Visual serial processing deficits explain divergences in human and VLM reasoning

Nicholas Budny[1*], Kia Ghods[1*], Declan Campbell[1*], Raja Marjieh[2], Amogh Joshi[1], Sreejan Kumar[1]
Jonathan D. Cohen[1,2†], Taylor W. Webb[3,4†], and Thomas L. Griffiths[2,5†]

[1]Princeton Neuroscience Institute
[2]Department of Psychology, Princeton University
[3]Department of Psychology, Université de Montréal
[4]Mila - Quebec AI Institute
[5]Department of Computer Science, Princeton University
[*]Equal contribution
[†]Equal advising

Abstract

Why do Vision Language Models (VLMs), despite success on standard benchmarks, often fail to match human performance on surprisingly simple visual reasoning tasks? While the underlying computational principles are still debated, we hypothesize that a crucial factor is a deficit in visually-grounded serial processing. To test this hypothesis, we compared human and VLM performance across tasks designed to vary serial processing demands in three distinct domains: geometric reasoning, perceptual enumeration, and mental rotation. Tasks within each domain varied serial processing load by manipulating factors such as geometric concept complexity, perceptual individuation load, and transformation difficulty. Across all domains, our results revealed a consistent pattern: decreased VLM accuracy was strongly correlated with increased human reaction time (used as a proxy for serial processing load). As tasks require more demanding serial processing—whether composing concepts, enumerating items, or performing mental transformations—the VLM-human performance gap widens reliably. These findings support our hypothesis, indicating that limitations in serial, visually grounded reasoning represent a fundamental bottleneck that distinguishes current VLMs from humans.

1 Introduction

Large-scale vision-language models (VLMs), such as GPT-4o (Achiam et al., 2023), Claude-Sonnet 3.7 (Anthropic, 2023), and Gemini 2.5 Pro (Team et al., 2023), demonstrate impressive capabilities, often matching or surpassing human-level accuracy on complex benchmarks like image captioning and visual question answering (Fu et al., 2023). Yet, paradoxically, they also exhibit surprising deficits in tasks that appear comparatively simple, such as counting and visual search (Campbell et al., 2024b). Here, we hypothesize that these discrepancies, and a large amount of VLM-human divergence more broadly, stem from deficits in visually-grounded serial processing. To investigate this hypothesis and better evaluate the extent of these deficits, we use tasks from cognitive science (Dehaene et al., 2006) to vary serial processing demands systematically.

To provide a proxy for serial processing load, we use human reaction time (RT) — a well-established behavioral marker of serial cognitive engagement. Humans adapt to complexity in visual reasoning by trading time for accuracy, using sequential attention to focus on different elements when analyzing complex scenes (Heitz, 2014; Stewart et al., 2020). We hypothesize that VLMs largely lack this

Preprint. Under review.

capability, as their reasoning is tethered to sequential generation of text (their chain of thought), not sequential analysis of the image. We therefore predict that VLM accuracy will be inversely correlated with human RT, indicating that as tasks require more extensive serial processing by humans, the VLM-human performance gap widens. This method allows us to identify where VLM performance diverges most significantly from human performance and to evaluate whether these divergences align with visual concepts and operations hypothesized to demand more intensive serial processing.

To test our hypothesis, we systematically manipulate the serial processing load in three distinct domains. First, in a geometric program induction task inspired by work modeling geometric concepts as programs (Hsu et al., 2022), we vary load by manipulating concept complexity, operationalized as the number of geometric primitives (Minimum Description Length; MDL) required to define a geometric concept (Sablé-Meyer et al., 2022). Second, for visual enumeration, drawing on work distinguishing subitizing from serial counting (Trick & Pylyshyn, 1994), serial load is varied by manipulating color distinctiveness and spatial overlap, known to increase demands on serial attention (Franconeri et al., 2013). Finally, in a mental rotation task based on (Shepard & Metzler, 1971), serial processing demands are modulated by varying the angular disparity between two stimuli, a factor shown to linearly increase human processing time (Cooper & Shepard, 1973). Across these diverse tasks we demonstrate that as task conditions necessitate greater serial processing the VLM-human performance gap widens reliably, providing convergent evidence for a serial processing deficit.

2 Background and Related Work

2.1 Vision Language Models

Vision Language Models (VLMs) are designed to jointly model visual and linguistic information, typically by learning multimodal representations that bridge inputs from both domains (Radford et al., 2021; Alayrac et al., 2022). These models commonly incorporate a vision encoder to extract features from images or videos, which are then interfaced with a large language model conditioned on these visual features alongside textual prompts (Achiam et al., 2023; Team et al., 2023; Anil et al., 2023; AI, 2024). This architectural paradigm has enabled significant progress on a variety of downstream tasks, including image captioning, visual question answering (VQA), and, more recently, complex multimodal reasoning (Liu et al., 2023; Zhu et al., 2023). Despite these advances, the precise mechanisms by which these models integrate and reason about visual information remain an active area of study. Of particular interest is the extent to which the reasoning strategies learned by VLMs correspond to, or diverge from, human visual reasoning strategies.

2.2 Measuring correspondences between human visual processing and VLMs

To gauge the alignment between human and VLM visual reasoning, recent studies have benchmarked VLMs against human performance on a wide range of visual reasoning tasks from cognitive science. These comparisons have uncovered both similarities and crucial differences in their behavioral biases (Campbell et al., 2024b; Dasgupta et al., 2022). In some geometric reasoning tasks, VLMs exhibit remarkably human-like biases (Campbell et al., 2023, 2024a), successfully performing tasks involving perceptual grouping and identifying partially obscured objects (Allen et al., 2025; Pothiraj et al., 2025). The relative difficulty of spatial reasoning tasks also aligns between VLMs and humans (Xu et al., 2025), suggesting some commonalities.

Despite these similarities, a significant gap persists between VLM and human performance across several core cognitive tasks (Schulze Buschoff et al., 2025). VLMs are known to have deficits in spatially grounding reasoning (Tong et al., 2024; Kamath et al., 2023; Xu et al., 2025; Ramakrishnan et al., 2025), feature integration (Campbell et al., 2024b), and multi-step reasoning (Campbell et al., 2024b; Greff et al., 2020). These manifest in specific tasks where VLM performance is near-chance: determining whether circles overlap, counting, identifying circled letters (Rahmanzadehgervi et al., 2025), estimating depth, establishing visual correspondence, performing multi-view reasoning (Fu et al., 2024), and even failing at simple visual analogies that are solvable by toddlers (Goddard et al., 2025)—all tasks where human accuracy is at ceiling.

What unites these seemingly diverse failures? This constellation of behavioral traits (Ramakrishnan et al., 2025) currently lacks a unifying explanation. While various frameworks have catalogued VLM failures across different benchmarks (Schulze Buschoff et al., 2025; Ramakrishnan et al., 2025;

Huang et al., 2025), effectively demonstrating *when* performance falters, the question of *why* they fail in these settings remains unresolved. Existing explanations range from training data constraints to deficits in core knowledge, but no single framework accounts for the breadth of observed failures. We propose that a deficit in serial processing provides this missing unifying framework.

3 The serial processing deficit hypothesis

We hypothesize that these diverse perceptual failures stem from a fundamental deficit in visually grounded serial processing. Three observations support this hypothesis: VLMs perform well when visual elements are well-separated but fail when elements require sequential analysis (Rahmanzadehgervi et al., 2025); vision encoders contain sufficient information to solve these tasks, yet language models fail to effectively decode this information (Rahmanzadehgervi et al., 2025); and tasks that cannot be easily decomposed into linguistic steps pose particular challenges (Fu et al., 2024).

Our contribution is threefold: We provide the first investigation of serial processing as a unifying explanation for VLM failures across diverse visual reasoning tasks. We demonstrate that human RT serves as a reliable predictor of VLM performance deficits, offering a way to predict model failures on novel tasks. And finally, we show that these limitations hold across different model architectures and prompting strategies, suggesting a fundamental limitation rather than a superficial training gap.

3.1 Serial processing in VLMs

Inductive biases encouraging serial, step-by-step processing have been instrumental in unlocking advances in the mathematical reasoning and coding capabilities of Large Language Models (LLMs). The Tree of Thoughts framework (Yao et al., 2023), for instance, enables LLMs to systematically explore, evaluate, and prune multiple reasoning pathways, tackling problems that were previously intractable. The success of such methods in language, where decomposing complex tasks dramatically boosts performance, naturally motivated exploration of analogous strategies for Vision-Language Models (VLMs). Initial attempts to instill serial processing in VLMs — through fine-tuning (Deitke et al., 2024) and tool use (Yang et al., 2023) — have been successful in specific task domains. However, whether these initial successes will translate to robust performance across the diverse landscape of real-world visual reasoning tasks remains an open question. We extend this work in two ways: First, we evaluate a tool-augmented VLM across all three task domains to assess whether external tool use can overcome serial processing deficits. Second, we propose training regimes that could instill more generalizable serial visual reasoning capabilities in VLMs.

4 Test Domain 1: Geometric Concepts

The first domain in which we investigate a potential serial processing deficit in VLMs is geometric reasoning. A long tradition in cognitive science has examined how humans reason over geometric patterns, demonstrating that response times (RTs) increase systematically with the structural complexity of the stimulus (Just & Carpenter, 1985). In particular, a growing body of work has argued that humans interpret geometric figures by inferring the underlying generative rules that produced them, rather than merely matching surface features. This *program induction* approach has been explored in experiments that manipulate the minimum description length (MDL) of symbolic programs used to generate visual stimuli – where MDL reflects the number of discrete compositional steps required to construct the pattern from a fixed vocabulary of geometric primitives and relations (Sablé-Meyer et al., 2022). Such studies suggest that human geometric reasoning is fundamentally compositional and serial: more complex patterns require more sequential operations, resulting in longer RTs but typically preserved accuracy. Recent work introduced the *Geoclidean* domain-specific language (DSL) to generate 2D geometric stimuli by composing geometric primitives (Hsu et al., 2022). This work revealed that while humans are adept at inferring abstract geometric rules, standard transformer and CNN-based vision encoders are insensitive to this structure. However, it remains unclear whether modern VLMs, which incorporate language-based reasoning, exhibit similar deficits — particularly for complex concepts that demand greater serial processing in humans. We test this by parametrically varying the complexity of geometric concepts generated using the Geoclidean DSL, predicting that model performance will degrade with increasing complexity, even as human accuracy remains stable.

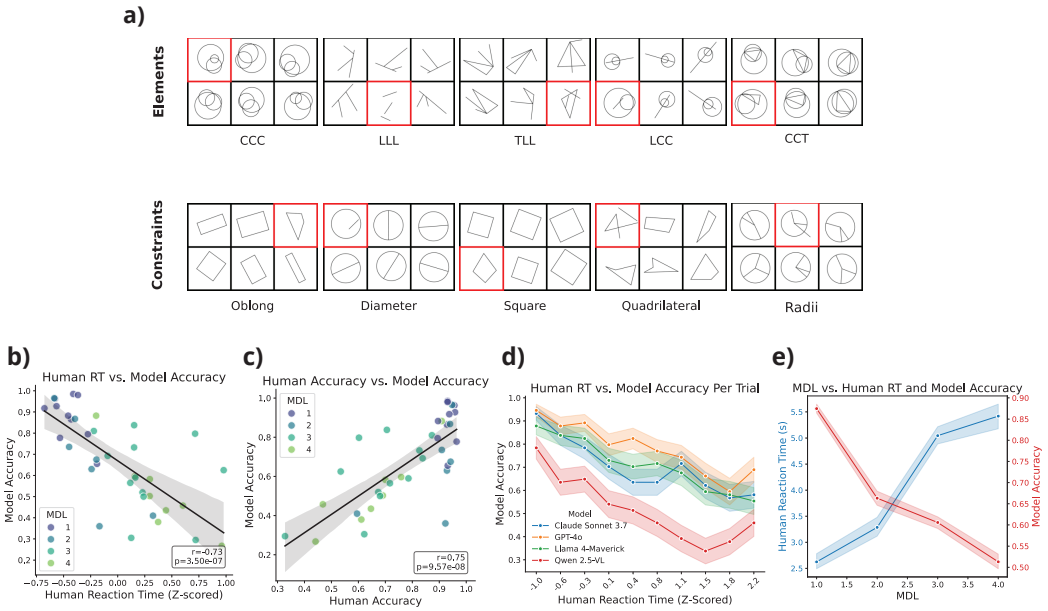

Figure 1: **Geometric Reasoning Task. a)** Example oddball detection trials from a subset of 37 geometric concepts spanning both primitive elements (left) and relational constraints (right), generated with Geoclidean DSL. **b)** Relationship between z-scored human reaction time and model accuracy; each point is one geometric concept and color denotes program complexity (Minimum Description Length, MDL). **c)** Correlation between human and model accuracy across concepts. **d)** Trial-level correlation between human reaction time (RT) and model accuracy. **e)** Human RT (blue, left axis) and model accuracy (red, right axis) as a function of MDL. Shaded regions are 95% confidence intervals.

4.1 Methods

To evaluate the sensitivity of VLMs to geometric concepts, we adapted the "oddball" detection paradigm from foundational studies of non-verbal human geometric intuition (Dehaene et al., 2006). Each trial consisted of an array of six images. Five of these images consistently instantiated a particular geometric concept, while the sixth item—the 'oddball'—violated this concept. We evaluated model and human performance on a geometric oddball task generated using the Geoclidean DSL (Hsu et al., 2022). Stimuli from 37 distinct concepts belonging to two categories — Geoclidean-Elements and Geoclidean-Constraints — were generated. (Figure 1a). These concepts spanned a range of geometric properties, including Euclidean geometry (e.g., parallelism, right angles, line bisecting), specific figures (e.g., squares, triangles, polygons), and constraints and relations (tangency, midpoint relations, angle constraints). We created 100 trials for each of the 37 concepts, yielding a total of 3,700 trials. Human participants each completed a randomly selected subset of 50 trials, drawn from a uniformly sampled 20% subset of the model evaluation trials. These trials were sampled such that each task condition was equally represented, and each trial received at least 20 independent human judgments. The oddball was constructed by removing two relational constraints from the reference concept while preserving its underlying components. Concepts ranged in program complexity from MDL 1 to 4, where Minimum Description Length (MDL) denotes the number of geometric primitives needed to generate the concept. Human participants and VLMs were tasked with identifying the position of the oddball within the array. (Appendix Figure 10). Human participants indicated their choice by clicking on the target image, and their reaction times were recorded. (Appendix Figure 9). VLMs were presented with the same array of images, each overlaid with a red index in the corner, and prompted to output the index corresponding to the oddball. (Appendix A.1.1). We evaluated four large multimodal models — GPT-4o (Achiam et al., 2023), Claude Sonnet 3.7 (Anthropic, 2023), Llama 4 Maverick (AI, 2024), Qwen2.5 VL (Bai et al., 2025) — on this task.

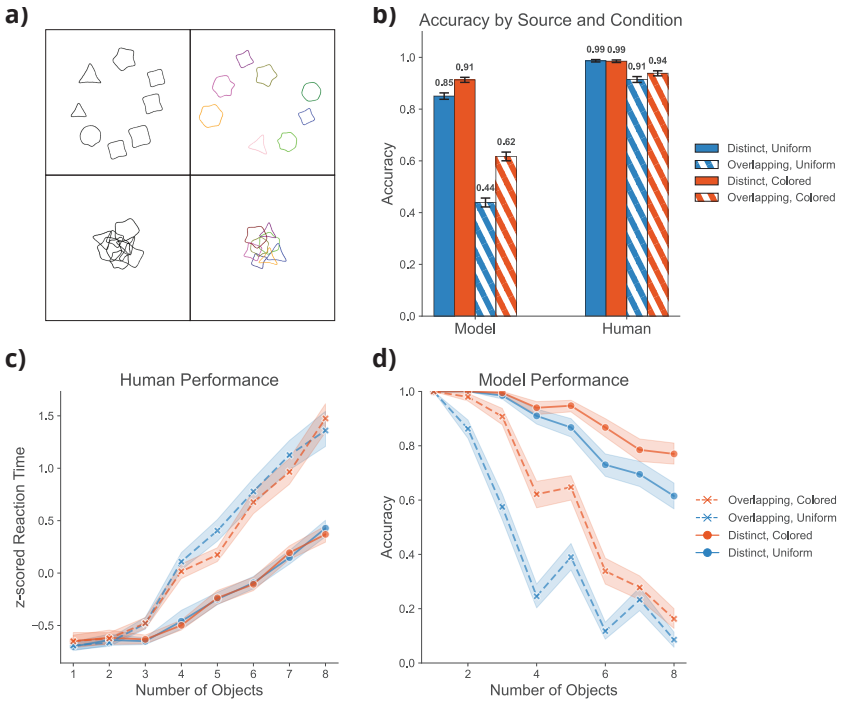

Figure 2: **Numerosity Task.** **a)** Example stimuli from the four experimental conditions: non-overlapping uniformly colored (top-left), non-overlapping uniquely colored (top-right), overlapping uniformly colored (bottom-left), and overlapping uniquely colored (bottom-right). **b)** Model accuracy as a function of numerosity across the four conditions. **c)** Mean accuracy for humans and models across each condition, averaged across numerosities. **d)** Human z-scored reaction times as a function of numerosity across the four conditions. Shaded regions indicate 95% confidence intervals.

4.2 Results

We first evaluated the alignment between human and model performance across geometric concepts. As shown in Figure 1c, human accuracy and model accuracy were positively correlated across concepts ($r = 0.75$, $p = 9.57 \times 10^{-8}$), suggesting that both humans and vision-language models (VLMs) are broadly sensitive to task difficulty – though potentially for different underlying reasons.

Importantly, we also observed an inverse relationship between human reaction time (RT) and model accuracy (Figure 1b). Concepts that took humans longer to solve were precisely those where models performed worst ($r = -0.73$, $p = 3.50 \times 10^{-7}$), implying that model failures are most acute on tasks requiring extended human deliberation. Moreover, at the level of trials, we observed a similar inverse relationship between human RT and model accuracy (Figure 1d). We formalized this pattern by analyzing performance as a function of minimum description length (MDL). As shown in Figure 1e, human RT increased with MDL (blue), while model accuracy declined (red). This divergence suggests that humans engage more cognitive effort as complexity rises, whereas models lack the structured inference mechanisms necessary to handle deeper compositional programs.

Together, these findings support our serial processing deficit hypothesis: while humans accommodate complexity via slower, sequential reasoning, VLMs exhibit declining accuracy as task complexity increases – despite equivalent visual input. In our subsequent experiments we show that similar patterns can be observed in other settings where serial processing is important.

5 Test Domain 2: Numerical Estimation

Numerical estimation — such as assessing how many objects appear in a scene — is a classic serial processing task. Human numerical estimation abilities are characterized by a transition between the near-instantaneous recognition of small quantities (up to 4–6 items), a process known as subitization,

and the serial enumeration of larger quantities (Trick & Pylyshyn, 1994; Kaufman et al., 1949). As humans enumerate larger sets or when visual discrimination becomes challenging, we deploy serial attention to identify each object — binding features to locations and maintaining distinct object representations (Mazza & Caramazza, 2015; Xu & Chun, 2009). Such serial attention-based enumeration becomes more challenging as objects become harder to distinguish (Franconeri et al., 2013; Lavie, 2005). Prior work has documented VLMs' difficulty in counting (Pothiraj et al., 2025; Campbell et al., 2024b); we predict that, in alignment with human behavior, VLM performance will be highest when the number of objects is low and the objects are non-overlapping. However, while humans maintain accuracy in challenging conditions (Vetter et al., 2008) (e.g. overlapping objects at higher numerosities) by deploying serial attention, we hypothesize that VLM performance will degrade rapidly. Furthermore, we predict that this degradation in performance would coincide with an increase in human reaction time. Critically, we anticipate an asymmetry in performance, where VLM performance increases when overlapping objects are uniquely colored compared to when they are uniformly colored, a pattern not mirrored in human performance. This prediction stems from our hypothesis that VLMs' serialization capabilities depend heavily on their ability to ground reasoning in their linguistic chain of thought, which is enabled when the objects can be uniquely identified.

5.1 Methods

We evaluated model and human performance on a visual enumeration task designed to systematically vary perceptual individuation load. Stimuli consisted of 2-dimensional shapes generated using Hermitian splines, featuring between 3 and 5 pointed contours with randomly jittered fixed points (Figure 2a). To parametrically manipulate serial processing demands, we varied both numerosity (1-8 objects) and object distinctiveness across two dimensions: spatial arrangement (overlapping vs. non-overlapping) and color (uniformly colored vs. uniquely colored). This factorial design created four distinct conditions: (1) non-overlapping, uniformly colored objects; (2) non-overlapping, uniquely colored objects; (3) overlapping, uniformly colored objects; and (4) overlapping, uniquely colored objects. These manipulations were designed to vary reliance on serial attention required for object identification while controlling for the total number of objects.

In the uniform color conditions, all objects had the same hue, requiring individuation based solely on their spatial arrangement. In the unique color conditions, each object was assigned a distinct hue, providing an additional dimension for differentiation. In the non-overlapping conditions, objects were positioned with clear spatial separation, while in the overlapping conditions, objects were arranged with 60-80% intersection between adjacent shapes so boundary resolution required attention.

Each of the four conditions was instantiated across 100 trials for each numerosity level (1-8), yielding 3,200 total examples. Each human participant completed a randomly selected set of 50 trials, drawn from a uniformly sampled 20% of the model evaluation set. Sampling was stratified to ensure balanced representation across task conditions, with each trial evaluated by at least 10 independent participants. Human participants and VLMs were tasked with identifying the number of objects in the scene (Appendix Figure 12). Participants performed the task while reaction times and accuracy were recorded (Appendix Figure 11). VLMs were prompted to describe the image and report the number of objects present (Appendix A.3.1). This allowed us to directly measure the impact of increasing perceptual individuation load on task performance, and to compare the results for humans and VLMs.

5.2 Results

We first evaluated the alignment between human and model performance across the four task conditions. As shown in Figure 2c, humans sustained accuracy across conditions, while model accuracy significantly degraded in the overlapping conditions.

We also observed an inverse relationship between model performance and numerosity (Figure 2b), indicating that models struggled both with high numerosities and perceptual individual loads. These same conditions took humans longer to solve ($r = -0.97$, $p = 8.22 \times 10^{-5}$) and we observed an inverse relationship between human RT and model performance. Performance was also consistently impaired on overlapping trials, with models showing substantially lower accuracy ($t = -8.43$, $p = 7.9 \times 10^{-10}$), and humans requiring more time ($t = 3.45$, $p = 0.005$) and exhibiting reduced accuracy ($t = -2.43$, $p = 0.023$).

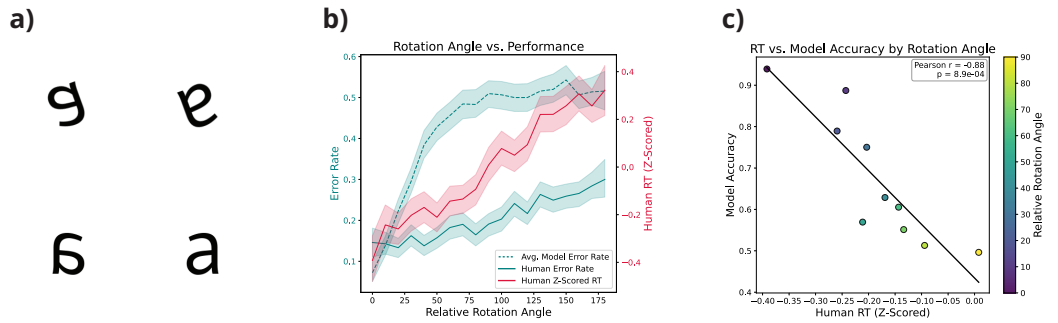

Figure 3: **Mental Rotation Task. a)** Example stimuli from the mental rotation task. Participants and models judged whether a rotated letter was the same or a mirror-reversed version of a reference. **b)** Human error rate (solid line), model error rate (dashed line), and z-scored human reaction time (right axis) plotted against relative rotation angle. Shaded regions denote 95% confidence intervals. **c)** Relationship between z-scored human reaction time and model accuracy across rotation angles ($\leq 90°$). Each point corresponds to a single rotation angle; colors indicate relative angular disparity.

Finally, we observed that model performance was significantly better in the colored overlapping condition compared with the uniform overlapping condition ($t = 5.90$, $p = 8.1 \times 10^{-7}$; Figure 2b). This model-specific improvement was not accompanied by a corresponding decrease in human reaction time ($t = -1.17$, $p = 0.28$; Figure 2d) or accuracy ($t = 2.06$, $p = 0.078$), suggesting that color aided model performance not by reducing perceptual difficulty, but by facilitating object individuation. This supports our hypothesis that VLM serialization is improved when objects can be uniquely referenced by their color in context, a capability that humans do not benefit from.

6 Test Domain 3: Mental Rotation

Mental rotation provides a final test case where the sequential nature of the task cannot be effectively substituted with linguistic reasoning. Unlike symbolic pattern recognition, mental rotation requires the manipulation of spatial representations through continuous transformations—operations that are difficult to express in discrete, language-based tokens. A standard mental rotation task presents participants with two images which are either two distinct objects or one object shown from different angles. Decades of cognitive science research demonstrate that humans solve these tasks by mentally rotating one object to align with another, an analog process captured by the robust linear relationship between angular disparity and reaction time (Shepard & Metzler, 1971; Cooper & Shepard, 1973). Functional imaging confirms that this process involves incremental spatial adjustments, engaging parietal regions known to support visuospatial transformation (Zacks et al., 1999; Just & Carpenter, 1985). These transformations rely on what (Kosslyn, 1994) termed "depictive representations" – mental images that preserve geometric structure and spatial relations. Because such analog operations are difficult to serialize into linguistic prompts, mental rotation offers a particularly stringent test of whether VLMs can support visual reasoning that is not scaffolded by language.

6.1 Methods

We evaluated human and model performance on a classic mental rotation task, adapted from prior cognitive science paradigms (Shepard & Metzler, 1971; Cooper & Shepard, 1973). Stimuli consisted of character-like shapes that were either identical or mirror-reversed versions of a reference, displayed at various rotation angles (Figure 3a). Each trial presented a pair of shapes, and the task was to determine whether they matched or were mirrored. (Appendix Figure 14).

Rotation angles varied systematically from 0° to 360° in 10° increments, yielding 36 distinct orientations. For each angle and letter stimulus (26 total), we generated four trial types crossing two factors: whether the pair was the same or mirror-reversed, and whether the first image was mirrored or not. This design resulted in $36 \times 4 \times 26 = 3744$ total trials. Trial order was fully randomized, and conditions were balanced across same/mirror status and initial image mirroring. As before, each participant completed a randomly selected set of 50 trials, drawn from a uniformly sampled 20%

subset of the full model evaluation set. Sampling ensured equal representation across task conditions, and each trial was judged by at least 10 independent human participants.

Human participants completed the task via mouse click, and RT was recorded from stimulus onset to response. (Appendix Figure 13). VLMs were shown images of the two shapes, and prompted to indicate whether the shapes are 'same' or 'different'. (Appendix A.2.1). Model performance was evaluated as overall accuracy across each numerosity and load condition.

6.2 Results

We first examined the effect of rotation angle on performance. As shown in Figure 3b, human reaction time increased monotonically with angular disparity, replicating the classic linear RT-angle relationship. This trend reflects the analog nature of human mental rotation: more extreme angles require longer sequences of transformations (Just & Carpenter, 1985; Zacks et al., 1999).

As predicted, VLM performance exhibited an inverse relationship. As rotation angle increased, model accuracy declined, with particularly sharp drops in accuracy beyond 75°, consistent with prior findings on spatial reasoning limitations in vision-language systems (Stogiannidis et al., 2025; Chen et al., 2025). This decline is reflected in the average model error rate (dashed line, Figure 3b), which rose substantially across angle bins.

Human accuracy also declined with rotation angle, but more modestly. While participants made more errors at higher angles, their performance degraded far less steeply than that of the models, indicating partial robustness to increased serial demands. This distinction reinforces the idea that humans can compensate for increasing task difficulty by allocating more cognitive resources—reflected in longer RTs—whereas VLMs currently lack such dynamic inference capabilities.

Critically, we observed a strong negative correlation between human RT and model accuracy across rotation angles (Figure 3c; $r = -0.88$, $p = 8.9 \times 10^{-4}$). Trials that demanded longer processing times from humans were precisely those where VLMs performed worst. This inverse relationship supports the serial processing deficit hypothesis: the tasks that elicit deeper, stepwise transformations in humans are the ones that reveal structural limitations in current VLMs.

Together, these findings suggest that mental rotation engages a form of non-linguistic serial reasoning that vision-language models do not replicate. Despite identical visual input, models falter where humans invoke continuous internal transformations – highlighting a fundamental divergence in the underlying mechanisms of reasoning.

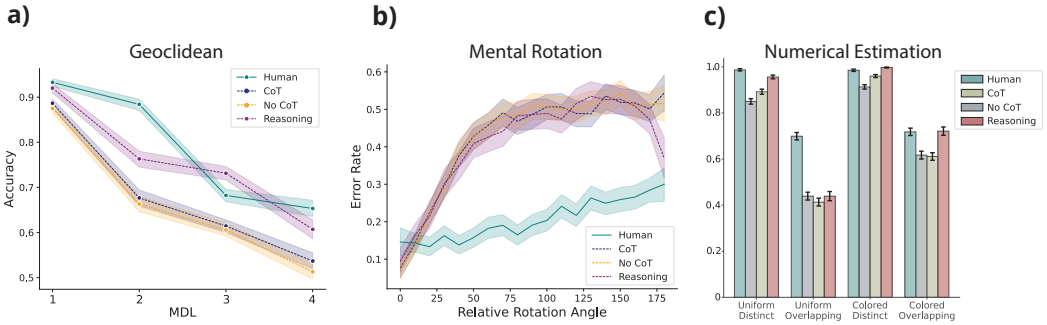

Figure 4: **Results for Augmented VLMs. a)** Accuracy as a function of MDL for humans and different model conditions. **b)** Error rate as a function of relative rotation angle for humans (green line) and models under different augmentation paradigms. **c)** Mean accuracy for humans and models across each condition, averaged across numerosities. **d)** Accuracy on a challenging subset of Geoclidean and rotation tasks, comparing humans, reasoning-augmented models, and GPT-o3 with tool use. **e)** Accuracy by condition for $n = 8$ items, across humans, reasoning models, and GPT-o3 with tool use. Shaded regions in **a)** and **b)** represent 95% confidence intervals.

7 Causal evidence for a serial processing deficit

One way to causally test the hypothesis that VLMs lack serial processing capabilities is to augment VLMs with a form of serial processing, and test whether this enables them to solve more complex visual tasks. To investigate this, we tested models augmented with three forms of serial processing: Chain-of-thought (CoT), reasoning training, and tool use (e.g., the ability to crop and rotate images).

We do not expect these forms of serial processing to be effective in all tasks. Improvements from CoT and reasoning training should be limited to cases where individual objects can be uniquely referenced in the models CoT. Our numerosity estimation task provides an ideal setting to test this prediction. In this task, when objects are both non-overlapping and distinctly colored, textual CoT can effectively be used to describe the objects one at a time (based on color), whereas textual CoT should be less effective when objects are overlapping or uniform in color. Consistent with this, we see that this condition (colored distinct) is the one where CoT ($t = 7.79$,$p < 0.005$) and reasoning training ($t = 13.81$, $p < 0.005$) help the most. Reasoning training results in nearly perfect performance in this condition (99.7% accuracy; Figure 4), but does not improve performance significantly in the overlapping conditions ($p > 0.5$).

Similarly, improvements from tool use should be limited to cases where objects can be isolated by cropping an image. This results in near-perfect accuracy on our rotation task (Table 2, $p < 0.05$) and perfect performance on conditions in which the objects are spatially separated (uniform distinct and colored distinct) in our numerosity task (100% accuracy, $p < 0.005$). However, tool use does not improve on baseline performance for either of the overlapping conditions (Table 2, $p > 0.1$).

Two major conclusions can be drawn from these results. First, they show that augmenting VLMs with serial processing enables them to solve visually complex tasks, thus providing causal evidence for the serial processing deficit hypothesis. Second, these results highlight the limitations of existing methods for serial processing in VLMs (textual CoT and tool use) as their application is limited to specific settings. This points to the need for developing intrinsically visual forms of serial processing to help VLMs deal with a broader range of visually complex tasks.

8 Discussion

We have shown that Vision-Language Models (VLMs) struggle with tasks that require visually grounded serial processing, revealing behavioral limitations that parallel specific constraints in human cognition. These results offer an explanatory account of why models so adept at many tasks struggle with activities as simple as counting. They also point toward concrete strategies for improving VLM performance by targeting improvements in visual serial processing. Across three experimental paradigms, we observed a consistent pattern: VLM performance inversely correlates with human reaction time. This relationship provides strong evidence for our serial processing deficit hypothesis, suggesting that operations requiring sequential processing in humans—indicated by longer reaction times—correspond precisely to operations that current VLMs fail to perform. The cross-domain consistency of this relationship indicates that serial processing capacity represents a general computational constraint rather than a domain-specific limitation.

In geometric reasoning, as program complexity (measured by MDL) increased, human reaction times rose, whereas VLM performance deteriorated. Previous work using similar tasks argued that human performance implies symbolic program induction, a capability that neural network models lack (Sablé-Meyer et al., 2022). Contrary to this finding, we show that neural networks (here, VLMs) *do* correlate with both human behavior and measures of task complexity. It is an open question whether to interpret this as evidence that VLMs perform program induction, or that the human behavioral measures do not reflect program induction. However, our results indicate that VLMs appear unable to flexibly compose visual operations in ways that scale with program complexity.

The numerical estimation task revealed that VLMs demonstrate subitizing-like behavior for small quantities under optimal viewing conditions but show performance degradation when object individuation becomes challenging. Humans maintain accuracy in these conditions by deploying serial attention—binding features to locations and maintaining distinct object representations—at the cost of increased processing time (Mazza & Caramazza, 2015; Xu & Chun, 2009). We also find that VLMs benefit from unique coloring in overlapping object conditions, while humans show no comparable advantage. This contrasts with some findings, such as those by (Rahmanzadehgervi et al., 2025),

who reported that coloring conditions in numerosity tasks made no significant difference for most of the VLMs tested. Here, we show that when objects overlap, unique coloring provides a distinct and significant advantage to VLMs. This advantage likely arises because distinct colors allow VLMs to individually reference and effectively count objects, thereby enabling them to linguistically ground features and perform feature binding in a way that is not possible in uniformly colored conditions.

Mental rotation provided a particularly stringent test of *visual* serial processing, as this analog transformation is difficult to decompose into discrete linguistic steps. While humans show a linear relationship between rotation angle and reaction time (Shepard & Metzler, 1971) (reflecting an incremental transformation process), and relatively limited effects of rotation angle on accuracy, VLMs exhibited sharp performance drops as rotation angles increased. This suggests a limitation in performing the continuous transformations that humans execute through depictive representations (Kosslyn, 1994). Importantly, we are not claiming that models are incapable of serial processing—indeed, autoregressive generation implicitly implements serial processing. However, this seems to be a *strictly linguistic* form of serialization that is difficult to apply to visual tasks like mental rotation.

8.1 Architectural Improvements

Recent advances in VLMs, such as V* (Yang et al., 2023) and tool-augmented models, are intended to mimic how humans process scenes during challenging visual reasoning tasks. Rather than processing entire images at once, humans sequentially focus their attention on different regions, using high-resolution central vision to examine each area in detail—a strategy that helps overcome the limitations of trying to process all visual information simultaneously (Stewart et al., 2020). Our work shows two fundamental limitations of these current augmentations. First, linguistic Chain-of-Thought (via prompting and reasoning fine-tuning) fails when input cannot be grounded in language. Second, while image tool use effectively solves well-defined transformations with explicit algorithms, it struggles with tasks requiring more general forms of serial reasoning that do not involve explicit transformations of the input images.

A more promising path forward involves inducing a more sequential and deliberate mode of processing directly, with a new class of models trained using Visually Grounded Reinforcement Learning (RL) showing particular promise. Rather than processing an image in a single pass, methods like ViGoRL, GRIT, UniVG-R1, and Ground-R1 use RL to train a policy that generates an explicit reasoning trace where each textual thought is anchored to a specific coordinate or bounding box in the visual input (Sarch et al., 2025; Wang et al., 2025; Li et al., 2025; Zhao et al., 2025). This process of generating a sequence of region-grounded thoughts functions as a computational analogue to the human saccade-and-fixate cycle, forcing the model to abandon its default holistic strategy and adopt a more structured, pseudo-serial approach to scene analysis. This also increases model interpretability, given the visual-attentional path to an answer can be explicitly examined.

Other post-training methods inspired by human serial vision show promise in specific domains. Molmo (Deitke et al., 2024) demonstrates improved performance on counting when trained to successively point to objects. Similar fine-tuning approaches for sequential attention could prove beneficial in other visual reasoning domains. The processing deficits we observed emerge from interference between visual elements, as shown in our numerosity experiments and by (Campbell et al., 2024b). This interference could be mitigated through causal masking around fixation points identified by models like Molmo, producing effects similar to the human distinction between foveal and peripheral vision (Stewart et al., 2020). This controlled, sequential attention could reduce the interference we observed and provide a path toward more human-like serial visual processing.

8.2 Limitations & Future Directions

This work has several limitations. First, our analysis is constrained to a limited set of VLMs. We chose models that are competitive with other available models, but our selection does not necessarily capture the full diversity of performance. Second, our tasks were selected to manipulate serial processing load in visual reasoning; future work could consider a broader range of tasks. Specifically, extending our analysis to tasks like visual search and spatial reasoning (Gilden et al., 2010; Reimer & Schubert, 2019) would increase the generality of our results. Moreover, evaluating spatial reasoning alongside a broader range of tasks with varying serial processing load may allow us to more effectively dissociate spatial reasoning deficits from serial reasoning deficits.

8.3 Conclusion

Our work provides a unifying explanation for the discrepancy between VLMs' impressive benchmark performance and their difficulty with seemingly simple visual reasoning tasks. The serial processing deficit hypothesis offers both a theoretical framework for understanding current model limitations and a roadmap for future architectural innovations. By identifying specific visual reasoning operations where VLMs diverge from human behavior, we contribute to the understanding of computational principles underlying visual intelligence. As the field develops the next generation of visual reasoning models, addressing the serial processing deficit may be key to achieving more human-like flexibility across diverse visual tasks—combining the parallel pattern recognition capabilities of contemporary VLMs with the serial reasoning characteristics that have been a hallmark of human visual cognition.

9 Ethics Statement

We conducted our human behavioral experiments under an IRB-approved protocol. Further information, such as about the potential risks of behavioral data collection and details about compensation, can be found in the Appendix.

We address potential societal impacts of our work in the Discussion. Our experiments do not pose a high risk for misuse since we do not release new models or use any proprietary assets.

10 Reproducibility Statement

All experimental details and methods for testing models are described in the main text and Appendix. The code base necessary to generate experimental stimuli, run model experiments, and human behavioral experiments will be open-sourced upon publication. We will include our full datasets and instructions to faithfully reproduce our work.

We include screenshots of instructions provided to participants in the Appendix, as well as details on compute needed to run model experiments. Our work does not involve any novel theory that would require formal justification. Finally, we report statistical significance for all results where appropriate in the main text and Appendix.

References

Josh Achiam, Simen Adler, Sandhini Agarwal, Ilge Ahmad, Ithihas Akkaya, Adriana Aleman, Christian Allen-Blanchette, Joel Apolinarski, Tasuku Arahata, Saul Arce-Cardenas, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Meta AI. Llama 4: Open foundation and multimodal models. *Meta AI Blog*, 2024. URL `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. URL `https://arxiv.org/abs/2204.14198`.

Kelsey Allen, Ishita Dasgupta, Eliza Kosoy, and Andrew K. Lampinen. The in-context inductive biases of vision-language models differ across modalities. *arXiv preprint arXiv:2502.01530*, 2025.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. Palm 2 technical report, 2023. URL `https://arxiv.org/abs/2305.10403`.

Anthropic. Claude: A next-generation ai assistant. *Anthropic Blog*, 2023. URL https://www.anthropic.com/index/claude-3-family.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Declan Campbell, Sreejan Kumar, Tyler Giallanza, Jonathan D. Cohen, and Thomas L. Griffiths. Relational constraints on neural networks reproduce human biases towards abstract geometric regularity, 2023. URL https://arxiv.org/abs/2309.17363.

Declan Campbell, Sreejan Kumar, Tyler Giallanza, Thomas L. Griffiths, and Jonathan D. Cohen. Human-like geometric abstraction in large pre-trained neural networks, 2024a. URL https://arxiv.org/abs/2402.04203.

Declan Campbell, Sunayana Rane, Tyler Giallanza, Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M. Frankland, Thomas L. Griffiths, Jonathan D. Cohen, and Taylor Webb. Understanding the limits of vision language models through the lens of the binding problem. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/cdcc6d47c1627350014a3076112ab824-Paper-Conference.pdf.

Shiqi Chen, Zhaofei Zhang, Siyuan Zhou, Zhipeng Li, Yikun Wang, Zihan Liu, Ziyi Wang, Yifei Wang, and Yisen Zhang. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas, 2025.

Lynn A. Cooper and Roger N. Shepard. *Chronometric studies of the rotation of mental images.*, pp. xiv, 555–xiv, 555. Visual information processing. Academic, Oxford, England, 1973. ISBN 0121701506.

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J Gershman, and Noah D Goodman. Language models show human-like content effects on reasoning. *Trends in Cognitive Sciences*, 26 (6):493–505, 2022.

Stanislas Dehaene, Véronique Izard, Pierre Pica, and Elizabeth Spelke. Core knowledge of geometry in an amazonian indigene group. *Science*, 311(5759):381–384, 2006.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and PixMo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.

Steven L Franconeri, George A Alvarez, and Patrick Cavanagh. Flexible cognitive resources: competitive content maps for attention and memory. *Trends in Cognitive Sciences*, 17(3):134–141, 2013.

Chaoyou Fu, Peixian Chen, Yunhang Zhang, Xinhong Xie, Haomiao Zhang, Haosen Shen, Abasifreke Bassey Chen, Soumyendu Chakraborty, Qian Zhang, Xuweiyi Xu, Yu Ding, Wei Ye, Yun Zhao, Shengfeng Tong, Gang Yu, Bingshuai Luo, Hao Dong, Xuefei Wang, Xiu Yu, Josiah Poon, Xiaowei Guo, Tao Chen, Zhijie Cui, Weipeng Wang, Taihao Sun, Zhifang Zhang, Meng Ding, Zhengyuan Yang, Aiming Wang, Haowei Wang, Liang Li, Yi Chang, Junnan Li, Ziheng Wu, Dahua Lin, Lingxi Xie, Wei Liu, Zhe Wei, Yuan Qi, Bing Cheng, Wen Li, Yang You, Zhenguo Lu, Hao Wang, and Zhou Zhao. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive, 2024. URL https://arxiv.org/abs/2404.12390.

David L Gilden, Thomas L Thornton, and Laura R Marusich. The serial process in visual search. *J. Exp. Psychol. Hum. Percept. Perform.*, 36(3):533–542, June 2010.

Taylor Goddard, Vanya Ayzenberg, Spandana Prasad, Kanishk Gandhi, Michael C Frank, Michael J Tarr, and Yevgeniya Yermolayeva. KiVA: Kid-inspired visual analogies for testing large multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vNATZfmY6R.

Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020.

Richard P Heitz. The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in Neuroscience*, 8:150, 2014.

Joy Hsu, Jiajun Wu, and Noah D. Goodman. Geoclidean: Few-shot generalization in euclidean geometry, 2022. URL https://arxiv.org/abs/2211.16663.

Kung-Hsiang Huang, Can Qin, Haoyi Qiu, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. Why vision language models struggle with visual arithmetic? towards enhanced chart and geometry understanding, 2025. URL https://arxiv.org/abs/2502.11492.

Marcel A. Just and Patricia A. Carpenter. Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability. *Psychological Review*, 92(2):137–172, 1985. doi: 10.1037/0033-295X.92.2.137. URL https://doi.org/10.1037/0033-295X.92.2.137.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=RN5KLywTll.

Edna L Kaufman, Michael W Lord, Thomas W Reese, and John Volkmann. The discrimination of visual number. *The American Journal of Psychology*, 62(4):498–525, 1949.

Stephen M. Kosslyn. *Image And Brain: The Resolution of the Imagery Debate*. The MIT Press, 06 1994. ISBN 9780262277488. doi: 10.7551/mitpress/3653.001.0001. URL https://doi.org/10.7551/mitpress/3653.001.0001.

Nilli Lavie. Distracted and confused?: Selective attention under load. *Trends in Cognitive Sciences*, 9(2):75–82, 2005.

Jiajun Li, Jiacheng Chen, Ziyun Li, Chong Zhang, Guangyi Chen, Chen Zhang, Shiliang Wang, Yi Yang, and Ya Zhang. UniVG-R1: Reasoning guided universal visual grounding with reinforcement learning, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Veronica Mazza and Alfonso Caramazza. Multiple object individuation and subitizing in enumeration: a view from electrophysiology. *Frontiers in Human Neuroscience*, 9:162, 2015.

Atin Pothiraj, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. Capture: Evaluating spatial reasoning in vision language models via occluded object counting, 2025. URL https://arxiv.org/abs/2504.15485.

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. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind: Failing to translate detailed visual features into words, 2025. URL https://arxiv.org/abs/2407.06581.

Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, and Vladlen Koltun. Does spatial cognition emerge in frontier models? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=WK6K1FMEQ1.

Christina B. Reimer and Torsten Schubert. More insight into the interplay of response selection and visual attention in dual-tasks: Masked visual search and response selection are performed in parallel. *Psychological Research*, 83(3):459–475, 2019. doi: 10.1007/s00426-017-0906-2. URL https://doi.org/10.1007/s00426-017-0906-2.

Kevin Sablé-Meyer, Adam Santoro, Brenden M Lake, Joshua B Tenenbaum, Jean-Louis Dessalles, Steven T Piantadosi, Radoslaw Martin Cichy, and Stanislas Dehaene. Language as a bottleneck: a cross-cultural study of the efficient coding of geometric shapes. *Proceedings of the National Academy of Sciences*, 119(26):e2123203119, 2022.

Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J Tarr, Aviral Kumar, and Katerina Fragkiadaki. ViGoRL: Visually grounded reinforcement learning for visual reasoning, 2025.

Luca M. Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz. Visual cognition in multimodal large language models. *Nature Machine Intelligence*, 7(1):96–106, Jan 2025. ISSN 2522-5839. doi: 10.1038/s42256-024-00963-y. URL https://doi.org/10.1038/s42256-024-00963-y.

Roger N. Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971. doi: 10.1126/science.171.3972.701. URL https://www.science.org/doi/abs/10.1126/science.171.3972.701.

Emma EM Stewart, Matteo Valsecchi, and Alexander C Schütz. A review of interactions between peripheral and foveal vision. *Journal of Vision*, 20(12):2–2, 2020.

Ilias Stogiannidis, Steven McDonagh, and Sotirios A Tsaftaris. Mind the gap: Benchmarking spatial reasoning in vision-language models, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Andrew Cai, Cătălina Cangea, Stephen Chang, Louis C Cohen, Souradeep Dey, Fartash Faghri, William Fedus, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs, 2024. URL https://arxiv.org/abs/2406.16860.

Lana M Trick and Zenon W Pylyshyn. Why are small and large numbers enumerated differently? a limited-capacity preattentive stage in vision. *Psychological Review*, 101(1):80–102, 1994.

Petra Vetter, Brian Butterworth, and Bahador Bahrami. Modulating attentional load affects numerosity estimation: evidence against a pre-attentive subitizing mechanism. *PLoS One*, 3(9):e3269, 2008.

Bosheng Wang, Chase Zhang, Yichuan Xu, Joseph E Gonzalez, Trevor Darrell, and Xin Wang. GRIT: Teaching MLLMs to think with images, 2025.

Wenrui Xu, Dalin Lyu, Weihang Wang, Jie Feng, Chen Gao, and Yong Li. Defining and evaluating visual language models' basic spatial abilities: A perspective from psychometrics, 2025. URL https://arxiv.org/abs/2502.11859.

Yaoda Xu and Marvin M Chun. Selecting and perceiving multiple visual objects. *Trends in Cognitive Sciences*, 13(4):167–174, 2009.

Zhengyuan Yang, Linjie Wang, Kevin Shen, Haoxuan Zhang, Jianfeng Zhou, Jianwei Sun, Jiebo Yu, Karthik Gholami, Peng Zhang, Chuang Gan, et al. The dawn of lmms: Preliminary explorations with gpt-4v(ision). *arXiv preprint arXiv:2309.17421*, 2023.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

J Zacks, B Rypma, J D Gabrieli, B Tversky, and G H Glover. Imagined transformations of bodies: an fMRI investigation. *Neuropsychologia*, 37(9):1029–1040, August 1999.

Yizhe Zhao, Yang Yang, Jiacheng Chen, Ziyun Li, Jiajun Li, Ya Zhang, and Shiliang Wang. Ground-R1: A reasoning-guided grounding model with a reinforcement learning fine-tuning framework, 2025.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. URL `https://arxiv.org/abs/2304.10592`.

A Appendix: Prompts for Vision-Language Model Experiments

A.1 Geoclidean

A.1.1 Baseline Prompt

```
Which of the 6 shapes is different from the others?

Return only the number of the odd one out in square brackets (e.g., [3]) without any
  additional reasoning or justification.
```

A.1.2 Chain-of-Thought Prompt

```
Which of the 6 shapes is different from the others?

Analyze the relations between the parts of each stimulus to identify the stimulus that
  is different from the rest. Return your reasoning followed by the integer label of
  the odd one out enclosed in square brackets (e.g., [3]).
```

A.2 Rotation

A.2.1 Baseline Prompt

```
You will be shown two objects that are rotated at different angles. Your task is to
  determine whether these objects are:

A) The **SAME** object shown at different rotations
B) **DIFFERENT** objects that are mirror images of each other (these mirror images may
  also be rotated at different angles)

Even when rotated, identical objects will have the same features in the same
  arrangement. Mirror images will have reversed features that cannot be matched by
  rotation alone.

## Response Format:
[1] - The objects are identical but rotated differently
[0] - The objects are mirror images of each other

Please just include your final answer in the format [1] or [0] without any additional
  reasoning or justification.
```

A.2.2 Chain-of-Thought Prompt

```
You will be shown two objects that are rotated at different angles. Your task is to
  determine whether these objects are:

A) The **SAME** object shown at different rotations
B) **DIFFERENT** objects that are mirror images of each other (these mirror images may
  also be rotated at different angles)

Even when rotated, identical objects will have the same features in the same
  arrangement. Mirror images will have reversed features that cannot be matched by
  rotation alone.

## Response Format:
[1] - The objects are identical but rotated differently
[0] - The objects are mirror images of each other
```

```
Please include your reasoning followed by your final answer in the format [1] or [0].
```

A.3 Numerosity

A.3.1 Baseline Prompt

```
The following images contains 1 or more objects that may or may not be overlapping.

Count the total number of objects in the scene, and return your final answer as an
  integer enclosed in square brackets (e.g., [10]).
```

A.3.2 Chain-of-Thought Prompt

```
The following images contain 1 or more objects that may or may not be overlapping.

Count the total number of objects in the scene one at a time. Once you have described
  all objects, return your final answer as an integer enclosed in square brackets (e.g
  ., [10]).
```

B Appendix: Individual Model Results

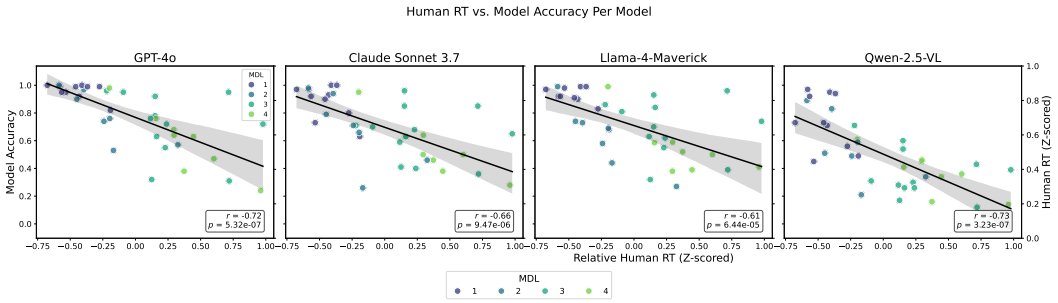

Figure 5: **Human RT vs. Model Accuracy by Model.** Z-scored human reaction time and model accuracy plotted across geometric concepts, shown separately for each VLM. Each point corresponds to one concept, colored by program complexity (MDL). Black lines show linear regression fits with 95% confidence intervals.

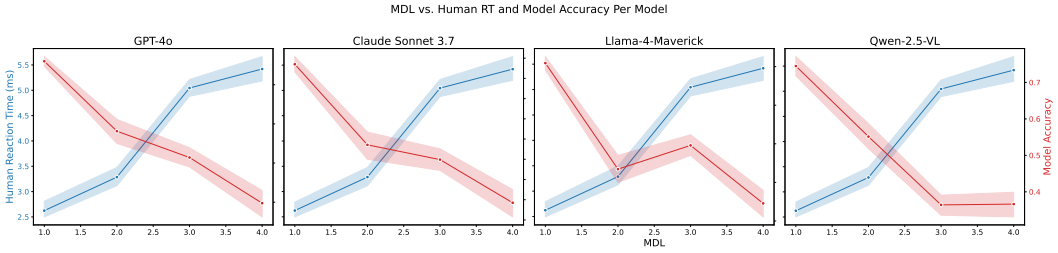

Figure 6: **MDL vs. Human RT and Model Accuracy by Model.** Human reaction time (blue, left axis) and model accuracy (red, right axis) plotted as a function of MDL, shown separately for each VLM. Shaded regions indicate 95% confidence intervals.

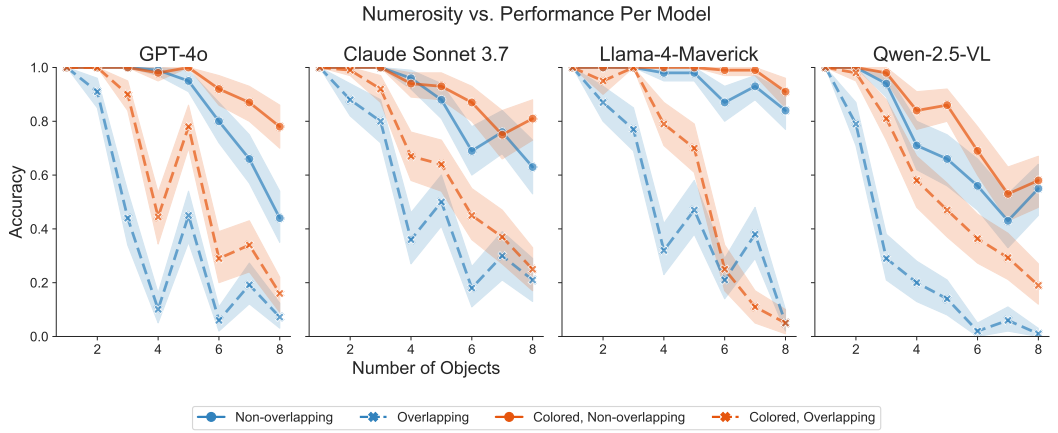

Figure 7: **Numerosity vs. Performance by Model.** Model accuracy in the overlapping condition (dashed), distinct condition (solid), colored condition (orange), and uniform condition (blue) plotted as a function of the number of objects in the scene, shown separately for each VLM. Shaded regions indicate 95% confidence intervals.

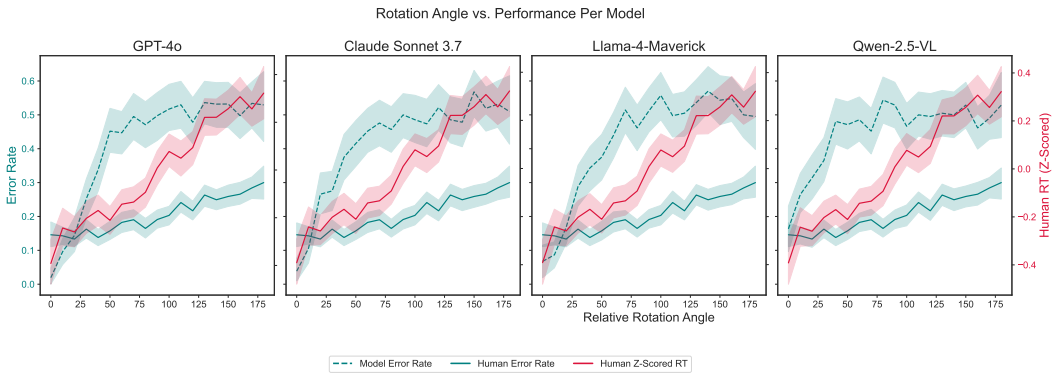

Figure 8: **Rotation Angle vs. Performance by Model.** Model error rate (dashed), human error rate (solid), and z-scored human reaction time (right axis) plotted as a function of relative rotation angle, shown separately for each VLM. Shaded regions indicate 95% confidence intervals.

C Appendix: Human Behavioral Experiments

Human behavioral data for the geometric complexity, numerical estimation, and mental rotation tasks were collected via Prolific under an IRB-approved protocol. To maintain double-blind review standards, we do not include the consent form. In the geometric complexity task, participants selected the oddball from an array of geometric objects. In the numerical estimation task, they counted the total number of objects presented. In the rotation task, they determined whether a pair of objects were mirrored versions of each other. Participants were compensated at an approximate rate of $12 per hour.

C.1 Geoclidean

Which of the following geometric shapes is the outlier? Once you press the cross, click on the outlier as quickly as you can.

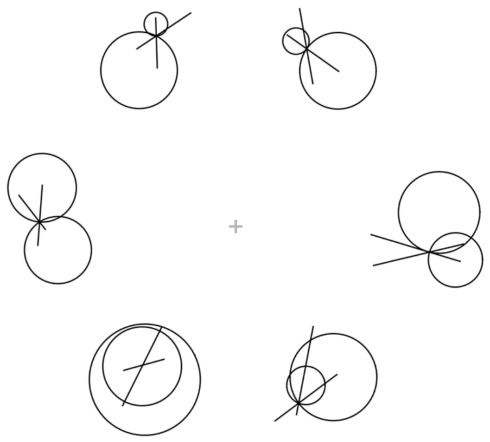

Figure 9: **Example of Geoclidean Task.**

Thank you for participating in our study!

Task Instructions

In this study, you will be shown six geometric shapes in each trial. Five shapes will be similar, while one shape will be different - this is the "outlier". Your task is to find and click on the outlier as quickly as possible.

Important Notes

Please note that while the shapes may appear in different sizes and orientations, these differences do not matter. Each shape is made up of the same basic parts, but what's important is how these parts relate to one another. The five reference shapes will share the same relationships amongst parts, while the outlier shape will have these same parts joined together differently. Focus only on identifying this outlier. Even if you are unsure, you must make a choice in each trial.

How to Respond

Speed is essential in this task. You can take your time to prepare before each trial begins, but once you click the central cross to start, you should identify and click the outlier shape as quickly as you can. After each trial you will receive feedback.

Next

Figure 10: **Instructions for Geoclidean Task.**

C.2 Numerosity

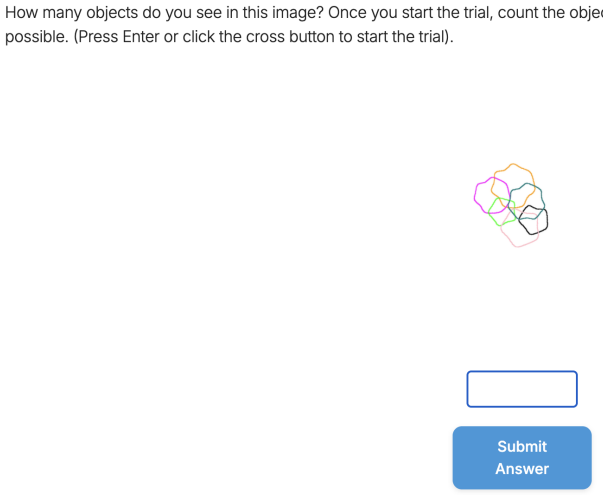

Figure 11: **Example of Numerosity Task.**

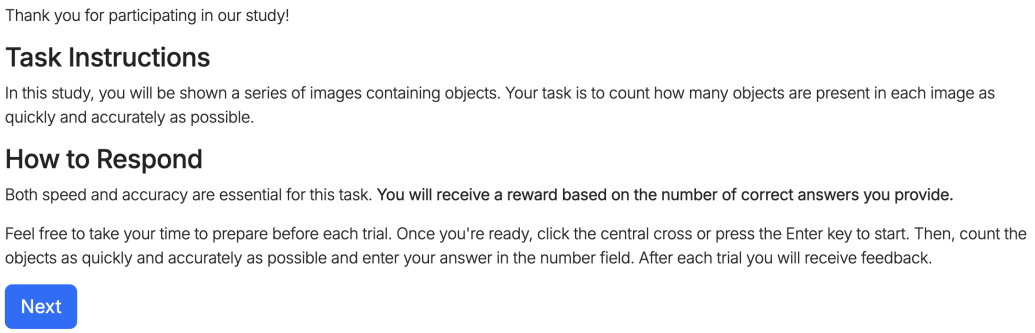

Figure 12: **Instructions for Numerosity Task.**

C.3 Rotation

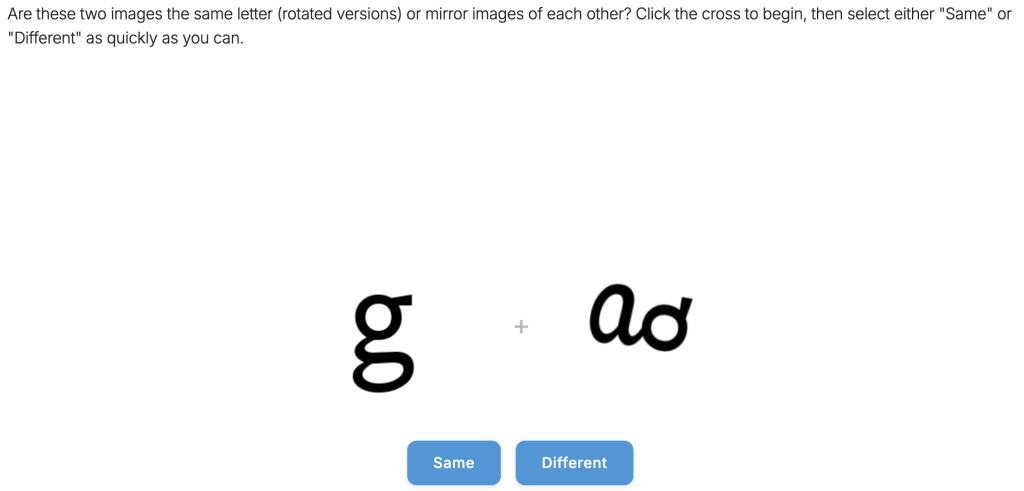

Figure 13: **Example of Rotation Task.**

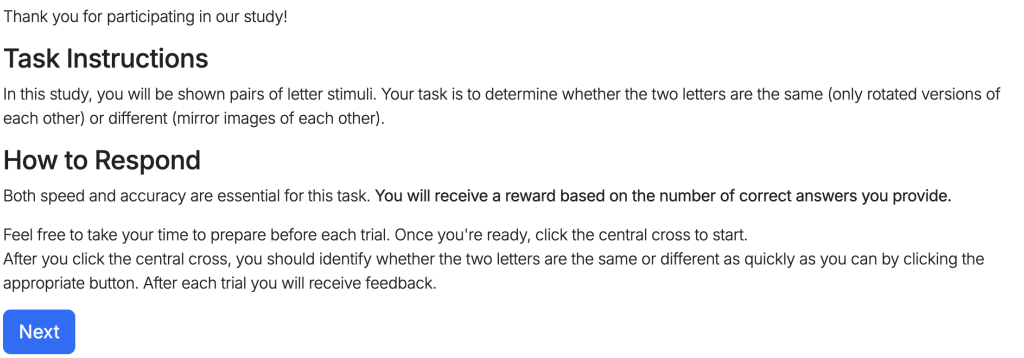

Figure 14: **Instructions for Rotation Task.**

D Appendix: Causal evidence for a serial processing deficit

D.1 Chain-of-Thought and Reasoning

Chain-of-Thought experiments were conducted with GPT-4o, Claude Sonnet 3.7, Llama 4 Maverick, and Qwen2.5 VL, each prompted with the Chain-of-Thought templates described in Appendix A. Reasoning experiments were carried out with GPT-o3, Gemini 2.5 Pro, and Claude Opus 4, also using the Chain-of-Thought prompts. All three experimental paradigms were completed in full under the conditions specified in the main paper. The full set of results for these models, compared against humans, are reported in Table 1.

D.2 Tool-Use

Tool-use experiments were conducted through the ChatGPT interface with the GPT-o3 model. Across three experiments, we performed a total of 400 trials, focusing on challenging task conditions: 100 trials on Geoclidean tasks (20 for each of five tasks, including two with MDL 3 and three with MDL 4), 100 trials on Rotation tasks (angles from 90° to 270° in 30° increments), and 200 trials on Numerosity tasks ($n = 8$ objects, 50 trials for each of the four conditions). The model was prompted with the Chain-of-Thought prompts, as detailed in Appendix A. Results on this restricted set of 400 challenging trials are shown in Table 2.

Model	Geoclidean (CI)	Numerosity (CI)	-	-	-	Rotation (CI)
		uniform_distinct	uniform_overlapping	colored_distinct	colored_overlapping	
Human	78.2 (77.5–78.9)%	98.7 (98.1–99.1)%	69.9 (68.3–71.5)%	98.5 (97.9–98.9)%	71.8 (70.2–73.4)%	79.6 (78.9–80.3)%
NoCoT	66.9 (66.2–67.7)%	85.0 (83.8–86.2)%	43.9 (42.2–45.6)%	91.3 (90.3–92.2)%	61.7 (60.0–63.4)%	56.7 (55.9–57.5)%
CoT	68.5 (67.7–69.2)%	89.2 (88.1–90.3)%	41.3 (39.6–43.0)%	96.0 (95.3–96.6)%	61.1 (59.4–62.7)%	56.9 (56.1–57.8)%
Reasoning	76.2 (75.4–77.0)%	95.6 (94.7–96.3)%	43.9 (41.9–45.9)%	99.7 (99.5–99.9)%	72.1 (70.3–73.9)%	58.1 (57.1–59.0)%

Table 1: Mean accuracy (95% CI) for humans and all model classes (NoCoT (Baseline), CoT, and Reasoning) across the full evaluation set, spanning Geoclidean, Numerosity (four conditions), and Rotation tasks.

Model	Geoclidean (CI)	Numerosity (CI)	-	-	-	Rotation (CI)
		uniform_distinct	uniform_overlapping	colored_distinct	colored_overlapping	
Human	58.0 (55.7–60.2)%	96.9% (92.3–98.8%)	73.5% (66.3–79.6%)	93.8% (88.7–96.7%)	73.9% (65.4–81.0%)	75.2 (71.5–78.6)%
NoCoT	45.5 (40.7–50.4)%	58.5% (51.6–65.1%)	9.5% (6.2–14.4%)	75.0% (68.6–80.5%)	14.5% (10.3–20.0%)	47.8 (43.1–52.5)%
CoT	43.5 (38.7–48.4)%	70.9% (64.2–76.7%)	6.0% (3.5–10.2%)	78.4% (72.2–83.5%)	18.1% (13.4–24.0%)	47.3 (42.6–52.2)%
Reasoning	49.3 (43.7–55.0)%	79.2% (72.0–84.9%)	8.7% (5.1–14.3%)	99.3% (96.3–99.9%)	29.3% (22.6–37.1%)	48.4 (43.0–53.9)%
Tool Use	66.0 (56.3–74.5)%	100.0% (92.9–100.0%)	10.0% (4.3–21.4%)	100.0% (92.9–100.0%)	28.0% (17.5–41.7%)	88.8 (81.0–93.6)%

Table 2: Mean accuracy (95% CI) for humans and all model classes (NoCoT (Baseline), CoT, Reasoning, and Tool Use) evaluated on the same 400 challenging trials spanning Geoclidean, Numerosity (four conditions), and Rotation tasks.

E Appendix: Computational Resources

All model queries were conducted via API (e.g., Together), including open-weight models such as Llama. No specialized hardware was required. Dataset generation and preprocessing were lightweight and completed on standard CPU machines.