each subdivision of the visual perception test.

Division	Subdivision	Control Parameters	Question Type	Unique Images	Number of Questions
Visual Discrimination	Shape Discrimination	Number of Shapes: $S \in [3, 7]$ Instances per Shape: $S_I \in [3, 6, 10]$ Overlap Factor: $\alpha \in [-40, -30, -20, 10]$	Integer	241	241
	Joint Shape-Color	Number of Shapes: $S \in [2, 4, 6]$ Number of Unique Colors: $C \in [2, 4, 6]$]	Integer	90	408
	Letter Discrimination	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]	Text	135	135
	Form Constancy	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: $\beta \in [0.8, 1.1, 1.4]$	MCQ	270	270
Visual Spatial	Spatial Grids	Grid Dimension: $D \subseteq [3, 6, 9] \times [3, 6, 9]$ Number of Grids: $G \in [1, 3, 5]$	Integer	270	806
Visual Figure-Ground	N.A.	Number of Shapes: $N \in [2, 6, 10]$ Background Density Factor: $bdf \in [0.1, 0.3, 0.5]$	MCQ	90	90
Visual Closure	N.A.	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: $\delta \in [0.1, 0.12, 0.14]$	MCQ	166	166

E Human Performance Benchmarking

We recruited 7 participants, consisting of 5 men and 2 women. The selection criterion was that participants had no history of vision-related disorders. All participants were between 25 and 35 years of age. Each test took approximately 10 minutes to complete, with a total testing time of 70 minutes. To prevent mental fatigue, we conducted the tests over two days.



(a) Information Collected in the Test

(b) Calibration Instructions

Figure 11: Human Performance Benchmarking

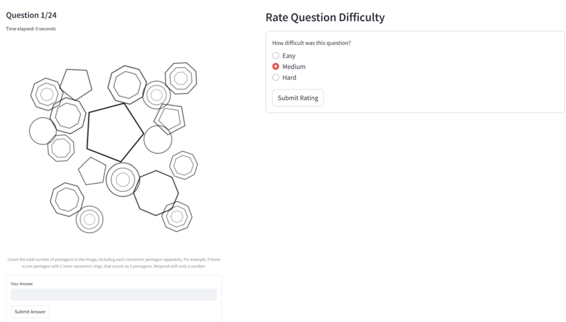


Figure 12: Actual Test and Human Difficulty Rating Collection

F Joint Perception-Reasoning Dataset

F.1 Correct Reasoning Incorrect Perception

Example 1

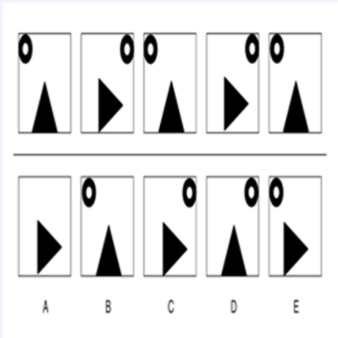


Figure 13

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

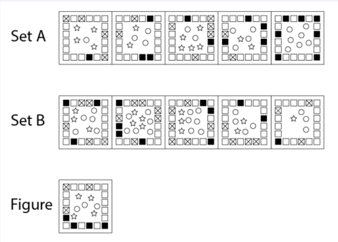


Figure 14

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.

1636

F.2 Qualitative Analysis of Reasoning Chain Errors

Example 1

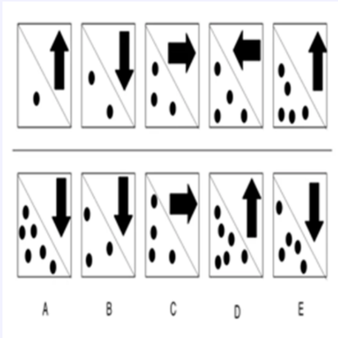


Figure 15

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

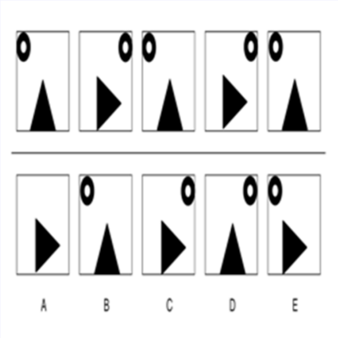


Figure 16

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*

1639

Example 3

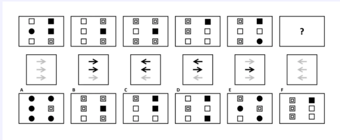


Figure 17

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

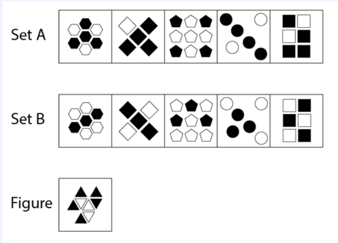


Figure 18

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.

G MLLM Prompts

Expert LLM Evaluator

You will be given three inputs consisting of: a question, the answer by a human giving the test, and the ground truth answer. The answer from the human can be a longform one, which can contain an explaination or reasoning. Your task is to determine if the final answer reached matches the ground truth answer.
Inputs
Question: {question}
Ground Truth Answer: {gt_answer}
Human Answer: {mllm_answer}
Your task is to reply in just YES (or) NO. If the answer matches your response should be YES, else NO.

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.